ECONOMETRIC SOCIETY MONOGRAPHS

Advances in Economics and Econometrics

Theory and Applications, Eighth World Congress, Volume II

Edited by Mathias Dewatripont, Lars Peter Hansen, and Stephen J.Turnovsky

CAMBRIDGE

more information - www.cambridge.org/9780521818735

This page intentionally left blank

Advances in Economics and Econometrics

This is the second of three volumes containing edited versions of papers and commentaries presented at invited symposium sessions of the Eighth World Congress of the Econometric Society, held in Seattle, WA, in August 2000. The papers summarize and interpret recent key developments, and they discuss future directions for a wide range of topics in economics and econometrics. The papers cover both theory and applications. Written by leading specialists in their fields, these volumes provide a unique survey of progress in the discipline.

Mathias Dewatripont is Professor of Economics at the Université Libre de Bruxelles where he was the founding Director of the European Centre for Advanced Research in Economics (ECARE). Since 1998, he has been Research Director of the London-based CEPR (Centre for Economic Policy Research) network. In 1998, he received the Francqui Prize, awarded each year to a Belgian scientist below the age of 50.

Lars Peter Hansen is Homer J. Livingston Distinguished Service Professor of Economics at the University of Chicago. He was a co-winner of the Frisch Prize Medal in 1984. He is also a member of the National Academy of Sciences.

Stephen J. Turnovsky is Castor Professor of Economics at the University of Washington and recently served as an Editor of the *Journal of Economic Dynamics and Control.* He is an Associate Editor and is on the Editorial Board of four other journals in economic theory and international economics.

Professors Dewatripont, Hansen, and Turnovsky are Fellows of the Econometric Society and were Program Co-Chairs of the Eighth World Congress of the Econometric Society, held in Seattle, WA, in August 2000.

Econometric Society Monographs No. 36

The Econometric Society is an international society for the advancement of economic theory in relation to statistics and mathematics. The Econometric Society Monograph Series is designed to promote the publication of original research contributions of high quality in mathematical economics and theoretical and applied econometrics.

Other titles in the series:

G. S. Maddala *Limited dependent and qualitative variables in econometrics*, 0 521 33825 5

Gerard Debreu *Mathematical economics: Twenty papers of Gerard Debreu*, 0 521 33561 2

Jean-Michel Grandmont *Money and value: A reconsideration of classical and neoclassical monetary economics*, 0 521 31364 3

Franklin M. Fisher *Disequilibrium foundations of equilibrium economics*, 0 521 37856 7

Andreu Mas-Colell *The theory of general economic equilibrium: A differentiable approach*, 0 521 26514 2, 0 521 38870 8

Truman F. Bewley, Editor *Advances in econometrics – Fifth World Congress (Volume I)*, 0 521 46726 8

Truman F. Bewley, Editor *Advances in econometrics – Fifth World Congress (Volume II)*, 0 521 46725 X

Herv́e Moulin *Axioms of cooperative decision making*, 0 521 36055 2, 0 521 42458 5

L. G. Godfrey *Misspecification tests in econometrics: The Lagrange multiplier principle and other approaches*, 0 521 42459 3

Tony Lancaster *The econometric analysis of transition data*, 0 521 43789 X

Alvin E. Roth and Marilda A. Oliviera Sotomayor, Editors *Two-sided matching: A study in game-theoretic modeling and analysis*, 0 521 43788 1

Wolfgang Härdle, *Applied nonparametric regression*, 0 521 42950 1

Jean-Jacques Laffont, Editor *Advances in economic theory – Sixth World Congress (Volume I)*, 0 521 48459 6

Jean-Jacques Laffont, Editor *Advances in economic theory – Sixth World Congress (Volume II)*, 0 521 48460 X

Halbert White *Estimation, inference and specification*, 0 521 25280 6, 0 521 57446 3

Christopher Sims, Editor *Advances in econometrics – Sixth World Congress (Volume I)*, 0 521 56610 X

Christopher Sims, Editor *Advances in econometrics – Sixth World Congress (Volume II)*, 0 521 56609 6

Roger Guesnerie *A contribution to the pure theory of taxation*, 0 521 23689 4, 0 521 62956 X

David M. Kreps and Kenneth F. Wallis, Editors *Advances in economics and econometrics – Seventh World Congress (Volume I)*, 0 521 58011 0, 0 521 58983 5

David M. Kreps and Kenneth F. Wallis, Editors *Advances in economics and econometrics – Seventh World Congress (Volume II)*, 0 521 58012 9, 0 521 58982 7

David M. Kreps and Kenneth F. Wallis, Editors *Advances in economics and econometrics – Seventh World Congress (Volume III),* 0 521 58013 7, 0 521 58981 9

Donald P. Jacobs, Ehud Kalai, and Morton I. Kamien, Editors *Frontiers of research in economic theory: The Nancy L. Schwartz Memorial Lectures, 1983–1997*, 0 521 63222 6, 0 521 63538 1

A. Colin, Cameron and Pravin K. Trivedi *Regression analysis of count data*, 0 521 63201 3, 0 521 63567 5

Steinar Strøm, Editor *Econometrics and economic theory in the 20th century: The Ragnar Frisch Centennial Symposium*, 0 521 63323 0, 0 521 63365 6

Eric Ghysels, Norman R. Swanson, and Mark Watson, Editors *Essays in econometrics: Collected papers of Clive W. J. Granger (Volume I)*, 0 521 77297 4, 0 521 80401 8, 0 521 77496 9, 0 521 79697 0

Eric Ghysels, Norman R. Swanson, and Mark Watson, Editors *Essays in econometrics: Collected papers of Clive W. J. Granger (Volume II)*, 0 521 79207 X, 0 521 80401 8, 0 521 79649 0, 0 521 79697 0

Cheng Hsiao, *Analysis of panel data*, second edition, 0 521 81855 9, 0 521 52271 4

Mathias Dewatripont, Lars Peter Hansen, and Stephen J. Turnovsky, Editors *Advances in economics and econometrics – Eighth World Congress (Volume 1)*, 0 521 81872 8, 0 521 52411 3

Mathias Dewatripont, Lars Peter Hansen, and Stephen J. Turnovsky, Editors *Advances in economics and econometrics – Eighth World Congress (Volume III)*, 0 521 81874 5, 0 521 52413 X

Advances in Economics and Econometrics

Theory and Applications, Eighth World Congress, Volume II

Edited by

Mathias Dewatripont
Université Libre de Bruxelles
and CEPR, London

Lars Peter Hansen
University of Chicago

Stephen J. Turnovsky
University of Washington

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo

Cambridge University Press
The Edinburgh Building, Cambridge CB2 2RU, United Kingdom

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org

Information on this title: www.cambridge.org/9780521818735

First published in print format 2003

ISBN-13 978-0-511-06990-1 eBook (EBL)
ISBN-10 0-511-06990-1 eBook (EBL)

ISBN-13 978-0-521-81873-5 hardback
ISBN-10 0-521-81873-7 hardback

ISBN-13 978-0-521-52412-4 paperback
ISBN-10 0-521-52412·1 paperback

Contents

Contributors

Franklin Allen
University of Pennsylvania

Manuel Arellano
CEMFI, Madrid

Richard Blundell
University College, London and Institute for Fiscal Studies

Raquel Fernández
New York University, CEPR, and NBER

Jean-Pierre Florens
University of Toulouse

John Geanakoplos
Yale University

Andrew W. Lo
Massachusetts Institute of Technology and NBER

Costas Meghir
University College, London and Institute for Fiscal Studies

James L. Powell
University of California at Berkeley

Patrick Rey
IDEI, University of Toulouse

Jiang Wang
Massachusetts Institute of Technology and NBER

Frank A. Wolak
Stanford University

Kenneth I. Wolpin
University of Pennsylvania

Preface

These volumes contain the papers of the invited symposium sessions of the Eighth World Congress of the Econometric Society. The meetings were held at the University of Washington, Seattle, in August 2000; we served as Program Co-Chairs. Volume 1 also contains an invited address, the "Seattle Lecture," given by Eric Maskin. This address was in addition to other named lectures that are typically published in *Econometrica*. Symposium sessions had discussants, and about half of these wrote up their comments for publication. These remarks are included in the book after the session papers they comment on.

The book chapters explore and interpret recent developments in a variety of areas in economics and econometrics. Although we chose topics and authors to represent the broad interests of members of the Econometric Society, the selected areas were not meant to be exhaustive. We deliberately included some new active areas of research not covered in recent Congresses. For many chapters, we encouraged collaboration among experts in an area. Moreover, some sessions were designed to span the econometrics–theory separation that is sometimes evident in the Econometric Society. We followed the lead of our immediate predecessors, David Kreps and Ken Wallis, by including all of the contributions in a single book edited by the three of us. Because of the number of contributions, we have divided the book into three volumes; the topics are grouped in a manner that seemed appropriate to us.

We believe that the Eighth World Congress of the Econometric Society was very successful, and we hope that these books serve as suitable mementos of that event. We are grateful to the members of our Program Committee for their dedication and advice, and to Scott Parris at Cambridge University Press for his guidance and support during the preparation of these volumes. We also acknowledge support from the officers of the Society – Presidents Robert Lucas, Jean Tirole, Robert Wilson, Elhanan Helpman, and Avinash Dixit – the Treasurer, Robert Gordon, and Secretary, Julie Gordon. Finally, we express our gratitude to the Co-Chairs of the Local Organizing Committee, Jacques Lawarree and Fahad Khalil, for a smoothly run operation.

CHAPTER 1

Sorting, Education, and Inequality

Raquel Fernández

1. INTRODUCTION

Individuals sort in a variety of fashions. The workplace, the school of one's child, the choice of neighborhood in which to reside, and the selection of a spouse are all important arenas in which a choice of peers and access to particular goods and networks is explicitly or implicitly made. The aim of this chapter is to review the subset of the literature in the rapidly growing field of education and inequality that is primarily concerned with how individuals sort and the consequences of this for the accumulation of human capital, equity, efficiency, and welfare.

At first blush, sorting may seem like a rather strange lens through which to examine education. After all, this field has been primarily concerned with examining issues such as the returns to education, the nature of the education production function, or, at a more macro level, the relationship between education and per capita output growth.[1] A bit more thought, though, quickly reveals that sorting is an integral component of these questions. With whom one goes to school or works, who one's neighbors are, and who is a member of one's household are all likely to be important ingredients in determining both the resources devoted to and the returns to human capital accumulation.

It is interesting to note that in all these spheres there is at least some evidence indicating that sorting is increasing in the United States. Jargowsky (1996), for example, examines the changing pattern of residential segregation in the United States over the past few decades. He finds that although racial and ethnic segregation has stayed fairly constant (with some small decline in recent years), segregation by income has increased (for whites, blacks, and Hispanics) in all U.S. metropolitan areas from 1970 to 1990. This increased economic segregation, and the fact that schools increasingly track students by ability, suggests that there is likely to be increased sorting at the school or classroom level by

[1] For a survey of the education production function literature, see Hanushek (1986); for returns to education, see, for example, Heckman, Layne-Farrar, and Todd (1996); for education and growth, see, for example, Benhabib and Spiegel (1994).

income and ability. Kremer and Maskin (1996) find evidence for the United States, Britain, and France that there is increased sorting of workers into firms, with some high-tech firms (e.g., silicon valley firms) employing predominantly high-skilled workers and low-tech firms (e.g., the fast-food industry) employing predominantly low-skilled workers. Lastly, there is also some indication of greater sorting at the level of household partner (or "marital" sorting). Although the correlation between spousal partners in terms of years of education has not changed much over the past few decades (see Kremer, 1997), the conditional probability of some sociological barriers being crossed – for example, the probability that an individual with only a high-school education will match with another with a college education – has decreased, indicating greater household sorting (see Mare, 1991).

This chapter examines some of the literature that deals with the intersection of sorting, education, and inequality. This review is not meant to be exhaustive but to give a flavor of some of the advances in the theory and quantitative evidence. Furthermore, it should be noted that there is no overarching theoretical framework in this field. Rather, different models are interesting because of how they illuminate some of the particular interactions among these variables and others – for example, the role of politics, the interaction between private and public schools, or the efficacy of different mechanisms (e.g., markets vs. tournaments) in solving assignment problems. Thus, rather than sketch the contribution of each paper, I have chosen to discuss a few models in depth. Furthermore, as a primary concern in this area is the magnitude of different effects, wherever possible I focus on the contributions that have attempted to evaluate these.

The organization of the chapter is as follows. I begin with the topic of residential sorting. Local schooling is prevalent in most of the world. This policy easily leads to residential sorting and may have important implications for education and inequality, particularly in countries such as the United States in which the funding of education is also largely at the local level. I also use this section to review the theory of sorting. Next, I turn to examining sorting at the school level. The papers here are different, as they are primarily concerned with the interaction of public and private schools and with the properties of different mechanisms. Lastly I turn to recent work on household sorting and its consequences for education and inequality.

2. SORTING INTO NEIGHBORHOODS

Neighborhoods do not tend to be representative samples of the population as a whole. Why is this? Sorting into neighborhoods may occur because of preferences for amenities associated with a particular neighborhood (say, parks), because of some individuals' desire to live with some types of people or not to live with some others (say, ethnic groups who wish to live together in order to preserve their culture, or who end up doing so as a result of discrimination), and in response to economic incentives. This chapter is primarily concerned with the latter, and in particular with the endogenous sorting that occurs in response to economic incentives that arise as a result of education policies.

Primary and secondary education is a good that is provided locally. In industrialized countries, the overwhelming majority of children attend public schools (in the United States, this number was a bit over 91 percent in 1996; similar percentages exist in other countries).[2] Typically, children are required to live in a school's district to attend school there, making a neighborhood's school quality a primary concern of families in deciding where to reside. Furthermore, in most countries at least some school funding (usually that used to increase spending above some minimum) is provided locally; this is particularly true in the United States, where only 6.6 percent of funding is at the federal level, 48 percent is at the state level, and 42 percent is at the local level.[3]

Does it matter that education is provided at the local level? How does local provision of education affect the accumulation of human capital, its distribution, and efficiency in general? What are the dynamic consequences of local provision? How do other systems of financing and providing education compare? These are some of the questions explored in this section. I start with a brief overview of the economics of sorting, much of which carries through to the other sections.

2.1. Multicommunity Models: The Economics of Sorting

Characterizing equilibrium in models in which heterogeneous individuals can choose among a given number of potential residences, and in which these choices in aggregate affect the attributes of the community, is generally a difficult task. Since Westhoff (1977), economists working with these often called "multicommunity models" have tended to impose a single-crossing condition on preferences in order to obtain and characterize equilibria in which individuals either partially or completely separate out by type.[4] As discussed in further detail in the paragraphs that follow, the single-crossing condition also has two other very useful implications. First, it guarantees the existence of a majority voting equilibrium. Second, in many models it allows one to get rid of "trivial" equilibria (e.g., those in which all communities are identical) when a local stability condition is employed.

A typical multicommunity model consists of a given number of communities, each associated with a bundle (q, p). These bundles consist of a good, or input that is provided in some quality or quantity q *at the community level* and of a community level price p of some (usually other) good or service. The latter

[2] Digest of Education Statistics (1999).

[3] The remaining percentages come from other miscellaneous sources. These figures are for 1996–1997 (Digest of Education Statistics, 1999).

[4] In games of asymmetric information (e.g., signaling models and insurance provision), the assumption of single-crossing indifference curves is used to obtain either partial or completely separating equilibria.

can simply be a price associated with residing in the neighborhood, such as a local property tax. Thus, we can assess the indirect utility of an individual from these residing in a given community as $V(q, p; y)$, where y is an attribute of the individual such as income, ability, parental human capital, wealth, or taste. We assume throughout that q is "good" in the sense that $V_q > 0$, whereas $V_p < 0$.

Individuals choose a community in which to reside. In these models, equilibria in which individuals sort into communities along their characteristic y are obtained by requiring the slope of indifference curves in (q, p) space,

$$\left.\frac{dp}{dq}\right|_{v=\bar{v}} = -\frac{V_q}{V_p}, \tag{2.1}$$

to be everywhere increasing (or decreasing) in y. This implies that indifference curves cross only once and that where they do, if (2.1) is increasing in y, then the slope of the curve of an individual with a higher y is greater than one with a lower y; the opposite is true if (2.1) is decreasing in y.

The assumption of a slope that increases (decreases) in y ensures that if an individual with y_i prefers the bundle (q_j, p_j) offered by community j to some other bundle (q_k, p_k) offered by community k, and $p_j > p_k$, then the same preference ordering over these bundles is shared by all individuals with $y > y_i$ $(y < y_i)$. Alternatively, if the individual with y_i prefers (q_k, p_k), then community k will also be preferred to community j by all individuals with $y < y_i$ $(y > y_i)$.

Either an increasing or a decreasing slope can be used to obtain separation.[5] Henceforth, unless explicitly stated otherwise, I assume that (2.1) is increasing in y; that is,

$$\frac{\partial\left(\left.\frac{dp}{dq}\right|_{v=\bar{v}}\right)}{\partial y} = -\frac{V_{qy}V_p - V_{py}V_y}{V_p^2} > 0. \tag{2.2}$$

We shall refer to equilibria in which there is (at least some) separation by characteristic as sorting or stratification.

Condition (2.2) is very powerful. Independent of the magnitude of the expression, the fact that it is positive implies that individuals have an incentive to sort. As we discuss in the next section, this will be problematic for efficiency because it implies that even very small private incentives to sort will lead to a stratified equilibrium, independent of the overall social costs (which may be large) from doing so.

There are many economic situations in which condition (2.2) arises naturally. Suppose, for example, that q is the quality of education, and that this is determined by either a lump sum or a proportional tax p on income. If individuals

[5] Note that although either assumption can be used to obtain separation, the economic implications are very different. If increasing, then in a stratified equilibrium, higher-y individuals would obtain a higher (q, p). If decreasing, the high (q, p) bundle would be obtained by lower-y individuals.

are, for example, heterogeneous in income (so y denotes the income of the individual), then this condition would imply that higher-income individuals are willing to pay more (either in levels or as a proportion of their income, depending on the definition of p) to obtain a greater quality of education. This can then result in an equilibrium stratified along the dimension of income. Alternatively, if the quality of education is determined by the mean ability of individuals in the community school, p is the price of housing in the community, and individuals are heterogeneous in ability y, then (2.2) will be met if higher-ability individuals are willing to pay a higher price of housing to obtain higher-quality (mean ability) schooling, allowing the possibility of a stratified equilibrium along the ability dimension.

It is important to note, given the centrality of borrowing constraints in the human capital literature, that differential willingness to pay a given price is not the only criterion that determines whether sorting occurs.[6] Suppose, for example, that individuals are unable to borrow against future human capital or, less restrictively, that individuals with lower income, or lower wealth, or whose parents have a lower education level, face a higher cost of borrowing. Then even in models in which there is no other incentive to sort (e.g., in which the return to human capital is not increasing in parental assets or, more generally, in which V_q is not a function of y), there will nonetheless be an incentive to sort if the cost of residing in communities with higher qs (i.e., the effective p that individuals face) is decreasing in y. So, for example, if individuals with fewer assets face a higher effective cost of borrowing (they are charged higher rates of interest by banks), then they will be outbid by wealthier individuals for housing in communities with a higher q.

In many variants of multicommunity models, not only does (2.2) give rise to stratified equilibria, but it also implies that all locally stable equilibria must be stratified.[7] In particular, the equilibrium in which all communities offer the same bundle, and thus each contains a representative slice of the population, is locally unstable.[8]

There are many local stability concepts that can be imposed in multicommunity models. A particularly simple one is to define local stability as the property that the relocation of a small mass of individuals from one community to another implies that under the new configuration of (q, p) in these communities, the relocated individuals would prefer to reside in their original community. More rigorously, an equilibrium is locally stable if there exists an $\varepsilon > 0$, such that, for all possible combinations of measure δ $(0 < \delta \leq \varepsilon)$ of individuals $y_i \in \Lambda_j^*$ (where Λ_j^* is the set of individuals that in equilibrium reside in community j),

[6] For human capital models in which imperfections in credit markets play a central role, see Fernández and Rogerson (1998), Galor and Zeira (1993), Ljungqvist (1993), and Loury (1981), among others.

[7] In many settings, this gives rise to a unique locally stable equilibrium.

[8] Note that this zero sorting configuration is always an equilibrium in multicommunity models, as no single individual has an incentive to move.

a switch in residence from community j to k implies

$$V(q_k(\delta), p_k(\delta), y) \leq V(q_j(\delta), p_j(\delta), y), \quad \forall y \in \Lambda_{jk}, \forall j, k, \tag{2.3}$$

where $(q_l(\delta), p_l(\delta))$ are the new bundles of (q, p) that result in community $l = j, k$. Thus, condition (2.3) requires that, for all individuals who switch residence (the set Λ_{jk}), at the new bundles they should still prefer community j. This condition is required to hold for all community pairs considered.[9]

To see why the equilibrium with no sorting is rarely locally stable, consider, for example, the relocation of a small mass of high-y individuals from community j to k. In models in which the provision of the local good is decided by majority vote, this will tend to make the new community more attractive to the movers (and the old one less attractive), because the median voter in community k will now have preferences closer to those of the high-y individuals whereas the opposite will be true in community j. In models in which q is an increasing function of the mean of y (or an increasing function of an increasing function of the mean of y), such as when q is spending per student or the average of the human capital or ability of parents or students, then again this move will make community k more attractive than community j for the high-y movers. Thus, in all these cases, the no-sorting equilibrium will be unstable.

In several variants of multicommunity models, existence of an equilibrium (other than the unstable one with zero sorting) is not guaranteed.[10] For example, in a model in which the community bundle is decided upon by majority vote and voters take community composition as given, a locally stable equilibrium may fail to exist. The reason for this is that although there will exist (often infinite) sequences of community bundles that sort individuals into communities, majority vote need not generate any of these sequences. Introducing a good (e.g., housing) whose supply is fixed at the local level (so that the entire adjustment is in prices), though, will typically give rise to existence.[11]

Condition (2.2) also has an extremely useful implication for the political economy aspect of multicommunity models. Suppose that p and q are functions of some other variable t to be decided upon by majority vote by the population in the community (say, a local tax rate). They may also be functions of the characteristics of the (endogenous) population in the community. An implication of (2.2) is that independent of whether p and q are "nicely" behaved functions of t, the equilibrium outcome of majority vote over t will be the value preferred by the individual whose y is median in the community.

The proof of this is very simple. Consider the (feasible) bundle $(\tilde{q}, \tilde{p})$ preferred by the median-y individual in the community, henceforth denoted $\tilde{y}$. An

[9] See, for example, Fernández and Rogerson (1996). If communities have only a fixed number of slots for individuals as in models in which the quantity of housing is held fixed, then this definition must be amended to include the relocation of a corresponding mass of individuals from community k to j.

[10] See Westhoff (1977) and Rose-Ackerman (1979).

[11] See, for example, Nechyba (1997).

implication of (2.2) is that any feasible (q, p) bundle that is greater than $(\tilde{q}, \tilde{p})$ will be rejected by at least 50 percent of the residents in favor of $(\tilde{q}, \tilde{p})$, in particular by all those whose y is smaller than $\tilde{y}$. On the other hand, any feasible bundle with a (q, p) lower than $(\tilde{q}, \tilde{p})$ will also be rejected by 50 percent of the residents, namely all those with $y > \tilde{y}$. Thus, the bundle preferred by $\tilde{y}$ will be chosen by majority vote.[12]

It is also important to note that even in the absence of a single-crossing condition, to the extent that education is funded in a manner that implies redistribution at the local level, wealthier individuals will have an incentive to move away from less wealthy ones. This is by itself a powerful force that favors sorting but often requires a mechanism (e.g., zoning) to prevent poorer individuals from chasing richer individuals in order to enjoy both a higher q and a lower p.

For example, a system of local provision of education funded by a local property tax implicitly redistributes from those with more expensive housing to those with less expensive housing in the same neighborhood. The extent of redistribution, though, can be greatly minimized by zoning regulations that, for example, require minimum lot sizes.[13] This will raise the price of living with the wealthy and thus greatly diminish the amount of redistribution that occurs in equilibrium. In several models, to simplify matters, it will be assumed that mechanisms such as zoning ensure perfect sorting.

2.2. The Efficiency of Local Provision of Education

The simplest way to model the local provision of education is in a Tiebout model with (exogenously imposed) perfect sorting. In this model, individuals with different incomes y_i, but with identical preferences over consumption c and quality of education q, sort themselves into homogeneous communities. Each community maximizes the utility of its own representative individual subject to the individual or community budget constraint. Let us assume that the quality of education depends only on spending per student (i.e., the provision of education exhibits constant returns to scale and there are no peer effects). Then, perfect sorting is Pareto efficient. Note that this system is identical to a purely private system of education provision.

The model sketched in the previous paragraph often guides many people's intuition in the field of education. This is unfortunate, as it ignores many issues central to the provision of education. In particular, it ignores the fact that education is an *investment* that *benefits the child* and potentially affects the *welfare of others* as well. These are important considerations, as the fact that education is primarily an investment rather than a consumption good implies that borrowing constraints may have significant dynamic consequences; the fact that

[12] See Westhoff (1977) and Epple and Romer (1991). Also see Gans and Smart (1996) for a more general ordinal version of single crossing and existence of majority vote.

[13] See Fernández and Rogerson (1997b) for an analysis that endogenizes zoning, sorting, and the provision of education.

education primarily affects the child's (rather than the parents') welfare raises the possibility that parents may not be making the best decisions for the child. Furthermore, the potential externalities of an agent's education raise the usual problems for Pareto optimality.

In the paragraphs that follow, I explore some departures from the assumptions in the basic Tiebout framework and discuss how they lead to inefficiency of the stratified equilibrium. This makes clear a simple pervasive problem associated with sorting, namely that utility-maximizing individuals do not take into account the effect of their residence decisions on community variables. I start by discussing the simplest modification to the basic Tiebout model – reducing the number of communities relative to types.

Following Fernández and Rogerson (1996), consider an economy with a given number of communities $j = \{1, 2, \ldots, N\}$, each (endogenously) characterized by a proportional income tax rate t_j and a quality of education q_j equal to per pupil expenditure, that is, $q_j = t_j \mu_j$. Individuals who differ in income y_i, where $i \in I = \{1, 2, \ldots, \bar{I}\}$ (with $y_1 > y_2 > \cdots > y_{\bar{I}}$), simultaneously decide in which community, C_j, they wish to reside. Once that decision is made, communities choose tax rates by means of majority vote at the community level. Individuals then consume their after-tax income and obtain education.[14]

Assume for simplicity that individual preferences are characterized by the following separable specification,

$$u(c) + v(q), \tag{2.4}$$

so that sorting condition (2.2) is satisfied if $-\{[u''(c)c]/[u'(c)]\} > 1, \forall c$. We henceforth assume that the inequality is satisfied, ensuring that individuals with higher income are willing to suffer a higher tax rate for higher quality.[15]

Suppose that the number of communities is smaller than the number of income types.[16] In such a case the equilibrium will generally not be Pareto efficient. The clearest illustration of this can be given for the case in which individuals have preferences such that an increase in the mean income of the community *ceteris paribus* decreases the tax rate that any *given* individual would

[14] Very often, the literature in this field has implicitly adopted a sequencing such as the one outlined here. Making the order of moves explicit as in Fernández and Rogerson (1996) allows the properties of equilibrium (e.g., local stability) to be studied in a more rigorous fashion. It would also be of interest to examine properties of models in which communities act more strategically and take into account the effect of their tax rate on the community composition. There is no reason to believe that this modification would generate an efficient equilibrium, however.

[15] Most assumptions here are for simplicity only; for example, preferences need not be separable, and introducing housing and property taxation rather than income taxation would allow a sorting equilibrium to be characterized by higher-income communities having lower tax rates (but higher tax-inclusive prices) and higher q. We forgo the last option, as it simply complicates matters without contributing additional insights.

[16] Note that type here is synonymous with income level. Hence the assumption that there are fewer neighborhoods than types is a reasonable one to make.

like to impose. As the preferred tax rate of an individual is given by equating $u'(c)y_i$ to $v'(q)\mu_j$, this is ensured by assuming $-\{[v''(q)q]/[v'(q)]\} > 1$ (note that this is the parallel of the condition on u that generates sorting).[17]

As discussed previously, the result of majority vote at the community level is the preferred tax rate of the median-income individual in the community. There are a few things to note about the characteristics of equilibrium. First, in equilibrium, no community will be empty. If one were, then in any community that contained more than one income type, those with higher income would be made better off by moving to the empty community, imposing their preferred tax rate, and engaging in no redistribution. Second, in a locally stable equilibrium, communities cannot offer the same bundles and contain more than one type of individual (as a small measure of those with higher income could move to one of the communities, increase mean income there, and end up with the same or a higher-income median voter who has preferences closer to theirs). Lastly, if communities have different qualities of education (as they must if the communities are heterogeneous), then a community with a strictly higher q than another must also have a strictly higher t (otherwise no individual would choose to reside in the lower-quality–higher-tax community).

In the economic environment described herein, all locally stable equilibria must be stratified; that is, individuals will sort into communities by income. In such equilibria, communities can be ranked by the quality of education they offer, their income tax rate, and the income of the individuals that belong to them. Thus, all stable equilibria can be characterized by a ranking of communities such that $\forall j, q_j > q_{j+1}, t_j > t_{j+1}$, and $\min y_i \in C_j \geq \max y_i \in C_{j+1}$.

To facilitate the illustration of inefficiency, assume for simplicity that there are only two communities, $j = 1, 2$, and $I > 2$ types of individuals.[18] A stratified equilibrium will have all individuals with income strictly greater than some level y_b living in C_1 and those with income strictly lower than y_b living in C_2 with $q_2 > q_1$ and $t_2 > t_1$.

Suppose that in equilibrium individuals with income y_b live in both communities. It is easy to graph the utility

$$W_b^j \equiv u(y_b(1 - t_j)) + v(t_j\mu_j) \tag{2.5}$$

of these "boundary" individuals as a function of the community in which they reside and as a function of the fraction ρ_b of these individuals that reside in C_1. Let ρ_b^* denote the equilibrium value of the boundary individuals residing in C_1. Note that a decrease in ρ_b from its equilibrium value that does not alter the identity of the median voter in either community will make individuals with income y_b better off in both communities, as mean incomes will rise, qualities of education increase, and tax rates fall in both communities. Thus,

[17] This assumption implies that an increase in the mean income of the community that does not change the identity of the median voter will result in a higher q and a lower t, ensuring that all residents are made better off.

[18] See Fernández and Rogerson (1996) for a generalization of this argument to many communities.

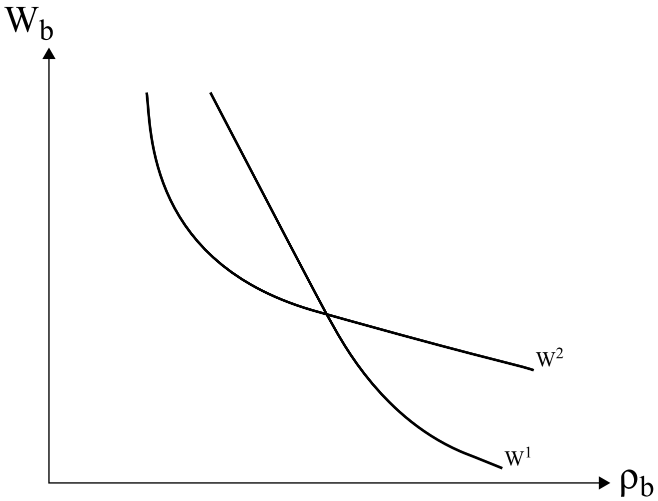


Figure 1.1. Equilibrium.

for this equilibrium to be locally stable, it must be that such a decrease makes y_b individuals even better off in C_1 relative to C_2, reversing the outward flow and reestablishing ρ_b^* as the equilibrium. Thus, as shown in Figure 1.1, the W_b^1 curve must cross the W_b^2 curve from above.[19]

This equilibrium is clearly inefficient. Consider a marginal subsidy of $s > 0$ to all individuals with income y_b who choose to reside in C_2.[20] Given that without a subsidy these individuals are indifferent between residing in either community, it follows that a subsidy will increase the attractiveness of C_2 relative to C_1. Consequently, some y_b individuals will move to C_2, thereby increasing mean income in both communities. For a small enough subsidy such that the identity of the median voter does not change in either community, the overall effect will be to decrease tax rates and increase the quality of education in both communities, thus making all individuals better off. Thus, it only remains to show that the subsidy can be financed in such a way to retain the Pareto-improving nature of this policy. A simple way to do so is by (marginally) taxing those y_b individuals who remain in C_1.[21] This tax will only further increase their outflow from C_1 to the point where they are once again indifferent between residing in both communities. As shown in Figure 1.1, the tax serves to further increase the utility of this income group (and consequently everyone else's). This last point suggests that a simpler way of producing the same

[19] Note that we are assuming for the range of ρ_b shown that neither of the communities' median voters are changing.

[20] If income is unobservable, then a small subsidy to all individuals who reside in C_2 would have to be paid.

[21] Again, if income is not observable, it is possible to preserve the Pareto-improving nature of this policy by (marginally) taxing all C_1 residents.

Pareto-improving results is a policy that forgoes the subsidy and simply taxes any y_b individual in C_1. This would again induce the desired migration and increase mean income in both communities.

Fernández and Rogerson (1996) examine these and other interventions in a model with many communities. The principle guiding the nature of Pareto-improving policies is not affected by the number of communities considered; policies that serve to increase mean income in some or in all communities by creating incentives to move relatively wealthier individuals into poorer communities will generate Pareto improvements.[22]

The possibility of Pareto improvements over the decentralized equilibrium in the model given here arises when individuals do not take into account the effect of their residence decisions on community mean income. In the next example, the inefficiency of equilibrium occurs when individual residence decisions do not internalize diminishing returns.

Consider a multicommunity model with two communities, C_1 and C_2, and a total population (of parents) of $N = 2$. Parents differ in their human capital, h_i, and potentially in their own income y_i. To simplify matters, we assume that the initial distribution is confined to two values h_1 and h_2 with $h_1 > h_2$, and total numbers of parents of each type given by n_1 and n_2, respectively, such that $n_1 + n_2 = 2$.

We assume that each community has a fixed number of residences, $N/2 = 1$, each available at a price p_j, $j = 1, 2$. Let λ_1 be the fraction of high-human-capital parents who choose to live in C_1 (and thus $\lambda_2 = n_1 - \lambda_1$), and let μ_j be the mean human capital of parents that reside in C_j. Thus, $\mu_j(\lambda_j) = \lambda_j h_1 + (1 - \lambda_j)h_2$.

Parents decide in which community to live, pay the price p_j of residing there, and send their children to the community school. Parents care about aggregate family consumption, which is given by the sum of their own income and the child's future income, I, minus the cost of residing in the community and a lump-sum transfer T.

The child's future income is an increasing function of the human capital he or she acquires. This depends on his or her parents' human capital and on local human capital q, which is assumed to be an increasing function of the mean human capital in the neighborhood. As the latter is simply a linear function of λ_j, we denote this function as $q_j = Q(\lambda_j)$, $Q' > 0$. Thus,

$$I_{ij} = F(h_i, Q(\lambda_j)), \tag{2.6}$$

with $F_h, F_q > 0$, and where I_{ij} indicates the income of a child with a parent of human capital h_i that resides in neighborhood j.

Hence, parents choose a community in which to reside that maximizes

$$u(y_i + I_{ij} - p_j + T) \tag{2.7}$$

[22] The exact specification of these policies, however, depends on the number of communities involved in a rather odd fashion, as explained in Fernández (1997).

subject to (2.6) and taking p_j, T, and q_j as given. Note that if parental and local human capital are complements in the production of a child's future income, then (2.7) obeys (2.2), and hence individuals will sort.[23] Henceforth, I assume this is the case; that is,

$$\frac{\partial(dp/d\lambda\ |_{u=\bar{u}})}{\partial h} = F_{hq}Q' > 0. \tag{2.8}$$

Given (2.8), the only locally stable equilibrium is that with maximal sorting. Individuals with human capital h_1 live in C_1, characterized by a higher p and a higher q than that in C_2; individuals with h_2 live in C_2. If the number of one of these types exceeds the space available in a community (i.e., 1), then that type is indifferent and lives in both communities. Thus, in equilibrium, $\lambda_1 = \min(1, n_1)$.

To close the model, we need to specify housing prices. Rather than determining the price by specifying the microfoundations of the housing market, as in many models in the literature, we simply solve for the price differential such that no individual would wish to move.[24] Depending on whether n_1 is greater than, smaller than, or equal to 1, there are three different possible configurations, as in the first case h_1 types must be made indifferent ($p_1 - p_2 = I_{11} - I_{12}$), in the second case h_2 types must be made indifferent ($p_1 - p_2 = I_{21} - I_{22}$), and in the third case each type must be at least as well off in its own community as in the other ($I_{11} - I_{12} \geq p_1 - p_2 \geq I_{21} - I_{22}$). Rather than include landlords or define the structure of house ownership by agents, we simply assume, as in de Bartolome (1990), that housing rents are rebated to individuals in a lump-sum fashion so that each individual receives $T = (p_1 + p_2)/2$ regardless of the community of residence.[25]

Is the decentralized equilibrium efficient? Rather than characterizing Pareto-improving policies, I confine my discussion here to investigating whether the unique locally stable decentralized equilibrium (that with maximum sorting) maximizes productive efficiency.

The tensions that exist in this model are easy to define. On one hand, parental and local human capital are complements, suggesting that future output is maximized by sorting; that is, efficiency requires concentrating high-human-capital

[23] See de Bartolome (1990) for a two-community fixed housing stock model in which there is complementarity between spending on education and ability but in which peer effects matter more for low-ability students.

[24] See Wheaton (1977) and de Bartolome (1990).

[25] See Bénabou (1993) for a multicommunity model in which individuals can acquire high or low skills or be unemployed. The costs of acquiring skills are decreasing in the proportion of the community that is highly skilled, but this decrease is larger for those acquiring high skills. This leads to sorting although ex ante all individuals are identical. As in the model discussed here, there will be maximal sorting by (ex post) high-skill individuals. The interesting question in this paper is how the decentralized equilibrium compares with one with no sorting given that neither is efficient (because in both cases individuals ignore the externality of their skill-acquisition decision on the costs faced by others).

parents in the same community, precisely what occurs in equilibrium. On the other hand, there is an externality to individual residence decisions that is not being taken into account, namely potentially decreasing returns to the concentration of high-human-capital individuals in the same neighborhood. In particular, individuals do not take into account whether an additional unit of high human capital on the margin increases local human capital more in the community with a high or low concentration of h_1. Similarly, they do not take into account whether a marginal increase in local human capital will add more to total output by being allocated to a community with a high or low concentration of h_1.

To see this more formally, consider the total future income Y generated by a community given that a fraction λ of high-human-capital parents live there:

$$Y(\lambda) = \lambda F(h_1, Q(\lambda)) + (1-\lambda)F(h_2, Q(\lambda)). \tag{2.9}$$

Note that if future income is concave in λ, then it is maximized by allocating high-human-capital parents so that they constitute the same proportion in both communities; that is, $\lambda_1 = \lambda_2$. If, in contrast, future income is convex in λ, then maximum sorting will maximize future income, that is, as in the decentralized equilibrium $\lambda_1 = \min(1, n_1)$.

Taking the appropriate derivatives yields

$$\begin{aligned} Y'' = {} & 2[F_q(h_1, Q(\lambda)) - F_q(h_2, Q(\lambda))]Q' + [\lambda F_q(h_1, Q(\lambda)) \\ & + (1-\lambda)F_q(h_2, Q(\lambda))]Q'' + [\lambda F_{qq}(h_1, Q(\lambda)) \\ & + (1-\lambda)F_{qq}(h_2, Q(\lambda))]Q'^2. \end{aligned} \tag{2.10}$$

Let us examine the terms in (2.10). The complementarity of parental and local human capital in the production of children's human capital guarantees that the expression in the first square brackets is positive. Thus, this factor pushes in the direction of convexity of Y and thus in favor of sorting. Recall from (2.8) that it is only on the basis of this factor that sorting occurs in equilibrium. If, however, there are decreasing returns to community mean human capital in the formation of local human capital, that is, if Q is concave (and thus $Q'' < 0$), then Q'' times the expression in the second square brackets will be negative, imposing losses from concentrating parents with high human capital in the community. Lastly, there will be an additional loss from sorting if there are decreasing returns to local human capital in the production of future income, that is, if $F_{qq} < 0$, as this implies that the term in the third square brackets is negative. Thus, decreasing returns to community mean human capital in the formation of local human capital and decreasing returns to local human capital in the production of children's future income suggest that Y is concave, and hence, that efficiency would be maximized by having parents with high human capital distributed in both communities in the same proportion (see Bénabou, 1996a).

It is important to recall that maximum sorting will take place as long as Fh_q is positive but otherwise independently of its magnitude. Hence a very small amount of complementarity (again, the expression in the first square brackets) and private gain could easily be swamped by the concavity of F and Q and social loss.

The model presented here is one in which all sorting is taking place because of peer effects – that is, people want to live with individuals with high human capital, as it increases the earnings of their children. As local human capital and parental human capital are complements, high-human-capital parents outbid others to live in a community where the level of local human capital is highest, leading to stratification by parental human capital levels. Note that income and the perfection or imperfection of capital markets actually played no role in producing the results shown here.[26]

This analysis also suggests that if spending on education E were an additional factor in the production of future income but not a factor that individuals sorted on, that is, $F(h, Q(\lambda), E(\lambda))$ with $F_E > 0$, $E' > 0$, and $F_{hE} = 0$, then sorting would occur for the same reasons as before, but even a policy of enforced equalization of spending across communities would not stop individuals from sorting.

Unfortunately, there has been very little work done to assess the significance of the inefficiencies discussed here. Although much work points, for example, to the importance of peer effects in learning, whether the appropriate cross partial is negative or positive remains in dispute (i.e., we do not even know whether it would be efficient, all considerations of diminishing returns aside, for children to sort by aptitude, say, or for them to mix).[27] Similarly, we do not know whether quality of education (say, spending) and parental human capitals are complements. This, to my view, makes models in which the main imperfection lies in the functioning of the capital market (and sorting on grounds of minimizing redistribution) relatively more attractive.[28]

2.3. Comparing Systems of Financing Public Education: Dynamic Considerations

The choice of education finance system matters for various reasons. First, and foremost, different finance systems tend to imply different levels of redistribution. In economies in which there is imperfect access to financing the acquisition of human capital, redistribution can play an important role in increasing the human capital levels of children from lower-income families. Different finance

[26] The fact that utility depends only on total net family income and that the latter is not influenced by spending allows us to abstract from issues of borrowing and lending as long as parents have sufficient income to bid successfully for housing.

[27] For example, Henderson, Mieszkowski, and Sauvageau (1978) argue for a zero cross partial and diminishing returns, whereas Summers and Wolfe (1977) argue for a negative cross partial.

[28] These borrowing constraints may not allow families to borrow to send their child to private school, for example. Alternatively, they may not allow poorer families to borrow to live in (wealthy) neighborhoods with higher-quality public education. The general failure of these credit markets is that parents are unable to borrow against the future human capital of their children.

systems may also may have important consequences for who lives where and thus for the identity of a child's peers and for the use of the land market.

There have been several papers written in this area that primarily examine the static consequences of different systems of financing education.[29] Fernández and Rogerson (1999b), for example, examine five different education-finance systems, and they contrast the equity and resources devoted to education across these systems by assuming that the parameters of the education-finance system are chosen by majority vote. They calibrate their benchmark model to U.S. statistics and find that total spending on education may differ by as much as 25 percent across systems. Furthermore, the trade-off between redistribution and resources to education is not monotone; total spending on education is high in two of the systems that also substantially work to reduce inequality. A political economy approach to the contrast of different education-finance systems has also been pursued by Silva and Sonstelie (1995) and Fernández and Rogerson (1999a), who attempt to explain the consequences of California's education-finance reform, whereas Nechyba (1996) and de Bartolome (1997) both study foundation systems. There is also a growing empirical literature devoted to examining how changes in state-level education-finance systems affect education spending, including work by Downes and Schoeman (1998), Loeb (1998), Hoxby (1998), Evans, Murphy, and Schwab (1997, 1998), and Manwaring and Sheffrin (1997).

The papers mentioned here, however, are only indirectly concerned with the consequences of sorting, and they are all static models. In this section, by way of contrast, we focus on dynamic consequences of sorting in response to different education-finance systems. To facilitate the theoretical analysis, we focus on two extreme systems: a pure local system with perfect sorting and a state system with uniform school spending per student across communities.[30]

This section presents two models.[31] The first, based on work by Fernández and Rogerson (1997a, 1998), uses a Tiebout model in which perfect sorting, from a static perspective, is efficient. It then examines the trade-off imposed by switching to a state-financed system. The model is calibrated to U.S. statistics, allowing one to determine whether these trade-offs are quantitatively significant. The main trade-off this analysis illustrates is that between a system that, loosely speaking, allows individuals to consume bundles that are "right" for them given their income versus a system that imposes a uniform education bundle across heterogeneous individuals, but allows for more efficient use of resources from the perspective of future generations. In particular, in an economy in which borrowing constraints prevent individuals from financing their education and missing insurance markets does not allow children (or parents)

[29] See Inman (1978) for an early quantitative comparison of education-finance systems in the context of an explicit model.

[30] See Fernández and Rogerson (2000a) for a dynamic analysis of a foundation system.

[31] Other dynamic analyses of education-finance systems include those by Cooper (1998), Durlauf (1996), Glomm and Ravikumar (1992), and Saint-Paul and Verdier (1993).

to insure against income or ability shocks, a state system may result in a more efficient production of next period's income (again in a sense that will be made rigorous in the paragraphs that follow) than in a local system in which the possibilities for redistribution are only at the local level.[32] The trade-offs are found to be quantitatively significant.

The second model is based on Bénabou (1996). This is a purely theoretical analysis that contrasts the short-run versus long-run consequences of a local system compared with a state system in which the main trade-off is between human capital being complementary in production at the economywide level but parental human capital and spending on education being complementary at the local level.

The simplest contrast between the dynamic consequences of these two extreme forms of education finance – local versus state – can be examined in the familiar Tiebout model of *perfect* sorting in which income is the only source of heterogeneity among individuals. This allows us to abstract away from complications that would be introduced by the political economy of tax choice at the local level when individuals are heterogeneous, by changes in residence over time with the dynamic evolution of the income distribution, by housing (and the inefficiencies that stem from taxing this good), peer effects, or simply from diversity in tastes.[33] Note that by considering a Tiebout system with perfect sorting, we can reinterpret what follows as contrasting a purely private system of education with a state-financed one.

Following Fernández and Rogerson (1997a), consider a two-period overlapping generations model in which each person belongs to a household consisting of one old individual (the parent) and a young one (the child). Parents make all the decisions and have identical preferences described by

$$U(c, y') = u(c) + Ew(y'), \tag{2.11}$$

where y' is next period's income of the household's child, and E is the expectations operator.

In the first period of life, the child attends school and obtains the quality of education q determined by her parent's (equivalently community's) spending. In the second period, the now old child receives a draw from the income distribution. A child's income when old is assumed to depend on the quality of schooling and on an *iid* shock ξ whose distribution $\Psi(\xi)$ is assumed to be independent of q. Thus,

$$y' = f(q, \xi). \tag{2.12}$$

[32] It may be objected that this analysis confounds two things – the amount of redistribution (or insurance) and the system of education. In reality, education always entails some redistribution, and a multidimensional political economy model would be required to allow one to differentiate between redistribution directly through income and through education.

[33] See, however, Fernández and Rogerson (1998) for a more complex dynamic model in which the sorting of individuals into communities endogenously evolves over time along with housing prices and the housing stock.

Once the adult's income is determined, so is the community of residence as an adult. The adult (now a parent) then decides how much of her income to consume and how much to spend on her own child's education. Letting $v(q) \equiv \int w(f(q, \xi))d\Psi$, we can now write preferences exactly as in Equation (2.4). Assuming that v is well behaved, we find that under a local system individuals will set spending on education to equate the marginal utility of consumption with the marginal utility of education quality; that is, $u'(c) = v'(q)$, implying a local tax rate $\tau(y)$ and $q = \tau(y)y$.

We next turn to the determination of spending on education in a state-financed system. We assume that all individuals face the same proportional income tax rate τ_s that is used to finance public education $q = \tau_s \mu$ and that individuals are unable to opt out of public education to a private system.[34]

The first-order condition for utility maximization now equates the ratio of the marginal utility of consumption and the marginal utility of education quality to the ratio of the mean relative to individual income; that is, $[u'(c_i)]/[v'(q)] = \mu/y_i$. Note that this condition reflects the fact that under a state-financed system, unlike in the local system, the relative price of a unit of education (in terms of forgone consumption) is not the same across individuals. Lower-income individuals face a lower price than higher-income individuals. In a local finance system, in contrast, this relative price equals 1 for all individuals. Under majority vote, concavity of u and v implies that the preferences of the individual with a median income in the population determine the choice of τ_s.

Letting $g_t(y)$ be the income distribution of old individuals at the beginning of period t, under either education-finance system an equilibrium at the end of period t generates a beginning-of-period income distribution for period $t + 1$, g_{t+1}. Let $F(g(y))$ be the income distribution that results in the following period given this period's distribution of $g(y)$. A steady state in this model then consists of an income distribution g^* such that $g^*(y) = F(g^*(y))$.

Calibrating this simple model involves making choices over the education quality technology and preferences. There is a large and controversial literature that surrounds the education production function, and there is no consensus on the form it should take.[35] Guided primarily by simplicity, a convenient specification is $y' = Aq^{\theta}\xi$, which yields an elasticity of future income with respect to education quality that is constant and equal to θ. Evidence presented by Card and Krueger (1992), Wachtel (1976), and Johnson and Stafford (1973) suggests an elasticity of earnings with respect to education expenditures close to 0.2. We assume that ξ is lognormally distributed such that log ξ has zero mean and standard deviation σ_{ξ}.

Our specification of preferences comes from noting that across U.S. states the share of personal income devoted to public elementary and secondary

[34] Introducing a private option into this system greatly complicates the analysis, as existence of equilibrium is not ensured. See Stiglitz (1974).

[35] See Coleman et al. (1966), Hanushek (1986), Card and Krueger (1992), and Heckman et al. (1996).

education has remained roughly constant over the 1970–1990 period.[36] This property will be satisfied if the indirect utility function takes the form $c^\alpha/\alpha + E(\Phi(\xi))(q^\alpha/\alpha)$, where $\Phi(\xi)$ is some function of ξ. This requires a utility function of the form

$$\frac{c^\alpha}{\alpha} + E\left(\frac{b}{\alpha}y'^\gamma\right), \tag{2.13}$$

with the restriction that $\theta\gamma = \alpha$.

Under local financing, the preferences given here imply a constant and identical tax rate across individuals, $\tau^* = 1/(1+\kappa)$, where $\kappa = (bA^\gamma E(\xi^\gamma))^{1/(\alpha-1)}$. If a parent's income in period 0 is y_0, it follows that the child's income, y_1, is given by $\log y_1 = \log A + \theta \log \tau^* + \theta \log y_0 + \log \xi_1$. Given $\theta < 1$, it follows that $\log y_t$ has a limiting distribution that is normal with mean and standard deviation:

$$\mu_\infty = \frac{\log A + \theta \log \tau^*}{1-\theta}, \quad \sigma_\infty = \frac{\sigma_\xi}{(1-\theta^2)^{1/2}}. \tag{2.14}$$

We calibrate the steady state of the local model to match U.S. statistics. We choose A and σ_ξ such that μ_∞ and σ_∞ match the mean and median of the U.S. family income distribution, respectively, \$23,100 and \$19,900 in the 1980 census. The remaining parameters to be set are b and α, as the value of θ is already determined by the elasticity of earnings with respect to q.

For any given α, we set b to match the fraction of personal income devoted to public elementary and secondary education (in 1980 equal to 4.1 percent), that is, to yield a tax rate $\tau^* = 0.041$. This determines, for a given value of α, a value of b given by $b = \{[(1-\tau^*)/\tau^*]^{\alpha-1}\}/[A^\gamma E(\xi^\gamma)]$. To set α, we draw upon two pieces of information. The first is the price elasticity of expenditures on education. In our model this can be computed at the equilibrium price (in terms of the consumption good), which here has been set to 1. A survey of the literature by Bergstrom, Rubinfeld, and Shapiro (1982) suggests an elasticity between -0.5 and -0.25, yielding α between -1 and -2. The second is from Fernández and Rogerson (1999a), who model a foundation education-finance system and use it to match the distribution of spending per student in California prior to the Serrano reform. They find an implied value for α equal to -0.2.

One of the main questions we are interested in asking is whether a local system will outperform a state system. Obviously, there is no reason to expect that individuals of all income levels will prefer one system over another, nor that different generations will agree on the relative merits of the two systems. To have a measure of aggregate welfare, we use the sum of individual utilities or, equivalently, the expected utility that an agent would obtain if she or

[36] See Fernández and Rogerson (2001a) for evidence.

Table 1.1. *Steady-state comparisons of local vs. state*

α	$\tau_s \times 10^2$	μ_∞	Median y	$1+\Delta_1$	$1+\Delta_\infty$
0	4.10	25,300	21,900	1.006	1.108
−0.2	4.01	25,100	21,800	1.007	1.104
−0.5	3.91	25,000	21,600	1.009	1.101
−1	3.82	24,000	21,500	1.011	1.101
−2	3.74	24,000	21,400	1.015	1.105

he were to receive a random draw from the equilibrium income distribution. Thus, we use

$$V_{rt} \equiv \int U_{rt} g_{rt}(y) dy \tag{2.15}$$

as our measure of aggregate welfare at time t, where $r = L, S$ (i.e., local L, state S) indicates the education-finance regime.[37]

To provide a measure of welfare change at time t that is unaffected by monotone transformations of the utility function, we examine the proportion by which the income distribution in the *steady state* of the *local* regime would have to be changed so that it provided the same aggregate welfare as the state system in period t. Given that the functional forms adopted are homogeneous of degree α in income, we see that this amounts to finding the value of Δ_t such that $(1 + \Delta_t)^\alpha V_L = V_{S_t}$, where the local system is evaluated at its steady state and the state system in period t.

If α is negative (as our calibration procedure suggests), then preferred tax rates under a state system are increasing in income (under a local system, as noted previously, they are independent of income) and only equal to the local tax rate for those individuals with income such that $y_i = \mu$. Because the median voter's income is lower than the mean income, it follows that the tax rate will be lower under the state system than the (identical) tax rate chosen by each income group under the local system. This implies that in the first period, given that the income distribution is the same as in the local system, aggregate spending on education will decrease.

Table 1.1 shows the tax rate, mean income, and median income in the steady state of the state-finance regime. The last two columns report the first-period gain in aggregate welfare (i.e., prior to the change in mean income), which is denoted by Δ_1, and the steady-state gain in aggregate welfare, denoted Δ_∞. Despite the fact that for α, strictly negative spending on education will decrease in the first period of reform (relative to its value in the local system), we find that

[37] Note that this is equivalent to a utilitarian welfare measure or one chosen "behind the veil of ignorance" (i.e., an individual's welfare if her or his parents were a random draw from the income distribution in that system).

steady-state mean income is always higher than in the local steady state. Furthermore, aggregate welfare increases in period 1 as well as in every subsequent period relative to the initial local-finance system steady state.[38]

As shown in Table 1.1, the first-period gains are relatively small (around 1 percent).[39] The steady-state gain of around 10 percent is surprisingly constant across parameter values, even though the tax rate is changing relative to the local steady state by as much as 10 percent.[40]

The more complicated analysis in Fernández and Rogerson (1998) gives rise to an even starker illustration of differences in short-run and long-run welfare. In that paper, spending on education affects the mean (but not the variance) of the lognormal distribution from which individual income is assumed to be a random draw. Comparing across steady states of a local system relative to a state system of financing education, we find that, *given* an individual's income, each individual prefers a local system to a state system. However, an individual's income is of course not the same across systems, because the probability with which any particular level is realized depends on spending on education, which in turn depends on the system of financing education. It is taking the new distribution of income that results into account that yields a higher steady-state welfare level under the state system.

Next let us turn to an analysis based primarily on Bénabou (1996b). Consider an economy populated by overlapping generations (OLG) dynasties indexed by i who spend some amount of time ν working and the remainder $1-\nu$ passing on education to their single child. The law of motion for the evolution of future descendants' human capital is given by

$$h_{t+1}^{i} = \kappa \xi_t^i \left((1-\nu)h_t^i\right)^{\delta} \left(E_t^i\right)^{1-\delta}, \tag{2.16}$$

reflecting an inherited portion as given by h_t^i (the parent's human capital) and an unpredictable portion given by an *iid* shock ξ_t^i. The shock is assumed to be distributed lognormally such that $\ln \xi_t^i \sim N(-s^2/2, s^2)$ and thus $E(\xi_t^i) = 1$. Formal schooling, E_t^i, is the other input into the production of next period's human capital. This is financed by taxing at rate τ the labor income of local residents. Hence,

$$E_t^i = \tau Y_t^i \equiv \tau \int_0^{\infty} y dm_t^i(y), \tag{2.17}$$

[38] The new steady state is typically reached in five periods.

[39] More generally, the "static" welfare gain might well be negative. In a model with housing, for example, the unbundling of the education and residence decision that a state system allows relative to a local system will generally imply an increase in housing prices in relatively poorer communities and a decrease in wealthier ones. Thus, lower-income individuals will end up paying higher property prices than previously, and the transition to the new steady state may well involve some losses in early periods. In the more complicated model studied by Fernández and Rogerson (1998), this change in housing prices, and the fact that agents' preferred tax rates differ, implies a small decrease (0.3 percent) in the first period of the policy reform.

[40] See Fernández and Rogerson (1997a) for a sensitivity analysis for other parameter values.

where m_t^i is the distribution of income (and Y_t^i is its average) in the community Λ_t^i to which family i belongs at time t.[41]

The production sector is made up of competitive firms with constant returns to scale (CES) technology given by $Y_t = (\int_0^\infty (x_t^r)^{(\sigma-1)/\sigma} dr)^{\sigma/(\sigma-1)}$, $\sigma > 1$, where x_t^r denotes intermediate input r. Each worker must specialize in an intermediate input. As there are an infinite number of inputs, and each faces a downward-sloping demand curve for its services, each worker will choose to specialize in a different intermediate input such that $r(i) = i$ and supply that input in the quantity $x_t^i = \nu h_t^i$. Thus aggregate output simplifies to

$$Y_t = \nu \left(\int_0^\infty h^{(\sigma-1)/\sigma} d\mu_t(h) \right)^{\sigma/(\sigma-1)} \equiv \nu H_t, \tag{2.18}$$

where μ denotes the distribution of human capital in the entire labor force Λ. Note that the complementarity between inputs in the production function implies that a worker's earnings depend on both his or her own human capital and an economywide index of human capital, H_t. That is, $y_t^i = \nu(H_t)^{1/\sigma} \times (h_t^i)^{(\sigma-1)/\sigma}$. This interdependence is also reflected in the per capita income of each community as $Y_t^i = \int_0^\infty y dm_t^i(y) = \nu(H_t)^{1/\sigma}(\int_0^\infty h^{(\sigma-1)/\sigma} d\mu_t^i(h)) \equiv \nu H_t^{1/\sigma}(L_t^i)^{(\sigma-1)/\sigma}$, where $\mu_t^i(h)$ is the distribution of human capital in the community Λ_t^i.

Incorporating the definitions given here into the law of motion for the evolution of human capital (2.16) yields

$$h_{t+1}^i = K\xi_t^i \left(h_t^i\right)^\alpha \left(L_t^i\right)^\beta (H_t)^\gamma, \tag{2.19}$$

where $K = \kappa(1-\nu)^\delta(\nu\tau)^{1-\delta}$, $\alpha = \delta$, $\beta = (1-\delta)(\sigma-1)/\sigma$, and $\gamma = (1-\delta)/\sigma$. Note that this function exhibits constant returns to scale, that is, $\alpha + \beta + \gamma = 1$, and that the law of motion incorporates a local linkage L_t^i because education is funded by local funds, and a global linkage H_t because workers (the inputs) are complementary in production.

The relative merits of a local versus a state system of education can be studied in this framework by comparing the benefits of a system in which individuals are completely segregated into homogeneous jurisdictions such that $L_t^i = h_t^i$ with one in which all communities are integrated and hence $L_t^i = H_t$.

Intuitively, the trade-off between the two systems is clear. On one hand, decreasing returns in human capital accumulation ($\alpha < 1$) and complementarity and symmetry of inputs in production suggest that total output is maximized if individuals are homogeneous, pointing toward the benefits of a more homogenizing system such as a state-financed one. On the other hand, the fact that parental human capital and community resources are complements (i.e., the

[41] In Bénabou (1996b), individuals choose how much time to spend working relative to educating their children so as to maximize the discounted value of future generations' log of consumption (the dynastic utility function). Given the assumption of log preferences, all individuals choose the same ν. They also choose a constant value of τ. See the appendix in Bénabou (1996b) for details.

marginal return to an extra dollar spent on formal education is increasing in the level of parental human capital) suggests that, at a local level, assortative grouping of families is beneficial. The relative merits of the two systems, as we shall see, depend on the time horizon.

To analyze the pros and cons of the two systems, we need to derive the dynamic path of the economy under each education-finance policy. We do this under the assumption that the initial distribution of human capital at time t is lognormal; that is, $\ln h^i_t \sim N(m_t, \Delta^2_t)$. The cost of heterogeneity at both the local and global levels then can be seen in that $H = (E[(h)^{(\sigma-1)/\sigma}])^{\sigma/(\sigma-1)} = e^{-(\Delta^2/2)}E[h] < E(h)$ and $L^i = e^{-(\Delta^2/2)}E[h^i] < E(h^i)$.[42]

Noting that $\ln\ H_t = m_t + (\Delta_2/2\sigma)(\sigma - 1)$, we find that the law of motion implied by (2.19) under a local-finance regime, that is, $h^i_{t+1} = K\xi^i_t(h^i_t)^{\alpha+\beta}(H_t)^\gamma$, implies that the distribution in the following period will also be lognormal with

$$\begin{aligned} m_{t+1} &= \ln K - \frac{s^2}{2} + m_t + \gamma\frac{(\sigma-1)}{\sigma}\frac{\Delta^2_t}{2}, \\ \Delta^2_{t+1} &= (\alpha+\beta)^2\Delta^2_t + s^2. \end{aligned} \tag{2.20}$$

Similarly, under a state-finance regime, (2.19) implies $\hat{h}^i_{t+1} = K\xi^i_t(\hat{h}^i_t)^\alpha \times (\hat{L}_t)^{\beta+\gamma}$. Thus, if the initial distribution of human capital is described by $\ln \hat{h}^i_t \sim N(\hat{m}_t, \hat{\Delta}^2_t)$, then $\ln \hat{L}_t = \hat{m}_t + (\Delta^2/2\sigma)(\sigma-1)$ and next period's distribution of human capital is also lognormal with

$$\begin{aligned} \hat{m}_{t+1} &= \ln K - \frac{s^2}{2} + \hat{m}_t + (\gamma+\beta)\frac{(\sigma-1)}{\sigma}\frac{\Delta^2_t}{2}, \\ \hat{\Delta}^2_{t+1} &= \alpha^2\hat{\Delta}^2_t + s^2 \end{aligned} \tag{2.21}$$

(where the caret is used to denote the state-finance regime).

We examine the implications of both regimes on per capita human wealth $A_t \equiv \int_0^\infty h d\mu_t(h)$. Under a local-finance regime, $A_{t+1} = K\int_0^\infty h^{\alpha+\beta}d\mu_t(h)H^\gamma_t$, which, with the use of (2.20), implies

$$\ln\frac{A_{t+1}}{A_t} = \ln K - \left((\alpha+\beta)(1-\alpha+\beta) + \frac{\gamma}{\sigma}\right)\frac{\Delta^2_t}{2}. \tag{2.22}$$

The first term represents the growth rate of a standard representative agent economy. When agents are heterogeneous in terms of their human capital, however, $\alpha+\beta < 1$, $\gamma < 1$, and Jensen's inequality imply $\int_0^\infty h^{\alpha+\beta}d\mu_t(h) < A^{\alpha+\beta}_t$, and $H^\gamma_t < A^\gamma_t$. These differences are reflected in the last term of (2.22), which captures the decrease in growth that is due to heterogeneity as a product of the current variance times a constant term that measures the economy's efficiency loss per unit of dispersion, $\Pi = (\alpha+\beta)(1-\alpha-\beta) + (\gamma/\sigma)$. These losses reflect the concavity of the combined education production function $h^{\alpha+\beta}$

[42] Recall that if $y \sim N(m, \Delta^2)$ and $y = \ln x$, then $x \sim$ lognormal with $E(x) = e^{m+(\Delta^2/2)}$ and $\mathrm{Var}(x) = e^{2m+2\Delta^2} - e^{2m+\Delta^2}$. Furthermore, $[E(x^{(\sigma-1)/\sigma})]^{\sigma/(\sigma-1)} = e^{-(\Delta^2/2\sigma)}E(x)$.

and the complementarity $1/\sigma$ of inputs in production that has weight γ in the economywide aggregate H.[43]

For the state-finance system, similar derivations yield

$$\ln \frac{\hat{A}_{t+1}}{\hat{A}_t} = \ln\ K - \left(\alpha(1-\alpha) + \frac{(\beta+\gamma)}{\sigma}\right)\frac{\hat{\Delta}_t^2}{2}, \tag{2.23}$$

and thus $\hat{\Pi} = (\alpha(1-\alpha) + (\beta+\gamma)/\sigma)$. The interaction of heterogeneous agents at the local level imposes a loss of β/σ, and the concavity of the parental human capital contribution to production function (i.e., $\alpha < 1$) implies losses from heterogeneity along with the usual losses stemming as before from the complementarity in production in the economywide aggregate H.

The analysis given here implies that for *given* rates of resource and time investment in education, τ and ν, in the short run a state-finance education system will lead to lower human capital accumulation than a local system. To see this, note that

$$\phi \equiv \Pi - \hat{\Pi} = \beta\left(1 - 2\alpha - \beta - \frac{1}{\sigma}\right) = -\delta(1-\delta)\left(1 - \frac{1}{\sigma^2}\right) < 0,$$

implying that the drag on growth from heterogeneity is greater in a state-financed system. That is, two economies that start out with the same distribution of human capital in the first period will have a greater level of human capital in the second period under a local regime than under a state-finance regime.

In the long run, however, the conclusion is different. The handicap to growth from heterogeneity under a state regime tends to be reduced, as individuals have access to the same formal education system, whereas this source of heterogeneity is maintained under a local system in which education funding depends on family human capital. Solving for the long-run variances of the two systems given the same initial conditions yields $\Delta_t^2 = \Delta_\infty^2 + (\alpha+\beta)^{2t}(\Delta^2 - \Delta_\infty^2)$ and $\hat{\Delta}_t^2 = \hat{\Delta}_\infty^2 + \alpha^{2t}(\Delta^2 - \hat{\Delta}_\infty^2)$, where $\Delta_\infty^2 = s^2/[1-(\alpha+\beta)^2]$ and $\hat{\Delta}_\infty^2 = s^2/[1-\alpha^2]$.

Note that we can write $\ln A_t$ as $\ln\ A_0 + t \ln K - \Pi/2\{(\Delta^2 - \Delta_\infty^2)[1 + (\alpha+\beta)^{2t}]/[1-(\alpha+\beta)^2] + t\Delta_\infty^2\}$ and similarly $\ln\ \hat{A}_t = \ln A_0 + t\ln K - \hat{\Pi}/2[(\Delta^2 - \hat{\Delta}_\infty^2)(1+\alpha^{2t})/(1-\alpha^2) + t\hat{\Delta}_\infty^2]$. Hence, taking the limit of these expressions as $t \to \infty$, we obtain that in the case of no uncertainty in which initial endowments are the only source of inequality (i.e., $s^2 = 0$), in the long run the two economies grow at the same rate (namely ln K) and converge to a constant ratio of per capita human capital levels,

$$\ln \frac{\hat{A}_\infty}{A_\infty} = \left(\frac{\Pi}{1-(\alpha+\beta)^2} - \frac{\hat{\Pi}}{1-\alpha^2}\right)\frac{\Delta^2}{2} = \Phi\frac{\Delta^2}{2},$$

[43] Note that the same reasoning implies that heterogeneity in human capital is a source of gain when agents are substitutes in the production function or when the inputs of the community do not consist solely of education funds but also, say, peer effects that on aggregate imply increasing returns to scale in human capital at the local level.

where $\Phi \equiv \{\Pi/[1-(\alpha+\beta)^2] - \hat{\Pi}/(1-\alpha^2)\} = [\sigma/(2\sigma+\delta-1)][(1-\delta)/(1+\delta)][(\sigma-1)/\sigma]^2 > 0$.

If there is uncertainty in the generation of human capital, then for t sufficiently large,

$$\ln \frac{\hat{A}_t}{A_t} \approx \Phi \frac{s^2}{2} t,$$

and the growth rate of the state-finance education system exceeds that of the local regime by $\Phi(s^2/2)$. Hence state financing raises the long-run levels of human capital by $\Phi(\Delta^2/2)$, when there is no uncertainty and raises the long-run growth rate of human capital by $\Phi(s^2/2)$ when there is uncertainty. Thus, in the long run, a state system always does better. Whether a local or state education system is preferable will depend on how we discount different generations' welfare. For a sufficiently patient social planner, the state education system will be preferred.

3. SORTING INTO SCHOOLS

At some level it is possible simply to repeat much of the analysis of the preceding sections but refer to schools rather than neighborhoods. Obviously, little additional insight would be gained by doing this. A topic that did not have a natural place in the previous section is how the possibility of attending a private rather than a public school matters.

Introducing private schooling in a model that includes public schooling is generally problematic because in these models the funding of public schools is usually decided by majority vote at the local level, making it difficult to obtain existence of majority vote equilibrium.[44] The problem lies in the fact that those individuals who opt out of public schooling prefer (in the absence of externalities) to provide zero funding for private schools.

Epple and Romano (1998) provide a model that allows us to study some of the interactions between the private and public provision of education in an economy in which the demand for education depends both on ability and on income. They sidestep the problem of funding for education by assuming that the quality of a school depends only on the mean ability of its students. Although their theoretical results are somewhat incomplete given the difficulty of characterizing equilibria in an economy in which individuals differ in more than one dimension, their model nonetheless provides an extremely useful framework to begin thinking about sorting into schools.[45] The rest of this section is primarily dedicated to a discussion of their model.[46]

[44] See, though, Epple and Romano (1996), Fernández and Rogerson (1995), and Glomm and Ravikumar (1998) for some related attempts.

[45] Furthermore, for the interesting case of Cobb–Douglas preferences, their characterization holds, as will be discussed later.

[46] See also Caucutt (forthcoming) for a discussion of how different policies matter when students sort (in a complex fashion) across schools by ability and income.

Consider an economy in which students are assumed to differ in ability b and in income y. A school's quality is determined solely by the mean ability, q, of the student body. Students care about the quality of the school, as their utility depends on their achievement a, a function of their own ability b and school quality. They also care about private consumption, which will equal their income minus the price p they pay for schooling. Public schools are free and financed (so that costs are covered) by proportional income tax rates, t. Letting y_t denote after-tax income, we see that individuals maximize

$$V = V(y_t - p, a(q, b)). \tag{3.1}$$

The authors characterize the equilibrium distribution of student types (y, b) across public and private schools, assuming that types are verifiable. Preferences are assumed to be single crossing in income in the (q, p) plane; that is, (2.2) holds. That implies that, for the same ability level, students with higher income will be willing to pay a higher price to attend a school with higher mean ability. Preference for quality is also assumed to be nondecreasing in ability; that is, $[\partial(dp/dq|_{V=\bar{V}})]/\partial b \geq 0$.

All schools have the same cost function consisting of a fixed cost and an increasing, convex variable cost in the number N of students $c(N)$. Public schools all offer the same quality of schooling. The number of public schools simply minimizes the cost of operating the public sector, which is financed by a proportional income tax on all households. Private-sector schools, in contrast, maximize profits and there is free entry and exit.

Private schools maximize profits, taking as given the competitive utility $V^*(y, b)$ the student could obtain elsewhere. Schools can condition prices on ability and income. Thus, the profit maximization problem of a private school is to choose prices as a function of ability and income and the proportion of each type of student it wishes to admit (recognizing that there is a limit to the number of students of each type), taking into account the effects that these choices have on school quality and on cost by means of the types and number of students admitted.

The solution to private schools j's maximization problem is characterized by a first-order condition that, for an interior solution for that student type, equates the effective marginal cost of admitting the additional student i of type (b_i, y_i) to its reservation price. Note that when a school admits a student with ability b_i, its quality changes by $(b_i - q_j)/N_j$. The effective marginal cost of admitting this student is thus the increase in cost $c'(N)$ resulting from the fact that an additional student is being admitted minus the change in marginal revenue that is due to that student's effect on the school's quality.[47] The reservation price of a particular type of student is given by the maximum price p_i^* the school can charge (given its quality) so as to leave the individual at her or his market utility.

[47] Thus the effective marginal cost can be negative for a relatively high-ability student, leading to the possiblity of negative prices (e.g., fellowships) in equilibrium.

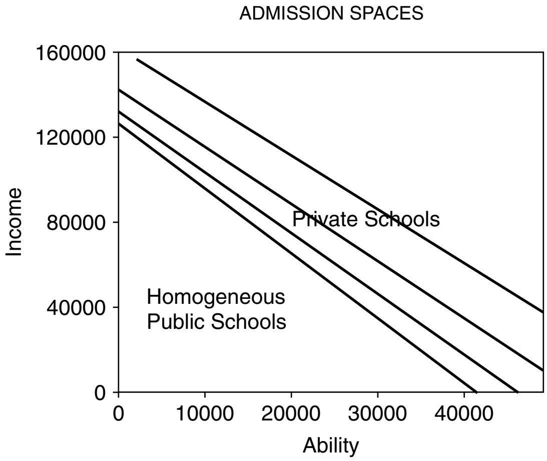


Figure 1.2. Sorting into public and private schools.

Note that this implies that some student types will not be admitted because their reservation price is too low to cover their effective marginal cost.

The equilibrium that emerges from this model has some nice properties.[48] As shown in Figure 1.2, there will be a strict hierarchy of school qualities $q_n > q_{n-1} > \cdots > q_0$, with the public sector (denoted by $j = 0$) having the lowest-ability peer group. Define the boundary loci between two schools as the set of types who are indifferent between the two schools (a curve with zero measure). Students who are on the boundary loci between two private schools will be charged their effective marginal costs; all other students will be charged strictly more than their effective marginal costs. This follows from the fact that students on the boundary are indifferent between attending either of the two schools competing for them, which drives down the price that each school can charge to that ability type's effective marginal cost. Furthermore, because a type's effective marginal cost is independent of income, their price will depend only on their ability. For students within the boundary loci, in contrast, the fact that they are not indifferent over which school they attend leaves the school with some monopoly power, which the school exploits by increasing the price. Hence, in general, the price charged to students within a school's boundary loci will depend on both ability and income. Note though that competition and free entry among schools implies that a school's profit is equal to zero.[49]

Lastly, it is also possible to characterize the type of students that will attend each school in equilibrium. The single-crossing condition in income ensures

[48] Epple, Newlon, and Romano (forthcoming) adapt this model to study ability tracking (or streaming) in public and private schools. Epple and Romano (1999) use a modified version of the model to study voucher design.

[49] As usual with models with fixed costs, free entry does not imply zero profits because of the integer problem. Ignore that qualification here.

that if an individual with income y_i attends a school with quality q_j, then all individuals with the same ability but greater income will attend schools of at least that level of quality and all individuals with lower income will attend schools with no greater quality.[50]

Thus, this model yields stratification by income. Stratification by ability need not follow, although the authors are able to find conditions (unfortunately on equilibrium variables) such that schools will also be stratified by ability.[51] Note that, as public schools have the lowest quality level, they will be composed of low-income individuals. If stratification by quality also holds, then public schools will consist of the lowest-income and lowest-ability students.

To understand the normative implications of the model, first suppose that no public option exists. Given the number of private schools, the allocation of types into schools is Pareto efficient. Private schools internalize the effect of peers via price discrimination. The equilibrium number of schools is not generally efficient, however, because the finite size of schools implies entry externalities. Furthermore, public-sector schooling in this model generally implies Pareto inefficiency, even given the equilibrium number of schools. Zero pricing by public schools independent of ability implies that the allocation of types among public and private-sector schools is inefficient.[52]

A very different issue in sorting into schools is studied by Fernández and Gali (1999). This paper is primarily interested in the properties of different assignment mechanisms under borrowing constraints. They examine a perfectly competitive model in which schools that vary in their (exogenous) quality each charge a market-clearing price to agents who vary in their ability and income. Schools have a fixed capacity and agents are assumed to be unable to borrow. In this model, the assumption that ability a and school quality q are complements in the production of output $x(a, q)$ implies that a social planner (or perfect capital markets) would assign the highest-ability student to the highest-quality school, the next-highest-ability student to the next-highest-quality school, and so forth. A perfectly competitive pricing mechanism does not produce this outcome. Instead, lower-ability but higher-income individuals are able to outbid higher-ability but lower-income agents for a place in a high-quality school.

This equilibrium outcome is contrasted with an exam mechanism, which assigns students to schools based on their performance on the exam. The exam score is assumed to be an increasing function of expenditures on education (e.g., better preparation or tutors) and innate ability. The exam technology is such

[50] This property of equilibrium does not follow immediately from single crossing, because schools can discriminate by types and thus a higher-quality school may charge an individual with higher income a higher price. This behavior, however, will not disrupt income stratification because effective marginal cost depends only on ability, and schools are sure to attract all types willing to pay more than effective marginal cost.

[51] In their working paper (1999), Epple and Romano show that for a Cobb–Douglas specification of utilty $U = (y_t - p)a(q, b)$, the equilibrium yields stratification by ability.

[52] See Epple and Romano (1998) for a fuller discussion.

that the marginal increment in expenditure required to increase a given score is decreasing in ability.

The authors find that an exam mechanism will always produce greater output. However, as expenditures under an exam system are wasteful, aggregate consumption need not be higher. The authors show, nonetheless, that for a sufficiently powerful exam technology (one that is sufficiently sensitive to ability relative to expenditures), the exam mechanism will always dominate the market mechanism for both aggregate production and consumption.

4. HOUSEHOLD SORTING

People sort not only into neighborhoods and schools; they also sort at the household level by deciding whom to marry or more generally whom to match with. Although there is a small body of literature that analyzes the economics of matching (e.g., Becker, 1973, and Burdett and Coles, 1997), there has been very little analysis, empirical or theoretical, of how this interacts with other general equilibrium variables such as growth and inequality.[53]

What are the consequences of household sorting for the transmission of education and inequality? Following Fernández and Rogerson (2001b), I set down a rudimentary model that allows us to examine this issue. This model will leave exogenous several important features of the decision problem (such as whom to match with and fertility), but it will simplify the analysis of key features of the transmission process.[54]

Consider an OLG model with two types of individuals – skilled (s) and unskilled (u) – in which the level of skill is also synonymous with the level of education (college and noncollege, respectively). These individuals meet, match, have children, and decide how much education to give each of their children.

Given a population at time t whose number is given by N_t and some division of that population into skilled workers, N_{st}, and unskilled workers, N_{ut}, where $N_t = N_{st} + N_{ut}$, let β denote the fraction of the population that is skilled; that is, $\beta_t = N_{st}/N_t$. Rather than endogenize matches, we assume an exogenous matching process in which a fraction θ of the population matches with probability one with someone of the same type, whereas the remainder match at random. As there are two types of individuals, this gives rise to three types of household matches indexed by j, which we shall denote by high (h) when it is between two skilled individuals, middle (m) when the match is between a skilled and an unskilled individual, and low (l) when it is between two unskilled individuals.

The matching process specified here yields $\lambda_{ht} \equiv \theta\beta_t + (1-\theta)\beta_t^2$ as the fraction of matches that are high, $\lambda_{mt} \equiv 2(1-\theta)\beta_t(1-\beta_t)$ as the fraction of

[53] Some exceptions are Cole, Mailath, and Postlewaite (1992) and Kremer (1997).

[54] See Fernández, Guner, and Knowles (2001) and Fernández and Pissarides (2000) for models that endogenize several of these features.

matches that are middle, and $\lambda_{lt} \equiv \theta(1-\beta_t) + (1-\theta)(1-\beta_t)^2$ as the fraction that are low. Of course, $\lambda_{ht} + \lambda_{mt} + \lambda_{lt} = 1$. Note that θ equals the correlation of partners' education levels.

Families have $n = \{0, 1, \ldots, \bar{n}\}$ children. We allow the probability ϕ_{nj} with which they have a particular number n to depend on the family type, so that average fertility f for a family of type j is given by $f_j = \sum_{n=0}^{\bar{n}} n\phi_{nj}$.

Children are either "college material" (whereupon if they went to college they would become skilled workers) or they are not (sending them to college would still produce unskilled workers). We denote these types as either high or low "aptitude" and allow the probability γ_j that a child is of high aptitude to depend on his or her parental type.[55] If a high-aptitude child is sent to college, he or she earns the skilled wage, w_s; otherwise, he or she earns the unskilled wage, w_u.

Lastly, we come to the education decision. We assume that the cost of college is given by $\nu > 0$. Capital and insurance markets are imperfect in that parents cannot borrow to finance the college education of their children but must finance it from their earnings. Insurance (as to which type of child a family might have) is also assumed not to be available. The assumption of not being able to borrow for a college education is not necessarily meant to be taken literally. Rather, we have in mind the local primary and secondary education system described earlier whereby education is financed to a large extent at the local level and minimum lot sizes (or higher borrowing costs), for example, constrain the quality of education that less wealthy parents are able to give to their children.

Parents choose per family member consumption level c and the number, r, of their high-aptitude children to educate so as to maximize the utility function:

$$U = \begin{cases} (c-\bar{c}), & \text{for } c < \bar{c} \\ (c-\bar{c}) + \frac{r}{(2+n)}w_s + \frac{(n-r)}{(2+n)}w_u, & \text{otherwise} \end{cases}. \tag{4.1}$$

This implies that, subject to a minimum per family member consumption level of $\bar{c}$, parents will send a high-ability child to college if they can afford to (and it is economically advantageous to do so). The family budget constraint is given by $(2+n)c + r\nu \leq I_j(\beta)$, $0 \leq r \leq a$, where a is the total number of high-aptitude children the family has, and

$$I_j(\beta) = \begin{cases} 2w_s(\beta) & \text{for } j = h \\ w_s(\beta) + w_u(\beta) & \text{for } j = m \\ 2w_u(\beta) & \text{for } j = l \end{cases}. \tag{4.2}$$

Lastly, wages are determined in a competitive market as the appropriate marginal revenue products of a constant returns to scale aggregate production

[55] One should consider aptitude to reflect family background in the sense of making it more probable that a child will obtain a college education. It should be noted that this is not really standing in for a genetically determined process, because in that case we would have to keep track of whether a particular match consisted of zero, one, or two high-aptitude individuals.

function, given by

$$F(N_s, N_u) = N_u F(N_s/N_u, 1) \equiv N_u F\left(\frac{\beta}{1-\beta}, 1\right) \equiv N_u f(\beta),$$
$$f' > 0,\ f'' < 0. \qquad (4.3)$$

Hence, wages are solely a function of β and given by $w_s(\beta) = (1-\beta)^2 f'(\beta)$ and $w_u(\beta) = f(\beta) - \beta(1-\beta)f'(\beta)$. Note that (4.3) implies that skilled wages are decreasing in the ratio of skilled to unskilled workers, whereas unskilled wages are increasing. Also note that no parents would want to send their child to college if the fraction of skilled workers exceeds $\bar{\beta}$, where $\bar{\beta}$ is defined by $w_s(\bar{\beta}) = w_u(\bar{\beta}) + \nu$.

To solve for the steady states, we need one additional piece of information, $\Gamma_j(z_j(\beta))$, the average proportion of children sent to college by families of type j. This will depend on how constrained each family is (this may differ according to family size and how many high-aptitude children the family has), z_{nj}, which in turn depends on wages and hence on β. Hence,

$$\Gamma_j(z_j(\beta)) \equiv \frac{1}{f_j}\sum_{n=1}^{\bar{n}} \phi_{nj}\left[\sum_{a=0}^{z_{nj}} \binom{n}{a} \gamma_j^a (1-\gamma_j)^{n-a} a + \sum_{a=z_{nj}+1}^{n} \binom{n}{a} \gamma_j^a (1-\gamma_j)^{n-a z_{nj}}\right], \qquad (4.4)$$

where the first summation term within the square brackets is the number of children that attend college from families of type j with n children that are not constrained (as the number of high-aptitude kids they have is fewer than z_{nj}), and the second summation is over the number of children that attend college from constrained families of type j with n children.[56]

The steady states of the economy are the fixed points of the dynamic system,

$$\beta_{t+1}(\theta) = \frac{N_{st}+1}{N_t+1} = \frac{\sum_j \Gamma_j(z_j(\beta_t)) f_j \lambda_{jt}(\beta_t;\ \theta)}{\sum_j f_j \lambda_{jt}(\beta_t;\ \theta)}, \qquad (4.5)$$

that is, a level of β, such that $\hat{\beta} = \beta_t = \beta_{t+1}$. We restrict our attention to those that are locally stable, that is, $\partial\beta_{t+1}/\partial\beta_t|_{\beta_t=\hat{\beta}} < 1$.

Note that in general there may be multiple steady states. To see why this is so, consider what may happen if we start out with a low level of β. In this case, low-type families (and perhaps middle types as well) will be relatively constrained because unskilled wages are low. Thus, this will tend to perpetuate a situation in which β is low next period as well, and thus a steady state with a low proportion of unskilled wages (and high inequality). If, in contrast, the economy started out with a high level of β, unskilled wages would be high

[56] If a family of type j with n children is not constrained, we simply indicate this by $z_{nj} = n$.

and hence low-type families would be relatively unconstrained, perpetuating a situation of high β (and low inequality).

How does the degree of sorting affect this economy? If the change in sorting is sufficiently small that the degree to which constraints are binding is unaffected (i.e., the Γ_js are constant), then

$$\begin{aligned}\frac{d\hat{\beta}}{d\theta} &= \frac{\hat{\beta}(1-\hat{\beta})[f_h(\Gamma_h-\hat{\beta}) - 2f_m(\Gamma_m-\hat{\beta}) + f_l(\Gamma_l-\hat{\beta})]}{D} \\ &= \frac{\hat{\beta}(1-\hat{\beta})[(f_h\Gamma_h - 2f_m\Gamma_m + f_l\Gamma_l) - \hat{\beta}(f_h - 2f_m + f_l)]}{D},\end{aligned} \tag{4.6}$$

where $D = \sum_j f_j\lambda_j(\hat{\beta};\theta) + \sum_j f_j[\partial\lambda_j(\hat{\beta};\theta)]/(\partial\beta)(\hat{\beta}-\Gamma_j)$. It is easy to show that local stability requires $D > 0$.

The expression in (4.6) is easy to sign for a few cases. Suppose all the Γ_js are the same, that is, $\Gamma_j = \Gamma$. In that case, $\hat{\beta} = \Gamma$ and the extent of sorting does not affect the personal income distribution (though it does the household income distribution), as wages are unchanged.[57]

Suppose next that average fertility is the same across all groups; that is, $f_j = f$. In this case, the sign of (4.6) is given by the sign of $\Gamma_h + \Gamma_l - 2\Gamma_m$. The intuition behind this is simple. Note that the effect of an increase in sorting is to destroy middle-type matches and replace these by high and low ones. In particular, for every two middle matches destroyed, one high and one low match are created. Because average fertility is the same across family types, the effect of increased sorting depends on whether the fraction of children sent to college on average by two middle-type marriages ($2\Gamma_m$) is smaller than the combined fraction of children that go to college on average in one high-type and one low-type family ($\Gamma_h + \Gamma_l$). Thus, if the relationship between parents' education and children's education is linear, changes in sorting will have no effect on $\hat{\beta}$; if concave, increased sorting will decrease $\hat{\beta}$; and the reverse if the relationship is convex.

Lastly, making no assumptions about fertility or the Γ_js, we see that a sufficient condition for an increase in sorting to decrease $\hat{\beta}$ is $f_h\Gamma_h - 2f_m\Gamma_m + f_l\Gamma_l \leq 0$ and $f_h + f_l - 2f_m \geq 0$ (with at least one inequality strict). The first expression is the counterpart of the expression in the preceding paragraph. That is, subject to no change in population growth, it ensures that there will be fewer skilled individuals in the following period. The second expression ensures that the population growth rate will not decline as a result of the increased sorting (thereby potentially giving rise to a larger *proportion* of skilled people despite the fall in their growth rate).[58]

[57] Recall that we are assuming that constraints are unaffected by the change in sorting.

[58] The opposite signs on the two expressions create a sufficient condition for increased sorting to increase β.

Table 1.2. *Aptitude profiles under various scenarios $z_{nm} = n$ and $z_{nh} = n$*

	$z_{nl} = n$	$z_{nl} = 2$	$z_{2l} = 2$, $z_{3l} = 1$	$z_{nl} = 1$
γ_h	0.81	0.81	0.81	0.81
γ_m	0.63	0.63	0.63	0.63
γ_l	0.30	0.303	0.334	0.401

This discussion assumed that the Γ_js remained invariant to the change in sorting. Note, however, that these may well change as constraints become more or less binding as a result of the change in wages.[59] Hence, even if fertility is exogenous, the sign of $f_h\Gamma_h - 2f_m\Gamma_m + f_l\Gamma_l$ is generally endogenous because the Γ_js are endogenous variables.[60] Thus, whether the expression is concave or convex may itself depend on β.

Fernández and Rogerson (2001b) explore the effect of increased sorting on inequality by calibrating the model given here to U.S. data. They use the Panel Study of Income Dynamics (PSID) to obtain a sample of parents and children and group all individuals with high school and below into the unskilled category and everyone who has had at least some college into the skilled one. The correlation of parental education (θ) equals 0.6. Average fertility is given by $f_h = 1.84$, $f_m = 1.90$, and $f_l = 2.26$ (from PSID and from Mare, 1997). For any average fertility number, the two integers that bracket the average are chosen as the only two possible numbers of children to have, with the appropriate weights used as the probabilities (e.g., $\phi_{1h} = 0.16$, and $\phi_{2h} = 0.84$).

To calibrate the model, we need to know the γ_js. These are not available in the data, but what are computable from the PSID are the Γ_js (i.e., the fraction of children of each family type that on average attend college). These are given by $\Gamma_h = 0.81$, $\Gamma_m = 0.63$, and $\Gamma_l = 0.30$. Note from (4.4) that any value of Γ_j can be decomposed into an assumption about how "inheritable" education is (the γ_js) and a corresponding assumption about how binding borrowing constraints are (the z_{nj}s). Table 1.2 shows various such decompositions for Γ_l (for the other Γ_js it is assumed that the constraints are not binding and hence $\Gamma_j = \gamma_j$).

Fernández and Rogerson (2001b) use the second column as their benchmark. Note that this implies the existence of very mild constraints. Only low-type families with three high-ability children are affected, and these are fewer than 1 percent of low-type families.

This information, along with the Γ_js, allows us to compute the steady state, yielding $\hat{\beta} = 0.60$. To obtain wages, we use a CES production function

[59] In a more general model in which household incomes were continuous, a change in θ that affected β for a constant set of Γs would also necessarily affect the Γ_js.

[60] In a more complex model in which fertility and/or matching are endogenized, one can perform a similar exercise by changing technology such that the skill premium for any β is higher or by changing the cost of search.

Table 1.3. *Effects of increased sorting on steady state*

	$\theta = 0.6$	$\theta = 0.7$	
	$\Gamma_l = 0.30$	$\Gamma_l = 0.30$	$\Gamma_l = 0.27$
Mean(e)	13.52	13.48	13.40
cv(e)	0.134	0.135	0.137
$\hat{\beta}$	0.600	0.589	0.568
w_s/w_u	1.900	1.95	2.07
std(log y)	0.315	0.330	0.361

$y = A[bN_s^\rho + (1-b)N_u^\rho]^{1/\rho}$ and match the steady-state ratio of skilled to unskilled wages to 1.9 (Katz and Murphy, 1992) and obtain $\rho = 0.33$ by matching an elasticity of substitution between skilled and unskilled workers of 1.5 (see the survey by Katz and Autor, 1999). Lastly, for ease of interpretation of our results, we choose a value of A to scale steady-state unskilled wages to some "reasonable" value, which we set to be 30,000. This is purely a normalization.

It is important to note that the steady state of the calibrated model fulfills the sufficient conditions such that an increased θ leads to a lower proportion of skilled individuals. Hence, from a theoretical perspective, we know that an increase in sorting will lead to higher skilled wages and lower unskilled ones. The quantitative impact is given in Table 1.3. The first row reports mean years of education (in which the skilled group and the unskilled group have been assigned the mean from their PSID sample). The second row gives the coefficient of variation of education. The last entry is the standard deviation in log income – our measure of inequality in the personal income distribution.

The first numerical column of the table reports the result of the calibration.[61] The second column reports the effect of an increase in sorting to 0.7, assuming that the values of Γ are unchanged. The third column does the same but assumes that the decrease in the unskilled wage means constraints are tightened for low-type families and that those with three children can afford to send only a maximum of one of them to college.

The main message of Table 1.3 is that changes in sorting can have large effects on inequality and that seemingly small changes in average years of education or in its coefficient of variation can underlie large changes in income distribution. As shown in the table, a change in sorting from 0.6 to 0.7 will increase the standard deviation of log income by a bit under 5 percent in the absence of any assumption about borrowing constraints.[62] If as a result of the

[61] Note that the standard deviation of log income is about half of what it is in reality for the United States. It is not surprising that our model is not able to produce as much variation as in the data, as there are only two wages.

[62] Note that these results, therefore, are independent of which column we choose from Table 1.2 as our benchmark.

approximately \$600 drop in w_u that occurs (and consequently a \$1,200 drop in low-type family income), constraints tighten, this leads to an increase in inequality of almost 15 percent. In both cases, the effect on the standard deviation of log family income is large: 8.3 percent and 19 percent, respectively.[63]

This analysis also points out the dangers with assuming that intergenerational processes are linear. Kremer (1997), for example, assumes that years of education a child acquires is a linear function of average parental years of education, as given by

$$e_{i,t+1} = \kappa + \alpha \frac{(e_{i,t} + e_{\acute{\imath},t})}{2} + \xi_i, \tag{4.7}$$

where $e_{i,t+1}$ is the education level for the child, $e_{i,t}$ and $e_{i',t}$ are the education levels of the two parents, and ξ is a normally distributed random shock that is iid across families, with mean 0 and standard deviation equal to σ_ξ. Parents are all assumed to have two kids, and an (exogenous) assortative matching of individuals takes place yielding θ as the correlation between the education levels of parents.

Note that, within the framework of Fernández and Rogerson (2001b), the assumptions of a linear transmission process and the same fertility across all parent types would yield no effect of an increase in sorting on inequality. In Kremer's model, this is not the case, because, although the mean of the distribution is unaffected, the inclusion of a shock implies that greater sorting will increase inequality. To see this, note that with constant parameter values the distribution of education converges to a normal distribution with steady-state mean and standard deviation given by $\mu_\infty = \kappa/1 - \alpha$, and

$$\sigma_\infty = \frac{\sigma_\xi}{[1 - \alpha^2(1+\theta)/2]^{0.5}}, \tag{4.8}$$

respectively. Thus an increase in θ, although not affecting the mean, increases the variance of the distribution of education.

To investigate the effects of sorting within this model, Kremer uses PSID data to run the regression suggested by (4.7), and finds α equals 0.4. Parents' correlation in years of education, as we saw previously, is 0.6. This implies, using (4.8), that even a large increase in the correlation of parental education, say from 0.6 to 0.8, will increase the standard deviation of the distribution of education by only about 1 percent. Furthermore, if we assume, as Kremer does, that log earnings are linear in years of education (i.e., $y_{i,t+1} = a + be_{i,t+1}$), then exactly the same conclusion applies to the distribution of earnings.

The very different conclusions obtained by Kremer relative to Fernández and Rogerson emphasize the importance of certain features of the data (i.e., fertility differentials and nonconvexities in the transmission process) as well as

[63] The results for the θ increase to 0.8 follow a pattern similar to the one given here. The changes in the mean and standard deviation of the education distribution are small, as before, but the changes in income distribution are large.

the endogeneity of wages. Furthermore, as shown in Fernández and Rogerson, borrowing constraints can greatly multiply the effect of increased sorting.

In light of this, it is of interest to ask how inequality, fertility, and sorting are related in a model in which these variables are endogenous. Fernández et al. (2001) developed a simple two-period search model in which individuals are given multiple opportunities to match with others. As before, there are two types of individuals (skilled and unskilled), distinguished only by their educational attainment. In the first period, we assume that agents meet others from the population in general. In the second period, agents meet only others who are similar to themselves in terms of skill level.[64] Agents' characteristics (income) are fully observable, as is the quality of the match. The latter is assumed to be a random draw from a quality distribution, and is fully match specific. If agents decide to keep their first-period match, they are unable to search in the second period.

Having matched, individuals decide how many children to have (at a cost per child t that is proportional to income I) and devote the rest of their income to consumption. Thus individuals maximize

$$\max_{c,n}[c + \gamma \log(n) + K + q], \tag{4.9}$$

subject to

$$c \leq I(1 - tn), \quad t > 0,$$

where n is the number of children, $I > \gamma$ is household income, q is the quality of the match, and K is a constant. Plugging in the optimal decisions for an individual (and choosing K such that the sum of the constants is zero) allows us to express the indirect utility function as $V(I, q) = I - \gamma \log I + q$.

Assuming a constant returns production function allows us, as before, to express wages solely as a function of the ratio of skilled to unskilled workers and to express household income as in (4.2). The cutoff match quality that a high-wage worker will accept in order to match with a low-wage individual in the first period is an increasing function of w_s and a decreasing function of w_u.[65]

Children face two costs to becoming a skilled worker. First, there is a constant monetary cost of d. Second, there is an individual-specific (additive) psychic cost (e.g., effort) of δ with a cumulative distribution $\Psi(\delta)$. The return to being a skilled worker is the probability of matching with a skilled worker and obtaining household income I_{ss} (in which wages are assumed to be net borrowing and repaying d) plus the probability of matching with an unskilled worker and obtaining household income I_{su}. These probabilities depend on the probability that in the first period a particular type of worker is met and

[64] One could just as easily simply assume that the first period one meets a more representative sample of the population relative to the second period in which it is biased toward individuals who are similar.

[65] The skilled worker will always be the one whose cutoff quality level is binding, as his or her income is greater.

on the cutoff quality of the match a skilled worker will accept (and hence on the fraction of individuals that are skilled in the population, i.e., β). A similar calculation holds for the return to being an unskilled worker.[66]

If there were no borrowing constraints, then all families would have the same fraction of children become skilled so that the net return to being a skilled worker equalled the return to being an unskilled worker plus $\delta^*(\beta)$ – the equilibrium psychic cost such that no worker with $\delta_i > \delta^*(\beta)$ is willing to become skilled. If, however, there are borrowing constraints such that the amount that an individual can borrow depends (positively) on family income, then families with higher household incomes will have a higher fraction of their children become skilled.

How does inequality matter? It is easy to show that, as family income increases, fertility declines. Thus fertility differentials are increasing with inequality. Furthermore, as wage inequality increases, skilled workers become pickier about the quality of the match required to make them willing to match with an unskilled worker.

As before, this model will generally have multiple steady states. If the economy starts out with a low proportion of skilled workers, the skill premium will be high, skilled workers will be very picky about matching with unskilled workers, and hence there will be a high level of sorting. Given borrowing constraints, only a small fraction of children from low-income households will become skilled implying that in the next period a similar situation will tend to perpetuate itself – a high level of inequality, high sorting, and high fertility differentials. The opposite would be true if instead the economy starts out with a high level of skilled workers. In this case inequality is low, high-skilled agents choose a low cutoff quality for matching with unskilled agents so sorting is low, fertility differentials are low, and borrowing constraints are not very binding. This leads again to a high proportion of skilled workers the following period.

We take the implications of this model to the data. Using a sample of thirty-three countries, we examine the relationship between sorting and inequality and find that, as the theory predicts, these are positively correlated. Countries with greater inequality exhibit greater sorting at the household level. Furthermore, as also predicted by the theory, fertility differentials are increasing in inequality and per capita GDP and sorting are negatively correlated.[67]

5. CONCLUDING REMARKS

This chapter has reviewed some of the principal contributions to the literature that examine the links between sorting, education, and inequality. Much work remains to be done in all of the areas discussed in this chapter: education-finance

[66] Note that, unlike Fernández and Rogerson (2001b), the return to being skilled or unskilled depends also on how this decision affects the type of match one will obtain.

[67] Kremer and Chen (1999) examine the relationship between fertility and inequality for a large sample of countries and find that fertility differentials and inequality are positively correlated.

systems and residential sorting, schools, and household sorting. In particular, it would be of interest to see more work that examined how different education systems matter, and that provided an empirical basis on which to assess different policy proposals. At the school level, very little is known about how parents, teachers, students, administrators, and the community interact in producing schooling of a particular quality. I think that the largest challenge here is the creation of a convincing multiple principal-agent model that endogenizes the quality of the school in response to information constraints, the availability of alternative options, and the system in which it is embedded. In addition, it would be of interest to study the incentive effects of external standards (e.g., national- or state-level exams) that allow schools to be "graded" against one another. Finally, work on household sorting is still at an embryonic level, both theoretically and empirically.[68] A notable omission from the models discussed herein is the role of gender: They do not distinguish between the education and income distributions of men and women. It would be of interest to examine how these matter and to investigate, empirically and theoretically, the role of women's large increase in labor force participation and educational attainment.

ACKNOWLEDGMENTS

This paper was prepared for the 2000 World Congress of the Econometric Society. I thank the National Science Foundation and the CV Starr Center for financial support. I also thank Costas Meghir, Jonathan Portes, Richard Rogerson, and Rich Romano for helpful comments. Author contact: Department of Economics, NYU, 269 Mercer St., New York, NY 10003. E-mail: Raquel.Fernandez@nyu.edu.

References

Becker, G. (1973), "A Theory of Marriage," *Journal of Political Economy*, LXXXI, 813–846.

Bénabou, R. (1993), "Workings of a City: Location, Education, and Production," *Quarterly Journal of Economics*, 108(3), 619–652.

Bénabou, R. (1996a), "Equity and Efficiency in Human Capital Investment: The Local Connection," *Review of Economic Studies*, 62, 237–264.

Bénabou, R. (1996b), "Heterogeneity, Stratification, and Growth," *American Economic Review*, LXXXVI, 584–609.

Benhabib, J. and M. Spiegel (1994), "The Role of Human Capital and Political Instability in Economic Development," *Journal of Monetary Economics*, 34, 143–173.

Bergstrom, T., D. Rubinfeld, and P. Shapiro (1982), "Micro-Based Estimates of Demand Functions for Local School Expenditures," *Econometrica*, 50, 1183–1205.

[68] See Greenwood, Guner, and Knowles (1999) for recent work in this field.

Burdett, K. and M. Coles (1997), "Marriage and Class," *Quarterly Journal of Economics*, CXII, 141–168.

Card, D. and A. Krueger (1992), "Does School Quality Matter? Returns to Education and the Characteristics of Public Schools in the U.S.," *Journal of Political Economy*, 100, 1–40.

Caucutt, E. (forthcoming), "Educational Vouchers when There Are Peer Group Effects," *International Economic Review*.

Cole, H., G. Mailath, and A. Postlewaite (1992), "Social Norms, Savings Behavior, and Growth," *Journal of Political Economy*, C, 1092–1125.

Coleman, J., et al. (1966), *Equality of Educational Opportunity*, Washington, DC: U.S. Government Printing Office.

Cooper, S. (1998), "A Positive Theory of Income Redistribution," *Journal of Economic Growth*, 3, 171–195.

de Bartolome, C. (1990), "Equilibrium and Inefficiency in a Community Model with Peer Group Effects," *Journal of Political Economy*, 98(1), 110–133.

de Bartolome, C. (1997), "What Determines State Aid to School Districts? A Positive Model of Foundation Aid as Redistribution," *Journal of Policy Analysis and Management*, 16, 32–47.

Digest of Education Statistics (1999), available at http://nces.ed.gov.

Downes, T. and D. Schoeman (1998), "School Finance Reform and Private School Enrollment: Evidence from California," *Journal of Urban Economics*, 43(3), 418–443.

Durlauf, S. (1996), "A Theory of Persistent Income Inequality," *Journal of Economic Growth*, 1, 75–93.

Epple, D., E. Newlon, and R. Romano (forthcoming), "Ability Tracking, School Competition, and the Distribution of Educational Benefits," *Journal of Public Economics*.

Epple, D. and R. Romano (1996), "Ends Against the Middle: Determining Public Service Provision when There Are Private Alternatives," *Journal of Public Economics*, 62, 297–325.

Epple, D. and R. Romano (1998), "Competition Between Private and Public Schools, Vouchers, and Peer-Group Effects," *American Economic Review*, 88(1), 33–62.

Epple, D. and R. Romano (1999), "Educational Vouchers and Cream Skimming," August, Working Paper, University of Florida.

Epple, D. and T. Romer (1991), "Mobility and Redistribution," *Journal of Political Economy*, 99, 828–858.

Evans, W., S. Murray, and R. Schwab (1997), "Schoolhouses, Courthouses and Statehouses after Serrano," *Journal of Policy Analysis and Management*, 16(1), 10–37.

Evans, W., S. Murray, and R. Schwab (1998), "Education-Finance Reform and the Distribution of Education Resources," *American Economic Review*, 88(4), 789–812.

Fernández, R. (1997), "Odd Versus Even: Comparative Statics in Multicommunity Models," *Journal of Public Economics*, 65, 177–192.

Fernández, R. and J. Gali (1999), "To Each According to . . . ? Tournaments, Markets and the Matching Problem under Borrowing Constraints," *Review of Economic Studies*, 66(4), 799–824.

Fernández, R., N. Guner, and J. Knowles (2001), "Matching, Fertility, and Inequality: A Cross-Country Comparison," NBER Working Paper 8580.

Fernández, R. and C. Pissarides (2000), "Matching and Inequality," work in progress.

Fernández, R. and R. Rogerson (1995), "On the Political Economy of Education Subsidies," *Review of Economic Studies*, LXII, 249–262.

Fernández, R. and R. Rogerson (1996), "Income Distribution, Communities, and the Quality of Public Education," *Quarterly Journal of Economics*, CXI, 135–164.

Fernández, R. and R. Rogerson (1997a), "Education Finance Reform: A Dynamic Perspective," *Journal of Policy Analysis and Management*, 16, 67–84.

Fernández, R. and R. Rogerson (1997b), "Keeping People Out: Income Distribution, Zoning and the Quality of Public Education," *International Economic Review*, 38, 23–42.

Fernández, R. and R. Rogerson (1998), "Income Distribution and Public Education: A Dynamic Quantitative Analysis of School Finance Reform," *American Economic Review*, 88, 813–833.

Fernández, R. and R. Rogerson (1999a), "Education Finance Reform and Investment in Human Capital: Lessons From California," *Journal of Public Economics*, 74, 327–350.

Fernández, R. and R. Rogerson (1999b), "Equity and Resources: An analysis of Education Finance Systems," NBER Working Paper 7111.

Fernández, R. and R. Rogerson (2001a), "The Determinants of Public Education Expenditures: Longer-Run Evidence from the States, 1950–1990," *Journal of Education Finance*, Summer, 27, 567–583.

Fernández, R. and R. Rogerson (2001b), "Sorting and Long-Run Inequality," *Quarterly Journal of Economics*. 116(4), 1305–1341.

Fernández, R. and R. Rogerson (2002), "School Voucher as a Redistributive Device: An Analysis of Three Alternative Systems" in *The Economic Analysis of School Choice*, (ed. by C. Hoxby), Chicago: University of Chicago Press.

Galor, O. and J. Zeira (1993), "Income Distribution and Macroeconomics," *Review of Economic Studies*, 60, 35–52.

Gans, J. and M. Smart (1996), "Majority Voting with Single-Crossing Preferences," *Journal of Public Economics*, 59, 219–237.

Glomm, G. and B. Ravikumar (1992), "Public Versus Private Investment in Human Capital: Endogenous Growth and Income Inequality," *Journal of Political Economy*, 100, 818–834.

Glomm, G. and B. Ravikumar (1998), "Opting Out of Publicly Provided Services: A Majority Voting Result," *Social Choice and Welfare*, 15(2), 187–199.

Greenwood, J., N. Guner, and J. Knowles (1999), "More on Marriage, Fertility, and the Distribution of Income," mimeo, University of Rochester.

Hanushek, E. (1986), "The Economics of Schooling: Production and Efficiency in Public Schools," *Journal of Economic Literature*, 24(3), 1147–1217.

Heckman, J., A. Layne-Farrar, and P. Todd (1996), "Does Measured School Quality Really Matter? An Examination of the Earnings-Quality Relationship," in *Does Money Matter? The Effect of School Resources on Student Achievement and Adult Success* (ed. by G. Burtless), Washington, DC: Brookings Institution Press, 192–289.

Henderson, J., P. Mieszkowski, and Y. Sauvageau (1978), "Peer Group Effects and Educational Production Functions," *Journal of Public Economics*, 10, 97–106.

Hoxby, C. (1998), "All School Finance Equalizations Are Not Created Equal," NBER Working Paper 6792.

Inman, R. P. (1978), "Optimal Fiscal Reform of Metropolitan Schools: Some Simulation Results," *American Economic Review*, 68(1), 107–122.

Jargowsky, P. (1996), "Take the Money and Run: Economic Segregation in U.S. Metropolitan Areas," *American Sociological Review*, LXI, 984–998.

Johnson, G and F. Stafford, (1973) “Social Returns to Quantity and Quality of Schooling,” *Journal of Human Resources*, Spring, 8(2), 139–55.

Katz, L. and D. Autor (1999), “Changes in the Wage Structure and Earnings Inequality,” in *Handbook in Labor Economics*, Vol. 3A, (ed. by O. Ashenfelter and D. Card), Amsterdam: North-Holland, 3309–3415.

Katz, L. and K. Murphy (1992), “Changes in Relative Wages, 1963–87: Supply and Demand Factors,” *Quarterly Journal of Economics*, CVII, 35–78.

Kremer, M. (1997), “How Much Does Sorting Increase Inequality,” *Quarterly Journal of Economics*, CXII, 115–139.

Kremer, M. and D. Chen (1999), “Income Distribution Dynamics with Endogenous Fertility,” *American Economic Review Papers and Proceedings*, May 89, 155–160.

Kremer, M. and E. Maskin (1996), “Wage Inequality and Segregation by Skill,” mimeo, Massachusetts Institute of Technology.

Ljungqvist, L. (1993), “Economic Underdevelopment: The Case of a Missing Market for Human Capital,” *Journal of Development Economics*, 40, 219–239.

Loeb, S. (1998), “Estimating the Effects of School Finance Reform: A Framework for a Federalist System,” mimeo, Stanford University.

Loury, G. (1981), “Intergenerational Transfers and the Distribution of Earnings,” *Econometrica*, 49, 843–867.

Manwaring, R. and S. Sheffrin (1997), “Litigation, School Finance Reform, and Aggregate Educational Spending,” *International Tax and Public Finance*, 4(2), 107–127.

Mare, R. (1991), “Five Decades of Educational Assortative Mating,” *American Sociological Review*, LVI, 15–32.

Mare, R. (1997), “Differential Fertility, Intergenerational Educational Mobility, and Racial Inequality,” *Social Science Research*, XXVI, 263–291.

Nechyba, T. (1996), “A Computable General Equilibrium Model of Intergovernmental Aid,” *Journal of Public Economics*, 62, 363–397.

Nechyba, T. (1997), “Existence of Equilibrium and Stratification in Local and Hierarchical Tiebout Economies with Local Property Taxes and Voting,” *Economic Theory*, 10, 277–304.

Rose-Ackerman, S. (1979), “Market Models of Local Government: Exit, Voting, and the Land Market,” *Journal of Urban Economics*, 6, 319–337.

Saint-Paul, G. and T. Verdier (1993), “Education, Democracy and Growth,” *Journal of Development Economics*, 42(2), 399–407.

Silva, F. and J. Sonstelie (1995), “Did Serrano Cause a Decline in School Spending?” *National Tax Journal*, 48, 199–215.

Stiglitz, J. (1974), “The Demand for Education in Public and Private School System,” *Journal of Public Economics*, 3, 349–385.

Summers, A. and B. Wolfe (1977), “Do Schools Make a Difference?” *American Economic Review*, 67, 639–652.

Tiebout, C. (1956), “A Pure Theory of Local Expenditures,” *Journal of Political Economy*, 65, 416–424.

Wachtel, P. (1976), “The Effects on Earnings of School and College Investment Expenditures,” *Review of Economics and Statistics*, 58, 326–331.

Westhoff, F. (1977), “Existence of Equilibria in Economies with a Local Public Good,” *Journal of Economic Theory*, 14, 84–112.

Wheaton, W. (1977), “A Bid-Rent Approach to Housing Demand,” *Journal of Urban-Economics*, 4(2), 200–217.

CHAPTER 2

Wage Equations and Education Policy

Kenneth I. Wolpin

1. INTRODUCTION

One of the most widely estimated regression equations in economics is that of the earnings or wage function, which relates a measure of market remuneration (e.g., hourly wage rates or annual earnings) to measures of human capital stocks such as completed schooling and market work experience.[1] In the first half of this paper, I address two related questions: (1) What interpretation should be given to the wage equation? and (2) Why should we care about estimating it?

Twenty-five years ago, Zvi Griliches, in his 1975 Presidential Address to the World Congress of the Econometric Society, raised the first question of interpretation, together with several others that were concerned with the appropriate specification and estimation of the wage function.[2] In that address, Griliches chose to adopt the competitive market–human capital production interpretation of Ben-Porath (1967), in which the wage an individual is offered is the product of the competitively determined skill rental price and the amount of human capital or skill units cumulatively produced by and thus embodied in the individual.[3] About a decade later, Heckman and Sedlacek (1985) and Willis (1986) extended the competitive market–human capital production interpretation of the wage function to a multidimensional skill setting, formally incorporating the self-selection model of Roy (1951) to the choice of employment sector in the former case and to the choice of occupation cum schooling in the latter.[4] More recently, Keane and Wolpin (1997) extended and generalized

[1] I will generally refer to "wages" as opposed to "earnings" because the former has the connotation of a market price for labor.

[2] See Griliches (1977).

[3] In the Ben-Porath model of human capital accumulation, individuals optimally choose a fraction of their skill units to produce additional skill, forgoing wage compensation for those units used in on-the-job investment. This contrasts with the notion that skills are acquired through a process of learning by doing, in which skills are produced jointly with output. Learning by doing is not necessarily a costless activity if there are alternative uses for the time spent on the job, for example, in the household sector or in alternative jobs.

[4] Heckman and Sedlacek did not model schooling choice. Willis did not model postschooling human capital investment. However, the work of Willis and Rosen (1979), which was the precursor, did allow for exogenous wage growth with age.

the Heckman and Sedlacek and Willis papers to a dynamic setting of schooling, employment, and occupational choice, adopting as well the competitive market–human capital production paradigm.

In contrast to the development and estimation of models that embed wage functions within a formal choice structure, as exemplified by these papers, over the past forty years or so there has been a continuing effort to pin down estimates of wage function parameters as an end in itself. Given this sustained effort, the existence of a substantial payoff to identifying the parameters of the wage function, such as the schooling coefficient in a log wage equation, would appear to be self-evident. I pose the second question because it seems to me useful to rethink that enterprise.

The short answer to both questions, that of the interpretation of wage equations and that of the value in estimating them, is that it depends on the underlying economic model. The most direct example is where the paradigms concerning the origin of human capital differences are orthogonal, as in the productivity-enhancing vs. signaling models of schooling.[5] What is (one would hope) less obvious is that, even within a human capital production paradigm, in which schooling for example directly augments skills, the schooling coefficient may have different interpretations depending on the economic environment that is assumed. In this paper, I demonstrate this point by contrasting the interpretation of wage equation parameters that arises in a single-skill model under the assumption of a competitive equilibrium in the skill market to that which arises in the wage-posting equilibrium search model of Mortensen (1990) and Burdett and Mortensen (1998). In the competitive model, wage parameters directly recover fundamental skill production parameters. In the wage-posting model, the same parameters recover composites of skill production and other fundamental parameters of the model, those of the job search technology and the opportunity cost of time; the mapping from skills to wages is not direct.

The contrast between the competitive and wage-posting models is also illuminating with respect to the second question concerning the value in knowing the parameters of the wage equation. In the simplest single-skill competitive-equilibrium model of wage determination, in which schooling produces skill and is subject to choice, given the interest rate, the only fundamental parameters that determine schooling are those that also determine wages. Thus, knowledge of wage equation parameters alone is sufficient to perform counterfactual experiments, such as determining the impact on schooling of subsidizing the borrowing interest rate or of providing a college tuition subsidy. In the equilibrium search model, as noted, the wage equation identifies only a composite of the fundamental parameters.[6] It is, therefore, not possible solely with

[5] For a recent creative attempt at empirically disentangling these seemingly observationally equivalent explanations, see Fang (1999).

[6] The fundamental choices in the search model context are with respect to unemployment and employment durations, and the prototypical policy experiment is to determine the impact on those durations of increasing unemployment benefits. The fact that the two models differ in these fundamental ways is, however, not relevant to the basic point.

knowledge of wage equation parameters to perform counterfactual experiments in that model.[7]

The simplicity of the early models of schooling choice, in which comparative static results depend only on the parameters of a single wage equation, provides a rationale for concentrating on wage equation estimation per se. A second factor that has led to the surfeit of wage equation estimates, and the recent resurgence of instrumental variables – natural experiment approaches to its estimation, stems from the extremely difficult problem of accounting in estimation for the existence of "ability" heterogeneity.[8] The goal of that literature has been to obtain estimates of wage equation parameters that are free of ability bias and the rationale of the natural experiment approach to uncover exogenous sources of variation in wage determinants. Yet, resolving the ability bias issue through the use of instruments, or otherwise, would, in richer choice models, still allow for the identification of only a part of the relevant structure necessary to perform counterfactual policy experiments.

The development and structural empirical implementation of richer choice models within the human capital production framework have been slow, despite the potentially large payoff in understanding behavior and in the estimation of the effects of policy interventions. When models that relax some of the assumptions of the simple schooling choice model have been estimated, they have generally found that the extensions are empirically important. Aside from their methodological contributions, an important result from the early structural empirical papers of Willis and Rosen (1979) and Heckman and Sedlacek (1985) was the demonstration of the empirical relevance of a comparative advantage in skills. Later papers by Michael Keane and I, building on these early contributions, confirm that finding and demonstrate its importance for understanding the impact of education policies aimed at increasing school attainment and improving labor market outcomes.

The second half of the paper is devoted primarily to the consideration of one such policy, changes in college tuition costs. A large literature exists that attempts to estimate the impact of increasing college tuition costs on college enrollment within a nonstructural empirical framework. In that literature, wage payoffs, and thus wage equation parameters, are only implicitly incorporated. It is not my purpose to survey the methodology and findings of that literature in detail, but rather to contrast that literature to the recent structural empirical literature mentioned herein. I review the findings of the structural literature, which formally embeds preferences and technology, that is, human capital production cum wage functions, and show how they help to interpret the behavioral

[7] Although my primary concern is not with estimation issues, it is true that prior knowledge of wage equation parameters would improve the efficiency of the estimates of the rest of the model's parameters.

[8] As I have pointed out in a recent paper with Mark Rosenzweig (1999), the interpretation of the estimates of wage equation parameters by using instrumental variables based on natural experiments, presumed to be random, requires a theoretical model that specifically spells out the assumptions of how such instruments are related to underlying behavior.

responses estimated in the nonstructural literature and how they can also be used to quantify the impact of alternative education policies.

Two examples, discussed in more detail in the paragraphs that follow, will suffice. In Keane and Wolpin (1997), school attainment is shown to respond significantly to college tuition subsidies, consistent with the nonstructural literature, but to have only a negligible impact on market wages. The reason for this result is due to (unobserved) heterogeneity in premarket skills, occupation-specific skills determined prior to age 16 (comparative advantages that are either innate or produced within the family or that are due to the differential quality of schools attended up to that age).

As a second example, consider the general finding in the nonstructural literature that tuition effects are larger for those from lower socioeconomic backgrounds. This result has led researchers to infer the existence of important borrowing constraints in the financing of higher education. Keane and Wolpin (2001), in an extension of their preceding paper that allows for multiple methods of financing higher education, through the use of one's own assets, through working while attending school, through parental transfers, and through external borrowing, show that such an inference is not necessarily warranted. Consistent with their previous research, their current research leads them to find that the responsiveness to changes in tuition costs declines with parental education and, seemingly consistent with the inference drawn from that result, that financing a significant portion of college costs through external borrowing is infeasible. However, because of the existence of heterogeneity in premarket skills, they also find that a relaxation of borrowing constraints leads to only a negligible impact on school attainment.

The structural literature reviewed up to this point measures the impact of tuition changes by assuming that skill prices are unaffected by them; that is, they measure partial rather than general equilibrium effects. However, policies that increase school attainment will reduce skill prices and thus reduce the incentive to acquire market skills. In two recent efforts, the first by Heckman, Lochner, and Taber (1998, 1999) and another by Lee (2000), researchers have developed and empirically implemented methodologies for the estimation of general equilibrium models of schooling choice, and they used them to assess the extent to which partial equilibrium effects of tuition changes are misleading. These efforts differ in important ways and lead to quite different results. On one hand, Heckman et al. (1999) find that the partial equilibrium effect of a change in college tuition overstates general equilibrium effects by about a factor of ten. On the other hand, Lee finds that partial equilibrium effects of a college tuition subsidy overstate general equilibrium effects by at most 10 percent.

2. INTERPRETING WAGE EQUATIONS

Essentially all empirical wage equations take the log-linear form

$$\log\, w = s(X;\beta) + \varepsilon, \tag{2.1}$$

where the Xs are observable wage "determinants," β is a vector of parameters, and ε accounts for unobservable wage determinants and for measurement error. The expectation of (2.1) for any X represents the mean of the distribution of wages *offered* to an individual with observable characteristics X. The human capital framework generates (2.1) as the equilibrium wage determined in a competitive and frictionless labor market. A quite different paradigm, one not usually associated with the human capital literature, generates (2.1) as the equilibrium outcome of a labor market search on the part of workers and firms.These two economic models demonstrate an interesting contrast in the interpretation of (2.1) and in the usefulness of estimating its parameters.

2.1. The Wage Equation in a Competitive Model

The simplest version of the competitive model assumes a single homogeneous productive skill in which workers are perfect substitutes in skill units.[9] Letting S_t be the aggregate number of units of skill used in production at calendar time t, that is, the sum of skill units over workers, and K_t the aggregate stock of physical capital at time t, we find that aggregate output at t is given by the constant returns to scale production function

$$Y_t = F(S_t, K_t; \theta_t), \tag{2.2}$$

where θ_t is a vector of production function parameters that may evolve with time and where (2.2) exhibits diminishing returns in each input. If aggregate skill over workers in the economy at time t is S_t^*, then for a given capital stock, the competitive assumption implies that the rental price per unit of skill at time t, r_t, is the marginal product of skill evaluated at S_t^*, that is,

$$r(S_t^*, K_t; \theta_t) = \left.\frac{\partial F}{\partial S_t}\right|_{S_t = S_t^*}. \tag{2.3}$$

An individual j who has s_{ja} units of skill at age a will receive a wage offer at time t, w_{jat}, given by the product of r_t and s_{ja}. Upon taking logs, the equilibrium wage offer function, that is, the wage offered to an individual with s_{ja} units of skill at time t, is given by

$$\log[w(s_{ja}; r_t)] = \log(r_t) + \log(s_{ja}). \tag{2.4}$$

To complete the competitive equilibrium model, one must specify a model for the supply of skill units to the labor market.[10] There are two related issues in

[9] Griliches (1977) explicitly derived the log wage equation in this single-skill setting. The multi-skill analog is found in Heckman and Sedlacek (1985) and in Willis (1986), both of which have their roots in Roy (1951). Many of the same points made here are contained in those papers, and I do not delineate individual contributions further. See, in addition, Heckman, Layne-Farrar, and Todd (1996).

[10] It is, of course, also necessary to specify a model of physical capital accumulation. The discussion assumes that physical capital evolves exogenously.

determining skill supply: (i) the amount of skill possessed by the individuals in the economy, and (ii) if skill amounts are heterogeneous, the identities of the individuals who participate in the labor market and thus supply their skills to the market.

Ben Porath (1967), following Becker (1967) and Mincer (1958, 1962), provided the first formalization of the acquisition of skills within an explicit optimization framework. The essential idea was that human capital is produced according to some technology that combines an individual's time with purchased goods. In the production function framework, the determinants of skill at any age would consist of all skill-producing activities and complementary goods up to that age, beginning at least from conception (and therefore including parental behaviors, such as the mother's smoking behavior during pregnancy, and the time parents spend reading to their children at various ages, as well as the quantity and quality of formal schooling and training on the job), plus a genetically determined skill endowment, usually referred to as ability.[11]

A basic result from Ben-Porath was that individuals would optimally specialize in human capital production early in the life cycle, when an individual's stock of human capital was low, but, after some point, human capital production would decline with age. Mincer (1974) used that result to derive a wage function of the form

$$\log(w_{ja}) = \beta_0 + \beta_1 E_{ja} + \beta_2 P_{ja} - \beta_3 P_{ja}^2 + \varepsilon_{ja}, \tag{2.5}$$

where E, school attainment, and P, work experience, are intended to capture all of the time and goods investments used in the production of skills given the difficulty of directly measuring inputs, as Ben-Porath noted.

The wage, and thus, indirectly, market-valued skill, feasibly can be observed only at the first age a market wage is offered, say a_0.[12] The amount of skill possessed at that age, premarket skills, is the sum of the individual's skill endowment and of the amount of skill produced up to that age. It is convenient to distinguish between premarket skills produced through formal schooling and premarket skills produced outside of school (e.g., through investments made by the parents and by the child). I denote the latter by s_0.

A wage function consistent with Mincer's formulation is easily derived from the competitive skill market model. If the skill production function takes the form $s_{ja} = s_{j0} \exp[\beta_1 E_{ja} + \beta_2 P_{ja} - \beta_3 P_{ja}^2 + \varepsilon_{ja}]$, (2.4) can be written as

$$\log(w_{jat}) = \log(r_t) + \log(s_{j0}) + \beta_1 E_{ja} + \beta_2 P_{ja} - \beta_3 P_{ja}^2 + \varepsilon_{ja}. \tag{2.6}$$

As (2.6) indicates, the constant term in the Mincer formulation can be interpreted as a composite of (i) the skill rental price, which depends on calendar

[11] Given the inherent nonobservability of skill, ability is assumed to be measured in skill units.

[12] This age may depend both on technology and on legal restrictions.

time through changes in the aggregate skill supply and in the capital stock, and through permanent changes in technology or transitory production shocks, and (ii) the level of the individual's (nonschool) premarket skills, which may differ among individuals because of either genetic endowment or family investment behavior.[13] Shocks to wages, ε_a, other than through aggregate production shocks incorporated in the rental price, reflect idiosyncratic shocks to skills (e.g., random mental or physical lapses or enhancements). In this framework, the effect of school attainment on the wage is fixed by the technology of the educational system in producing market-valued skills and does not have the interpretation of an internal rate of return.[14] Likewise, the effect of work experience on the wage is fixed by the technology of skill acquisition on the job, that is, by learning by doing.[15]

To complete the specification requires an explicit model of schooling, work decisions, and premarket skill acquisition. The incorporation of work decisions is particularly important for estimation because wage offers are observed only for individuals who work. I will return to this task, and to the implications of this modeling for estimation, after presenting a wage determination model in which there are labor market search frictions and that leads to a quite different interpretation of the parameters of the wage function.

2.2. The Wage Equation in the Burdett–Mortensen Equilibrium Search Model

In contrast to the interpretation of (2.6) as the outcome of a frictionless competitive equilibrium in unobservable labor quality or skill units, wage equations that arise from equilibrium search models are based on the existence of search frictions. In the Burdett–Mortensen model, firms post wages that are revealed to potential workers when they meet. Workers engage in search both while unemployed and while working.[16] All potential workers and all firms are identical. A steady-state equilibrium wage distribution is generated as firms that

[13] School-produced premarket skills are accounted for in $\beta_1 E_a = \beta_1 E_0 + \beta_1(E_a - E_0)$, where E_0 is schooling at a_0. Endowed ability, in addition to being a component of premarket skills, may also augment the amount of skill produced by additional schooling, work experience, or premarket nonschool investments, leading to interaction terms in (2.5). This point is explicit in the specification of the human capital production function in Ben-Porath (1967) and was also made by Griliches (1977) in referring to a wage specification such as (2.5). The existence of these interaction effects has become an important issue in the "treatment effects" literature; see, for example, Manski (1997), Heckman (1997), and Heckman and Vytlacil (1998).

[14] Heckman and Sedlacek (1985) use this observation to rationalize the relative constancy of the schooling coefficient over time in the face of large changes in aggregate schooling levels.

[15] Economic theory is not a useful guide for the manner in which inputs, schooling, and work histories should enter the skill production function.

[16] There are a number of excellent surveys of this literature; for example, see Mortensen and Pissarides (2000) or Van den Berg (1999).

post higher wages nevertheless earn the same monopsony rents as firms that post lower wages because they attract more workers; that is, rent per worker times the number of workers is equalized across firms. The model generates a unique continuous wage offer density.

Unemployed workers solve a standard (infinite horizon–discrete time) job search model with search on the job, which is known to satisfy the reservation wage property.[17] The reservation wage, w^*, is given by the implicit solution to

$$w^* = b + (\lambda_0 - \lambda_1) \int_{w^*}^{\infty} \frac{1 - F(w)}{\delta + \lambda_1 [1 - F(w)]} dw, \tag{2.7}$$

where b is unemployment income (net of the cost of search plus the monetized value of leisure), λ_0 is the per period probability that an unemployed worker receives a wage offer, λ_1 is the per period probability that an employed worker receives a wage offer from another firm, δ is the exogenous per period layoff or separation rate while employed, and F is the cumulative distribution function of wage offers.[18] The reservation wage while employed is the worker's current wage.

Firms maximize steady-state profits given by

$$\pi = [p - w] n(w; w^*, F), \tag{2.8}$$

where p is the common level of worker–firm productivity, and where $n(w)$, the number of workers employed at firms offering wage w, is increasing in w.[19] A unique equilibrium exists for this model with the wage offer distribution function given by

$$F(w) = \frac{\delta + \lambda_1}{\lambda_1} \cdot \left(1 - \sqrt{\frac{p - w}{p - w^*}} \right), \tag{2.9}$$

with support $[w^*, \bar{w}]$ and $\bar{w} < p$.[20] Given (2.9), the reservation wage is a weighted average of unemployment income and productivity,

$$w^* = \theta b + (1 - \theta) p, \tag{2.10}$$

where $\theta = (\delta + \lambda_1)^2 + (\lambda_0 - \lambda_1)\lambda_1$.[21] Given (2.10), the mean wage offer is also

[17] Burdett (1978) first introduced on-the-job search in a partial equilibrium setting. See also Mortensen (1986).

[18] See Mortensen and Neumann (1988). The discount rate is set to zero in order to simplify the firm's problem (see Van den Berg and Ridder, 1998).

[19] Note that productivity is independent of the number of workers in the firm. For a model that allows for declining marginal productivity, see Robin and Roux (1997).

[20] The density function that corresponds to (2.9) is monotonically increasing in the wage, which is counterfactual. As a way to fit the wage data, empirical implementations have been conducted allowing for heterogeneity in firm productivities. See Bontemps, Robin, and Van den Berg (1999) and Bowlus, Kiefer, and Neumann (1998).

[21] The upper bound of the support of the wage offer distribution, $\bar{w}$, is also a weighted average of b and p.

a weighted average of b and p, namely

$$E(w) = \gamma b + (1 - \gamma)p, \tag{2.11}$$

where γ is a function of the two offer probabilities, λ_0 and λ_1, and the separation probability, δ.[22]

As noted, the model assumes workers to be homogeneous in skill levels. Suppose, however, that the labor market is perfectly segmented by education; that is, worker–firm productivity levels vary by education and the productivity within an education segment is independent of the size of other segments, and the productivity for workers of education level E is $\beta_1 E$.[23] Then, the wage offer to individual j is

$$w_j = \gamma b + (1 - \gamma)\beta_1 E_j + \varepsilon_j, \tag{2.12}$$

where, by definition, $E(\varepsilon_j|E_j) = 0$.[24] It is also possible to obtain a closed-form representation for the mean of the log wage from (2.9). However, it turns out to be a highly nonlinear function of p. To maintain the parallel with the competitive model, as well as to conform to the empirical literature on estimating wage functions, consider a first-order approximation to the mean of the log wage, which takes the general form

$$\log(w_j) = \gamma_0'(\lambda_0, \lambda_1, \delta, b) + \gamma_1'(\lambda_0, \lambda_1, \delta, b)\beta_1 E_j + \varepsilon_j'. \tag{2.13}$$

If approximation error is ignored, analogously $E(\varepsilon_j'|E_j) = 0$.

2.3. The Schooling Coefficient

There is an enormous literature whose goal is to estimate the schooling coefficient in Mincer-type wage equations. As is evident from a comparison of (2.6) and (2.13), however, its interpretation is quite different depending on which of the two labor market models is adopted. In the competitive model, knowledge of the schooling coefficient in (2.6), that is, of the (percentage) effect of incremental increases in schooling on the wage, recovers the effect of school attainment on skill production, a fundamental structural parameter. In the wage-posting model, however, knowledge of the schooling coefficient recovers a composite of the effect of school attainment on skill production (normalizing product price to unity), of the parameters describing the worker–firm meeting technology, of the job separation rate, and of the (common) level of unemployment income plus monetary equivalent leisure value. It is thus necessary to combine wage data with other information in order to recover the schooling–skill relationship.

[22] Specifically, $\gamma = \theta\{1 - [(\lambda_1/\delta)(3 + 2\lambda_1/\delta)]/[3(1 + \lambda_1/\delta)^2]\}$. See Bowlus and Eckstein (1999).

[23] Van den Berg and Ridder (1998) assume that labor markets are segmented by education and occupation in their empirical implementation of the Burdett–Mortensen equilibrium search model.

[24] To make the point most simply, unemployment income, offer probabilities, and layoff rates are assumed not to differ across education-specific employment sectors.

In contrast to the human capital literature, the estimation of wage offer equations, and in particular, the estimation of the schooling–skill relationship, has not been a singular goal of the equilibrium search literature. Indeed, given the identification problem associated with estimating the schooling–skill relationship from knowledge only of the wage equation parameters, estimating the wage offer function by itself is clearly fruitless for that purpose.[25]

In the competitive-equilibrium model, the schooling coefficient identifies a single structural parameter of the skill production function. In early schooling models, it was the case that knowledge of that parameter was sufficient to draw policy implications. In the prototypical model in which (i) individuals maximize their present value of lifetime (nonstochastic) earnings, (ii) individuals work in all periods after completing school (and for a fixed period independent of schooling), (iii) the only cost of schooling is forgone earnings ($c = 0$), and (iv) there is a fixed borrowing rate of interest; the optimal level of schooling is simply given by the condition that the marginal percentage effect of schooling on wages (equal to the percentage increase in skill) is equal to the borrowing rate of interest. Thus, knowledge of the schooling coefficient is sufficient in those models to perform policy experiments that vary the borrowing rate.

To illustrate, consider the following stylized model of schooling choice.[26] Suppose each individual enters school at a mandated school entry age a_e and must remain in school until a mandated minimum school leaving age that is perfectly enforced, assumed to be the initial age at which wage offers are received, a_0. Thus, school attainment at a_0, initial schooling, is $E_0 = a_0 - a_e$, and is not subject to choice. Assume also that the individual decides on whether to attend school for only one period beyond the school leaving age.[27] School attendance in that decision period (at age a_0) is denoted by $e_1 = 1$ and nonattendance by $e_1 = 0$; completed schooling, at age $a_0 + 1$, E_1, is therefore either $E_0 + 1$ or E_0. An individual who decides not to attend school is assumed to work in that period and in all subsequent periods until the end of working life, $a = A$. An individual who attends school is precluded from working in that period, but is assumed to work in all subsequent periods, that is, from $a_0 + 1$ to $A + 1$.[28]

[25] The empirical implementation of equilibrium search models is a recent literature and has been structural in its methodology. The goal of that research has been twofold: (i) to determine whether the dispersion observed in wages for seemingly observationally equivalent individuals is consistent with the degree of dispersion that can be generated by equilibrium search models; and (ii) to determine the magnitude of the effects of wage and unemployment insurance policies on labor market transitions. See Eckstein and Wolpin (1990) and Van den Berg and Ridder (1998) for examples.

[26] The model is in the spirit of the formulation in Mincer (1974). See also Becker's (1967) Woytinsky lecture, Rosen (1977), Willis (1986), and, more recently, Heckman (1997).

[27] The discrete nature of the choice parallels the sequential decision models discussed in the text that follows.

[28] Allowing the working life to be independent of schooling simplifies the school attendance decision rule.

In addition, there is assumed to be a direct cost of attending school, c. The skill rental price is assumed to be constant and idiosyncratic shocks to skills are ignored.

The individual is assumed to make the choice of whether or not to attend school according to which option maximizes the present discounted value of lifetime earnings. The present value of each alternative to individual j, assuming the wage equation takes the form of (2.6), but ignoring for simplicity the stochastic component, $V_1(e_1 = 1|E_0, s_0)$ and $V_1(e_1 = 0|E_0, s_0)$, is given by

$$\begin{aligned} V_1(e_1 = 1|E_0, s_0) &= \exp[\log r + \log s_0 + \beta_1(S_0 + 1)] \\ &\quad \times \sum_{a=1}^{A+1} \delta^a \exp[\beta_2(a-1) - \beta_3(a-1)^2] - c, \\ V_1(e_1 = 0|E_0, s_0) &= \exp[\log r + \log s_0 + \beta_1 S_0] \\ &\quad \times \sum_{a=0}^{A} \delta^a \exp[\beta_2 a - \beta_3 a^2], \end{aligned} \tag{2.14}$$

where $\delta = 1/(1+i)$ is the discount factor and i is the interest rate. The decision rule is to attend school if $V_1(e_1 = 1|E_0, s_0) \geq V_1(e_1 = 0|E_0, s_0)$, which reduces to

$$\begin{aligned} e_1 &= 1 \quad \text{if } \beta_1 \geq i + \frac{c}{V_1(e_1 = 0|E_0, s_0)}, \\ e_1 &= 0 \quad \text{otherwise.} \end{aligned} \tag{2.15}$$

Thus, the individual attends school if the percentage increase in the wage from attending school is sufficiently greater than the interest rate, or, equivalently, if the incremental income flow per unit time obtained from attending school net of the cost of tuition exceeds the income flow obtained from investing initial wealth at the market rate of interest.[29]

The schooling decision in this model depends on the level of premarket skills (as defined, inclusive of "ability") and initial schooling, as well as on the interest rate. Individuals with a greater level of either premarket skills or initial schooling, and thus greater initial wealth, $V_1(e_1 = 0|E_0, s_0)$, will be more likely to obtain the additional period of schooling. Indeed, given that the present value of earnings is increasing monotonically in premarket skills, there exists a unique cut-off value below which $e_1 = 0$ is optimal and above which $e_1 = 1$ is optimal (for any value of initial schooling). Further, under the maintained assumption that individuals work in each period after completing school up to an exogenous retirement age (that is independent of completed schooling), work experience is independent of premarket skills.

[29] In deriving (2.15), we use the approximation $\ln(1 + x) \approx x$ for x equal to the interest rate and for x equal to the ratio of the direct cost of schooling to the present value of earnings when $e_0 = 0$.

2.4. A Brief Digression: Estimating the Schooling Coefficient by Using Natural Experiments

This simple theory has important implications for estimation. Clearly, an estimation of (2.6) founders on the problem of measuring premarket skills. The search for methods of obtaining a consistent estimate of the schooling coefficient in this setting has been ongoing for several decades. One class of such methods that has recently gained renewed attention relies on the use of so-called natural experiments.[30] The general argument is that naturally occurring events that are by their nature random can be used as instrumental variables to identify important economic parameters, such as the schooling effect on wages. Angrist and Krueger (1991) use information on month of birth together with state-level school entry and school leaving age requirements to estimate the schooling coefficient. The idea is that two, otherwise identical, individuals residing in the same state whose birthdays differ by as little as one day may nevertheless have entered school a year apart. Given the difference in age at entry, their school completion levels would likely differ given that they must differ by a year at the minimum school leaving age.

The schooling model implies that there will indeed be a behavioral response to the additional schooling completed at a_0. From the decision rule (2.15) it can be seen that an individual who entered at an earlier age, and thus had completed more schooling at a_0, would not only have that additional year but also would be more likely to attend school after the minimum school leaving age.[31] Thus, month of birth should be related to completed schooling through its impact on age at entry. It would seem, then, that the presumably inherent randomness in month of birth would make it an ideal instrument for estimating the schooling coefficient.

A critical assumption, however, is that the variation created in age at entry caused by differences in month of birth is unrelated to variation in premarket skills (inclusive of the ability endowment) and to variation in work experience. With respect to premarket skills, it is plausible that children whose entry is delayed will not have the same skills upon entry as those who were not delayed. Parental investment behavior would presumably respond to the delayed age at entry. If, for example, skills that would have been acquired in school are acquired in the home or in preschool, the instrumental variables estimator of the schooling coefficient based on month of birth would potentially understate the true schooling effect on skills.[32]

[30] Rosenzweig and Wolpin (2000) review the literature, restricting their attention to the use of random events that arise in nature; they call these “natural” natural experiments.

[31] As noted, the reason for this is that, as with premarket skills, initial schooling increases initial wealth, which increases the propensity to obtain additional schooling. This result is obviously sensitive to the assumption that the impact of schooling on log skill is linear.

[32] It is possible that the schooling–skill production function is such that the additional s skills at school entry are irrelevant.

The model also assumed that individuals supply their labor inelastically in all postschooling years. If, instead, individuals make labor supply decisions over their life cycles, it will generally be true that initial schooling, and thus age at school entry, will affect labor supply at postschooling ages and thus be related to work experience. The reason is simply that because age at entry affects completed schooling, which itself affects wage offers, age at entry will also affect work decisions at all postschooling ages.[33] Thus, the instrumental variables estimator based on month of birth (m) is

$$\frac{\partial \ln w_a/\partial m}{\partial E_a/\partial m} = \beta_1 + \frac{\partial s_0/\partial m + \beta_2 \partial P_a/\partial m}{\partial E_a/\partial m}, \tag{2.16}$$

where, for convenience, we have assumed that the experience effect on wages is linear. As seen in (2.16), the interpretation of instrumental variables estimators, even those based on arguably random events, requires the researcher to specify a behavioral model inclusive of the role that the instrument plays.

3. EXTENSIONS OF THE COMPETITIVE SKILL MARKET EQUILIBRIUM MODEL

The competitive skill market equilibrium model has been the foundation for many applications in labor economics, although it is often not explicitly stated. There is surprisingly little empirical work on estimating wage equations that has appealed to models beyond the simple version of the schooling–skill model presented herein. Yet, when models that relax some of the assumptions of that model have been estimated, they have generally found those extensions to be empirically important. The primary extension has been to incorporate self-selection in a multidimensional skill setting along the lines of the Roy (1951) model.

3.1. Willis and Rosen

The first important extension, by Willis and Rosen (1979), allows for heterogeneous skills in a schooling choice model. Following the interpretation in Willis (1986), suppose that there are two occupations in the economy, white-collar (wc) and blue-collar (bc) occupations, and that to enter the wc occupation one must have, using previous notation, $E_0 + 1$ years of schooling. Thus, an individual j who decides not to attend school in the remaining period is concomitantly deciding to enter the bc occupation. An individual has premarket skills given

[33] As Rosenzweig and Wolpin (2000) show, the bias due to the schooling–experience relationship changes with age. At ages close to the school leaving age, those with more schooling have less work experience, while at later ages the more schooled, who are more likely to participate and work more hours when they do participate, have more experience. In addition, Bound and Jaeger (1996) provide same empirical evidence calling into question the randomness of month of birth.

by the pair $(s_{0j}^{\mathrm{wc}}, s_{0j}^{\mathrm{bc}})$.[34] It is natural in this setting to assume that the marginal product of incremental schooling may differ for wc and bc skills. Thus, abstracting from wage growth arising from the accumulation of work experience, which is allowed for as exogenous in Willis and Rosen, an individual faces the following schedule of wage offers, depending on the schooling decision the individual makes:

$$\begin{aligned} \log\left(w_j^{\mathrm{wc}}\right) &= \log(r^{\mathrm{wc}}) + \log\left(s_0^{\mathrm{wc}}\right) + \beta_1^{\mathrm{wc}}(E_0 + 1) + \varepsilon_j^{\mathrm{wc}} \quad && \text{if } e_1 = 1, \\ \log\left(w_j^{\mathrm{bc}}\right) &= \log(r^{\mathrm{bc}}) + \log\left(s_0^{\mathrm{bc}}\right) + \beta_1^{\mathrm{bc}} E_0 + \varepsilon_j^{\mathrm{bc}} && \text{if } e_1 = 0, \end{aligned} \tag{3.1}$$

where the individual's premarket skills have been decomposed into a population mean level and an individual-specific component, that is, $s_{0j}^k = s_0^k + \varepsilon_j^k$ for $k =$ wc, bc, and where the r^ks are the respective competitively determined skill rental prices.[35]

The multiskill model creates an important distinction between the impact of additional schooling on skills, and its impact on wage offers. In the single-skill model, the schooling parameter in the wage equation is equal to both the percentage increase in skills and the percentage increase in the wage caused by an additional period of schooling. In the multiskill setting, the former is occupation specific and is given by the schooling parameters in (3.1), β_1^{wc} and β_1^{bc}, whereas the latter, reflecting also the return to the option of working in the wc occupation if one attends school, depends on all of the wage determinants in both occupations. In particular, the percentage change in the wage offer to a person randomly chosen from the population to attend school is

$$\frac{1}{w}\frac{\Delta w}{\Delta E} = \log\left(\frac{w^{\mathrm{wc}}}{w^{\mathrm{bc}}}\right) = \log\left(\frac{r^{\mathrm{wc}} s_0^{\mathrm{wc}}}{r^{\mathrm{bc}} s_0^{\mathrm{bc}}}\right) + \left(\beta_1^{\mathrm{wc}} - \beta_1^{\mathrm{bc}}\right) E_0 + \beta_1^{\mathrm{wc}}. \tag{3.2}$$

As (3.2) reveals, the schooling effect on wages in this case is a composite of the structural parameters of both wage offer functions (those of the underlying skill production functions), as well as the occupation-specific rental prices determined in general equilibrium. Knowledge of (3.2), obtained, say, from variation in initial schooling as in the Angrist–Krueger natural experiment, would not recover the impact of schooling on either wc or bc skills.

Because individuals differ in their wc and bc premarket skills, the sample analog of (3.2) depends on how individuals sort themselves into schooling groups. The choice of occupation, and thus schooling, as in the single-skill model is based on comparing the respective present values of wages. Assuming

[34] Premarket skills are, in their terminology, abilities.

[35] The aggregate production function in this case is $Y = F(S^{\mathrm{wc}}, S^{\mathrm{bc}}, K)$, with skill rental prices given by their respective marginal products evaluated at the equilibrium aggregate skill quantities and capital. Willis and Rosen did not explicitly embed the model within a market equilibrium setting.

the direct cost of schooling is zero ($c = 0$) as in Willis and Rosen, the decision by individual j to attend school is given by

$$\begin{aligned} e_{1j} &= 1, \quad \text{iff } \varepsilon_{1j}^{\text{wc}} - \varepsilon_{1j}^{\text{bc}} > i + \overline{\log(w^{\text{wc}})} - \overline{\log(w^{\text{bc}})}, \\ &= 0, \quad \text{otherwise.} \end{aligned} \tag{3.3}$$

Given this decision rule, the sample analog of (3.2) is simply

$$\begin{aligned} \left\{\frac{1}{w}\frac{dw}{dE}\right\}_{\text{SAMPLE}} = \left\{\frac{1}{w}\frac{dw}{dE}\right\}_{\text{POP}} &+ E\left(\varepsilon_{1j}^{\text{wc}} | e_{1j} = 1\right) \\ &- E\left(\varepsilon_{1j}^{\text{bc}} | e_{1j} = 0\right). \end{aligned} \tag{3.4}$$

As (3.4) reveals, the schooling effect that is estimated from sample wage differences can either overstate or understate the effect under random selection. For example, it will be overstated if those who select wc have higher than average wc premarket skills and those who select bc have lower than average bc premarket skills. Willis and Rosen find that those who attend school (college) have higher than average wc premarket skills and those that do not attend have higher than average bc premarket skills. From (3.4), that implies that the sign of the difference in the sample and population schooling effects is indeterminate in theory.[36] On the basis of their empirical estimates, the random selection effect (3.2) is 9 percent whereas the sample effect (3.4) is 9.8 percent. Their estimates are the first to indicate empirically the existence of heterogeneity in premarket skills, that is, comparative advantage in the labor market, a finding that has been replicated in more recent studies and that appears to have critically important implications for education policy.

3.2. Heckman and Sedlacek

Heckman and Sedlacek (1985) provide a second important extension of the Roy model, although they focused on sectoral rather than occupational choice and their direct concern was not with the schooling–skill relationship. Unlike Willis and Rosen, they embed the sectoral choice model within a market equilibrium setting. In addition, they allow for a nonmarket sector and, given that, assume that individuals maximize utility rather than wealth. As in Heckman and Sedlacek, consider the choice of working in two different sectors, manufacturing

[36] Willis and Rosen empirically implement the model by using data from the NBER–Thorndike sample. They employ a two-step estimation procedure with the additional assumptions that family background characteristics affect the borrowing rate of interest but not premarket skills and that the idiosyncratic components of premarket skills and the borrowing rate of interest are distributed as joint normal. The "ability" tests available in their data are taken to be observable measures of premarket skills and are assumed to be unrelated to borrowing costs.

It will not serve my purpose to repeat identification arguments in Willis and Rosen. Identification in selection models has been the subject of intensive investigation for the past 25 years. Without minimizing its importance, my concern is rather with the substantive economic interpretations and implications of these models.

(m) and nonmanufacturing (n), in postschooling periods. In each such period, an individual j of age a at calendar time t receives a wage offer from each sector k, $k = m, n$, that takes the form[37]

$$\log\left(w^k_{jat}\right) = \log\left(r^k_t\right) + \log\left(s^k_0\right) + \beta^k_1 E_{ja} + \beta^k_2 P_{ja} - \beta^k_3 P^2_{ja} + \varepsilon^k_{ja}. \tag{3.5}$$

Notice that work experience in one sector is assumed to be a perfect substitute for experience in the other sector.

Individuals are assumed to be static optimizers.[38] The utilities associated with the three choices, denoting the home sector as h, are assumed to be given by

$$\begin{aligned} u^m_{jat} &= \log\left(w^m_{jat}\right) + b^m + \xi^m_{ja}, \\ u^n_{jat} &= \log\left(w^n_{jat}\right) + b^n + \xi^n_{ja}, \\ u^h_{ja} &= b^h + \xi^h_{ja}, \end{aligned} \tag{3.6}$$

where the bs reflect the nonpecuniary values attached to each of the market sectors and to the nonmarket sector and the ξs are their associated time-varying shocks.[39] At any age a, individuals choose the option with the highest utility. Heckman and Sedlacek substitute the log wage function into (3.6) and normalize the utility of the home alternative to zero.[40]

Following arguments in Heckman and Sedlacek, estimating (3.5) by using data from repeated cross sections, and accounting for sectoral choice recovers the sum of the sector-specific skill rental prices and premarket skills.[41] With the assumption that premarket skills do not vary over cohorts, the time series of skill rental prices for each sector can be estimated up to a normalization of one skill rental price in each sector to unity.[42] This insight avoids having to solve the model for aggregate skill quantities that are needed to determine equilibrium marginal skill products, and thus skill rental prices, which in turn, as

[37] This formulation of the sector-specific skill production functions differs from that in Heckman and Sedlacek, who adopt a more general Box–Cox formulation that nests the exponential form as used in obtaining (2.6) and that allows for nonnormal errors.

[38] This assumption is not innocuous because, if individuals did not completely discount the future, they would take into account that working (in either sector) at any age increases work experience at all future ages.

[39] Heckman and Sedlacek jointly estimate the reduced-form choice model together with the wage function, but they do not impose the parameter restrictions arising from the fact that the wage parameters also enter the choice model. The reduced-form approach to estimation, although more robust, does not allow the recovery of sector-specific nonpecuniary values. The structure given by (3.6) is consistent with their reduced-form specification.

[40] With the assumption of a general variance–covariance structure, estimation also requires that one of the variances of the nonpecuniary preference plus wage shock be normalized to unity.

[41] They use March Current Population Surveys from 1968 to 1981.

[42] As a validation of this approach, they empirically demonstrate that rental prices behave as one would expect from theory.

they note, would require estimating the model on demographic groups that differ in any fundamental parameters.[43]

Like Willis and Rosen, Heckman and Sedlacek find evidence of comparative advantage. They estimate that a reduction in the demand for manufacturing sector workers increases the average skill level in manufacturing, as those with less of that sector-specific skill leave manufacturing. In addition, however, the average skill level in nonmanufacturing is reduced as the new entrants from the manufacturing sector have lower nonmanufacturing skills than those in the sector before the demand reduction.[44]

3.3. Keane and Wolpin

Keane and Wolpin (1997) extend the Roy model to a dynamic setting, combining features of both the Willis and Rosen and the Heckman and Sedlacek models. In their model, at each age an individual chooses among five mutually exclusive alternatives: attend school (e), work in a white-collar (wc) occupation, work in a blue-collar occupation (bc), work in the military (ml), or remain at home (h). The following current period utility payoffs capture their essential features:

$$\begin{aligned}
u_{ja}^{k} &= w_{ja}^{k} + b^{k} \quad \text{for } k = \text{wc, bc},\\
u_{ja}^{\text{ml}} &= w_{ja}^{\text{ml}},\\
u_{ja}^{e} &= b_{j}^{e} - tc_1 I(S_{ja} \geq 12) - tc_2 I(S_{ja} \geq 16) + \xi_{ja}^{e},\\
u_{ja}^{h} &= b_{j}^{h} + \xi_{ja}^{h},
\end{aligned} \tag{3.7}$$

where, as before, the bs reflect the nonpecuniary components of the alternatives, tc_1 and tc_2 are parameters indicating tuition and other direct costs associated with college attendance and the incremental cost of graduate school attendance, and $I(\cdot)$ is the indicator function set equal to one if the expression in the parentheses is true. Occupation-specific skill rental prices are assumed to be constant over time.[45]

Skill production functions are modified from the single-skill model so as to account for their occupation-specific nature, namely, $s_{ja}^{k} = s_{j0}^{k} \exp[\beta_1^k E_{ja} + \beta_2^k P_{ja}^k - \beta_3^k (P_{ja}^k)^2 + \varepsilon_{ja}^k]$. In attempting to fit the wage (and choice) data, Keane and Wolpin found it necessary to augment the set of inputs of the occupation-specific skill production functions to include cross-experience effects (work experience accumulated in the wc and bc occupations enter in all three skill

[43] Estimates are based on white males aged 16–65 years. Skill rental prices are assumed to be identical for all demographic groups.

[44] In this context, comparative advantages are not those related only to premarket skills, but reflect differences in schooling and work experience at the time of the demand shift.

[45] With longitudinal data that do not span the entire working lives of the sample, it is not possible to estimate the time series of occupation-specific skill rental prices by using year dummies as in Heckman and Sedlacek. With perfect foresight, all future rental prices enter into any contemporaneous decision.

production functions), skill depreciation effects (whether worked in the same occupation in the previous period), a first-year experience effect (whether or not the individual ever worked previously in the occupation), age (maturation) effects, and separate high school graduation and college graduation effects.[46]

Individuals are assumed to maximize the expected present discounted value of their remaining lifetime utility at each age starting from age 16, the assumed initial decision-making age, and ending at age 65, the assumed terminal age. Letting $V(\Omega_{ja})$ be the maximum expected present discounted utility given the state space Ω_{ja} and letting the choice set be ordered from 1 to 5, in the order as given in (3.7), with d^k_{ja} equal to one if alternative k is chosen and zero otherwise, one sees that

$$V(\Omega_{ja}) = \max_{d^k_{ja}} E\left[\sum_{\tau=a}^{65} \delta^{\tau-a} \sum_{k=1}^{5} u^k_{ja} d^k_{ja} | \Omega_{ja}\right]. \tag{3.8}$$

The state space at age a consists of all factors, known to the individual at a, that affect current utilities or the probability distribution of future utilities.

The value function can be written as the maximum over alternative-specific value functions, $V^k(\Omega_{ja})$, each of which obeys the Bellman equation (Bellman, 1957):

$$V(\Omega_{ja}) = \max_k \left\{V^k(\Omega_{ja})\right\}, \tag{3.9}$$

where

$$\begin{aligned} V^k(\Omega_{ja}) &= u^k_{ja} + \delta E\left[V(\Omega_{j,a+1})|\Omega_{ja}, d^k_{ja} = 1\right], \quad \text{for } a < 65, \\ V^k(\Omega_{j,65}) &= u^k_{j,65}. \end{aligned} \tag{3.10}$$

The expectation in (3.10) is taken over the distribution of the random components of the state space at $a + 1$ conditional on the state-space elements at a, that is, over the unconditional distribution of the random shocks given serial independence. The predetermined state variables such as schooling and occupation-specific work experience evolve in a Markovian manner that is (conditionally) independent of the shocks. There is no closed-form representation of the solution. Keane and Wolpin use a numerical solution method based on an approximation method developed in an earlier paper (Keane and Wolpin, 1994).[47]

[46] In addition, they incorporate the following features (parameters) into the model they estimate: a monetary job-finding cost for civilian occupation employment that depends on whether the individual ever worked previously in the occupation, age effects on the net (of effort) consumption value of school attendance, separate reentry costs associated with returning to high school and to college, age effects on the utility of the home alternative, psychic values associated with earning high school and college diplomas, and a cost of prematurely leaving the military, that is, before completing two years of service.

[47] For surveys of methods used in solving and estimating discrete choice dynamic programming problems, see Eckstein and Wolpin (1989) and Rust (1994, 1996).

Individuals are assumed to be heterogeneous in their occupation-specific premarket civilian skills (s^k_{j0} for $k = \text{wc}, \text{bc}$) and in their school and home nonpecuniary valuations (b^e_j and b^h_j). It is assumed that such population heterogeneity can be completely characterized by a fixed number (J) of types of individuals each occurring in the population with probability π_j. Each type is uniquely described by a vector of premarket skills and nonpecuniary valuations. In addition to a difference at age 16 in these characteristics, there are also differences in completed schooling levels at age 16. As is well known (Heckman, 1981), to the extent that this observable initial condition is the result of the same or closely related optimization problem that determines future choices, the initial condition will generally not be exogenous. Keane and Wolpin make the assumption that initial schooling is exogenous conditional on the unobservable persistent initial conditions. The method of estimation is by simulated maximum likelihood, and the data they use are from the 1979 youth cohort of the National Longitudinal Surveys of Labor Market Experience. Their estimates are restricted to white males.

As noted, a consistent empirical finding of the Willis and Rosen and Heckman and Sedlacek papers was that comparative advantage plays an important allocative role in the labor market. Workers self-select into occupations and into sectors based on their relative productivities. This is also the result in Keane and Wolpin.

Comparative advantages determined by age 16 lead to large differences in school attainment, in later labor market outcomes, and in lifetime welfare. As shown in Table 2.1, of the four distinct types that are identified in the estimation, one type, comprising 17 percent of the sample, completes about four more years of schooling than any of the others, over 16 years on average, and specializes in white-collar employment. A second type, comprising another 23 percent of the sample, completes, on average, 12 years of schooling and specializes in blue-collar employment. Expected lifetime utility (not shown) as measured from

Table 2.1. *Selected characteristics at age 24 by type: Nine or ten years of initial schooling*

	Initial Schooling ≦ 9				Initial Schooling ≧ 10			
	Type 1	Type 2	Type 3	Type 4	Type 1	Type 2	Type 3	Type 4
Mean schooling	15.6	10.6	10.9	11.0	16.4	12.5	12.4	13.0
Proportion								
White collar	0.509	0.123	0.176	0.060	0.673	0.236	0.284	0.155
Blue collar	0.076	0.775	0.574	0.388	0.039	0.687	0.516	0.441
Military	0.000	0.000	0.151	0.010	0.000	0.000	0.116	0.005
School	0.416	0.008	0.013	0.038	0.239	0.024	0.025	0.074
Home	0.000	0.095	0.086	0.505	0.050	0.053	0.059	0.325
Sample Propor.	0.010	0.051	0.103	0.090	0.157	0.177	0.289	0.123

Source: Keane and Wolpin (1997), Tables 11 and 13.

age 16 is about the same for the two types, although it is 20 percent higher for the first type as measured from age 26, as lifetime utility from that age becomes dominated by life-cycle earnings paths. The third type, 39 percent of the sample, is the only type to enter the military. Individuals of this type also complete about 12 years of schooling and, given the shortness of the length of military service, spend most of their life in civilian employment, specializing as do those of type 2 in blue-collar employment. However, as compared with those of the second type, who have much the same schooling and employment patterns, because type 3 individuals have considerably lower age 16 levels of premarket skills, their expected lifetime utility is only about three-fifths as great. Those of the last type, the remaining 21 percent, also complete about 12 years of schooling on average, but spend about 40 percent of their lifetime in the home sector, about five times as much as any other group.[48]

As also seen in Table 2.1, the difference in school attainment that has already emerged by age 16, a difference of about 1.4 years, that is due to exogenous events unrelated to endowments, for example, illnesses that occurred in the past or date of birth that affects age at entry given the minimum school entry age, is magnified over time for two of the four types. Although the initial difference is reduced to about 0.8 years for type 1 individuals and is unchanged for type 3 individuals, the completed schooling difference increases to about 2 years for types 2 and 4.

Under the assumption that the four types capture all of the unobserved differences in premarket endowments, it is possible to determine the extent to which conventional measures of family background are able to account for differences in lifetime utilities that arise from age 16 endowments.[49] In the baseline, the difference in expected lifetime utility between the first type (the highest) and the third type (the lowest) is estimated to be 188,000 (1987) dollars. In comparison, the difference for individuals whose mothers were college graduates relative to those whose mothers were high school dropouts is estimated to be $53,000, the difference for those who lived with both parents at age 14 relative to those who lived in a single-parent home is $15,000, and the difference between those whose parental income was more than twice the median relative to those less than one-half the median is $66,000. Although the association of expected lifetime utility with family income and mother's schooling appears to be strong, a regression of expected lifetime utility on those family background characteristics is reported to explain only 10 percent of the variation. Coupled with the fact that 90 percent of the total variation in expected lifetime utility is determined to be due to between-type variation, these family background characteristics actually are rather imperfect proxies for premarket skills.

[48] Given the age 16 endowment levels of type 4 individuals, however, their expected lifetime utility exceeds that of type 3 individuals, who spend more time working in the market.

[49] These correlations may arise from family investment behavior as well as from the intergenerational transmission of innate talents.

4. USE OF STRUCTURAL ESTIMATION OF SCHOOLING CHOICE MODELS FOR THE EVALUATION OF EDUCATION POLICIES

Here I present evidence on the impact of three alternative policies to increase educational attainment: performance-based bonus schemes, college tuition subsidies, and the relaxation of borrowing constraints. For each of these policy interventions, I present evidence obtained from papers that are based on structural estimates of the competitive-equilibrium skill model. Where available, I also provide estimates from studies based on a nonstructural estimation approach. I also present estimates from the structural studies on the longer-term impact of these policies on labor market outcomes. A common theme is that although education policies can significantly influence school completion levels, they cannot have large impacts on labor market success unless they also influence the level of premarket skills.

These papers perform education policy experiments, ignoring general equilibrium feedbacks. I briefly review two recent papers that have made a start at developing solution and estimation methods that can account for the impact of policy-induced skill supply responses on equilibrium skill rental prices.

4.1. Graduation Bonuses

In the case of bonus schemes that provide monetary payments for high school and/or college graduation, which have been implemented in only a few instances and only on a small scale, available estimates come solely from the structural implementation of schooling choice models.[50] Keane and Wolpin (2000) evaluate the effect of two graduation bonus schemes using the previous model. One such plan, which has been advocated recently by Robert Reich, the former Secretary of Labor, provides a $25,000 bonus for high school graduation to all youths whose parents earn less than 120 percent of the median household income. A second scheme seeks to equalize the schooling distributions of black and white males through the combined use of a high school graduation bonus and a college graduation bonus.

Keane and Wolpin estimated the model pooling data on black and white males from the National Longitudinal Survey of Youth (NLSY). Interestingly, they found that observed differences in schooling and employment behavior by race could be accounted for solely by differences in the constant terms in the white- and blue-collar wage functions and by differences in type proportions. The simulated effects of the two graduation bonus experiments on schooling

[50] As I have discussed elsewhere in the context of the policy uses of discrete choice dynamic programming models (Wolpin, 1997), a major advantage of structural estimation is the ability it provides to perform counterfactual policy experiments that entail extrapolations outside of the current policy regime.

Table 2.2. *Effect of school performance-based bonuses on completed schooling and LM success by race*

	Baseline		Reich Proposal[a]		Keane and Wolpin[b]
	White	Black	White	Black	Black
Mean HGC	13.0	12.0	13.5	12.8	12.8
%HGC					
<12	26.2	37.9	7.6	12.1	26.0
=12	29.6	31.5	39.6	48.8	31.6
≥13 and ≤15	19.1	17.9	26.2	26.4	16.2
≥16	25.1	12.8	26.7	15.8	26.1
Age 40 LM outcomes					
Ann. earnings	32,635	24,626	32,687	26,623	25,414
% Employed					
White collar	39.5	26.1	42.0	32.4	30.0
Blue collar	58.7	68.3	56.0	63.2	65.1
PDV life. earnings[c]	285,400	210,258	290,393	228,850	213,885

Note: HGC = highest grade completed; PDV = present discounted value; LM = labor market.

[a] $25,000 bonus for high school graduation restricted to family income ≤120% of the median.

[b] $6,250 bonus for high school graduation, $15,000 for college graduation.

[c] Discounted to age 16.

Source: Keane and Wolpin (1999a), various tables. LM effects for the Reich proposal were computed by the author.

and labor market outcomes, based on the estimates from the pooled data, are reported in Table 2.2.

As seen in the first two columns of the table, in the baseline simulation, black males complete one less year of schooling than white males, 12.0 vs. 13.0, and they earn 75 percent of what white males earn at age 40 (74 percent in present discounted value over their lifetimes).[51] The second two columns report the effect of implementing the Reich proposal. The effect of that proposal is estimated to increase average schooling substantially, for white males by 0.5 years and for black males by 0.8 years. However, the effect of that plan is concentrated on high school graduation rates, reducing the proportion of white males who do not graduate from 26.7 percent to 7.6 percent and that of black males from 37.9 percent to 12.1 percent.[52]

[51] The model is shown to fit the data well. The fit is contrasted to that of a model in which individuals solve a static optimization problem. Although the within-sample fit of the static model is only slightly worse than the fit of the forward-looking model, the static model out-of-sample forecasts are incredible; for example, the average white-collar wage is forecasted to rise to $250,000 by age 50.

[52] Notice, however, that almost half of those who graduate from high school as a result of the bonus go on to complete some college.

Although schooling is increased appreciably, the impact of the bonus on labor market outcomes is small for both races, although somewhat larger for blacks. For example, white-collar employment of white males increases at age 40 by only 2.5 percentage points, earnings at age 40 increase by less than 1 percent, and the present value of lifetime earnings by less than 2 percent. Even though the skill production function estimates imply that an additional year of schooling increases white-collar wages by 7 percent and blue-collar wages by only 2 percent, those induced to obtain the additional schooling by the bonus are nevertheless, given their premarket skills, still better off in blue-collar occupations and those that do enter the white-collar occupation do so with a lesser amount of white-collar premarket skill.[53]

In the second experiment, reported in the last column of Table 2.2, both high school and college graduation bonuses are provided to black males at levels that are sufficient to essentially equalize the black and white distributions of schooling (80 percent of the gap in mean school attainment is closed).[54] However, analogous to the Reich proposal, although the schooling gap is closed, the effect on labor market outcomes is quite small. For example, white-collar employment among whites still exceeds that of blacks at age 40 by 9.5 percentage points, the black–white male earnings gap at age 40 is reduced only from 25 to 22 percentage points, and the difference in the present discounted value of lifetime earnings only by 4.8 percent. Race differences in age 16 endowments account for much of the remaining gap.[55]

4.2. Tuition Effects

In contrast to a graduation bonus scheme, which rewards individuals for years of schooling that are completed, tuition subsidies are based only on attendance. It is more conventional in the nonstructural literature to estimate tuition effects because tuition levels vary over time and cross sectionally. Keane and Wolpin (1997) simulate the effect of an experiment that provides a tuition subsidy of approximately 50 percent for each year of college attendance.[56] They report an

[53] This result is consistent with the findings in Cameron and Heckman (1998). Using a quite different framework, they find that those who are induced to attend college by a subsidy policy are of less ability than those who attend without the subsidy. They conclude that ignoring heterogeneity will lead to an overstatement of the impact of such policies on labor market outcomes.

[54] The exact bonuses are, in 1987 dollars, $6,250 for high school graduation and $15,000 for college graduation.

[55] Because the constant terms in wage functions are the sum of occupation-specific skill rental prices and age 16 endowments of premarket skills, it is not possible to disentangle the role of discrimination as distinct from differences in premarket skills. However, it is possible to establish an upper bound estimate of the component caused by discrimination, which turns out to be around 30 percent.

[56] Keane and Wolpin treat college tuition as a parameter. It is estimated to be $4,186 (in 1987 dollars).

increase in mean schooling of 0.5 years, from 13.0 to 13.5 years, an increase in the proportion graduating from high school (without attending college) by 3.5 percentage points, and an increase in college graduation rates of 8.4 percentage points.[57]

The model on which these estimates are based is silent with respect to the financing of college costs. The assumption of linear utility implies that the capital market environment is irrelevant, essentially permitting individuals to make school attendance decisions independent of financing considerations. In another paper, Keane and Wolpin (2001) drop the linear utility assumption and account explicitly for borrowing constraints. Uncollateralized borrowing is permitted, but is constrained to be less than some given amount, the net borrowing limit, that varies with age and other characteristics. The function describing the net borrowing limit is estimated. In addition, the borrowing rate of interest is not restricted to be equal to the lending rate. Several other potential mechanisms for financing college costs are modeled, through one's own savings, parental transfers, and current earnings.

More realistically, in this model, individuals are assumed to make part- and full-time work and school attendance decisions and may work while attending school. Unobserved heterogeneity, as before, takes the form of there being a fixed number of types who differ in their premarket skills, their net consumption value of school attendance, and the value they attach to the home sector. Although there is only one occupation option, individuals may exhibit comparative advantages across work, school, and home sectors. Parents make monetary transfers to coresident children that depend on their own schooling and that may be greater when youths are attending college.

Although this model differs considerably from the previous one, the impact of a college tuition subsidy is roughly the same magnitude. Subsidizing tuition costs fully would increase completed schooling by 1.1 years.[58] Interestingly, these estimates agree with those from the nonstructural literature. Table 2.3 presents a summary of estimates of the percentage change in college enrollment rates induced by a tuition increase of $100 per year, the standard metric used for calibrating tuition effects in the literature. As seen, Leslie and Brinkman (1987), in a survey of twenty-five empirical studies based on cross-state or time-series variation in college costs, report the modal estimate of the response to a $100 tuition increase (in 1982–1983 dollars) to be a 1.8 percent decline in the enrollment rate of 18- to 24-year-olds. Other estimates range from a low of 0.8 percent for 18- to 24-year-old whites to a high of 2.2 percent for 18- to 24-year-old blacks (Kane, 1994). In Keane and Wolpin (1999b), the comparable effect is a decline of 1.2 percent.

[57] The increased high school graduation rate occurs because of the forward-looking nature of the model.

[58] The college tuition cost is estimated to be (in 1987 dollars) $3,673.

Table 2.3. *Effects of college tuition changes on schooling*

	Tuition Change	Treatment Group	Schooling Effect
Partial Equilibrium			
1. Completed schooling			
Keane and Wolpin (1997)	50% subsidy	white males: NLSY cohort	0.5 year increase
Keane and Wolpin (1999b)	80% subsidy	white males: NLSY cohort	1.1 year increase
Lee (2000)	50% subsidy	males, females: 1957–1964 cohort	0.5, 0.9 year increase
2. College enrollment rates			
Leslie and Brinkman (1987)	$100 increase in tuition[a]	18- to 24-year-olds: var. samples	1.8% decline
St. John (1990)	$100 increase in tuition[a]	18- to 19-year-olds	0.9% decline
Kane (1994)	$100 increase in tuition[a,b]	18- to 19-year-old males: white, black	0.8%, 2.2% decline
Keane and Wolpin (1999b)	$100 increase in tuition[a,b]	18- to 19-year-old white males	1.3% decline
		18- to 24-year-old white males	1.2% decline
Heckman et al. (1999)	$100 increase in tuition[a,b]	white males: NLSY cohort	1.6% decline
Lee (2000)	$100 increase in tuition[a]	18- to 19-year-olds: males, females	1.12%, 1.66% decline
		18- to 24-year-olds: males, females	1.34%, 1.95% decline
		1957–1964 cohort	
General Equilibrium			
College enrollment rates			
Heckman et al. (1999)	$100 increase in tuition[a,b]	18- to 19-year-old white males	0.16% decline
Lee (2000)	$100 increase in tuition[a]	18- to 19-year-olds: males, females	1.05%, 1.52% decline
		18- to 24-year-olds: males, females	1.27%, 1.86% decline
		1957–1964 cohort	

[a] 1982–1983 dollars.

[b] Adjusted for comparability.

4.3. Relaxing Borrowing Constraints

As discussed in Keane and Wolpin (2001), it is also common to report effects of tuition increases separately by the income quantiles of the youth's parents. These studies typically find much larger tuition effects for low-income youths. As seen in Table 2.3, for instance, St. John (1990) estimates that a $100 tuition increase (in 1982–1983 dollars) lowers the enrollment rate of 18- to 19-year-old high school graduates by roughly 0.85 percent. However, he also reports that the enrollment rate would drop by 1.1 percent for youths from families with income below $40,000, but only by 0.4 percent for youths from families with higher income. Manski and Wise (1983) find, for the same age group, that a $100 tuition increase (in 1982–1983 dollars) leads to a large decline, by 3.6 percent, in the enrollment rate among youths whose parents are in (roughly) the bottom income quintile, while they find much smaller effects for youths from higher-income families. On the basis of more recent data, Kane (1994) estimates that a $100 increase (in 1982–1983 dollars) leads to a decline in college enrollment rates of roughly 1.4, 1.0, 0.5, and 0.2 percent for 18- to 19-year-old white male high school graduates whose parents are in the first through fourth income quantiles, respectively.

The finding that tuition effects are inversely related to parental income has often been interpreted as evidence for the existence of borrowing constraints that have adverse consequences for college attendance (see, e.g., Kane, 1999, p. 63). In the model estimated in Keane and Wolpin (1999b), borrowing constraints are found to be quite binding. The (net) debt limit is estimated to be very low, at most $1,000, and to differ only little by the individual's human capital. Financing college tuition through uncollateralized borrowing is therefore not feasible. In addition, consistent with the pattern found in the nonstructural literature, Keane and Wolpin report that a simulated $100 annual tuition increase (in 1982–1983 dollars) leads to declines in enrollment rates (for 18- to 19-year-old high school graduates) of 2.2, 1.9, 1.5, and 0.8 percent, respectively, if the youth's parents are in each of four ascending education categories (both are high school dropouts, at least one is a high school graduate, at least one has some college, at least one is a college graduate). Thus, the model generates a pattern of larger percentage declines in enrollment for youths whose parents have lower socioeconomic status.

On the surface, it would appear that the inference drawn in the nonstructural literature, that borrowing constraints exist and limit college attendance of youths from less affluent families, is validated by the congruence of these two findings. However, when Keane and Wolpin simulate the impact of relaxing the borrowing constraints, by allowing youths to borrow the full tuition cost, they find that there is only a negligible increase in college attendance.[59] Instead,

[59] Keane and Wolpin (2001) also find that, on average, youths receive a transfer from their parents, in excess of what is received when not attending college, sufficient to fully subsidize college tuition costs. The subsidy ranges from about one-half of the tuition cost for youths whose parents are the least educated (neither is a high school graduate) to almost twice the tuition cost for youths whose parents are the most educated (at least one parent is a college graduate). It might appear that it is because of the largesse of parents that relaxing borrowing constraints has only a minimal impact on college attendance. However, simulating the impact of relaxing the borrowing constraint in a regime where parents are assumed to provide no additional transfers to children who attend college leads to the same result.

allowing college attendees to borrow the full amount of their tuition costs leads to a reduction in their propensity to work while attending school and to an increase in their consumption.[60] College attendance is limited not by borrowing constraints, but rather primarily by age-16 endowments of premarket skills and/or preferences.[61]

5. GENERAL EQUILIBRIUM

The policy effects described herein assume that general equilibrium feedbacks are negligible. However, policies that serve to increase schooling levels, and thus the aggregate skill level, will reduce the skill rental price in equilibrium. Partial equilibrium estimates of the impact of those policies on schooling would, therefore, overstate the general equilibrium impact. There is at this point only limited evidence on the extent of the overstatement.

Consider the Keane and Wolpin (1997) framework in which occupation-specific skill rental prices are constant over time. A setting consistent with that assumption is one in which there are no aggregate production shocks and in which there is a stationary population with an unchanging age distribution. In that case, aggregate skill in an occupation at any calendar time is simply the sum of the skill amounts of individuals choosing to work in the occupation at that time. Rental prices are then given by the skill marginal products evaluated at aggregate skill levels. Equilibrium rental prices are those that yield the aggregate skill levels consistent with individual choices.

More concretely and for simplicity, assume that there is only a single occupation and that production is Cobb–Douglas, that is, $Y = AS^{\alpha}K^{1-\alpha}$. Aggregate skill is given by the sum of the skill levels over individuals, namely $S = \sum_a \sum_{j=1}^{N_a} s_{ja} = \sum_a \sum_{j=1}^{N_a} s_{0j} \exp[\beta_1 E_{ja} + \beta_2 P_{ja} - \beta_3 P_{ja}^2 + \varepsilon_{ja}]$, where N_a is the number of individuals of age a in the population. As noted, it is not possible to separately identify, from data on choices and wages, both the skill rental price and the level of premarket skills. Moreover, given the nonobservability of skill, the technology shifter, A, also cannot be identified. A baseline case can

[60] Cameron and Heckman (1998) reach a similar conclusion based on a sequential statistical decision model of schooling choice. They find that family income has only a small effect on college attendance conditional on long-term factors such as family background and permanent income. They conclude that short-term credit constraints are therefore unimportant in explaining schooling decisions. Cameron and Taber (2000) develop tests for the importance of borrowing constraints in schooling decisions based on the proposition that direct schooling costs would have a greater impact than opportunity costs if borrowing constraints were operative. As is consistent with the findings of Keane and Wolpin and Cameron and Heckman, they find no evidence that borrowing constraints play an important role.

[61] That is not to say that differentials in parental transfers are unimportant. However, as Keane and Wolpin report, equalizing parental transfers would reduce the difference in the mean schooling of youths whose parents are in the highest vs. the lowest education group only from 3.8 years to 2.6 years.

be established by normalizing either A or the rental price. With the assumption that $A = 1$, a baseline equilibrium rental price can be established that induces a level of aggregate skill that equates the marginal product of skill to that rental price. The general equilibrium impact of a tuition subsidy on schooling is determined by resolving jointly for the equilibrium rental price and optimal choices.

Donghoon Lee (2000) has implemented this procedure in the multiskill setting, allowing as well for nonstationarity in population size, that is, for changing cohort size. Lee adopts an overlapping generations version of the occupational and schooling choice model of Keane and Wolpin (1997), assuming that individuals have perfect foresight about equilibrium outcomes, but imperfect foresight about their idiosyncratic shocks to preferences and skills. Each cohort therefore faces a known, but different, sequence of occupation-specific skill rental prices. Individuals are assumed to face exogenous cohort and age-specific fertility rates and a constant college tuition cost. Aggregate production at any calendar time depends on the aggregate level of white-collar skill, the aggregate level of blue-collar skill, and capital. Capital grows exogenously at a known rate. The production function is Cobb–Douglas with time-varying factor shares (technology parameters), about which individuals also have perfect foresight.

Lee estimates the model by using aggregate data from the 1968–1993 Current Population Surveys (CPSs).The model is fit to the empirical moments of school enrollment rates, occupation-specific employment rates, and educational attainment by age, year, and sex and to wage moments by age, year, sex, and educational attainment. As seen in Table 2.3, the partial equilibrium effects are in line with those of other studies. Similar to Keane and Wolpin (1997), Lee finds that a 50 percent subsidy to college costs would increase completed schooling by 0.6 years for males (and by 0.9 for females). Further, the partial equilibrium effect of a $100 tuition increase reduces college enrollment rates by 1.12 percent for 18- to 19-year-old males and by 1.34 percent for 18- to 24-year-old males. General equilibrium effects on completed schooling for the 50 percent subsidy are 87 percent of the partial equilibrium effect for males and 93 percent for females, whereas they range from 92 to 95 percent of partial equilibrium effects on college enrollment rates for the $100 tuition increase.

In an earlier series of papers, Heckman et al. (1998, 1999) performed a similar comparison based on a general equilibrium model with different characteristics. The data and estimation methods they adopted also differ significantly from Lee. As with Lee, I only briefly describe the essential features of their model and the estimation method. Specifically, in their model, (1) skill classes are defined by educational attainment; (2) in addition to making an initial discrete schooling decision, in each postschooling period individuals explicitly decide on their human capital investment time on the job à la Ben-Porath; (3) individuals make savings decisions in each period; and (4) labor is inelastically supplied in each period.

As they noted, accounting in estimation for the restrictions that general equilibrium imposes given the structure of their model is not feasible with available data. In particular, estimation is hindered by the lack of microdata on consumption and on investment time on the job. For that reason the procedure adopted for recovering the parameters of the model involves a combination of calibration based on results from existing studies and new estimation using both aggregate CPS data (over the period from 1963 to 1993) and longitudinal data from the NLSY.[62] As seen in Table 2.3, the partial equilibrium effect of a $100 increase in college tuition costs is 1.6 percent, well within the range of other studies and not much different from that of Lee or Keane and Wolpin. However, the general equilibrium effect is reduced by an order of magnitude, to 0.16 percent.

Clearly, the bound established by the Heckman et al. and the Lee papers is extreme. Accounting for the divergence in their results is problematic because of the major differences in their modeling and estimation approaches. Clearly, further study is warranted.

6. CONCLUSIONS

Within the competitive market–human capital production paradigm, in a setting in which there is a single homogeneous skill, the wage equation is a fundamental structural relationship. Knowledge of its parameters, therefore, provides information useful in understanding human capital investment behavior and in analyzing the impact of policy interventions. Indeed, in early pioneering models of schooling choice, knowledge of the parameters of the wage equation was both necessary and sufficient to perform education policy analysis. However, as I have discussed, in richer models of schooling and on-the-job skill acquisition, such knowledge is necessary, but not sufficient.

Although it is not obvious that richer models are indeed necessary for evaluating specific policy interventions, models of schooling and employment choices that embed wage equations as part of a larger structure have yielded important empirical results and allow the evaluation of a wider range of policies. In those models, wage equation estimation is not an end in itself. It would seem to me fruitful to turn attention to refining and expanding behavioral models of human capital accumulation in both partial and general equilibrium contexts.

ACKNOWLEDGMENTS

This paper was prepared for the 2000 World Congress of the Econometric Society.

[62] It is not possible to describe in limited space the estimation procedure, and I would simply refer the reader to the paper for details.

References

Angrist, J. and A. B. Krueger (1991), "Does Compulsory School Attendance Affect Schooling and Earnings," *The Quarterly Journal of Economics*, November, 106(4), 979–1014.

Becker, G. (1967), *Human Capital and the Personal Distribution of Income*. Ann Arbor: University of Michigan Press.

Bellman, R. E. (1957), *Dynamic Programming*. Princeton, NJ: Princeton University Press.

Ben-Porath, Y. (1967), "The Production of Human Capital and the Life Cycle of Earnings," *Journal of Political Economy*, August, 75, 352–365.

Bontemps, C., J.-M. Robin, and G. G. J. Van den Berg (1999), "An Empirical Equilibrium Search Model with Search on the Job and Heterogeneous Workers and Firms," *International Economic Review*, November, 40, 1039–1074.

Bound, J. and D. Jaeger (1996), "On the Validity of Season of Birth as an Instrument in Wage Equations: A Comment on Angrist & Krueger's 'Does Compulsory School Attendance Affect Schooling and Earnings,'" November, NBER Working Paper W5835.

Bowlus, A. J. and Z. Eckstein (1999), "Discrimination and Skill Differences in an Equilibrium Search Model," mimeo, University of Western Ontario.

Bowlus, A. J., N. M. Kiefer, and G. R. Neumann (forthcoming), "Equilibrium Search Models and the Transition from School to Work," *International Economic Review*.

Burdett, K. (1978), "Employee Search and Quits," *American Economic Review*, March, 68, 212–220.

Burdett, K. and D. T. Mortensen (1998), "Wage Differentials, Employer Size, and Unemployment," *International Economic Review*, May, 39, 257–273.

Cameron, S. V. and J. J. Heckman (1998), "Life Cycle Schooling and Dynamic Selection Bias: Models and Evidence for Five Cohorts of American Males," *Journal of Political Economy*, April, 108, 262–334.

Cameron, S. V. and C. Taber (2000), "Borrowing Constraints and the Returns to Schooling," May, mimeo, Northwestern University.

Eckstein, Z. and K. I. Wolpin (1989), "The Specification and Estimation of Dynamic Stochastic Discrete Choice Models," *Journal of Human Resources*, Fall 24, 562–598.

Eckstein, Z. and K. I. Wolpin (1990), "Estimating a Market Equilibrium Search Model from Panel Data on Individuals," *Econometrica*, July, 58, 783–808.

Fang, H. (1999), "Disentangling the College Wage Premium: Estimating a Model with Endogenous Education Choices," mimeo, University of Pennsylvania.

Griliches, Z. (1977), "Estimating the Returns to Schooling: Some Econometric Problems," *Econometrica*, January, 45, 1–22.

Heckman, J. J. (1981), "The Incidental Parameters Problem and the Problem of Initial Conditions in Estimating a Discrete Time–Discrete Data Stochastic Process and Some Monte Carlo Evidence," in *Structural Analysis of Discrete Data with Econometric Applications* (ed. by C. F. Manski and D. McFadden), Cambridge, MA: MIT Press.

Heckman, J. J. (1997), "Instrumental Variables," *Journal of Human Resources*, Summer, 32, 441–462.

Heckman, J. J., L. Lochner, and C. Taber (1998), "Explaining Rising Wage Inequality: Explorations with a Dynamic General Equilibrium Model of Labor Earnings with Heterogeneous Agents," *Review of Economic Dynamics*, January, 1, 1–58.

Heckman, J. J., L. Lochner, and C. Taber (1999), "General-Equilibrium Cost-Benefit Analysis of Education and Tax Policies." in *Trade, Growth and Development: Essays in Honor of T. N. Srinivasan*, (ed. by G. Ranis and L. K. Raut), Amsterdam: Elsevier Science.

Heckman, J. J., A. Layne-Farrar, and P. Todd (1996), "Human Capital Pricing Equations with an Application to Estimating the Effect of Schooling Quality on Earnings," *Review of Economics and Statistics*, November, 78, 562–610.

Heckman, J. J. and G. Sedlacek (1985), "Heterogeneity, Aggregation, and Market Wage Functions: An Empirical Model of Self-Selection in the Labor Market," *Journal of Political Economy*, December, 93, 1077–1125.

Heckman, J. J. and E. Vytlacil (1998), "Local Instrumental Variables," mimeo, University of Chicago.

Kane, T. J. (1994), "College Entry by Blacks Since 1970: The Role of College Costs, Family Background, and the Returns to Education," *Journal of Political Economy*, October, 102(5), 878–911.

Kane, T. J. (1999), "Where Should Federal Education Initiatives Be Directed?" in *Financing College Tuition: Government Policies and Educational Priorities*, (ed. by M. H. Kosters), Washington D.C: The ABI Press.

Keane, M. P. and K. I. Wolpin (1994), "The Solution and Estimation of Discrete Choice Dynamic Programming Models by Simulation and Interpolation: Monte Carlo Evidence," *Review of Economics and Statistics*, November, 76, 648–672.

Keane, M. P. and K. I. Wolpin (1997), "Career Decisions of Young Men," *Journal of Political Economy*, June, 105, 473–522.

Keane, M. P. and K. I. Wolpin (2000), "Eliminating Race Differences in Educational Attainment and Labor Market Success," *Journal of Labor Economics*, October, 18, 614–652.

Keane, M. P. and K. I. Wolpin (2001), "The Effect of Parental Transfers and Borrowing Constraints on Educational Attainment," *International Economic Review*, November, 42, 1051–1103.

Lee, D. (2000), "An Estimable Dynamic General Equilibrium Model of School, Work and Occupational Choice," July, mimeo, University of Pennsylvania.

Leslie, L. L. and P. T. Brinkman (1987), "Student Price Response in Higher Education: The Student Demand Studies," *Journal of Higher Education*, March/April, 55(3), 181–204.

Manski, C. F. (1997), "Monotone Treatment Response," *Econometrica*, November, 65, 1311–1334.

Manski, C. F. and D. A. Wise (1983), *College Choice in America.* Cambridge, MA: Harvard University Press.

Mincer, J. (1958), "Investment in Human Capital and Personal Income Distribution," *Journal of Political Economy*, March/April, 281–302.

Mincer, J. (1962), "On-the-Job Training: Costs, Returns and Some Implications," *Journal of Political Economy*, January/February, 70, 50–79.

Mincer, J. (1974), *Schooling, Experience, and Earnings*. New York: NBER.

Mortensen, D. T. (1990), "Equilibrium Wage Distributions: A Synthesis," in *Panel Data and Labor Market Studies*, (ed. by G. Ridder and J. Theeuwes), Amsterdam: North Holland.

Mortensen, D. T. (1986), "Job Search and Labor Market Analysis," in *Handbook of Labor Economics*, Vol. 2, (ed. by O. Ashenfelter and R. Layard), Amsterdam: North Holland.

Mortensen, D. T. and G. Neumann (1988), “Choice or Chance? A Structural Interpretation of Individual Labor Market Histories,” in *Labor Market Dynamics*, (ed. by G. Neumann and N. Westergaard-Nielsen), Heidelberg: Springer-Verlag.

Mortensen, D. T. and C. Pissarides (2000), “New Developments in Models of Search and the Labor Market,” in *Handbook of Labor Economics*, Vol. 3, (ed. by O. Ashenfelter and D. Card), Amsterdam: North Holland.

Robin, J. M. and S. Roux (1993), “Random or Balanced Matching: An Equilibrium Search Model with Endogenous Capital and Two-Sided Search,” CREST-INSEE Working Paper 9838.

Rosen, S. (1977), “Human Capital: A Survey of Empirical Research,” in *Research in Labor Economics*, Vol. 2, (ed. by R. Ehrenberg), Greenwich, CT: JAI Press.

Rosenzweig, M. R. and K. I. Wolpin (2000), “‘Natural’ Natural Experiments in Economics,” *Journal of Economic Literature*, 38, no. 4, 827–871.

Roy, A. D. (1951), “Some Thoughts on the Distribution of Earnings,” *Oxford Economic Papers*, June, 3, 135–146.

Rust, J. (1994), “Structural Estimation of Markov Decision Processes,” in *Handbook of Econometrics*, Vol. 4, (ed. by R. Engle and D. McFadden), Amsterdam: North Holland.

Rust, J. (1996), “Numerical Dynamic Programming in Economics,” in *Handbook of Computational Economics*, Vol. 1, (ed. by H. M. Amman, D. A. Kendrick, and J. Rust), Amsterdam: North-Holland.

St. John, E. P. (1990), “Price Response in Enrollment Decisions: An Analysis of the High School and Beyond Cohort,” *Research in Higher Education*, 31, 161–176.

Van den Berg, G. G. J. (1999), “Empirical Inference with Equilibrium Search Models,” *Economic Journal*, 109, F283–306.

Van den Berg, G. G. J. and G. Ridder (1998), “An Empirical Equilibrium Search Model of the Labor Market,” *Econometrica*, September, 66, 1183–1221.

Willis, R. J. (1986), “Wage Determinants: A Survey and Reinterpretation of Human Capital Earnings Functions,” in *Handbook of Labor Economics*, Vol. 1, (ed. by O. Ashenfelter and R. Layard), Amsterdam: North-Holland, 525–602.

Willis, R. J. and S. Rosen (1979), “Education and Self-Selection,” *Journal of Political Economy*, October, Supplement, 87, S7–S36.

Wolpin, K. I. (1997), “Public-Policy Uses of Discrete-Choice Dynamic Programming Models,” *Papers and Proceedings of the American Economic Association*, May, 427–432.

Empirical and Theoretical Issues in the Analysis of Education Policy

A Discussion of the Papers by Raquel Fernández and by Kenneth I. Wolpin

Costas Meghir

1. INTRODUCTION

The papers by Raquel Fernández and Ken Wolpin form an excellent overview of important theoretical and empirical issues relevant for the analysis of policies designed to promote the development of human capital. However, their angle is quite different, with the former concentrating on sorting and the funding of schools (public vs. private) and the latter discussing ways of specifying empirical models that would allow a systematic evaluation of policies toward human capital development.

The paper by Raquel Fernández focuses on how best to finance education in view of the fact that individuals may sort on income, ability, or both. She seeks to establish a framework for analyzing the basic question of whether publicly- or privately-funded education is best from a welfare point of view. She discusses the impact of these different designs on efficiency and on inequality under different assumptions about complementarities in the human capital and goods production function as well as under the presence of liquidity constraints and sorting restrictions. In her context, sorting may not lead to efficient outcomes. She examines conditions under which public education may lead to improvements in efficiency.

The paper by Ken Wolpin emphasizes the importance of clearly specifying the framework within which empirical analysis is undertaken. He shows that in the absence of a well-defined framework, policy conclusions can be confusing and misleading. He then discusses empirical models of wages and education choice under different assumptions about the labor market. The ultimate aim of such a research agenda is to come up with a suitable framework for analyzing education policy.

Although the papers are broadly related by the underlying theme of education policy, they raise sufficiently different issues to merit separate discussion. I discuss the Fernández paper first and the Wolpin paper second.

2. SORTING, EDUCATION, AND INEQUALITY

Fernández uses a number of different models to address the issue of funding education when individuals sort by ability or income. The prototype is a static model of sorting in which preferences can be thought of as a reduced form of an altruistic model, education is locally financed, and there is an externality, in that the quality of education is an increasing function of the average income in the group. There are fewer communities than types of individuals or income levels. This last point implies that the sorting equilibrium is not Pareto efficient. In this context, a redistributive tax inducing middle-income individuals to move from the high-income area to the low-income one (thus raising average income in both) can lead to improvements in welfare. Of course, one has to ask why the number of school or community types would be limited in the first place and whether there is something in the education production function that requires some minimum breadth in the coverage of schools – in other words, some heterogeneity may be good. However, the main point of the author is that in a world in which perfect sorting cannot take place, a public intervention can be Pareto improving.

The rest of the paper is along the same theme but considers richer models with different types of imperfections that justify a preference for public rather than private funding of education.

The model by Fernández and Rogerson, for example, highlights the difference between dynamic and static considerations for efficiency. In that model, there are liquidity constraints preventing borrowing from future earnings, which would allow funding an efficient level of education. Parents differ in terms of income, a difference that is perpetuated across generations, but otherwise individuals are homogeneous, in particular in terms of ability. In that model, from a static point of view, private education is efficient because individuals obtain the amount of education that is optimal given parental preferences. However, from a dynamic point of view, public education can be efficient, because it allows a redistribution of resources from families with low marginal utility of education to those with high. In a sense the rich are "overeducated" in the sorting equilibrium.

A richer, more general model discussed by Fernández is that of Bénabou (1996). This combines liquidity constraints, complementarities at the local level in the production of human capital (peer effects), and complementarities at the aggregate level in the production of output. Initial endowments are the only source of heterogeneity. In that model, there is a tension between (a) the positive effects that assortative mating has on the production of human capital (leading to the local education model being more efficient) and (b) the complementarities in production that require sharing of all resources, leading to efficiency without any segregation. In the short run, local funding of education produces more human capital and is more efficient. However, in the long run, the growth that is due to the complementarities in production outweigh this consideration: State financing raises the level of long-run human capital. When we include uncertainty in the production of human capital, the additional insurance provided by the state system raises the growth rate as well.

The paper describes a number of further interesting models. It would be wasteful to describe them here when the author has done such a clear job of presenting them in her paper. Instead I focus on the empirical questions raised by this rich set of theories.

The main theme of almost all models presented here is either the presence of liquidity constraints or the presence of peer effects in education.

Both of these issues have recently been the center of attention. Consider peer effects first. There are not that many convincing empirical studies that establish the importance or otherwise of peer effects. Of course, the difficulty in obtaining such results is the endogeneity of the peer group and the absence of any credible exogenous variation that shifts the peer group without affecting the educational outcomes directly. In fact, the paper by Fernández in a sense illustrates the difficulties of identifying the importance (or otherwise) of peer effects, given the endogenous sorting that takes place. A recent paper by Hoxby (2001) measures the impact of the gender and race mix in the classroom by exploiting differences in sex and race composition across cohorts of pupils from year to year. The differences in composition are large, and hence the potential for identification is quite good. In fact, some of her results do show significant impacts of such composition. The gender and the race mix are of course two important dimensions of peer effects, and in particular the race mix is a dimension along which households may sort in practice, perhaps as a result of cultural differences. However, it is not clear that it measures the impact of changes in the ability mix of a class on one individual, which underlies the theory developed in Fernández. Beyond this study there is little hard evidence based on exogenous variation that peer effects matter. Brock and Durlauf (2001) have made some important progress in specifying and estimating a structural model of sorting and peer effects. The model is identified by exploiting restrictions that originate from the structure of the equilibrium. Obviously, a research program that would manage to embed sources of exogenous variation on peer effects with an equilibrium sorting model would be particularly valuable. Finally, Dale and Krueger (1999) compare students admitted to the same set of colleges but who for (assumed) random reasons ended up in different ones. They conclude from this matching exercise that attendance at a more selective college does not lead to significant improvements in performance; perhaps this may be interpreted as implying that at that age, at least, peer effects do not matter as much. In my view, the importance of peer effects of the sort assumed in the theoretical work, however plausible, remains an open empirical question. Given the central importance it has in the design of education policy, it should constitute a research priority.

The other major and possibly quite controversial issue is the presence and relative importance of liquidity constraints in education. There are very difficult identification and conceptual issues to be resolved before this issue can be put to rest. The literature tends to focus on the impact of liquidity constraints on college attendance and generally late educational outcomes, such as staying on at school beyond the age of compulsory schooling. In a recently published paper, Cameron and Heckman (2001) presented empirical evidence showing quite persuasively that liquidity constraints do not prevent college attendance.

It is ability as measured in the teenage period that seems to be the key factor in obtaining college education. Cameron and Taber (2000) reach a similar conclusion. They use an indirect test that compares results obtained by using instruments that affect opportunity cost of attending school with results obtained by using instruments that affect direct costs of attendance. Keane and Wolpin (1997), in contrast, find liquidity constraints but also conclude that they do not have a large impact on college attendance.

Much of the evidence relates to the United States. It is, of course, important to collect evidence on a broader scale from both industrialized countries (with different institutional frameworks) and from developing countries.

Meghir and Palme (2001) present evidence that, in Sweden, individuals from poorer backgrounds were induced into further education after a reform was made that increased compulsory schooling and offered subsidies to education. Importantly, among those individuals who were induced into extra schooling, some were of higher than median ability, although the majority were low ability (as measured at 12 years of age). However, only those of higher ability benefited substantially and significantly (in terms of earnings) from the reform. This could be interpreted in a number of ways. However, one plausible interpretation is that these high-ability individuals from poorer backgrounds were somehow constrained prereform and did not pursue further education, which was clearly profitable to do (unless they were also facing high costs). Nevertheless, the number of high-ability individuals from poorer backgrounds who were "held back" and stopped school early before the reform was relatively small (but significant).

As Cameron and Heckman emphasize, results generally point to the importance of early educational investments and outcomes as the single most important factor in determining future attainment. This is precisely where we need to focus if we are to identify the reasons for low investments and low educational outcomes, that is, at the early formative years. And this is where the difficulty arises: It is very difficult to disentangle long-term financial hardship of parents from other factors that lead to low investments in human capital. Finding long-term exogenous increases in income (e.g., caused by the introduction of a welfare program) unrelated to ability may be one key way of achieving identification. The Progresa welfare program in Mexico may provide some information in that direction because it provides unconditional (as well as school-attendance contingent) grants to a randomly selected set of villages. Testing for liquidity constraints would involve identifying the impact of the unconditional grant on schooling. However, the presence of the school-contingent grant changes the education price in the eligible villages, making the use of these excellent data for this particular purpose harder (see Attanasio, Meghir, and Santiago, 2001).

In conclusion, although there is evidence that liquidity constraints at later ages have little impact on college attendance in the United States, it is still possible that low income leads to lower investments in human capital at early ages, which in itself may have a long-term impact by shaping ability. This can be interpreted as a form of liquidity constraint. Again, this is a very important empirical issue central to evaluating the importance of the points raised by Fernández in her article.

The next issue that is raised by the Fernández paper relates to what the key inputs are that define quality of education (class size, quality of teachers, etc.). Card and Krueger (1992), Heckman, Layne-Farrar, and Todd (1996), Angrist and Lavy (1996), Hanushek (1996), Krueger (1999), Krueger and Whitmore (2001), Dearden, Ferri, and Meghir (2002; for the UK), and many others discuss these issues and present empirical evidence, but no consensus arises (see the special issue of the *Review of Economics and Statistics* for details). Although most papers looking at labor market outcomes show little impact of class size, for example (within a reasonable range), some do show large impacts on tests scores and college attendance (Angrist and Lavy, 1996; Krueger, 1999; and Krueger and Whitmore, 2001). Hanushek, Kain, and Rivkin (1998) have argued that by far the most important input is the quality of teachers. A key question here is whether incentive contracts for teachers help either by leading to a better (self) selection of teachers or by inducing more effort. Ultimately, what we need to evaluate and reach a consensus on is the impact of school inputs on labor market outcomes. Again this seems to be an open question.

Finally, an important issue relates to whether a good public school system is sustainable in the presence of a private sector. If individuals sort by income and if ability and income are correlated (say, because of early investments in education), then as shown in the paper by Epple and Romano referred to in Fernández, public schools will include the lowest-income individuals and possibly the lowest-ability ones. Consequently, optimal public policy should take this into account.

3. WAGE EQUATIONS AND EDUCATION POLICY

There are two principal themes in the Wolpin paper. The first relates to the interpretation of wage equations under different assumptions on the labor market. This issue is often ignored by the vast literature on the measurement of the returns to education. Wolpin uses different examples to show how the coefficient on the education variable changes as the assumption on the structure of the labor market changes. For example, if the labor market is competitive and human capital is homogeneous and supplied at different quantities by individuals, then the coefficient of education in a log wage equation identifies a parameter of the human capital production function. Alternatively, if there are labor market frictions, then the same coefficient reflects both the parameters of the human capital function as well as the parameters reflecting the search frictions. More generally, the parameters of estimated wage equations are complex functions of the human capital production function, the bargaining structure with the firm, and the distribution of unobserved heterogeneity. This becomes particularly important in the context of match-specific complementarities between workers and firms. A number of papers by Burdett and Mortensen, Ridder, van den Berg, and Robin, among others, have shown the difficulties of characterizing equilibria in this context (see van den Berg, 2000, for a review). In these cases, the returns to experience and tenure acquire quite complex interpretations, all of which have different policy implications.

An additional important issue discussed by Wolpin and raised in a number of papers by Heckman and others is whether we should view human capital as a homogeneous factor with one price or as many factors (see Heckman and Sedlacek, 1985). In this case, the returns to education will be changing over time as the relative demand for one type of education changes vis-à-vis the other types. Ignoring this issue can lead to misleading interpretations of the results.

These issues are important for a number of reasons. First, it affects our interpretation of the impacts policy. If the returns to education we measure are not a basic parameter of the human capital production function but a complex function of bargaining parameters and/or layoff and arrival rates, we have little guidance on whether and how increases in education will ultimately affect earnings. Second, it affects our understanding of returns to education and the reasons it may change from period to period. Third, our understanding of how wages should evolve over time within and across jobs depends on the context and on the equilibrium wage policies of the firm.

Of course, one important question here is how to decide which model is the best description of the labor market. Constructing credible tests for this purpose is difficult, to say the least. There is value in laying out the assumptions of the exercise, but ultimately we need to be able to choose between models when analyzing policy.

The second theme of the paper relates to the estimation of models that would be useful for the analysis of policies that encourage the acquisition of human capital. The discussion is built around the structural models of Keane and Wolpin (1997). In that work, individuals have unobserved skills, but there are no match-specific effects. Occupational-specific rental prices are fixed over time. Individuals choose between several occupations and schooling. Finally, the labor market is assumed to be competitive. There are two model versions. In the first, there are no liquidity constraints. This is relaxed in the second. As already mentioned, the authors do find evidence of liquidity constraints, but relaxing them does not much change college attendance, which is driven to a large extent by unobserved heterogeneity.

The benefits of such models are that the assumptions underlying the results are clear and the mechanism through which education policy may act is made clear. Moreover, there is an explicit treatment of unobserved heterogeneity and ultimately the data determine how important the incentive effects are relative to self-selection induced by unobserved skills. However, one has to be aware that such models rely on what some would consider strong identifying "functional form" assumptions. Nevertheless, one should not underestimate the power of models with a clear set of assumptions that lead to clear interpretations.

A major issue for policy analysis, discussed in the Wolpin review, is that of general equilibrium effects. This issue has been largely ignored, probably because it is particularly difficult to develop tractable empirical general equilibrium models suited to microeconomic data with a reasonable degree of realism.

The work by Heckman, Lochner, and Taber (1998) is probably the first attempt at this; they clearly illustrated the importance of the exercise, because relatively large partial equilibrium effects can be almost completely reversed by general equilibrium effects. In fact, the impact of a tuition subsidy is reduced to 10 percent of its original first-round direct effect when wages are allowed to adjust in response to changes in the supply of skills. Heckman et al. allow for schooling choices, on-the-job training, and saving. Generally, this leads to difficult estimation problems, straining to the limit the available data and illustrating both the importance and the difficulty of such an exercise.

General equilibrium results are bound to be sensitive to assumptions made about the production and the supply side. The presence of inadequate data increases the sensitivity of the exercise. Perhaps a good illustration of the sensitivity of general equilibrium results to specification is provided by the contrasting results obtained by Lee (2000) and presented in the paper by Wolpin: Lee finds only small impacts of allowing for general equilibrium effects. Here is not the place to evaluate the precise source of the differences between these studies. However, it seems very important to pursue research in general equilibrium models with heterogeneous agents and based on microeconomic data, with the specific aim to analyze important policy issues. Without a good understanding of long-term "second-round" effects, it will always be hard to design effective policies for human capital development. This is probably as much true in developing countries (where human capital acquisition is of central importance) as in the industrialized world.

I conclude this part of my discussion with some remarks on the structural modeling approach. Frequently, researchers attempt to estimate aspects of a model, such as the returns to education, by using exogenous events as instruments. A very useful literature has grown out of this, clarifying what is and is not identified under different assumptions on the statistical nature of the models. It is often the case that the underlying economic structure is left unspecified and the attraction is precisely that. The advantage of structural models is the clarity of the assumptions and of the explicit statement of the context under which the estimates are interpretable. It is often suggested that structural models suffer from identification problems. However, this is a misleading way of looking at the problem; identification issues do not disappear simply by ignoring them. Structural models bring to the open the identification problems that exist in estimating relationships of policy interest. Wolpin's paper makes this abundantly clear and demonstrates why this is the case with clear examples. Nevertheless, it is always important when estimating structural models to address the issue of identification and to make clear the identifying assumptions that drive the results. These assumptions are often not founded in any theory, and yet the results may not be robust to their relaxation. A good strategy would consist of remaining as nonparametric as the data and the problem would allow and to estimate structural models by exploiting as much as possible genuine exogenous variation in the estimation of the structural model. Two recent examples of such an approach (among others) are by Blundell, Duncan, and Meghir (1998) in their study of labor supply and by Attanasio et al. (2001) in their study of education choices in poor Mexican villages. In the former case, the authors exploit long-term changes in skill differentials and tax changes to identify a labor supply model. In the latter case, the authors exploit a randomized experiment providing education subsidies to identify the impact of incentives in educational attendance. Alternatively, one can consider conditions under which interpretable policy effects can be identified semiparametrically by using variation induced by other similar types of policies or variations; for example, we can identify the impact of an indirect tax by using estimated price variation and exploiting the theoretical restriction that consumers should react in the same way to a tax and price change. This idea, which has its origin in Marschak's work, was developed recently by Ichimura and Taber (2001). Finally, the treatment effects literature (see Heckman, LaLonde, and Smith, 1999) attempts to minimize the assumptions required to identify particular policy impacts and does not specify the complete structural model. Obviously, these ideas are related and can be viewed as a response to structural models. The robustness of the results obtained in this case comes at a cost because the models are not complete and hence do not allow simulation of policies out of sample or an analysis of the mechanisms by which a policy may operate.

4. CONCLUSION

These two well-written papers address critically important questions, which have an impact on the design of public policy. They also raise a number of issues and in a way define the empirical agenda. The paper by Raquel Fernández clearly calls for more emphasis on education policy and emphasizes the efficiency gains that could follow. The paper by Ken Wolpin in a sense illustrates the difficulty of delivering such a policy: Much of the educational outcomes are attributed to unobservables; alleviating liquidity constraints seems to have little or no impact. General equilibrium effects may mitigate or completely neutralize such policies. The key question is how to design education policy, how we should intervene, and at what age we should do so.

ACKNOWLEDGMENTS

I thank Hide Ichimura for very helpful comments. All errors are my own. Author contact: University College, London and Institute for Fiscal Studies. E-mail: c.meghir@ucl.ac.uk.

References

Angrist, J. and V. Lavy (1996), "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement," *Quarterly Journal of Economics*, 40, 533–575.

Attanasio, O., C. Meghir, and A. Santiago (2001), "Education Choices in Mexico: Using a Structural Model and a Randomised Experiment to Evaluate Progresa," mimeo, University College London.

Bénabou, R. (1996), "Equity and Efficiency in Human Capital Investment: The Local Connection," *Review of Economic Studies*, 62, 237–264.

Blundell, R., C. Meghir, and A. Duncan (1998), "Estimating Labour Supply Responses Using Tax Reforms," *Econometrica,* 66, 827–862.

Brock, W. A. and S. Durlauf (2001), "Discrete Choice with Social Interactions," *Review of Economic Studies*, 68, 235–260.

Cameron, S. and J. J. Heckman (2001), "The Dynamics of Educational Attainment for Blacks, Hispanics, and Whites," *Journal of Political Economy*, 109, 455–499.

Cameron, S. and C. Taber (2000), "Borrowing Constraints and the Return to Schooling," mimeo, Northwestern University.

Card, D. and A. Krueger (1992), "Does School Quality Matter? Returns to Education and the Characteristics of Public Schools in the United States," *Journal of Political Economy*, 100, 1–40.

Dale, S. B. and A. Krueger (1999), "Estimating the Payoff to Attending a More Selective College: An Application of Selection on Observables and Unobservables," CHLS.

Deardon, L., J. Ferri, and C. Meghir (2002), "The Effect of School Quality on Educational Attainment and Wages," *Review of Economics and Statistics*, 84, 1–20.

Hanushek, E. A., J. F. Kain, and S. G. Rivkin (1998), "Teachers, Schools, and Academic Achievement," Working Paper W6691, National Bureau of Economic Research.

Heckman, J. J., A. Layne-Farrar, and P. Todd (1996), "Human Capital Pricing Equations with an Application to Estimating the Effects of Schooling Quality on Earnings," *Review of Economics and Statistics*, 78, 562–610.

Heckman, J. J., L. Lochner, and C. Taber (1998), "Explaining Rising Wage Inequality: Explorations with a Dynamic General Equilibrium Model of Labor Earnings with Heterogeneous Agents," *Review of Economic Dynamics*, 1, 1–58.

Heckman, J. J., R. LaLonde, and J. Smith (1999), "The Economics and Econometrics of Active Labor Market Programs," in *Handbook of Labor Economics*, Vol. 3 (ed. by O. Ashenfelter and D. Card), New York: Elsevier Science.

Heckman, J. J. and G. Sedlacek (1985), "Heterogeneity, Aggregation and Market Wage Functions: An Empirical Model of Self-Selection in the Labor Market," *Journal of Political Economy*, 93, 1077–1125.

Hoxby, C. (2001), "Peer Effects in the Classroom: Learning from Gender and Race Variation," Mimeo, Harvard University.

Ichimura, H. and C. Taber (2001), "Direct Estimation of Policy Impacts," Working Paper WP0005, IFS.

Keane, M. and K. Wolpin (1997), "Career Decisions of Young Men," *Journal of Political Economy*, 105, 473–522.

Krueger, A. (1999), "Experimental Estimates of Education Production Functions," *Quarterly Journal of Economics*, 114, 497–532.

Krueger, A. and D. Whitmore (2001), "The Effect of Attending a Small Class in the Early Grades on College-Test Taking and Middle School Test Results: Evidence from Project STAR," *Economic Journal*, 111, 1–28.

Lee, D. (2000), "An Estimable Dynamic General Equilibrium Model of School Work and Occupational Choice," mimeo, University of Pennsylvania.

Meghir, C. and M. Palme (2001), "The Effect of a Social Experiment in Education," Working Paper WP0111, IFS, available at www.ifs.org.uk

van den Berg, G. (2000), "Empirical Inference with Equilibrium Search Models of the Labor Market," *Economic Journal*, 78, 526–610.

CHAPTER 3

Toward a Theory of Competition Policy

Patrick Rey

1. INTRODUCTION

Decades ago, one of the main regulatory debates concerned opposing marginal-cost pricing rules to average-cost or more sophisticated Ramsey-like pricing rules.[1] Since then, a huge effort has been made to account for the fact that a regulator must not only *choose* the pricing rule, but also *implement* it.[2] For example, cost-based pricing rules require information about costs that is not usually readily available. In addition, firms are better informed than regulators about their costs; it is all the more unfortunate that, in general, firms have little incentive to report this information truthfully, knowing that the information will be used to determine their prices. Policymakers must therefore take into consideration the information-acquisition problem when designing the regulation.

Consider the following example. A firm has a linear cost $C(q) = cq$, and its regulator seeks to maximize consumers' net surplus, given by

$$U(q) - t,$$

with $U' > 0 > U''$, subject to the firm's budget constraint

$$t - cq \geq 0.$$

The first-best regulation consists in equating marginal cost to marginal utility, which leads to a marginal-cost pricing rule: The first-best level of production, $q^{\text{FB}}(c)$, and the first-best price, p^{FB}, are defined by

$$p^{\text{FB}} = U'(q^{\text{FB}}) = c.$$

However, the regulator cannot implement this rule if he or she does not know the value of the marginal cost. Suppose, for example, that the marginal cost can take two values, a low value $\underline{c}$ or a high value $\bar{c} > \underline{c}$. If only the firm knows the

[1] See Laffont (2000), Chapter 6, for an overview of this debate.

[2] This is not to say that this issue had not been recognized before. For example, an argument – which goes back to Smith (1776) – in favor of average-cost or budget-balance pricing was the lack of information about the desirability of the projects. However, it is only since the 1980s that this and related points have been built into theory.

true value, it would have an incentive to report a high cost $\bar{c}$, even if it has a low cost $\underline{c}$: By doing so, it would secure a positive gain, equal to $(\bar{c} - \underline{c})D(\bar{c})$. In other words, the first-best marginal-cost pricing rule does not induce the firm to truthfully reveal its cost.

Recognizing this issue, regulation theorists have refined the analysis by modeling the constraints imposed by the firm's incentives to manipulate the information on its costs and by incorporating these informational constraints in the design of the regulator's problem.[3] In the simple case just described, assigning an output level $\bar{q}$ to a high-cost firm requires leaving a low-cost firm an informational rent $(\bar{c} - \underline{c})\bar{q}$: Because the high-cost firm must get a transfer $\bar{t}$ sufficient to cover its cost ($\bar{t} \geq \bar{c}q$), by reporting a high cost, a low-cost firm can indeed secure itself

$$\underline{\pi} = \bar{t} - \underline{c}\bar{q} \geq (\bar{c} - \underline{c})\bar{q}.$$

The "second-best" regulation accounts for this informational rent, which leads to lowering the production required from a high-cost firm. Formally, the regulator's problem can be summarized as

$$\max_{\underline{q},\bar{q},\underline{t},\bar{t}} \Pr(c = \underline{c})[U(\underline{q}) - \underline{t}] + \Pr(c = \bar{c})[U(\bar{q}) - \bar{t}],$$

subject to

$$\underline{t} - \underline{c}\underline{q} \geq \bar{t} - \underline{c}\bar{q} \quad \text{(IC)},$$
$$\bar{t} - \bar{c}\bar{q} \geq 0 \quad \text{(IR)},$$

where the (IC) constraint represents the low-cost firm's incentive condition and the (IR) constraint represents the high-cost firm's participation constraint. These are the only two relevant conditions,[4] and they are both binding. When those conditions are used to express the required transfers as a function of desired quantities, the regulator's problem can be rewritten as

$$\max_{\underline{q},\bar{q}} \Pr(c = \underline{c})[U(\underline{q}) - \underline{c}\underline{q} - (\bar{c} - \underline{c})\bar{q}] + \Pr(c = \bar{c})[U(\bar{q}) - \bar{c}\bar{q}].$$

The second-best level price and production for a high-cost firm are thus given by

$$\bar{p}^{\text{SB}} = U'(\bar{q}^{\text{SB}}) = \bar{c} + \frac{\Pr(c = \underline{c})}{\Pr(c = \bar{c})}(\bar{c} - \underline{c});$$

the price is therefore higher – and the production lower – than for the first-best level.

[3] Baron and Myerson (1982) provide the first analysis of this refined regulatory problem. Baron and Besanko (1984a) study the role of audits; Baron and Besanko (1984b) consider a multitemporal extension of the Baron–Myerson model with commitment. Caillaud et al. (1988) offer a review of this literature.

[4] Intuitively, a high-cost firm cannot gain by pretending to be more efficient than it is, and a low-cost industry is willing to participate, because it can get at least as much as a high-cost one by mimicking it.

This simple example contains two lessons. First, recognizing the implementation problem, generated here by the asymmetric information between the firm and the regulator, leads to modification of the regulation – and the merit of the approach is to indicate how the regulation should be modified. Second, the appropriate way to design the regulation consists of first characterizing the set of "implementable" rules – here, the incentive condition (IC) – in order to find out the second-best rule – among those that are implementable. The same approach has since then been adopted to account for additional implementation issues, such as the regulator's limited commitment, regulatory capture, multitasking, and the like.[5]

Compared with regulation theory, the theory of competition policy is still in its early stages of development, with little attention devoted to implementation issues. To be sure, understanding how oligopolistic industries work already constitutes a challenging task; this certainly explains why much effort has been devoted to the analysis of firms' interaction, leaving less room for the study of the supervision of these industries. As a result, most of the work on competition policy has focused on the analysis of firms' strategic interaction, under alternative (and often ad hoc) restrictions on their strategy spaces, meant to reflect different competition policy rules. More recently, however, some attention has been paid to implementation issues. This paper offers an outlook on recent advances in this direction, in various areas of competition policy, and advocates a fuller recognition of implementation problems in policy design.

The paper is organized as follows. Section 2 focuses on the enforcement of cartel laws, where implementation is the main issue. Section 3 turns to merger control. Finally, Section 4 offers some suggestions for further research, based on a better understanding of competition authorities' implementation problems.

2. PRICE-FIXING AGREEMENTS

This facet of competition policy is a natural candidate for discussing the role of implementation in policy design. First, there is a consensus, at least in practice, that price-fixing agreements should be forbidden[6]; the main issue is therefore

[5] See Laffont and Tirole (1993) for an extensive review of these developments.

[6] Selten (1984), however, points out that intensifying price competition may deter some firms from entering the market; fighting price-fixing agreements may thus backfire through a decrease in the number of participants. (A similar observation applies to the size of endogenous sunk costs, as emphasized by Sutton, 1991, 1998.)

D'Aspremont and Motta (2000) note, however, that if intensifying competition reduces the number of participants, it also selects the most efficient ones; in their model, a moderate softening of competition may have a positive impact on welfare, but allowing full collusion (price-fixing agreements) always reduces welfare. (Symeonidis, 2000, studies the impact of cartel policy on firms' profits, using a panel data set of UK manufacturing industries. He finds that the introduction of cartel laws in the UK in the late 1950s caused an intensification of the price competition and had a strong effect on the structure of the markets that were previously cartelized, but little impact on firms' profits.)

not what to do but, precisely, how to do it. Second, some effort has already been made to account for implementation problems in this area.

The economic analysis of price cartels has initially focused on the stability of *explicit cartels* and on the sustainability of *tacit collusion*. The first strand of literature studies the formation of explicit cartels: Assuming that firms can operate a cartel as a joint profit-maximizing entity, what is the optimal or the equilibrium size of the cartel? The main insights build on two observations: First, cartel members may not always gain from forming a cartel – except if it includes all firms in the industry[7]; second, nonmember firms usually gain from the creation of a cartel – and actually gain even more by staying outside the cartel than by joining it.[8] The second strand of literature characterizes the set of prices that can be sustained in a noncooperative equilibrium when competition is repeated over time.

The first literature assumes that firms can enter in (long-term) binding agreements.[9] The analysis is therefore more useful for understanding the formation of explicit cartels such as OPEC[10] than for the design of an appropriate policy against those cartels. In particular, declaring such agreements illegal and void will have little impact if firms can rely on hidden or implicit ways to enforce their agreements; the analysis may, however, help to identify good "target" industries for investigation.

More promising is the second literature, which identifies factors and practices that facilitate tacit collusion, that is, that allow the emergence of equilibria with higher prices in situations of repeated competition. Short of attacking tacit collusion per se, competition authorities can then fight those facilitating these practices.

Fighting collusion per se is a different issue. In particular, fighting tacit collusion comes close to regulating prices,[11] something that competition authorities

[7] See Selten (1973) and Salant, Switzer, and Reynolds (1983).

[8] See D'Aspremont et al. (1983). For example, in standard Cournot models, the creation of a cartel always benefits outsiders, because it leads cartel members to reduce their aggregate supply; however, a "small" cartel will typically not be profitable – for instance, in the symmetric case with linear demand and costs, a cartel between two firms is never profitable when there are at least three firms.

Attention has also been devoted to the design of (binding) cartel contracts in the presence of asymmetric information; see, for example, Roberts (1985), Cramton and Palfrey (1990), and Kihlstrom and Vives (1992).

[9] The analysis relies on the assumption that, once cartels are formed, all entities (cartels and remaining individual firms) maximize their (joint or individual) profits and play some oligopolistic competition game. This supposes that the cartel structure remains durably fixed, as compared with the strategic variables involved in the competition game.

[10] Whether OPEC actually operates as a binding cartel is a debated issue. For an opposing view, see Crémer and Salehi-Isfahani (1989).

[11] For example, in a repeated Bertrand duopoly in which firms have the same unit cost and use the same discount factor $\delta > \frac{1}{2}$, any price between the competitive and monopoly ones can be sustained as a noncooperative, Subgame–Perfect equilibrium. In principle, the equilibria involving supracompetitive prices could be ruled out by restricting the set of admissible strategies to, say, Markov strategies; in practice, however, this may be as difficult as directly dictating the level of the prices.

and courts are generally reluctant to do.[12] However, in practice, collusion often leaves significant pieces of evidence: notes of meetings, compromising fax exchanges, e-mails, and the like.[13] Competition authorities can thus try to identify cartels and recover those pieces of evidence. Building on this insight, a significant effort has been developed in the past decade to explicitly model the implementation of the ban on price-fixing agreements.

I now discuss in turn enforcement policies against both collusion and facilitating practices.

2.1. Fighting Collusion Per Se

Starting with Becker's (1968) seminal paper, a large effort has been devoted to the economic analysis of enforcement policies against illegal behavior.[14] It is only recently, though, that this literature has addressed enforcement policies against *concerted* illegal actions, involving several participants. Here I present some of the advances in this domain, first in a static and then in a dynamic framework.

2.1.1. Fighting Collusion in a Static Setting

Assuming that the law forbids collusion on prices, competition authorities' main problem is to detect and prove such collusion. Besanko and Spulber (1989) were the first to study the implications of this informational problem formally for cartel policy. In their model, firms have the same constant unit cost c and choose whether to compete, in which case they set their price at a competitive level $p^c(c) = c$,[15] or collude, in which case they jointly determine their supply decisions – because they are symmetric, there is always a consensus as to which price is best for them. Besanko and Spulber capture the implementation problem by assuming that the competition authority observes prices, quantities, or both, but neither the cost of the industry nor the occurrence of collusion. More

[12] In the United States, Section 2 of the Sherman Act condemns "monopolization," not the exploitation of a rightfully acquired market power. An inventor, say, can thus sell the product of his or her invention at a monopoly price if he or she wishes to. In the European Union, charging an excessive price can constitute an "abuse of dominant position," sanctioned by Article 81 of the Treaty of Amsterdam (formerly Article 86 of the Treaty of Rome). However, apart from some early cases, no firm has been fined for such abuses. (Whether competition authorities should attack tacit collusion, where by definition firms set prices noncooperatively, is itself a debated issue.)

[13] The appropriate model may thus be neither of the explicit cartel type, nor of the purely tacit collusion type; it may involve features from both paradigms: firms rely on secret contracts, which are not legally enforceable and are instead tacitly enforced thanks to repeated interaction. The next section develops a framework along those lines.

[14] See Polinski and Shavell (2000) for a recent survey.

[15] Baniak and Phlips (1996) extend the analysis to the case in which the competitive outcome is the Cournot equilibrium.

precisely:

1. The authority does not observe the cost c, which can take two values, a low value $\underline{c}$ or a high value $\bar{c}$. Observing prices thus does not suffice to detect collusion, but the authority can use prices or quantities to identify "suspect" industries.
2. The authority can audit the industry; an audit costs C but determines whether there is collusion, in which case a maximal fine F can be imposed on the firms.[16]

The industry audit can be interpreted as a "hunt for evidence" (dawn raids etc.) that allows the authority to discover formal proofs of price-fixing agreements, but not the actual level of costs; in particular, if the price is above c, the audit reveals collusion but does not allow cost-contingent fines.

Besanko and Spulber characterize the policy that maximizes expected total welfare, net of audit costs. Feasible policies consist, for each level of the price or output, in an audit probability μ and a fine f in case the audit reveals that the price was above the competitive level. The optimal policy boils down to picking two prices or, equivalently, two levels of output, one for the high-cost industry, $\bar{p} = P(\bar{q})$, and a lower one for the low-cost industry, $\underline{p} = P(\underline{q})$, attached with two probabilities of audit, $\underline{\mu}$ and $\bar{\mu}$, and two levels for the fine, $\underline{f}$ and $\bar{f}$, in case of detected collusion. As usual, the relevant conditions are the high-cost firms' participation constraint (IR) and the low-cost firms' incentive constraint (IC). In addition here, "transfers" are limited to fines, which can be imposed only when an audit reveals collusion. The authority's problem can then be summarized as

$$\max_{\underline{q},\bar{q},\underline{\mu},\bar{\mu},\underline{f},\bar{f}} \quad \Pr(c = \underline{c})[U(\underline{q}) - \underline{c}\underline{q} - \underline{\mu}C] \\ + \Pr(c = \bar{c})[U(\bar{q}) - \bar{c}\bar{q} - \bar{\mu}C],$$

subject to

$$[P(\underline{q}) - \underline{c}]\underline{q} - \underline{\mu}\underline{I}\underline{f} \geq [P(\bar{q}) - \underline{c}]\bar{q} - \bar{\mu}\bar{f} \quad \text{(IC)},$$

$$[P(\bar{q}) - \bar{c}]\bar{q} - \bar{\mu}\bar{I}\bar{f} \geq 0 \quad \text{(IR)},$$

where I is an indicator variable for collusion: for each type of industry, I equals 0 if $P(q) = c$ and 1 otherwise.

Besanko and Spulber show that the optimal policy has the following features.

Proposition 2.1. *Audit takes place only for high prices*: $\underline{\mu} = 0$.

Because high-cost firms have no incentive to adopt the lower price $\underline{p}$, there is no point undertaking costly audits when the price is $P(\underline{q})$.

[16] The upper bound on the fine, F, can be derived from firms' limited liability; it is supposed large enough to offset any gain from collusion.

Proposition 2.2. *The fine is set at its maximum for the case in which an audit reveals that low-cost firms have adopted a high price*: $\bar{f} = F$.

This is a standard property of optimal policies with costly audits: The low-cost firms' incentives to mimic high-cost ones are driven by the expected fine $\bar{\mu}\bar{f}$, and the least costly way to generate a given expected fine is to make the actual fine $\bar{f}$ as large as possible, in order to reduce the probability of audit $\bar{\mu}$.

Proposition 2.3. *Low-cost firms are allowed to charge supracompetitive prices*: $P(\underline{q}) > \underline{c}$.

A small departure from the competitive price $\underline{p} = \underline{c}$ generates only a second-order negative impact on total welfare, but it allows a first-order reduction in the audit probability (and thus its expected cost) in the event of the higher-price $P(\bar{q})$.[17] It is therefore optimal to tolerate a limited collusion in low-cost industries.

Proposition 2.4. *It may be optimal to allow high-cost firms to charge supracompetitive prices as well.*

This more surprising result comes from the low-cost industry's incentive constraint (IC). If the difference in the two costs is very large (e.g., if the low-cost monopoly price is lower than high cost firm's price), an increase in the price assigned to the high-cost industry makes it less attractive for the low-cost industry and relaxes (IC), thereby allowing the authority to reduce the audit probability. This result, however, relies somewhat on the assumption that the number of types is assumed to be discrete.[18]

This analysis yields several insights for policy design. For example, investigations must be launched when prices are "high," not because these prices are necessarily collusive[19] but, rather, to deter low-cost firms from unduly adopting those high prices. Several assumptions are, however, worth discussing.

First, some of the constraints imposed on the authority seem rather arbitrary. For example, the absence of transfers between the firms and the authority (or more generally, with the collectivity), other than fines in the event of proven

[17] This probability is determined by (IC):

$$\bar{\mu}\bar{f} = (\bar{p} - \underline{c})D(\bar{p}) - (\underline{p} - \underline{c})D(\underline{p}),$$

implying

$$\left.\frac{\partial(\bar{\mu}\bar{f})}{\partial \underline{p}}\right|_{\underline{p}=\underline{c}} = -D(\underline{c}) < 0.$$

[18] See Souam (1997) for an analysis of the case of a continuum of values.

[19] In the realistic case in which it is best to forbid any higher price than $\bar{c}$, it is optimal to audit precisely the firms that do not collude in equilibrium – and only those.

collusion, is meant to reflect a restriction commonly observed in practice, but the underlying reason for this restriction is not explicitly modeled. In particular, the reason cannot be found in the authority's prior lack of information: In their model, the authority could, in general, perform better if allowed to use additional transfers.[20] A first reason may be found in a prohibitive social cost of transfers. Another reason, which deserves further exploration, is that competition authorities and courts are ill fitted to manage such transfers. There may exist a suspicion about the authority's use of public funds (risk of capture), and this concern may be exacerbated by the lack of control by taxpayers and the fact that taxpayers' representatives ("advocates") are not sufficiently involved in the management of those transfers. Yet another explanation might be found in the authority's lack of commitment; ruling out transfers could, for example, be seen as a (drastic) way to impose a "hard budget constraint" on the industry.[21] In Section 4, I revisit those underlying reasons, but note here that little progress has been made in the analysis of how those reasons contribute to shaping cartel law enforcement.

Second, Besanko and Spulber assume that firms can perfectly collude if they wish to, and that competition authorities cannot affect the sustainability of collusion. This leads them to treat the industry as a single entity; their analysis would formally be the same if there was only one firm in the industry. This rules out some important means of intervention, such as playing firms against each other. More generally, the multiplicity of participants in the industry might allow for the movement from a "police patrol" to a "fire alarm" mode of operation, to use the terminology of McCubbins and Schwartz (1984).

In particular, Besanko and Spulber rule out any communication between the firms and the competition authority. Formally, the authority is confronted with an information-acquisition problem: Firms know whether they collude or not, whereas the authority does not. The authority could therefore try to devise revelation mechanisms à la Maskin, in order to induce the firms to report this information.[22]

Indeed, in practice, competition authorities often design leniency programs to allow cartel members to benefit from a favorable treatment if they bring information that helps competition authorities to dismantle the cartel. In the United States, firms bringing information before an investigation is opened have benefited from such a leniency program since 1978. Since 1993, a colluding firm can also avoid those sanctions if it reveals information once the investigation has been opened, as long as the Department of Justice has not yet been able to

[20] The problem of the competition authority studied here can be viewed as a standard regulation problem with audit, as studied by Baron and Besanko (1984a), with the additional restriction that the fines constitute the only allowed transfers.

[21] See Laffont and Tirole (2000), Chapter 2, for a discussion of restrictions on transfers in regulatory contexts.

[22] Building on the pioneering work of Maskin (1977), the literature on Nash and Subgame–Perfect implementation has confirmed the intuition that "a secret is no longer a secret when it is shared by several agents." See Moore (1992) for a very nice survey.

prove collusion.[23] The European Union adopted a leniency program in 1996, which allows firms that bring information to benefit from reduced fines.[24] In the UK, when a new Competition Policy Act was implemented two years ago, the Director General of the Office of Fair Trading introduced a leniency program close to the U.S. model. In France, following a bill passed last year, a leniency program is currently being implemented.

Revelation mechanisms may not be very effective in the case of "soft" information, when no evidence is left to be discovered by competition inspectors. In particular, as long as this information relates to past behavior and has no direct effect on firms' capabilities or objectives in the future, little can be done to extract this information. Things change, however, when collusion produces pieces of "hard" information that can be transmitted to the competition authority, as implicitly assumed by Besanko and Spulber. The competition authority could then encourage firms to report any collusion (i.e., provide hard evidence of it), reward informants, and use the information against the other firms.

Of course, the effectiveness of such mechanisms depends on the extent to which the industry can again collude at the revelation stage. If firms cannot collude at all at the revelation stage, the authority may be able to deter collusion *at almost no cost*. For example, consider the following mechanism. Once supply decisions have been undertaken, the competition authority randomly selects one firm and confronts it with the following choice:

- either it reports collusion, in which case all other firms are heavily fined, but the reporting firm is exempted from any fine;
- or it denies collusion, in which case the competition authority audits with an arbitrarily small probability ε, and all firms are fined if collusion is detected.

If firms cannot collude at this stage, it is then a strictly dominant strategy for the selected firm to report collusion whenever it takes place, and ex ante the threat of being fined then deters collusion.[25]

[23] See U.S. Department of Justice (1993). Thanks to this new leniency program, on average two cartels are now disclosed every month, and the fines often exceed \$100 million (not to mention jail for some managers). In 1999, only the Antitrust Division secured more than \$1 billion in fines, which is more than the total sum of fines imposed under the Sherman Act since its adoption more than a century ago.

[24] See European Union (1996).

[25] If there are n firms in the industry, the maximal fine would only need to be large enough to offset the gains from collusion when applied with probability $1 - 1/n$. In addition, as pointed out by Kaplow and Shavell (1994) in the context of single-party law enforcement, the selected firm would be induced to report as long as the reduced fine is lower than the *expected fine* it would face otherwise. A difference with single-party enforcement contexts, however, is that here the competition authority can use the information brought by one cartel member against the other members; hence, if ρ denotes the probability of being caught and F the level of the fine, collusion can be deterred whenever $(1 - 1/n + \rho)F$ offsets the gains from collusion.

If firms can instead perfectly collude and behave as a single entity at the revelation stage, this revelation mechanism would not work anymore: Nonselected firms would have an incentive to bribe the selected one and induce it not to report collusion. But then, the competition authority could in theory try to elicit information on this second form of collusion, and so on. A relevant analysis of this issue requires a deep understanding of how firms organize the collusion.[26]

Of course, in practice, competition between firms is rarely of the "one-shot" type analyzed herein: Colluding firms know that they will be competing again in the future and will therefore be reluctant to report past collusion, as this will likely reduce the scope for collusion in the future. To tackle this issue, however, one needs to develop a dynamic framework.

2.1.2. Fighting Collusion in a Dynamic Setting

Analyzing the struggle against collusion in a dynamic setting[27] raises, of course, additional intricacies. To keep things tractable, I will therefore highly simplify the just-described framework and assume that the authority does not observe any relevant information in the absence of audit. This assumption is meant to capture the fact that, in practice, the competition authority has very little information about supply and demand conditions and cannot, in general, infer collusion from the mere observation of prices.[28]

Two firms ($i = 1, 2$) play an infinitely repeated game: In each period, they can either compete or collude; gross profits are given by Table 3.1.

[26] A large effort has been devoted to the study of collusion in organizations, with both hard information and soft information. Whereas the earlier literature assumed that colluding parties shared their information, attention has recently been devoted to the incentive constraints that colluding parties may face when they have private information. See Tirole (1992) for a first extensive survey, Laffont and Rochet (1997) for an updated overview, and Laffont and Martimort (1997, 2000) for recent advances in the modeling of collusion between privately informed agents.

[27] This section draws on the work of Aubert, Kovacic, and Rey (2000).

[28] Competition authorities could, in principle, try to infer collusion not only from the current level of prices, but also from the pattern of their evolution. In the famous Woodpulp case, for example, the European Commission observed a parallel evolution of the prices quoted in dollars, despite substantial variations of exchange rates between the producing countries, as well as a remarkable stability over time, except for a six-month price war. It asserted that this pattern was conclusive evidence of collusion, but this decision was overruled in appeal by the European Court of Justice, which concluded that the Commission failed to establish that this pattern of prices was not compatible with competitive behavior.

Kühn (2000) offers a detailed discussion of the difficulties in detecting collusive behavior from observable behavior, beyond the lack of information on cost. First, information on actual prices and quantities may be unavailable. Second, even in the ideal case when reliable price and output data would allow quantitative studies, the conclusions may be too sensitive to functional form specifications. This is illustrated by the divergence of the findings of two studies of the U.S. railroad cartel of the 1880s, based on the same data set. Whereas Porter (1983) concludes that firms' observed markups were consistent with Cournot behavior, by allowing for autocorrelation on the demand side, Ellison (1994) instead obtains an estimate close to full collusion.

Table 3.1. *Basic stage game*

	Firm 1	
Firm 2	Compete	Collude
Compete	(π^C, π^C)	$(\pi^D, \underline{\pi})$
Collude	$(\underline{\pi}, \pi^D)$	(π^M, π^M)

Here, $\underline{\pi} \leq \pi^C < \pi^M < \pi^D$ and $\underline{\pi} + \pi^D < 2\pi^M$: Firms gain from collusion, but each firm may benefit at the expense of the other from "cheating," that is, from competing when the other colludes.[29] Collusion moreover generates hard information:

1. Whenever collusion is successful (i.e., both firms collude), it generates a piece of evidence that is found with probability ρ by the competition authority; ρ can be thought of as the exogenous probability of a successful audit.
2. In addition, each of them can then bring a piece of hard information to the competition authority.[30]

To keep the analysis simple, I will assume that any piece of hard information disappears after one period. This limits the scope for revelation mechanisms, which can apply only to "current" collusive behavior. The maximal fine F that can be imposed in case of proven collusion is large enough to deter collusion if collusion is detected with certainty, but not sufficient if collusion is detected only with probability ρ:

$$F > \pi^M - \pi^C > \rho F.$$

In the absence of any revelation mechanism, the net expected payoffs of the stage game are thus given by Table 3.2.

If both firms use the same discount factor δ, collusion is sustainable if

$$\pi^D - (\pi^M - \rho F) \leq \frac{\delta}{1-\delta}(\pi^M - \rho F - \pi^C). \tag{2.1}$$

Indeed, the most profitable collusive strategy is to collude in each period and to punish deviations by returning forever to the static competitive equilibrium.[31]

[29] This assumption is, for example, relevant for situations in which "colluding" amounts to maintaining the monopoly price. If instead collusion were to involve an agreement that "takes two to tango," one might expect $\pi^D = \underline{\pi} = \pi^C$.

[30] This assumption allows the competition authority to design revelation games, as suggested in the previous section, while ruling out trivial strategies in which a firm would unilaterally collude just to report collusion and get a reward for doing so.

[31] The competitive outcome is both the Nash equilibrium and the minmax of the stage game. Therefore, an indefinite reversal to the competitive situation constitutes the harshest credible punishment that can be imposed on deviators.

Table 3.2. *Audit and fines*

	Firm 1	
Firm 2	Compete	Collude
Compete	(π^C, π^C)	$(\pi^D, \underline{\pi})$
Collude	$(\underline{\pi}, \pi^D)$	$(\pi^M - \rho F, \pi^M - \rho F)$

Let us now introduce revelation mechanisms. The competition authority could ask the firms to reveal their choice (compete or collude), both before and after it audits the industry. The scope for such revelation mechanisms depends on whether the messages sent by one firm are observed by the other firm. Inducing reporting is clearly easier when it is not observed by rivals, but keeping such reports secret may be difficult in practice. I consider both situations in turn.

Secret Reporting. Let us start with the case in which firms can report collusion secretly. More precisely, suppose that firms observe only whether the competition authority has evidence of collusion, but not the origin of its information nor the fines or rewards possibly paid by or to the rival. Then, the competition authority can easily prevent "perfect collusion." The following mechanism would, for example, do the trick. At the beginning of each period, offer each firm the chance to report collusion in order to benefit from a slightly reduced fine in the event of a successful audit. Then, it cannot be the case that firms plan to collude in every period, even after a successful audit by the competition authority. Indeed, in that case it would be a strictly dominant strategy for the selected firm to report collusion whenever it takes place, as doing so reduces the expected fine it will have to pay (because of the possibility of a successful audit) and does not trigger any retaliation by the rival.[32]

To deter such reporting and sustain some collusion, firms will therefore have to plan periods of "competition" whenever the competition authority imposes a fine on them.[33] Furthermore, an increase in the amount of the reward exacerbates the temptation to deviate and report, and thus requires longer periods of competition. To see this, suppose that in each period the competition authority offers a (secret) reward R for (secret) reports; the firms can then adopt the following strategy: "collude and do not report as long as the authority imposes no fine; when the authority imposes a fine, compete for T periods before returning to collusion; furthermore, whenever a firm competes when it is supposed to collude, revert to competition forever." If both firms adopt this strategy, they obtain

$$V = \pi^M - \rho F + \delta[(1 - \rho)V + \rho \hat{V}],$$

[32] The threat of being fined would therefore deter collusion whenever the maximal fine F is large enough to offset the gains from collusion when applied with probability $\frac{1}{2}$.

[33] The analysis is reminiscent of that given in Tirole's (1988) version of Green and Porter (1984), for the case of unobservable demand shocks.

where

$$\begin{aligned}\hat{V} &= (1 + \cdots + \delta^{T-1})\pi^C + \delta^T V\\ &= \frac{1-\delta^T}{1-\delta}\pi^C + \delta^T V,\end{aligned}$$

so that

$$V = \frac{\pi^M - \rho F + \delta\rho[(1-\delta^T)/(1-\delta)]\pi^C}{1 - \delta + \delta\rho(1-\delta^T)}.$$

This value thus decreases when the number of competitive periods (T) increases. Adopting this strategy constitutes an equilibrium if the following are true.

1. Firms have no incentive to compete when they are supposed to collude, that is,

   $$V \geq \pi^D + \frac{\delta\pi^C}{1-\delta}, \tag{2.2}$$

   or

   $$\delta\frac{1-\rho(1-\delta^T)}{1-\delta+\delta\rho(1-\delta^T)}(\pi^M - \rho F - \pi^C) \geq \pi^D - (\pi^M - \rho F).$$

   The left-hand side of this condition, which I denote $A_1(T)$, satisfies $A_1'(T) < 0$. This condition is thus of the form

   $$T \leq T_1\left(\frac{\pi^M - \rho F - \pi^C}{\pi^D - (\pi^N - \rho F)}\right),$$

   with $T_1' > 0$: to prevent firms from deviating in this way, the value of collusion must be sufficiently large, and thus the duration of competition sufficiently short. Note that this condition is independent of the reward R.
2. Firms have no incentive to secretly report collusion, that is,

   $$V \geq \pi^M + R + \delta\hat{V}, \tag{2.3}$$

   or

   $$\delta\frac{(1-\rho)(1-\delta^T)}{1-\delta+\delta\rho(1-\delta^T)}(\pi^M - \rho F - \pi^C) \geq R + \rho F.$$

   The left-hand side of this condition, $A_2(T)$, increases with T. This condition is thus of the form

   $$T \geq T_2\left(\frac{\pi^M - \rho F - \pi^C}{R + \rho F}\right),$$

   with $T_2' < 0$: the larger the reward, the longer the duration of the competitive phases needed to prevent firms from reporting collusion.

Figure 3.1 illustrates the situation. Note that A_1 and A_2 converge toward the same limit, $\{[\delta(1-\rho)]/[1-\delta(1-\rho)]\}(\pi^M - \rho F - \pi^C)$, when T goes to

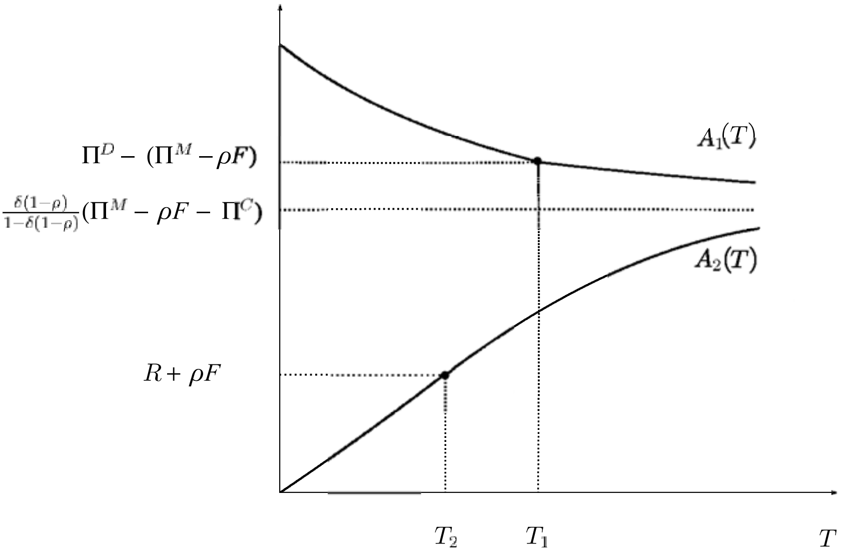


Figure 3.1. Scope for collusion with secret reports.

infinity. Thus, if

$$R + \rho F > \frac{\delta(1-\rho)}{1-\delta(1-\rho)}(\pi^M - \rho F - \pi^C),$$

that is, if the reward is large enough,

$$R > \hat{R} \equiv \frac{\delta(1-\rho)(\pi^M - \pi^C) - \rho F}{1-\delta(1-\rho)},$$

then this collusion is not sustainable.

Furthermore, even if $R < \hat{R}$, only some limited collusion is sustainable and the minimal duration of competition phases, given by $T_2(\cdot)$, increases with the amount of the reward for reporting collusion. Of course, increasing the probability of successful audit (ρ) or the maximal fine (F) further contributes to make collusion more difficult to sustain; in particular, a simple leniency program ($R = 0$) suffices when

$$\frac{\rho F}{1-\rho} > \delta(\pi^M - \pi^C).$$

Public Reporting. Things are more difficult when reports cannot be kept secret. In particular, in that case a leniency program that offers only a reduced fine cannot help the competition authority.

Proposition 2.5. *Suppose that the competition authority can only impose fines (i.e., it can "reward" informants through lower fines, but cannot make* *positive transfers to them). Then, revelation mechanisms do not help to prevent collusion.*

Proof. The proof is straightforward. First, it is not possible to make firms reveal past decisions, because hard information disappears after one period – and past decisions do not affect firms' preferences over their future choices, so soft information cannot be acquired. Therefore, revelation mechanisms can serve only to induce firms to report collusion in the current period.

Second, the competition authority cannot induce firms to report collusion ex post, once the result of its audit is known. If the authority does not detect collusion, firms have no incentive to report it and get (even moderately) fined. If the authority does detect collusion, there is no longer any need to induce firms to cooperate with the authority[34]; as stressed by Motta and Polo (2000) and by Spagnolo (2000a), offering a reduction of the fine would then actually erode the deterrence power of the authority's audit. In the case of a successful audit, firms would have an incentive to report ex post collusion, in order to benefit from the reduced fine, so that the right-hand side of (2.1) would be further increased.[35]

Last, whenever condition (2.1) holds, firms have no incentive to report collusion ex ante either, before they know the outcome of the authority's audit, even if they benefit then from a reduced fine $f < F$: Such reporting would trigger retaliation (no collusion in the future) and would thus not be a profitable strategy, as (2.1) implies[36]

$$\pi^D - \rho f - (\pi^M - \rho F) \leq \frac{\delta}{1-\delta}(\pi^M - \rho F - \pi^C). \qquad \blacksquare$$

To convince a firm to report collusion, the competition authority must therefore promise a *reward* R, sufficient to reverse the incentive condition given

[34] This may not be true anymore if firms' cooperation enhances the authority's ability to prosecute the cartel – see below.

[35] Spagnolo (2000a) notes that offering reduced fines could, however, help deter collusion by limiting the punishments that could be imposed on deviating firms. Suppose, for example, that "optimal penal codes" à la Abreu would allow firms to sustain expected profit levels *below* the static competitive equilibrium; then, introducing a leniency program might allow deviators to benefit from the authority's action and avoid such punishments, thereby making collusion more difficult to sustain.

Spagnolo (2000b) stresses instead that badly designed leniency programs may actually help collusion. This will, for example, be the case if the leniency program eliminates criminal sanctions but does not prevent civil damages. Because profits are reduced when the other firm deviates from the collusive agreement, the nondeviating firm may have higher incentives to report collusion in that case. Such a partial leniency program may therefore make credible the threat to denounce collusion whenever the other firm deviates from the cartel agreement, thereby contributing to enforcement of the agreement.

[36] That is, if collusion is sustainable in the absence of reporting, then, as long as firms do not get any additional information, they have no incentives to report collusion and get fined, even moderately. Malik and Schwab (1991) make a similar point for the enforcement of laws against single-party crimes (tax amnesty programs).

here; that is, this reward R should be such that[37]

$$\pi^M + R + \frac{\delta\pi^C}{1-\delta} \geq \frac{\pi^M - \rho F}{1-\delta},$$

which is equivalent to

$$R + \rho F \geq \frac{\delta}{1-\delta}(\pi^M - \rho F - \pi^C),$$

or

$$R \geq \bar{R} \equiv \frac{\delta(\pi^M - \pi^C) - \rho F}{1-\delta}. \tag{2.4}$$

The minimum reward $\bar{R}$ required to induce a firm to report collusion may thus be quite large[38]; it goes to infinity when δ gets close to 1. This raises several issues. First, granting large rewards may not be credible: The competition authority may be limited in the amount it can promise for reporting collusion. Second, large rewards may exacerbate social or political issues: The public opinion may be particularly reluctant to grant large amounts to colluding firms; more generally, granting rewards involves some transaction costs, which are likely to increase with the magnitude of the rewards.[39] Third, large rewards may have the perverse effect of generating additional incentives to collude, or different ways to do so. For example, it may become profitable for the firms to collude and "take turns" for reporting collusion.[40]

Extensions. Many lines of research are still open. In particular, consideration of the possibility of long-lasting pieces of hard information opens new perspectives. It would also be useful to further analyze the determinants of hard

[37] It is clearly more effective to ask firms to report collusion before an audit of the industry: There is no need for acquiring this information if the audit brings evidence of collusion, and if it does not, the information would be more costly to acquire (the term ρF would disappear from the left-hand side). Also note that, in this formulation, firms must collude in order to report. I discuss in the paragraphs that follow the (possibly more realistic) case in which firms can cheat on collusion at the time they report it.

[38] In particular, it is larger than the minimal reward $\hat{R}$ required to deter collusion in the case of secret reports.

[39] Spagnolo (2000a) points out that restricting eligibility to the first informant may help to limit those transaction costs while achieving the same deterrence effect.

[40] If the reward is made available to any reporting firm, the two firms would optimally collude, and this would be sustainable whenever condition (2.1) is satisfied.

To counter this, the authority may restrict the reward to the case in which only one firm reports. Then consider the strategy that consists in colluding and randomly selecting one firm (with equal probability) for reporting the collusion to the authority. The value of such a strategy would be

$$\frac{1}{1-\delta}\left(\pi^M + \frac{R-F}{2}\right),$$

and such collusion would thus be sustainable under (2.1) as soon as $R > (1+2\rho)F$.

information generation. For example, following Spagnolo (2000a) and Motta and Polo (2000), I have assumed that collusion leaves evidence only when it is successful, that is, when no firm deviates. This can be an appropriate assumption in some instances, but in other situations firms may wish to denounce collusion precisely when they decide to cheat, and may well be able to bring convincing evidence even in that case. Here is a two-step collusion stage that captures this idea:

1. Step 1: Firms choose simultaneously to either consider collusion or refuse collusion. Whenever at least one firm refuses collusion, the outcome is the competitive one (π^C, π^C). If, instead, both firms are open to collusion, an agreement is signed, say, which generates the hard information described herein; in that case the game proceeds to step 2.
2. Step 2: Firms choose simultaneously to compete or collude; payoffs are those given by Table 3.1.

There are thus now three relevant strategies for the firm, which can be interpreted as compete, collude, and cheat; the net expected payoffs for these strategies are given in Table 3.3.

Collusion is less fragile than before because deviating does not guarantee escaping the fines; it is now sustainable if

$$\pi^D - \pi^M \leq \frac{\delta}{1-\delta}(\pi^M - \rho F - \pi^C), \tag{2.5}$$

which is less restrictive than condition (2.1). However, leniency programs (i.e., here, offering a reduced fine f when reporting collusion) can now help deter collusion even in the case of public reporting. Because a deviating firm abandons any hope of collusion for the future, it is willing to report collusion as long as the leniency program is sufficiently attractive, that is, if $f < \rho F$. A leniency program is therefore useful to deter collusion when

$$\pi^D - \pi^M \leq \frac{\delta}{1-\delta}(\pi^M - \rho F - \pi^C) < \pi^D - \pi^M + (\rho F - f). \tag{2.6}$$

After some learning period, firms will likely adapt to the use of leniency programs and other bounty mechanisms, and they will try to minimize the amount

Table 3.3. *The role of collusion evidence*

	Firm 1		
Firm 2	Compete	Collude	Cheat
Compete	(π^C, π^C)	(π^C, π^C)	(π^C, π^C)
Collude	(π^C, π^C)	$(\pi^M - \rho F, \pi^M - \rho F)$	$(\underline{\pi} - \rho F, \pi^D - \rho F)$
Cheat	(π^C, π^C)	$(\pi^D - \rho F, \underline{\pi} - \rho F)$	$(\pi^C - \rho F, \pi^C - \rho F)$

of information that could be found by the competition authorities. However, hard information may still be required to implement collusion in an effective way. For example, memos may be needed when collusive agreements are too complex to be left to the sole accuracy of individuals' memories. Furthermore, if a firm delegates the negotiation of the collusive agreement to an agent, this agent will need to report to the head of the firm; there again, hard information about the details of the agreement may be required to minimize the agency problems associated with delegation. Cartel policies could take advantage of these intrinsic agency problems and try to exacerbate them.[41]

Suppose, for example, that n employees have access to the hard information generated by collusion. Then, even if reporting is public, introducing a reward r for individual informants would force colluding firms to compensate each employee for preventing them from becoming whistleblowers, thereby reducing the benefits from collusion. If employees are protected from retaliation by the industry (i.e., they can leave the firm in any period and cannot be threatened about their future job opportunities), then, in order to discourage whistleblowing, the firms must promise *each employee* the equivalent, of r; the best way to do this would be to grant the employee in each period, as long as that whistleblowing does not occur, a bonus

$$b = (1 - \delta)r.$$

The minimal reward $\underline{r}$ needed to deter perfect collusion would then be such that[42]

$$\pi^D - \pi^M = \frac{\delta}{1-\delta}(\pi^M - \rho F - \pi^C) - n\underline{r}, \tag{2.7}$$

and it is likely to be lower than the minimal reward $\underline{R}$ needed to extract the information directly from the firm, given by (2.4) set to equality:

$$\underline{R} + \rho F = \frac{\delta}{1-\delta}(\pi^M - \rho F - \pi^C) = \pi^D - \pi^M + n\underline{r}.$$

Such a bounty mechanism would be a fortiori more effective if the employees did not stay forever within the firm. For example, if employees work for only

[41] Exacerbating internal agency problems may, however, have a cost, as firms delegate decisions not only for collusive purposes, but also for "good" efficiency-enhancing reasons.

In the United States, individuals, too, benefit from a leniency program, which shields them from criminal sanctions (including jail). There are, however, no positive incentives to report information. In contrast, under the Civil False Claim Act, individuals that inform the government on fraud in procurement contracts can get up to 30 percent of the damages paid by convicted suppliers. (See Kovacic, 1996, and Tokar, 2000, for a discussion of this whistleblowing mechanism.)

[42] In practice, firms may try to retaliate – in particular, the employee's job opportunities would become more uncertain. However, the reward r could easily represent a huge sum compared with the employee's expected discounted lifetime salary (think of a sizeable fraction stream of the fines imposed on the firms, as in the U.S. bounty device for procurement fraud). Furthermore, it may be easier to keep reporting secret in the case of a single individual, as compared with that of a firm.

one period, the firm should grant in each period a bonus $B = r$. The minimal reward $\underline{\hat{r}}$ needed to deter perfect collusion would then be given by

$$\pi^D - \pi^M = \frac{\delta}{1-\delta}(\pi^M - \rho F - \pi^C - n\underline{\hat{r}}), \tag{2.8}$$

and would thus be much lower than $\underline{R}$ when δ is close to 1 because

$$\underline{R} = \pi^D - \pi^M - \rho F + \frac{\delta}{1-\delta}n\underline{\hat{r}}.$$

Rewards do not come without a cost. In particular, granting secret rewards for secret reports is likely to exacerbate the enforcer's temptation to abuse public funds. This cost has been discussed at length by political scientists, but how this cost – and, more generally, the underlying reasons for limiting secret or public rewards – interacts with the design of antitrust enforcement still remains an open issue.

Leniency Programs. In practice, competition authorities rarely offer rewards for reporting collusion. Instead, the leniency programs seek to encourage defection from cartel agreements by giving amnesty from criminal prosecution. Amnesty may be offered only to the first informant, or it may be extended to later informants if they bring additional evidence that increases the likelihood of success of the prosecution. Likewise, the program may apply only to information given before an enquiry is launched, or it may be extended to information given after an investigation has been started.[43] These leniency programs have been successful, particularly in the United States, both in terms of the number of successful prosecutions and in the record levels of the fines. This may sound surprising in the light of the analysis given here, as in practice reporting has hardly been kept secret – and the analysis suggests that leniency makes collusion rather more appealing in the case of public reporting.[44]

As emphasized by Motta and Polo (2000), leniency programs may, however, help competition authorities to successfully prosecute a case once an investigation has been launched. This may be particularly useful if the competition

[43] In the United States, amnesty is offered to the first informant only. The European Union program offers a 75–100 percent reduction of the fine to the first cartel member to inform the European Commission before an official investigation is started, as well as a lower 50–75 percent reduction to the first cartel member that would bring information once an investigation has started. In addition, cartel members that "cooperate" with the Commission during the prosecution can benefit from a 10–50 percent reduction of the fine. This applies not only to any member that provides "evidence that materially contributes to establishing the existence of the infringement," but also to a member that "informs the Commission that it does not substantially contest the facts on which the Commission bases its allegations."

[44] There are, of course, various reasons why a cartel member may wish to report collusion in practice. For example, a maverick firm or a new and efficient entrant may want to shake out existing arrangements in order to gain a better position in the market. Also, some individuals may act as whistleblowers because they disagree with their employers' participation in a cartel, or because they seek revenge for having been badly treated – for bad or good reasons – by their employers.

authority can impose some restriction on firms' conduct once a cartel has been exposed. I have assumed thus far that competition authorities could impose fines but not dictate firms' conduct; Spagnolo (2000a) supposes instead that, once a cartel has been exposed, the competition authority can discipline the industry and force the competitive outcome for ever. In practice, a successful prosecution is likely to have a discipline effect for at least a limited time. In that case, improving the chances of detection and successful prosecution has two benefits: It contributes as before to discourage firms from colluding, and it helps competition authorities to break up existing cartels.

To study this further, change the framework as follows. In each period, the competition authority audits with probability α; if it opens an investigation and a firm reports collusion, the authority imposes a reduced fine f and can force the competitive outcome in that period; otherwise, it must prosecute and can force the competitive outcome – and impose the maximal fine F – only with some probability p.[45]

Three strategies are relevant for the firms: compete (always compete), collude (always collude and never report collusion), and report (always collude and report collusion whenever an investigation is opened). The value attached to both firms following the first strategy is

$$V^{\text{compete}} = \frac{\pi^C}{1-\delta}.$$

If firms instead collude and never report, in each period they get the cartel profit, π^M, except when there is an audit followed by a successful prosecution, which happens with probability αp, in which case the competitive outcome is implemented:

$$V^{\text{collude}} = \frac{(1-\alpha p)\pi^M + \alpha p(\pi^C - F)}{1-\delta}.$$

Last, if firms choose to collude but cooperate with the authority when an investigation is launched, they lose the cartel profit whenever there is an audit but benefit in that case from the reduced fine:

$$V^{\text{report}} = \frac{(1-\alpha)\pi^M + \alpha(\pi^C - f)}{1-\delta}.$$

Intuitively, a report equilibrium exists when the probability of investigation is not too large, because the firms' payoff to colluding is then large – and whenever $f < F$, each firm has an incentive to report collusion when an investigation is opened, if it expects the others to do so. A collude equilibrium also exists when α is small, and this equilibrium is more profitable for the firms if the prosecution effort p is not too large. Last, the competition authority can deter collusion if

[45] In addition, Motta and Polo suppose that firms' cooperation also speeds up the prosecution stage. In their model, when the authority opens an investigation in period t, it can force the competitive outcome in period $t+1$ if firms cooperate; otherwise, it must prosecute in period $t+1$ and can force the competitive outcome (with probability p) only in period $t+2$.

it is very active (only the compete equilibrium exists when α and p are both close to 1). Welfare (gross of enforcement costs) is highest in the compete regime: In the report equilibrium, it is lower both because collusion occurs with some probability $(1-\alpha)$ and because collusion deterrence occurs ex post, with a lag of one period. In the collude regime, welfare is further reduced by the additional prosecution delay (the competition authority needs two periods to enforce competition in one period).

To determine the optimal enforcement policy, Motta and Polo suppose that the authority faces a budget constraint of the form

$$B(\alpha, p) \leq \bar{B},$$

where $B(\alpha, p)$ denotes the budget expenses associated with the audit probability α and the quality of prosecution p, and $\bar{B}$ represents the authority's maximal budget. Motta and Polo first characterize the best policy for each type of equilibrium. As usual, to deter collusion as effectively as possible, it is always best to impose the largest fine in the absence of cooperation from the firms. It is further shown that, to induce cooperation with the competition authority (the report regime), the best policy is "full leniency" ($f = 0$): To induce firms to report collusion, the competition authority can either reduce the fine ($f < F$) or increase the quality of prosecution. However, the latter has a cost – it requires increasing teams' size, thereby lowering the probability of investigation.

Conversely, the best policy in a collude regime consists in avoiding any leniency ($f = F$). This has no impact on the equilibrium path (because firms never cooperate with the authority), but it eliminates the perverse procollusive effect of leniency programs by reducing the expected fines imposed in the case of an audit.

Similarly, the best way to deter collusion completely (i.e., to impose the compete regime) is to grant no leniency. This is not completely straightforward here, because in principle the competition authority could try to destabilize collusion by inducing reports once an audit is launched. However, one must bear in mind that introducing leniency generates additional ways of collusion. For example, firms may choose to always collude and report whenever there is an audit. This is an equilibrium whenever pure collusion (without reporting) is an equilibrium and reporting is profitable.[46] Therefore, leniency could help to

[46] To see this, note first that no firm has an incentive not to report collusion if it expects the other to report it. Thus the only relevant constraint is

$$\pi^D - \pi^M \leq \delta\left(\hat{V}^R - \frac{\pi^C}{1-\delta}\right),$$

which is less restrictive than the one for the collude regime,

$$\pi^D - \pi^M \leq \delta\left(\hat{V}^M - \frac{\pi^C}{1-\delta}\right),$$

whenever the report regime is more profitable (implying $\hat{V}^R > \hat{V}^M$).

prevent collusion only if the following were true.

1. The collude regime was sustainable in the absence of leniency, which in particular implies

$$\pi^C \leq (1 - \alpha p)\pi^M + \alpha p(\pi^C - F). \tag{2.9}$$

2. In addition, this collude regime is more profitable than the report regime, which is the case only if $V^{\text{collude}} > V^{\text{report}}$:

$$(1 - \alpha p)\pi^M + \alpha p(\pi^C - F) > (1 - \alpha)\pi^M + \alpha(\pi^C - f)$$

or

$$(1 - p)\pi^M + p(\pi^C - F) > \pi^C - f;$$

that is, if the audit has a low probability of success,

$$p < \bar{p}(f) \equiv \frac{\pi^M - \pi^C + f}{\pi^M - \pi^C + F}. \tag{2.10}$$

However, when the two conditions (2.9) and (2.10) are satisfied, a firm has no incentive to deviate from the collude regime even once an audit is launched. Such deviation would yield

$$\pi^C - f + \frac{\delta\pi^C}{1 - \delta} < (1 - p)\pi^M + p(\pi^C - F) + \frac{\delta}{1 - \delta}[(1 - \alpha p)\pi^M + \alpha p(\pi^C - F)],$$

where the right-hand side is precisely the firm's expected profit once an audit is launched, in the absence of reporting.

Last, Motta and Polo determine the optimal enforcement policy, as a function of the resources of the competition authority. It is, of course, optimal to deter collusion if the authority has sufficient resources to do so, and, as we just saw, the best way to achieve this result is to rule out any leniency. If the authority has less resources, it must choose between two colluding regimes: a regime without reporting – in which case it is best not to introduce leniency – and one in which firms report once an investigation is launched – thanks to a "full-length" leniency program, together with a high prosecution effort.[47] The trade-off between those two regimes is as follows. Inducing reporting through a leniency program improves the success of an audit but requires a commitment to a high prosecution effort ($p > \bar{p}$) in order to convince firms that it is in their interest to report.[48] In contrast, ignoring the possibility of reporting allows the competition authority to save on prosecution and devote more resources for launching additional investigations.

[47] As pointed out herein, firms must find it profitable to agree on reporting in the case of an audit, which is the case only if $p > \bar{p}$.

[48] Note that the quality of prosecution plays no role in fines if firms report collusion. Still, to induce reporting, the competition authority must maintain a credible, high probability of successful prosecution and thus spare the needed resources.

2.2. Fighting Facilitating Practices

Short of fighting collusion directly, competition policy can try to attack practices that facilitate firms' coordination. These practices may, for example,

- help firms to "agree" on the terms of the collusive agreement (e.g., informal gatherings giving the opportunity to engage in "cheap talk" leading to the selection, or negotiation, of the desired terms of agreements)[49];
- serve its implementation (e.g., communication devices allowing firms to exchange hard or soft information and better adapt their strategies to the environment); and
- contribute to its enforcement (e.g., practices that enhance the detection of deviations from a collusive agreement and/or the severity of the punishments that can be inflicted on deviators).

Here I discuss some illustrations of the latter two points.

2.2.1. Communication Devices

It has long been recognized that communication can facilitate collusion.[50] First, communication may help firms to coordinate themselves on a particular equilibrium. This coordination problem is particularly acute in the context of repeated games, where the set of equilibria can be very large.[51] Second, information about rivals' past behavior – be it direct observation of their strategies, or enhanced information on the environment that allows a more accurate inference of their behavior – allows firms to better detect deviations and trigger punishments. For example, in a standard repeated Bertrand duopoly with linear costs, if prices are publicly observed, firms can sustain collusion if and only if

$$\frac{\pi^M}{2} \geq (1-\delta)\pi^M + \delta \times 0,$$

that is, whenever $\delta \geq \frac{1}{2}$. However, if firms observe their rival's price only every T periods, then collusion becomes sustainable only if

$$\frac{\pi^M}{2} \geq (1-\delta^T)\pi^M + \delta^T \times 0,$$

that is, whenever $\delta \geq \delta^*(T) \equiv (\frac{1}{2})^{1/T}$. Thus, collusion is more difficult to sustain when firms observe each other's price less often; $\delta^*(T)$ increases with T and

[49] McCutcheon (1997) points out that, by facilitating the renegotiation of the continuation equilibria once a deviation has occurred, meetings in "smoke-filled rooms" can actually make collusion more difficult to sustain (because they reduce the magnitude of the punishments that can be inflicted on deviators).

[50] See Kühn (2000) for a recent discussion. There is also a substantial empirical literature (see, e.g., Van Huyck, Battalio, and Beil, 1990, and Cooper et al. 1989).

[51] Farrell (1987) was among the first to formalize the idea that "cheap talk" can help resolve this "strategic uncertainty" problem.

tends to 1 when T grows infinitely. Relatedly, Green and Porter (1984) have shown that collusion is more difficult to sustain when firms cannot observe or infer each other's behavior.[52]

Last, communication may help in devising more efficient collusive agreements, particularly when firms have private information about cost or demand conditions. Communication may, for example, allow firms to allocate production to the most efficient firm. Athey and Bagwell (2001) analyze this issue by considering a duopoly in which firms' unit costs can take two values, high or low,[53] with independent draws in each period; each firm knows the realization of its costs but does not observe that of its rival. In addition, demand is stationary and inelastic; the monopoly price is thus equal to consumers' reservation price, which is constant over time. In the absence of communication between the firms, productive efficiency, which requires the low-cost firm to serve the entire market whenever the two costs differ, cannot be achieved if firms want to maintain the monopoly price in each period.[54]

In this context, to be sustainable collusion must meet two types of constraints:

1. Off-schedule constraints: These are the standard constraints that must be met even in the absence of asymmetric information; a firm may be tempted to cheat and undercut its rival if the short-term gains from such a deviation outweigh the loss of profits in the following periods.
2. On-schedule constraints: These constraints derive from the asymmetry of information between the firms; among the prices or quantities that are compatible with the collusive strategy, a firm must be induced to select the one that corresponds to the realization of its cost.

Athey and Bagwell point out that, when firms are patient, only the latter constraints are relevant. In that case, firms will maintain the monopoly price, with or without communication; forbidding communication can thus have the adverse effect of simply preventing firms from achieving productive efficiency. However, if firms are more impatient, preventing collusion may force them the charge lower prices.

2.2.2. Resale Price Maintenance

Competition authorities' attitude toward vertical restraints generally treats price restrictions more severely than nonprice restrictions, such as exclusive territories, selective distribution, and so on. In particular, resale price

[52] From a different perspective, Compte (1998) and Kandori and Matsushima (1998) argue that in games in which players have private information about the history of play, introducing public communication gives a recursive structure to the game and thus allows the use of dynamic programming techniques.

[53] Athey, Bagwell, and Sanchirico (1998) consider the case of a continuum of types.

[54] This feature of the model captures the essence of the role of communication. In richer models, in the absence of formal communication, firms could use additional variables (e.g., prices) to reveal part of their information.

maintenance (RPM) is often viewed as illegal per se or, at the very least, as most probably undesirable. This consensus against RPM contrasts with the economic analysis of vertical restraints, which shows that both price and nonprice vertical restrictions may either improve or harm economic efficiency – and often provide alternative ways to achieve the same objective. Furthermore, many arguments used in court in favor of nonprice restrictions would apply as well to RPM. One argument made in practice against RPM, however, is that it could facilitate horizontal agreements. For example, in *Sylvania and Business Electronics*, the Supreme Court repeatedly relied on that argument to justify the per se illegality of RPM:

> Our opinion in GTE Sylvania noted a significant distinction between vertical non-price and vertical price restraints. That is, there was support for the proposition that vertical price restraints reduce inter-brand price competition because they 'facilitate cartelizing.' . . . The authorities cited by the Court suggested how vertical price agreements might assist horizontal price fixing at the manufacturer level (by reducing the manufacturer's incentive to cheat on a cartel, since its retailers could not pass on lower prices to consumers) or might be used to organize cartels at the retailer level. Similar support for the cartel-facilitating effect of vertical non-price restraints was and remains lacking.[55]

As stated, the argument that RPM can facilitate collusion by reducing the manufacturer's incentive to deviate and lower its wholesale price is not very convincing. A manufacturer could "cheat" on the cartel agreement by modifying both the retail price and the wholesale price at the same time; RPM might actually make such a deviation more appealing, by ensuring that the cut in wholesale price is not partially appropriated by retailers. Jullien and Rey (2000) explore the issue further, starting from the idea that, under RPM, retail prices are centrally set by the manufacturer and thus do not fully adjust to local variations on retail costs or demand; as a result, retail prices are more uniform under RPM, and deviations from a tacit agreement are thus more easily detected. It follows that RPM, although being less efficient because it generates less flexible prices, can be adopted to facilitate interbrand collusion.

Consider the following framework. Two infinitely lived producers ($i = 1, 2$) sell to short-sighted retailers. In each period, each producer signs an exclusive contract with a retailer. Demand is linear in prices and stochastic:

$$D_i(p_i, p_j) = d + \varepsilon_i - p_i + \sigma p_j, \qquad i \neq j = 1, 2,$$

where shocks ε_1 and ε_2 are independently and uniformly distributed on the interval $[-\bar{\varepsilon}, \bar{\varepsilon}]$. Production and retail costs are normalized to 0, and manufacturers face a fixed cost k. At the contracting stage, shocks ε_1 and ε_2 are unknown to all parties and manufacturers have all the bargaining power. Then, ε_i is observed by retailer i but not by the manufacturers nor the other retailer.[56]

[55] 485 U.S. 717 (1988) at 725–6.

[56] A similar analysis applies to the case of independent shocks on retailers' costs.

In the absence of RPM, each manufacturer i offers a contract composed of a franchise fee A_i and a wholesale price w_i; under RPM, it can moreover set the retail price p_i. The timing of the stage game is thus as follows:

1. First, each manufacturer i secretly offers a contract (A_i, w_i) or (A_i, w_i, p_i) to a retailer, who accepts it or not.
2. Second, each retailer i observes ε_i and, if it has accepted the contract, sets the retail price p_i (at the level chosen by the manufacturer under RPM).
3. Third, demands and profits are realized; each manufacturer further observes the retail prices and the nature of the contract signed by its competitor.

In the absence of RPM, if it accepts the contract (A_i, w_i), retailer i will set its price to[57]

$$p_i = p_i^e + \frac{\varepsilon_i}{2}, \tag{2.11}$$

where the expected price, p_i^e, is the best response to retailer j's expected price p_j^e:

$$p_i^e \equiv \frac{d + w_i + \sigma p_j^e}{2}. \tag{2.12}$$

By setting the franchise fee A_i so as to recover expected retail profits, producer i can thus get an expected profit equal to

$$\pi\left(p_i^e, p_j^e\right) \equiv p_i^e\left(d - p_i^e + \sigma p_j^e\right) - k.$$

Furthermore, through its wholesale price, each producer can perfectly monitor the expected retail price of its good.[58] Therefore, the stage game is formally identical to one in which each producer i "chooses" p_i^e – thereby generating a distribution of retail prices given by $p_i = p_i^e + (\varepsilon_i/2)$ and expected profits $\pi(p_i^e, p_j^e)$. If both producers pick the same expected price p^e, their expected profit is then given by

$$E\left[\left(p^e + \frac{\varepsilon_i}{2}\right)\left(d + \varepsilon_i - \left(p^e + \frac{\varepsilon_i}{2}\right) + \sigma p^e\right) - k\right] = \pi(p^e, p^e) + v, \tag{2.13}$$

where

$$v \equiv \frac{\bar{\varepsilon}^2}{12}$$

denotes the variance of the retail prices.

[57] Only the rival's expected price matters, because of the linearity of the demand function and the independence of the two demand shocks.

[58] This applies both along an equilibrium path and along a unilateral deviation, because retailer i's expected price depends only on w_i and the anticipated value of p_j^e.

When attention is restricted to symmetric equilibria, and when the fact that shocks are uniformly distributed is used, it can be shown that the best collusive strategy is a trigger strategy of the following form: "choose an expected price p^e; stick to p^e as long as realized prices are compatible with this agreement, that is, $p_i, p_j \in [p^e - \bar{\varepsilon}/2, p^e + \bar{\varepsilon}/2]$; otherwise play $\underline{s}$ (the strategy that supports the equilibrium with the lowest payoff, $\underline{\pi}$)."

If both manufacturers follow this strategy, their expected profit is equal to $\pi(p^e, p^e) + v(\bar{\varepsilon})$. If a deviation was always detected with probability 1, no deviation would thus be profitable if

$$\max_p \pi(p, p^e) - \pi(p^e, p^e) \leq \frac{\delta}{1-\delta}[\pi(p^e, p^e) + v - \underline{\pi}]. \qquad (2.14)$$

However, a small deviation, $p_i^e \in [p^e - \bar{\varepsilon}, p^e + \bar{\varepsilon}]$, will be detected only with probability $|p^e - p_i^e|/\bar{\varepsilon} < 1$. In particular, it can be checked that a small deviation (p_i^e slightly different from p^e) will not be profitable only if

$$\bar{\varepsilon}|d - (2-\sigma)p^e| \leq \frac{\delta}{1-\delta}[\pi(p^e, p^e) + v - \underline{\pi}]. \qquad (2.15)$$

It turns out that the two conditions (2.14) and (2.15) are actually necessary and sufficient for the sustainability of collusion. In the absence of RPM, the most profitable collusive strategy thus consists in maintaining in each period the expected price, p^F, which maximizes $\pi(p^e, p^e)$ subject to (2.14) and (2.15).

When producers adopt RPM and impose the same rigid price $p_i(\varepsilon_i) = p^e$, they ignore their retailers' information about the demand shocks and thus generate lower expected profits, that is, $\pi(p^e, p^e)$, instead of $\pi(p^e, p^e) + v$ as before. Adopting RPM thus facilitates collusion by making deviations easier to detect, but it also hurts collusion by making deviations more attractive – deviating with franchise contracts generates an additional profit v compared with RPM.[59] Collusion on a rigid price p^e is sustainable if and only if

$$\max_p \pi(p, p^e) + v(\bar{\varepsilon}) - \pi(p^e, p^e) \leq \frac{\delta}{1-\delta}[\pi(p^e, p^e) - \underline{\pi}]. \qquad (2.16)$$

RPM thus allows one to ignore condition (2.15) – that is, it facilitates detection – but it results in a more stringent condition than (2.14) – because price rigidity, by itself, hurts profits.

Jullien and Rey (2000) show that if the noise is not too important (the adverse effect of price rigidity on profits otherwise dominates), there exists a range of

[59] As long as the nature of the contract (RPM or not) is public, allowing RPM increases the set of equilibria; in particular, the equilibria described herein resist deviations involving RPM, because (i) any such deviation is detected with probability 1; and (ii) profits from those deviations are lower than those achieved with simple franchise contracts, as RPM makes no use of the retailer's information.

Besides improving the detection of deviations, RPM can also facilitate collusion by allowing for tougher punishments. I will ignore this aspect here.

values for the discount factor in which the most profitable collusive strategy uses RPM. This is the case for "intermediate" values of the discount factor: When it is close to 1, it becomes possible to sustain the expected prices at the monopoly level even with franchise contracts, which is more profitable because this makes use of the retailers' information; and for very low values of the discount factors, only prices close to the level of the static Nash equilibrium can be sustained both with and without RPM, so again it is better not to impose rigid prices on retailers. There exists a middle range, however, in which RPM allows producers to sustain higher prices, which more than compensate the loss of profitability attaching to price rigidity.

This analysis has welfare implications. Note first that imposing price rigidity may be good for consumers: Consumer surplus is a convex function of the demand $(d + \varepsilon_i - p_i + \sigma p_j)$, so that, for a given expected price, consumers prefer prices that do not adjust to demand shocks – they would, however, favor prices that adjust to cost shocks. Furthermore, in the case of demand shocks, and despite its negative impact on profits, price rigidity increases total welfare.[60] However, firms will find it profitable to adopt RPM in equilibrium only if it leads to an increase in prices, sufficient to offset the adverse impact of price rigidity. Building on this, it can be shown that whenever the scope for collusion is substantial, banning RPM is socially optimal.[61]

3. MERGER CONTROL

Merger control differs significantly from cartel law enforcement. Whereas the latter case focuses on the sanction of past behavior, merger control requires the assessment of future conduct. Furthermore, although there is a general consensus against price-fixing agreements, there is more divergence in the competition authorities' policies toward proposed mergers. The general principle is to

[60] With the use of $\partial S/\partial p_i = -D_i$ and $\pi_i = p_i D_i$, the total expected surplus is given by

$$\begin{aligned}\frac{\partial W}{\partial p_i}(p_1, p_2; \varepsilon_1, \varepsilon_2) &= \left(\frac{\partial S}{\partial p_i} + \frac{\partial \pi_i}{\partial p_i} + \frac{\partial \pi_j}{\partial p_i}\right)(p_1, p_2; \varepsilon_1, \varepsilon_2)\\ &= -(d + \varepsilon_i - p_i + \sigma p_j) + (d + \varepsilon_i - 2p_i + \sigma p_j) + (\sigma p_j)\\ &= -p_i + \sigma p_j.\end{aligned}$$

With integration, the total surplus is thus given by

$$W(p_1, p_2; \varepsilon_1, \varepsilon_2) = C(\varepsilon_1, \varepsilon_2) - \tfrac{1}{2}\left(p_1^2 + p_2^2 - 2\sigma p_1 p_2\right).$$

If the two prices have the same expected value p^e, the expected total surplus is thus

$$W^e = C^e - (1 - \sigma)(p^e)^2 - \tfrac{1}{2}(\mathrm{Var}[p_1] + \mathrm{Var}[p_2]).$$

[61] Banning RPM is actually always socially optimal in the case of shocks on retail costs. In the case of demand shocks, banning RPM is still desirable, for instance, when the best collusive price with franchise contracts is above half the monopoly level, as well as when the best collusive price with RPM is larger than the difference between the monopoly and static competitive levels.

balance the efficiency gains that can be generated by the merger (which are due to economies of scale and scope, sharing of know-how, synergies, etc.) against the increase of market power in the industry. There is less agreement, however, on how to solve this trade-off.[62]

A substantial literature has been devoted to this trade-off. I argue in the next section that insufficient attention has been devoted to the implementation aspects in this area.

3.1. The Efficiency–Market Power Trade-Off

Several papers have explored the trade-off between efficiency gains and increased market power.[63] Farrell and Shapiro (1990) analyze this issue in the context of a general Cournot oligopoly (without a priori restrictions on costs and demand). They first analyze the impact of the merger on consumer surplus, summarized in this context by the evolution of prices, and show in particular that, under reasonable conditions (i.e., when supply decisions are strategic substitutes[64] and the equilibrium is stable[65]), a merger necessarily raises prices in the absence of synergies – even taking into account the reallocation of production from less to more efficient technologies. The intuition is as follows. First, strategic substitutability and stability imply that the "aggregate response" of any k firms to the quantity supplied by the remaining $n - k$ firms is itself a decreasing function, with a slope lower than unity.[66] It thus suffices to show that the merging firms will reduce their supply, assuming that the others do not change their own decisions. This would be obvious if the merging firms were

[62] The U.S. merger guidelines allow for some efficiency defense. European Union merger control focuses instead on the creation or reinforcement of a dominant position (more than a 50 percent market share, say), which has led in the past to efficiency offenses: Efficiency gains that enhance the merged entity's competitive edge may contribute to create a dominant position. For example, in its first negative decision since the adoption of the Merger Regulation in 1989, the European Commission argued that the merger would have given the combined Alenia-de Haviland the unique ability to offer a full range of commuter airplanes (from small to large ones), thereby creating a competitive edge over its competitors (because airlines benefit from dealing with a unique supplier for all their needs). Similarly, in the ATT–NCR case, the Commission mentioned that the venture could benefit from potential synergies between the parents' know-how in telecommunications and terminal equipment, but cleared the merger on the basis that these synergies were too unlikely to materialize (previous similar attempts had failed).

These, however, were early cases in European merger control. Furthermore, the fact that mergers that do not create a dominant position are more easily accepted than interfirm agreements can be interpreted as accepting an efficiency defense for structural changes, as long as firms' combined market share does not exceed 40–50 percent.

[63] Salant et al. (1983) and Perry and Porter (1985) consider a symmetric Cournot model with linear demand and, respectively, linear and quadratic costs, whereas Deneckere and Davidson (1985) analyze a model of Bertrand competition.

[64] That is, firm i's best response $q_i = R_i(q_i) = \max[P(q_i + q_{-i})q_i - C_i(q_i)]$ is a decreasing function of the rivals' aggregate quantity q_{-i}.

[65] Namely, Dixit's (1986) stability condition $C_i''(q_i) > P'(q)$.

[66] See Dixit (1986).

equally efficient, because they would internalize the negative externality that an increase in one firm's supply imposes on the other's margin. A nice revealed preference argument shows that this is still the case when the merging firms have different technologies. Farrell and Shapiro also address the impact of a merger on total welfare. They propose to focus on "external effects" (the impact on consumer surplus and outsiders' profits), for which they provide a simple test: A small reduction in the merging firms' output has a net positive external effect on outsiders and consumers if and only if[67]

$$\Sigma_{i \in \text{outsiders}} \lambda_i s_i > \Sigma_{j \in \text{insiders}} s_j,$$

where $s_i = q_i/q$ denotes firm i's market share and λ_i is related to the slope of firm i's best response function

$$\lambda_i = \frac{-R_i'}{1 + R_i'}.$$

The apparent simplicity of this test makes it very appealing. In particular, a merger between small firms ($\Sigma_{j \in \text{insiders}} s_j$ small) is likely to have a positive external effect, implying that any proposed such merger should be accepted.

This paper and the related literature on the efficiency–market power tradeoff have been criticized for relying excessively on specific assumptions (e.g., Cournot model with homogeneous good vs. Bertrand competition with differentiated products, prespecified functional forms for cost or demand, and no tacit collusion).[68] But even within these specifications, the analysis given here is in practice only moderately useful for merger policy. For example, evaluating the λ_i parameters given here is not straightforward. With the use of equilibrium conditions, λ_i can be expressed as

$$\lambda_i = -\frac{P'(q) + q_i P''(q)}{C_i''(q_i) - P'(q_i)},$$

and thus involves a detailed knowledge of both demand and supply conditions – a knowledge that would probably allow the competition authority to directly compute the postmerger Cournot equilibrium. In practice, unfortunately, it is unlikely that such information would be readily available. Besides, if merger control policy was based on the λ_i parameters, firms might wish to modify their behavior so as to manipulate the policy, an issue not addressed in the aforementioned analysis.

From a theoretical viewpoint, the competition authority again faces an information problem: Firms have privileged information about their motivations for the merger (efficiency gains, increased market power, enhanced scope for

[67] Farrell and Shapiro also show that this test applies to mergers (i.e., noninfinitesimal changes) under some additional assumptions – namely, P'', $P''' \leq 0 \leq C_i'''$.

[68] Building on the development of econometric models with differentiated products (see Berry, 1994 and Berry, Levinsohn, and Pakes, 1995), Hausman, Leonard, and Zona (1994) have advocated for relying more on existing data in the design of the appropriate framework.

collusion, etc.), which the authority should try to extract. For example, suppose that merging firms have private information only over the efficiency gains generated by the merger. In that case, the firms could have been asked to "pay" in one form or another for any negative external effect of the merger, so as to ensure that only socially desirable mergers are proposed. More generally, the competition authority could try to screen merger proposals by using transfers or quasi-transfers in the form of concessions, undertakings, restrictions on future behavior, and the like.

In many cases, however, extracting information may be difficult because of the inherent conflict of interest between the merging parties and the merger control office. For example, suppose instead that firms have private information about their ability to monopolize the market or about the impact of the merger on the scope for collusion in the industry. The authority would prefer mergers that generate only a small increase in market power, but it is precisely the firms with the highest ability to exert market power that will be the most eager to get their merger accepted. As a result of this conflict of interest, the best policy is likely to be a "pooling" policy, which does not try to extract any information from the merging parties.[69] This issue, however, is still an avenue for further research. Of particular interest is the possibility of relying on information brought by third parties (be it customers or rivals); a careful analysis of these parties' incentives (e.g., under which conditions rivals and customers are likely to benefit or lose from the merger) would probably be very useful.

Short of modeling this informational problem explicitly, an alternative approach consists of providing practical guidelines for evaluating the impact of the merger on "market power."

There has been some success along this line for assessing the impact of mergers on market power, using static Cournot models. Dansby and Willig (1979) show, for example, that the average markup in the industry is related to the Herfindahl index, defined as the sum of the squares of firms' market shares.[70] Relatedly, Cowling and Waterson (1976) point out that, when marginal costs

[69] This situation is an example in which preferences are "nonresponsive" in the terminology of the principal-agent literature; see Caillaud et al. (1988). It can occur in regulatory contexts when a welfare-maximizing regulator supervises a labor-managed firm exhibiting the so-called Ward pathology. A more efficient firm then seeks to reduce the number of workers and thus its output; see Ward (1958) and Vanek (1970).

Caillaud and Tirole (2000) study a similar phenomenon for funding infrastructure projects, when an incumbent operator has private information about market profitability. The infrastructure owner may try, for example, to screen projects by requiring from the incumbent a higher capital contribution in exchange for protection from competition; but the incumbent is unfortunately willing to pay more for protection precisely when competition would yield the highest benefits.

[70] In equilibrium, each firm maximizes its profit, of the form

$$P(q_{-i} + q_i)q_i - C_i(q_i),$$

where q_{-i} denotes the aggregate supply of the rivals. The first-order condition leads to (with p denoting the equilibrium price) are constant, the Herfindahl index is also linked to the industry profit (gross of fixed cost),[71] whereas Dansby and Willig (1979) show that it is related to total welfare.[72] This prominent role of the Herfindahl index has been translated into the U.S. merger guidelines, which use the Herfindahl index as a "filter."[73]

Beyond this success, no practical quantitative criteria have been proposed when it comes to evaluating the trade-off between efficiency and market power. Furthermore, by focusing mainly on static models, this literature has ignored

$$p - C_i' = -P'_{qi} = p\frac{s_i}{\varepsilon(p)},$$

where $s_i = q_i/q$ denotes firm i's market share and $\varepsilon(p)$ represents the demand elasticity. Therefore, in equilibrium, firm i's Lerner index, $L_i \equiv (p - C_i)/p$, is equal to $s_i/[\varepsilon(p)]$. The industry average markup or Lerner index is thus equal to

$$L = \Sigma_i s_i L_i = \frac{\sum_i s_i^2}{\varepsilon(p)} = \frac{H}{\varepsilon(p)},$$

where $H \equiv \sum_i s_i^2$ is the Herfindahl index for the industry.

[71] If $C_i(q_i) = c_i q_i$, each firm i's equilibrium variable profit is given by

$$\pi_i = (p - c_i)q_i = \frac{p - c_i}{p}\frac{q_i}{q}pq = s_i^2\frac{pq}{\varepsilon(p)}.$$

The aggregate variable profit is thus equal to

$$\pi = \sum\nolimits_i \pi_i = H\frac{pq}{\varepsilon(p)},$$

and can thus be expressed as a function of the total revenue (pq), the elasticity of demand (ε), and the Herfindahl index (H).

[72] Consider a small change of a firm's output, $(dq_i)_i$, starting from the Cournot outcome. The impact on the total welfare, defined as the sum of consumer surplus and firms' profits,

$$W = \int_{P(q)}^{+\infty} D(p)dp + \sum\nolimits_i (P(q) - c_i)q_i,$$

is given (using again the first-order equilibrium conditions) by

$$\begin{aligned} dW &= \sum\nolimits_i (p - c_i)dq_i = \sum\nolimits_i (-P'(q))q_i dq_i \\ &= -P'(q)d\left(\left(\frac{1}{2}\right)q^2 H\right) = -P'(q)\left(qHdq + \left(\frac{1}{2}\right)q^2 dH\right). \end{aligned}$$

The impact thus depends on total production and on the change of concentration, as measured by the Herfindahl index. In particular, the impact of a change in production dq is larger when the industry is more concentrated. A second effect arises from allocative efficiency: Because in equilibrium firms with lower marginal costs have larger market shares, an increase in concentration $(dH > 0)$ tends to move production to the larger and thus more efficient firms.

[73] The first guidelines issued in 1968 considered the four-firm ratio of concentration – the sum of the market shares of the four biggest firms in the industry. This index was replaced in 1982 by the Herfindahl–Hirshman index (HHI) – the Herfindahl index with market shares expressed as percentages: Each market share thus varies from 0 to 100, and the HHI lies between 0 (no concentration, a multitude of infinitesimal firms) and 10,000 (monopoly).

The HHI is used as a screening device, as follows:

- postmerger HHI is below 1,000 – clearance;
- postmerger HHI above 1,800 – in-depth review; and
- postmerger HHI between 1,000 and 1,800 – clearance if the merger increases the HHI by less than 50, in-depth review if the HHI increases by more than 100 (other factors apply when the increase lies between 50 and 100).

important dynamic issues, such as the possibility of entry and predation, or the risk of collusive behavior.[74] As a result, those concerns are in practice treated in a harsh way. For example, in the United States they are mentioned but rarely analyzed in practice. In the European Union, the Commission has adopted a rather rigid attitude that prevents any merger from creating a dominant position even if it generates huge efficiency gains.[75] This can be interpreted as reflecting the belief that a dominant firm can successfully prevent the entry of more efficient competitors, which is in contrast with the static analyses such as those just described, which tend to predict that mergers encourage entry. Any help identifying relevant factors that are likely to be available and could contribute to assessing the risk that a dominant firm could so abuse its position would certainly be welcome. Similarly, the risk of collusion is not likely to be taken into account in the absence of practical guidelines, based on accessible information.[76] The following section describes a recent attempt to make progress in that direction.

3.2. Assessing the Collusion Concern: The Role of Capacity Constraints

Capacity constraints affect the scope for collusion in two opposite ways: They reduce firms' incentives to deviate from a collusive agreement, but also limit the ability to punish such deviations. Most analyses of these two effects have focused on *symmetric* situations, in which all firms have the same capacity,[77] or on duopolistic industries, which is not very helpful for merger analysis.[78]

[74] The extent to which using econometric models can solve this issue is still unclear; for example, if a merger provokes a change of structure that allows a previously competitive industry to sustain collusion, observing past behavior may not suffice to predict the change of behavior. Relatedly, to predict the impact of the merger on the outcome of market competition, empirical studies must make assumptions on the size of efficiency gains.

[75] To take an extreme example, according to the current policy, a merger creating a monopoly would be banned even if "drastic" efficiency gains allowed the new entity to reduce its cost to the point that the monopoly price, based on that reduced cost, was lower than the premerger cost.

[76] In the United States, for example, the merger guidelines mention the concern about future collusive behavior, but this concern is rarely evaluated and does not contribute in practice to the decision. The European Union merger policy developed in the past years the concept of a "collective dominant position" that could be interpreted as an attempt to account for this risk; however, in the recent Airtours case, the Commission starts with a list of relevant factors affecting the scope for collusion, but eventually bases its decision on a static Cournot-like framework.

[77] See, for example, Abreu (1986) for an analysis of symmetric Cournot supergames and Brock and Scheinkman (1985) for a first analysis of symmetric Bertrand supergames, later extended by Lambson (1987).

[78] Capacities are unlikely to be symmetric both before and after the merger, and the collusion concern is relevant for merger policy only when initially there are at least three competitors. Davidson and Deneckere (1984) provide a first exploration of the issue, using standard trigger strategies and exogenous market-sharing rules and starting from a situation with symmetric capacities.

Analyzing tacit collusion in oligopolistic industries with asymmetric capacities is not an easy task. Lambson (1994) provides partial characterizations and shows, for example, that the optimal punishments are such that the firm with the largest capacity gets no more that its minmax profit, whereas smaller firms get more than their respective minmax profits (except if firms are very patient).[79] A few studies, however, have suggested that asymmetry in firms' capacities hurts tacit collusion. Mason, Phillips, and Nowell (1992) note, for example, that in experimental duopoly games, cooperation is more likely when players face symmetric production costs.[80] In a Bertrand–Edgeworth setting, Lambson (1995) shows that introducing a slight asymmetry in capacities hurts tacit collusion; and Davidson and Deneckere (1984, 1990) and Pénard (1997) show that asymmetric capacities make collusion more difficult in duopolies.[81]

3.2.1. A Simple Model

Compte, Jenny, and Rey (2002) further explore the issue, by simplifying the demand and cost sides but allowing for an arbitrary number of firms and asymmetric capacities.[82] The model is a repeated Bertrand–Edgeworth competition game between n firms with zero cost but fixed capacities. It is useful to distinguish the firms' actual capacities, denoted by $k = (k_i)_i$, from their *relevant capacities*, given by $\hat{k}_i \equiv \min\{k_i, M\}$. The demand is inelastic and of size M as long as the price does not exceed a reservation price (thus equal to the monopoly price), normalized to 1. In each period, firms simultaneously set their prices, which are perfectly observed by all buyers and firms; then, buyers go to the firm with the lowest price and decide whether or not to buy; if they are rationed they go to the next lowest priced firm, and so forth, as long as the price offered does not exceed their reservation price.[83] If several firms charge the same price, consumers divide themselves as they wish between those firms. Competition is

[79] Lambson also provides an upper bound on the punishments that can be inflicted on small firms by using penal codes proportional to capacities.

[80] Relatedly, Bernheim and Whinston (1990) show that, in the absence of capacity constraints, tacit collusion is easier when firms have symmetric costs and market shares.

[81] Davidson and Deneckere study the use of grim-trigger strategies in a Bertrand setting, whereas Pénard relies on minmax punishments (which can be sustained if the asymmetry is small) in a linear Cournot setting; both papers also address capacity investment decisions, whereas this paper focuses on the distribution of exogenous capacities. In a duopoly with sequential capacity choices, Benoît and Krishna (1991) show that the second mover cannot enhance its gains from collusion by choosing a capacity different from the first mover's capacity – however, their analysis relies on the assumption that firms share demand equally when charging the same price. Gertner (1994) develops a framework of "immediate responses," in which firms can react at once to each other's price cuts, and shows that asymmetric capacities may prevent firms from colluding perfectly.

[82] Fershtman and Pakes (2000), too, study the interaction between collusion and the industry structure, allowing for entry and exit as well as asymmetric sizes, and using a particular class of (Markovian) pricing policies.

[83] Because demand is inelastic, there is no need for being more specific about rationing schemes.

assumed to be effective, which is the case if firms' aggregate capacity is larger than the market size ($\sum_i k_i > M$), and $\underline{\pi}_i \equiv \max\{0, M - \sum_{j \neq i} k_j\}$ denotes firm i's minmax profit.

Last, all firms use the same discount factor $\delta \in (0, 1)$ and maximize the expected sum of their discounted profits, $\sum_{t \geq 1} \delta^{t-1} \pi_i^t$. To define collusion, define the *value* of an equilibrium as the normalized expected sum of discounted profits that firms obtain along the equilibrium path: $v = (1 - \delta) E[\sum_{t \geq 1} \delta^{t-1} \sum_i \pi_i^t]/M$.[84] *Collusion* is sustainable if a subgame–perfect equilibrium of the infinitely repeated game generates a higher value than the expected aggregate profit generated by any Nash equilibrium of the stage game, and *perfect collusion* is sustainable if there exists a subgame–perfect equilibrium with a value equal to 1. The goal is to characterize, for any distribution of capacities k, the lowest discount factor $\delta(k)$ for which (perfect) collusion is sustainable.

The difficulty in characterizing the set of collusive equilibria comes from the fact that maximal punishments also depend on capacities. A simple case is when small firms are not "too small," namely, when any subset of $(n - 1)$ firms can serve the entire market; the static Nash equilibrium then yields zero profits and obviously constitutes the optimal punishment. Denoting by $\alpha = (\alpha_i)_i$ the distribution of market shares (with $\alpha_i \leq k_i$ and $\sum_i \alpha_i \leq M$), one sees that collusion can then be sustained if and only if

$$\alpha_i \geq (1 - \delta)\hat{k}_i, \qquad i = 1, \ldots, n.$$

Hence collusion is sustainable if and only if $\delta \geq 1 - \max_i \{\alpha_i / \hat{k}_i\}$; the market shares that are most favorable to collusion are thus proportional to the relevant capacities[85] and, for those market shares, collusion is sustainable if and only if

$$\delta \geq \delta(k) = 1 - \frac{M}{\hat{K}}. \tag{3.1}$$

The sustainability of collusion thus depends in that case only on the aggregate relevant capacity, not on its distribution.[86]

3.2.2. α-Equilibria

The analysis is more difficult when small firms are indeed small, that is, when the $(n - 1)$ smallest firms cannot serve the entire market. A simple analysis can, however, be made in that case for a particular class of equilibria, in which firms maintain constant market shares: Define an *α-equilibrium* as a subgame–perfect

[84] Note that v can vary from zero (perfect competition without capacity constraints) to one (complete monopoly or collusion).

[85] Note that $\max_i \{\alpha_i / \hat{k}_i\}$ is smallest when $\alpha_i / \hat{k}_i$ is the same for all firms; that is, $\alpha_i = \hat{k}_i M / \hat{K}$.

[86] A redistribution of capacity may, however, affect the sustainability of collusion if it modifies firms' relevant capacities.

equilibrium such that, on any equilibrium path, each firm i obtains the same share α_i.[87]

The analysis is made easy by the following lemma.

Lemma 3.1. *Fix $\alpha = (\alpha_i)_{i=1,\dots,n}$ satisfying $0 \leq \alpha_i \leq k_i$ for $i = 1, \dots, n$ and $\sum_i \alpha_i \leq M$.*

(i) If there exists a collusive α-equilibrium, then there exists a per period value v satisfying, for $i = 1, \dots n$,

$$\begin{aligned} \alpha_i v &\geq \underline{\pi}_i && (P_i) \\ \alpha_i &\geq (1-\delta)\hat{k}_i + \delta \alpha_i v. && (E_i) \end{aligned}$$

(ii) If there exists v satisfying conditions $\{(E_i), (P_i)\}_{i=1,\dots,n}$, then there exists an α'-equilibrium with value v' for any value $v' \in [v, 1]$ and any market shares $\alpha' = (\alpha'_i)_{i=1,\dots,n}$ satisfying $\alpha_i \leq \alpha'_i \leq k_i$ for $i = 1, \dots, n$ and $\sum_i \alpha'_i \leq M$. In particular, perfect collusion ($v' = 1$, $\sum_i \alpha'_i = M$) is sustainable with any such market shares α'.

By construction, in a collusive α-equilibrium, firm i's continuation payoff is proportional to its market share and thus of the form $\alpha_i v$, for some continuation value v. Condition (P_i) asserts that firm i's continuation payoff cannot be worse than its minmax, whereas condition (E_i) asserts that the threat of being "punished" by $\alpha_i v$ deters firm i from deviating from the collusive path. These conditions are clearly necessary; the lemma establishes that, together, they ensure that the value v (and any larger value) can be sustained as an α-equilibrium. To see this, consider the following path (reminiscent of Abreu's optimal codes for symmetric firms), where p_t denotes the price charged in the tth period of the punishment:

$$p_t = \begin{cases} 0 & \text{for } t = 1, \dots, T \\ p & \text{for } t = T+1, \\ 1 & \text{for } t = T+2, \dots \end{cases}$$

where $T \geq 0$ and $p \in [0, 1]$ are chosen so that $\delta^T[(1-\delta)p + \delta] = v$. No deviation from this path is profitable if it is punished by returning to the beginning of the path. This is obvious for the first T periods, because a deviating firm cannot then get more than its minmax payoff. Condition (E_i) ensures that it is also true in the periods following $T+1$. In period $T+1$, the best deviation consists either in charging the monopoly price (if p is low) or in undercutting the rivals (if p is high, i.e., if $\alpha_i p > \hat{k}_i$). In the former case, the no-deviation condition is given by

$$\delta^{-T}\alpha_i v \geq (1-\delta)\underline{\pi}_i + \delta\alpha_i v \tag{3.2}$$

[87] The restriction applies to all continuation equilibrium paths, including those that follow a deviation, but not to possible deviations.

and is thus implied, too, by (P_i). In the latter case, the no-deviation condition is

$$(1-\delta)\alpha_i p + \delta\alpha_i \geq (1-\delta)\hat{k}_i p + \delta\alpha_i v. \tag{3.3}$$

Because $\alpha_i \leq \hat{k}_i$, it is most restrictive for $p = p^c$, in which case it is equivalent to (E_i).

A similar reasoning applies to higher values $v \in [\underline{v}, 1]$. For the lemma to be established, it suffices to note that the conditions (E_i) and (P_i) are relaxed by an increase in the market shares α.[88]

An implication of this lemma is that the conditions $((P_i), (E_i))_i$ characterize the set of equilibrium values: For given market shares $(\alpha_i)_i$, the set of equilibrium values is an interval of the form $[\underline{v}(k, \alpha, \delta), 1]$, where $\underline{v}(k, \alpha, \delta)$ is the smallest value satisfying conditions $((P_i), (E_i))_i$. In particular, *perfect* collusion $(v = 1)$ is sustainable whenever *some* collusion is sustainable.

This lemma also allows a simple characterization of the sustainability of perfect collusion. Note first that the lowest v satisfying conditions $(P_i)_i$ is

$$\tilde{v}(k, \alpha) \equiv \max_i \frac{\pi_i}{\alpha_i}.$$

Rewriting conditions $(E_i)_i$ as

$$(1-\underline{v})\frac{\delta}{1-\delta} \geq \max_i \frac{\hat{k}_i}{\alpha_i} - 1,$$

one sees that perfect collusion can therefore be sustained only if

$$\frac{\delta}{1-\delta} \geq \frac{\max_i(\hat{k}_i/\alpha_i) - 1}{1-\underline{v}} \geq \tilde{\delta}(k, \alpha) \equiv \frac{\max_i(\hat{k}_i/\alpha_i) - 1}{1-\tilde{v}(k, \alpha)}.$$

Conversely, if $\delta \geq \tilde{\delta}(k, \alpha)$, then perfect collusion is sustainable, using the lemma, with $\underline{v} = \tilde{v}(k, \alpha)$, and, for any δ satisfying this condition, the set of α-equilibrium values is $[\tilde{v}(k, \alpha), 1]$.

Building on this insight, one sees that identifying the market shares that most facilitate collusion amounts to minimizing $\tilde{\delta}(k, \alpha)$ with respect to the α. The denominator in $\tilde{\delta}(k, \alpha)$ is maximal when $\tilde{v}(k, \alpha)$ is minimized, that is, for market shares that are proportional to minmax profits. The numerator is instead minimal when market shares are proportional to (relevant) capacities. Because minmax profits are not generally proportional to capacities, there is a conflict between decreasing $\tilde{v}(k, \alpha)$ (to allow tougher punishments) and decreasing the numerator (to minimize the gains from deviations).[89] Compte et al. (2002) show that the concern for deviations dominates the concern for punishments,[90]

[88] This is obvious for condition (P_i), and it is also true for condition (E_i) because $\delta v \leq v < 1$.

[89] This conflict disappears only when firms are symmetric (same capacity).

[90] In particular, $\max_i(\hat{k}_i/\alpha_i)$ is not differentiable at its maximum; any change away from $\alpha = \alpha^*$ thus generates a first-order increase, which moreover dominates any possible benefit from a higher punishment $1 - \tilde{V}(\beta, \alpha)$.

so that the market shares that are best for collusion are proportional to relevant capacities; this result also determines the minimal threshold for the discount factor, above which collusion is sustainable.[91]

Proposition 3.6. *The threshold $\tilde{\delta}(k, \alpha)$ is minimized for $\alpha = \alpha^*(k)$ defined by*

$$\alpha_i^*(k) \equiv \frac{\hat{k}_i}{\sum_j \hat{k}_j} M,$$

and

$$\delta^*(k) \equiv \tilde{\delta}(k, \alpha^*(k)) = \frac{\hat{k}_n}{\hat{K}}.$$

3.2.3. Applications to Mergers

The threshold $\delta^*(k)$ can be used to assess the impact of capacity transfers and mergers on the scope for collusion. Denoting by $\hat{K}_L$ and $\hat{K}_S$, respectively, the relevant capacity of the largest firm and the sum of the other firms' relevant capacities, one can write this threshold as

$$\begin{aligned} \delta^*(k) &= 1 - \frac{M}{\hat{K}_L + \hat{K}_S} \quad \text{if } \hat{K}_S > M, \\ &= \frac{\hat{K}_L}{\hat{K}_L + \hat{K}_S} \qquad \text{if } \hat{K}_S \leq M. \end{aligned}$$

In particular, when small firms are "really" small (i.e., when they cannot serve the entire market), exacerbating asymmetry, by transferring capacity from a small firm to the largest one, makes collusion more difficult to sustain (δ^* increases): This is because this reduces small firms' retaliation ability (because $\hat{K}_S < M$) and moreover exacerbates the large firm's incentive to deviate if $\hat{K}_L < M$. Building on this insight, Compte et al. (2002) show that, when capacity constraints really matter, the distribution of capacity that most facilitates collusion, for a given total capacity, is the symmetric one.[92]

[91] Focusing on α-collusive equilibria a priori restricts the scope for collusion, by limiting the punishments that can be inflicted on deviating firms; $\delta^*(k)$ thus provides a lower bound for the values of the discount factor for which collusion is sustainable. Compte et al. (2002) show, however, that punishments achieved with α-equilibria are at least as effective as those generated by reverting to a Nash equilibrium of the competitive stage game. Thus, the threshold $\delta^*(k)$ is – weakly – lower than for standard trigger-strategy equilibria.

[92] This result remains valid when the most general class of equilibria is considered. When $\delta(k)$ is defined as the lowest discount factor for which perfect collusion would be sustainable in any subgame–perfect equilibrium (allowing for flexible market shares, mixed strategies, etc.), then, whenever $K < [n/(n-1)]M$, the symmetric distribution k^s minimizes $\delta(k)$; moreover, for $\delta = \delta^*(k^s)$, perfect collusion cannot be sustained whenever the largest capacity is $n/(n-1)$ larger than the smaller one.

Proposition 3.7. *If the total capacity K is sufficiently small, that is, $K \leq [n/(n-1)]M$, then the set of capacity distributions for which $\delta^*(k)$ is minimal is $\{k^s \equiv (K/n, \dots, K/n)\}$.*

Mergers reduce the number of competitors, which is often thought to facilitate collusion. In particular, keeping punishments constant, it reduces the incentives to deviate.[93] This effect dominates when capacity constraints are not too severe, because in that case a merger would have little impact on punishments profits, which in any case are close to zero. However, a merger may exacerbate the asymmetry in capacities when it involves the largest firm. This tends to hurt tacit collusion, and this effect dominates when the capacity constraints are more severe or their distribution is very asymmetric.[94]

Policy Implications.. This analysis suggests merger guidelines that substantially differ from those inspired by static analyses. In particular, for a given number of firms, the Herfindahl test presumes that a more symmetric configuration is more likely to be competitive (the Herfindahl index is minimal for a symmetric configuration). Similarly, the static Nash equilibrium industrywide profits often decrease with symmetry.[95] The aforementioned analysis instead suggests that asymmetry may be procompetitive, as it may hurt tacit collusion.[96] A sufficiently asymmetric configuration may even more than compensate for a reduction in the number of firms: If $\hat{K}_S < M$, *any* merger involving the large firm hurts collusion and may thus benefit competition because, although it reduces the number of competitors, it exacerbates the asymmetry between them (the Herfindahl test would in contrast advise that the premerger situation is more favorable to competition).[97] Last, the analysis provides a sufficient statistic, based on the distribution of capacities (which in practice are generally

[93] Because the gains from a unilateral deviation come at the expense of all rivals, the gains from a joint deviation are lower than the sum of the gains that each deviator could get through a unilateral deviation.

[94] When small firms are not too small ($\hat{K}_S \geq M$), the reasoning applies both to α-equilibria and to more general ones. When $\hat{K}_S < M$, the discussion that follows is based on the analysis of α-equilibria; however, the robustness of the result on the impact of asymmetry suggests that, even for more general collusive equilibria, a merger involving the largest firm is still likely to hurt collusion.

[95] For example, in the framework given here, each firm i gets max $\{0, M - [(n-1)/n]K\}$ if the distribution is symmetric and at least $\max\{0, (M - \hat{K}_S)\hat{k}_i/\hat{k}_n\}$ otherwise. The industrywide profits are thus minimized when the capacity K is distributed evenly among the firms.

[96] Kühn and Motta (2000) emphasize the same insight in a different context, in which "size" refers to the number of varieties that a firm offers.

[97] The analysis here also casts some doubt on standard merger remedies, which consist of divesting some of the capacity of the merged firm and transferring it to other competitors: Such a remedy tends to maintain a reasonable amount of symmetry between the competitors – in order to avoid the creation of a dominant position – but may help tacit collusion.

relatively easy to evaluate), for the assessment of the impact of a merger on the scope for collusion.[98]

4. RESEARCH AGENDA

The thrust of this paper is that more attention should be devoted to implementation problems when the theory of competition policy is built. I have tried to illustrate this point in the context of price-fixing agreements and merger control, but it applies as well, with perhaps even more force, to other areas of competition policy, such as predation cases or the treatment of vertical restraints.

I have briefly described some recent advances, but much remains to be done. The works I have mentioned suggest a few lines of research, including technical topics such as the equilibrium analysis of repeated games for "moderate" discount factors. (A research agenda is helpful for the analysis of factors and practices that affect collusion. Currently, the papers that have made advances in this area often have to restrict their attention to particular classes of equilibria, e.g., symmetric equilibria in Jullien and Rey or constant market shares in Compte et al. It would be nice to have appropriate tools for characterizing more general classes of equilibria.)

Many interesting topics are related to implementation problems. In particular, we need a better understanding of the underlying reasons for the various limitations that we observe in practice:

- conduct supervision rather than price regulation;
- absence of transfers, except fines for prespecified conducts; and
- intervention, mainly ex post.

In practice, we observe many forms of industry supervision, such as regulation, antitrust, or compulsory arbitration, and we can learn a few things from studying what works and when. With this idea in mind, I now briefly sketch a comparison of "regulation" and "antitrust" along various dimensions: procedures and control rights, timing of oversight, information intensiveness and continued relationship, and independence vis-à-vis the political environment. I use this rough comparison to discuss how the institutional features may contribute to an effective supervision of industry, taking into consideration the

[98] Compte et al. (2002) discuss in this light the well-known Nestlé–Perrier merger case. An interesting feature of this merger was the parties' proposal to transfer an important spring (Volvic) to the main other player in the industry (BSN, now Danone). This transfer could be seen as a remedy to avoid the creation of the dominant position, allowing the market share of the merged entity to remain below 50 percent. However, the Volvic spring also had huge unused capacity, and BSN did not have much excess capacity before the merger. Hence, the merger would have created a large asymmetry in capacities between the two remaining players while the proposed transfer restored perfect symmetry (both players could serve the entire market). According to the analysis given here, the merger would have actually made collusion more difficult to sustain absent the transfer of Volvic, but easier to sustain with the transfer.

overseer's imperfect knowledge of the cost and demand structure in the industry, the risk that the overseer may be captured by (collude with) specific interest groups, and his or her limited commitment ability.

4.1. Procedures and Control Rights

Antitrust authorities generally assess the lawfulness of conduct. In contrast, regulators have more extensive powers and engage in detailed regulation; they set or put constraints on wholesale and retail prices, determine the extent of profit sharing between the firm and its customers (as under cost-of-service regulation or earnings-sharing schemes), oversee investment decisions, and control entry into segments through licensing of new entrants and line-of-business restrictions for incumbents.[99]

There is some convergence of regulatory and competition policy procedures. For example, in the United States, regulatory hearings are quasi-judicial processes in which a wide array of interested parties can expose their viewpoints. The enlisting of advocates is prominent in both institutions and is a key factor in the reduction of the informational handicap faced by the industry overseer.[100] There are, however, a couple of differences relative to the role of intervenors. In antitrust enforcement, private parties, although they are more constrained in their access to the oversight process, play a bigger role than in a regulatory process. Competition policy officials occasionally conduct independent industry studies, but the vast majority of cases brought to courts are private suits. Another difference is that interest groups are motivated to intervene in the regulatory process solely by the prospect of modifying policy while they go to court either to modify industry conduct (through a court injuction) or to obtain monetary compensation (e.g., treble damages). Yet another difference between the two institutions is that courts have less control over the agenda than regulators. Although courts can throw out a case, they most often examine it first and may easily become overloaded. Conversely, courts can take only cases that are brought to them – competition authorities have, however, more flexibility.

Another distinction between the two institutions is the separation between investigation and prosecution in antitrust. In contrast, regulators conduct

[99] This general picture of a large number of instruments and of potentially high discretionary power held by regulators is, of course, to be qualified by the many constraints they face in their decision making: procedural requirements, lack of long-term commitment, safeguards against regulatory takings, constraints on price fixing or cost reimbursement rules (cost-of-service regulation, price caps, etc.), cost-based determination of access prices, and so forth.

Also, antitrust authorities and courts sometimes exercise regulatory authority by imposing line-of-business restrictions or forcing cost-of-service determination of access prices. A case in point is when judge Greene became a regulator of the American telecommunications industry. In Europe, where there has been a growing interest in essential facility and market access issues, the European Commission has tried to develop both antitrust and regulatory competencies and methods. Still, the pattern described in the text seems fairly general.

[100] See Dewatripont and Tirole (1999) for a formal analysis.

regulatory hearings and adjudicate on their basis. However, one should not overemphasize this distinction. First, some competition policy makers, such as the European Commission, can both investigate and take action against specific behaviors (subject to the possibility of court appeal). Second, regulatory decisions may be appealed in court in the same way a court decision may be overruled by a higher court.

A last point of departure between the two institutions relates to the consistency requirements. Regulators and courts are both required to apply relatively consistent reasoning. However, whereas precedents play some role in the two institutions, regulators are mainly bound to be somewhat consistent with their previous decisions for the industry they oversee. Courts, in contrast, must also refer to decisions of other courts – particularly in common law systems – as well as to decisions pertaining to other industries. In particular, the uniformity of interventions across industries imposes substantial constraints on the courts' discretion.

4.2. Timing of Oversight

An important difference between regulation and antitrust is that the former operates mainly ex ante and the latter operates ex post. Antitrust authorities assess conduct after the fact, whereas regulators define the rules for price setting, investment, and profit sharing ex ante. Again, some qualifiers are in order. Merger control by European and American competition policy officials requires notification for large mergers and is a quasi-regulatory process.[101] Conversely, an agency's decision of disallowing ex post "imprudent investments," that is, of excluding them from the rate base in a cost-of-service process, is an illustration of ex post decision making in a regulated environment. Nevertheless, the broad picture is that the timing of regulatory decision making differs from that of antitrust enforcement.

Concomitantly, the regulatory process must be more expedient. The necessity not to halt productive decisions as well as rules constraining the length of investigations often put pressure on regulators (or quasi-regulators such as merger control officers) to converge on rapid decisions. In contrast, the ex post nature of antitrust intervention does not call for a similar expediency (with the possible exception of predatory cases, in which interim provisions may be necessary to prevent irreversible damages).

Another implication of the timing of government intervention is that the uncertainty about the overseer's decision making differs between the two institutions. Ex ante intervention removes most of the uncertainty about this intervention (although not necessarily about its consequences). It may thus facilitate financing of new investment by alleviating the lenders' informational

[101] See Neven, Nuttall, and Seabright (1993) for a very relevant discussion of institutions in the context of merger control. Except for some licensing agreements, firms may, but are not required to, submit vertical agreements for approval to the European Commission.

handicap with respect to this intervention (to the extent that the lenders may have insufficient expertise in the industry and may thus be concerned about the borrower's superior knowledge about this intervention) and by sharpening the measurement of the borrower's performance (by eliminating extraneous noise not controlled by its management).[102]

Ex ante intervention also provides some commitment by the regulator toward the firm.[103] This commitment is desirable whenever the regulator has the opportunity to exploit the firm's demonstrated efficiency or investment by becoming very demanding.

Ex ante intervention may be particularly valuable when coordination problems are important, as for the design of the articulation between urban and intercity transport networks, or between different modes (rail and buses) of urban transport.

Finally, ex ante intervention may force the firm to disclose information that it would not disclose ex post. Intuitively, it is less risky for the firm to conceal or manipulate information ex post when it knows the state of nature than ex ante when it does not; for instance, the firm may know ex post that a lie about ex ante information that conditioned some business decision will not be discovered, but it may have no such certainty ex ante.[104]

A drawback of ex ante intervention is that it may foster collusion between the industry and the supervisor. The industry knows whom it is facing, whereas it is much more uncertain about whether it will be able to capture the (unknown) overseer in a context in which the oversight takes place ex post. This uncertainty about the possibility of capture increases the firm's cost of misbehaving.

A second benefit of ex post intervention is, of course, the opportunity to take advantage of information that accrues "after the fact." For example, over time it may become clearer what constitutes acceptable conduct. To be certain, ex ante decisions could in principle be flexible enough to allow for ex post adjustments that embody the new information; but properly describing ex ante the nature of future information that will be brought to bear on the determination of acceptability may prove difficult and not generate much gain relative to a pure ex post intervention.

Examples. These various differences suggest as many topics of research. For example, Bergès et al. (2000) develop a framework to study the choice between

[102] That is, the removal of uncertainty may reduce both adverse selection and moral hazard. Note that the removal of regulatory risk per se need not reduce the risk faced by risk-averse investors: to the extent that the regulatory risk in the industry is idiosyncratic, it should be diversified away under perfect capital markets.

[103] To be sure, competition authorities can publish guidelines to preannounce their policy. However, those guidelines need not be followed by the courts.

[104] That is, incentive constraints ex ante are pooled, because they are expressed in expectations. It is therefore easier to elicit information ex ante than ex post, because there are fewer incentive constraints.

ex ante notification and ex post audit; they show that notification is preferable when the competition authority has less knowledge about the industry, whereas ex post audit is more effective when the decisions of the authority are more accurate and, by the same token, more predictable. Aubert and Pouyet (2000) have started to study how the regulatory and antitrust modes of intervention could ideally be combined.

4.3. Information Intensiveness and Continued Relationship

Another useful distinction between antitrust and regulation is that regulatory decisions rely on superior expertise. The regulatory advantage in this respect is threefold. First, a regulatory agency specializes in a specific industry whereas antitrust enforcers have a fairly universal mandate. Second, regulators are usually involved in a long-term relationship with regulated industries whereas judges are not. Third, regulators usually have larger staffs than judges and monitor the firms' accounts on a continuous basis rather than on an occasional one. They can also insist on specific accounting principles (such as account separation) as well as disclosure rules. This information superiority can clearly be more or less important according to the context. It is, for instance, more likely to be substantial in the case of a single-industry firm regulated by a national agency, as for electricity in the UK or in France, than in the case of a multiactivities firm regulated by local agencies, as for the German *Stadtwerke* or the Italian *Aziende*. Furthermore, this superior sectorial expertise may be nuanced by a more limited experience with mechanisms or solutions applied in other sectors (limited benchmarking).

Superior expertise is, of course, a benefit in that it allows better-informed decision making. For example, for a long time, regulators have used cost-based rules for retail and wholesale prices even though the determination of costs is often a difficult task. It is also not surprising that antitrust enforcers are more at ease with cases based on qualitative evidence (price discrimination, price fixing, vertical restraints, etc.) than with those requiring quantitative evidence (predation, tacit collusion, access pricing, etc.)

However, superior expertise may be a handicap when regulators have limited commitment powers. When a regulated firm lowers its marginal cost through efficiency measures or investment, it is tempting for regulators (or politicians) to confiscate the efficiency gains through lower prices. This "ratchet effect," which is strengthened by the regulator's access to cost information, is an impediment to efficiency. Similarly, excessive attention (motivated by superior expertise) may inhibit the firm's initiative. An arm's length relationship may entail more commitment power and help provide better incentives.[105]

A second drawback derives from the way expertise is acquired. Part of the regulatory agencies' expertise stems from the long-term nature of their

[105] See, for example, Crémer (1995) and Aghion and Tirole (1997).

relationship with the industry. But, as is well known, long-term relationships are, in any organization, conducive to collusion. Indeed, regulators have often been more captured by interest groups than judges. This may also be related to the fact that, because regulators have deeper knowledge of a particular industry, a natural career evolution is more likely to involve close links with this industry (i.e., the regulators' expertise may well reinforce "revolving door" problems). Also, the need for such industry-focused expertise may impose some constraints on the recruitment of regulators.

Political scientists have repeatedly pointed out that agencies tend to start by behaving in the public interest, and then become increasingly inefficient, bureaucratized, and more eager to please private interests. For example, Bernstein (1955) contends that the life cycle of an independent agency can be divided into four periods: gestation (production of a regulatory statute), youth (lack of experience, unclear mandate, and creative, aggressive, and crusading spirit), maturity (focus on protection of the agency's existence and power, switch from policing to managing the industry, higher concern with the health of the industry, loss of vitality, and desire to avoid conflicts), and old age (extreme conservatism, loss of creativity, growing backlogs, and apathetic approach to the public interest in regulation). Martimort (1999) provides a very nice analysis of this issue, using a dynamic model of capture. The idea is that when capture is implemented through repeated interaction, the regulatory stake of any period affects the scope for collusion between the regulator and the industry in all previous periods. Therefore, to reduce the social cost of collusion, regulatory stakes must be more and more reduced over time; that is, the regulatory agency must have less discretion and behave more and more like a bureaucrat.

4.4. Independence vis-à-vis the Political Environment

The final dimension along which I compare regulation and antitrust is their relationship to political power. Antitrust authorities are traditionally described as being more independent than regulatory agencies. Although this view is generally correct, it is important to distinguish among forms of regulation and competition policy: An antitrust authority located within a ministry is more likely to be influenced by politics than an independent regulatory agency.

The Anglo-Saxon institution of regulation by an independent commission seeks to emulate the benefits of an independent judicial system. Independence can be partially obtained by offering long, staggered terms to commissioners and by limiting the impact of legislative bodies on the agency's budget and jurisdiction.

The benefits of independence are well known. First, the politicians' concern about public opinion and their taste for campaign contributions make them prone to give in to interest-group pressure. Relatedly, an independent agency may be less sensitive to alternative motivations (such as favoring domestic or public operators), which may reduce regulatory uncertainties and offer a better commitment to fair treatment of all competitors.[106] Independent agencies are less vulnerable to interest groups, although their officers are not immune to the influence of the revolving door and sometimes of bribes; decisions can then be reached more on efficiency grounds and less on the basis of the relative power of pressure groups. This is, of course, a substantial advantage of independence. Relatedly, independence may strengthen the agency's commitment power by limiting both opportunistic captures of the firm's rents and "soft-budget constraint" problems. Second, independence allows for more transparency. In France, for instance, the only European country where the air traffic control is directly managed by the State, through the DGAC, airline companies have argued that the accounting system does not provide a clear enough basis for the fees charged to the companies. Many countries have chosen to give the air traffic control to either an independent agency or a nonprofit organization, and some countries such as the United States are even considering privatizing it.

The cost of independence is also well known. Independent agencies and courts may lack accountability and follow their own agenda instead of the nation's agenda. For example, in the case of the complex oversight of network industries, the public has an especially ill-informed opinion and often no opinion at all. In such circumstances, the public cannot verify whether the agency really acts in its interests, which calls for limiting their discretion: procedural requirements, limited commitment power, possibility of appeal, and so on.

There has been remarkably little work in economics on the costs and benefits of independence. Lenoir's model (Lenoir, 1991) depicts a three-party hierarchy: a political principal (the legislative body, or more realistically its committees and subcommittees in charge of overseeing the industry), a regulatory agency, and the industry (e.g., a monopolist). Lenoir focuses on a particular version of accountability, in which the agency does not waste resources: The political principal can adjust the resources of a dependent agency to the latter's real needs, according to circumstances, while an independent agency's budget is protected from political intervention. However, other versions of accountability would be consistent with the overall argument.

The cost of dependency in Lenoir's model is the influence of politics on regulatory decisions. The influence of the interest group (the industry) on the regulatory agency flows through the political principal. Namely, the industry can offer campaign contributions to the political principal, who can threaten to reduce a dependent agency's budget and thus its rent; the political principal can then offer not to ratchet down its budget to the efficient level in exchange for the agency's lenient treatment of the industry. Thus, a dependency relationship

[106] Examples of such concerns can be found in the allocation of airport slots or rail slots. For instance, French private airlines have repeatedly complained about the allocation of slots, charging the State agency (the Direction Générale de l'Aviation Civile, or DGAC) with favoritism toward Air France and Air Inter. (For instance, at some point Euralair had authorization for Toulouse–Orly flights but no slots allocated to operate such flights.)

creates a quid pro quo and allows the industry to have a indirect impact on regulatory decisions.[107]

ACKNOWLEDGMENTS

I thank my colleagues in Toulouse very much for their support, and most particularly Jean-Jacques Laffont and Jean Tirole for their encouragement and suggestions. A first draft of this paper benefited from comments and feedback from Cécile Aubert, Bruno Jullien, Massimo Motta, Michele Polo, Giancarlo Spagnolo, and Jean Tirole.

References

Abreu, D. (1986), "Extremal Equilibria of Oligopolistic Supergames," *Journal of Economic Theory*, 39(1), 191–225.

Aghion, P. and J. Tirole (1997), "Formal and Real Authority in Organizations," *Journal of Political Economy*, 105, 1–29.

Athey, S. and K. Bagwell (2001), "Optimal Collusion with Private Information," *Rand Journal of Economics*, 32(3), 428–465.

Athey, S., K. Bagwell, and C. Sanchirico (1998), "Collusion with Price Rigidity," Department of Economics Working Paper 98-23, Massachusetts Institute of Technology.

Aubert, C. and J. Pouyet (2000), "Ex Ante Regulation and Ex Post Antitrust Intervention," mimeo, University of Toulouse.

Aubert, C., W. Kovacic, and P. Rey (2000), "The Impact of Leniency Programs on Cartels," mimeo, University of Toulouse.

Baniak, A. and L. Phlips (1996), "Antitrust Enforcement with Asymmetric Information and Cournot Strategies," mimeo, European University Institute.

Baron, D. and D. Besanko (1984a), "Regulation, Asymmetric Information, and Auditing," *Rand Journal of Economics*, 15, 447–470.

Baron, D. and D. Besanko (1984b), "'Regulation and Information in a Continuing Relationship," *Information, Economics, and Policy*, 1, 447–470.

Baron, D. and R. Myerson (1982), "Regulating a Monopolist with Unknown Cost," *Econometrica*, 50, 911–930.

Becker, G. (1968), "Crime and Punishment: An Economic Approach," *Journal of Political Economy*, 76(2), 169–217.

Benoît, J.-P. and V. Krishna (1991), "Entry Deterrence and Dynamic Competition," *International Journal of Industrial Organization*, 54, 23–35.

Bergès, F., F. Loss, E. Malavolti, and T. Vergé (2000), "Modernisation de la Politique Europénne de la Concurrence: Régime d'Autorisation ou d'Exception Légale?" mimeo, University of Toulouse.

Bernheim, B. D. and M. D. Whinston (1990), "Multimarket Contact and Collusive Behavior," *Rand Journal of Economics*, 21, 1–26.

[107] This is an illustration of the more general point that collusion is enhanced by a mutual power relationship; see Laffont and Meleu (1996).

Bernstein, M. (1955), *Regulating Business by Independent Commission*. Princeton, NJ: Princeton University Press.

Berry, S. (1994), "Estimating Discrete Choice Models of Product Differentiation," *Rand Journal of Economics*, 25, 242–263.

Berry, S., J. Levinsohn, and A. Pakes (1995), "Automobile Prices in Market Equilibrium," *Econometrica*, 63(4), 841–890.

Besanko, D. and D. Spulber (1989), "Antitrust Enforcement under Asymmetric Information," *Economic Journal*, 99, 408–425.

Brock, W. A. and J. Scheinkman (1985), "Price Setting Supergames with Capacity Constraints," *Review of Economic Studies*, 52, 371–382.

Caillaud, B., R. Guesnerie, J. Tirole, and P. Rey (1988), "Government Intervention in Production and Incentives Theory: A Review of Recent Contributions," *Rand Journal of Economics*, 19(1), 1–26.

Caillaud, B. and J. Tirole (2000), "Infrastructure Financing and Market Structure," mimeo, University of Toulouse.

Compte, O. (1998), "Communication in Repeated Games with Imperfect Private Monitoring," *Econometrica* 66(3), 597–626.

Compte, O., F. Jenny, and P. Rey (2002), "Collusion, Mergers, and Capacity Constraints," *European Economic Review*, 46(1), 1–29.

Cooper, R., D. Dejong, R. Forsyte, and T. Ross (1989), "Communication in the Battle of Sexes Game: Some Experimental Results," *Rand Journal of Economics*, 20(4), 568–587.

Cowling, K. and M. Waterson (1976), "Price–Cost Margins and Market Structure," *Economica*, 43, 267–274.

Cramton, P. and T. Palfrey, "Cartel Enforcement with Uncertainty about Costs," *International Economic Review*, 31(1), 17–47.

Crémer, J. (1995), "Arm's-Length Relationships," *Quarterly Journal of Economics*, 2, 275–295.

Crémer, J. and D. Salehi-Isfahani (1989), "The Rise and Fall of Oil Prices: A Competitive View," *Annales d'Economie et de Statistique*, 15/16, 427–454.

Dansby, E. and R. Willig (1979), "Industry Performance Gradient Indexes," *American Economic Review*, 69, 249–260.

D'Aspremont, C., A. Jacquemin, J. Gabszewicz, and J. Weymark (1983), "On the Stability of Collusive Price Leadership," *Canadian Journal of Economics*, 16, 17–25.

D'Aspremont, C. and M. Motta (2000), "Tougher Price Competition or Lower Concentration: A Trade-off for Antitrust Authorities?" in *Market Structure and Competition Policy: Game-Theoretic Approaches*, (ed. by G. Norman and J.-F. Thisse), New York: Cambridge University Press.

Davidson, C. and R. J. Deneckere (1984), "Horizontal Mergers and Collusive Behavior," *International Journal of Industrial Organization*, 2, 117–132.

Deneckere, R. and C. Davidson (1985), "Incentives to Form Coalitions with Bertrand Competition," *Rand Journal of Economics*, 16, 473–486.

Dewatripont, M. and J. Tirole (1999), "Advocates," *Journal of Political Economy*, 107(1), 1–39.

Dixit, A. (1986), "Comparative Statics for Oligopoly," *International Economic Review*, 27, 107–122.

Ellison, G. (1994), "Theories of Cartel Stability and the Joint Executive Committee," *Rand Journal of Economics*, 25, 37–57.

European Union (1996), "Commission Notice on the Non-Imposition of Reduction of Fines in Cartel Cases," July, 96/C207/04, *Official Journal*, C.207.

Farrell, J. (1987), "Cheap Talk, Coordination, and Entry," *Rand Journal of Economics*, 18, 34–39.

Farrell, J. and C. Shapiro (1990), "Horizontal Mergers: An Equilibrium Analysis," *American Economic Review*, 80(1), 107–126.

Fershtman, C. and A. Pakes (2000), "A Dynamic Oligopoly with Collusion and Price Wars," *Rand Journal of Economics*, 31(2), 207–236.

Gertner, R. (1994), "Tacit Collusion with Immediate Responses: The Role of Asymmetries," mimeo, University of Chicago.

Green, E. and R. Porter (1984), "Noncooperative Collusion under Imperfect Price Information," *Econometrica*, 52, 87–100.

Hausman, J., G. Leonard, and J. D. Zona (1994), "Competitive Analysis with Differentiated Products," *Annales d'Economie et de Statistique*, 34, 159–180.

Jullien, B. and P. Rey (2000), "Resale Price Maintenance and Collusion," mimeo, University of Toulouse.

Kandori, M. and H. Matsushima (1998), "Private Information, Communication, and Collusion," *Econometrica*, 66(3), 627–652.

Kaplow, L. and S. Shavell (1994), "Optimal Law Enforcement with Self-Reporting of Behavior," *Journal of Political Economy*, 102(3), 583–606.

Kihlstrom, R. and X. Vives (1992), "Collusion by Asymmetrically Informed Firms," *Journal of Economics & Management Strategy*, 1(2), 371–396.

Kovacic, W. (1996), "Whistleblower Bounty Lawsuits as Monitoring Devices in Government Contracting," *Loyola Law Review*, 29(4), 1799–1857.

Kühn, K.-U. (2000), "Fighting Collusion by Regulating Communication between Firms," mimeo, University of Michigan.

Kühn, K.-U. and M. Motta (2000), "The Economics of Joint Dominance," mimeo, European University Institute.

Laffont, J.-J. (2000), *Incentives and Political Economy*. New York: Oxford University Press.

Laffont, J.-J. and D. Martimort (1997), "Collusion Under Asymmetric Information," *Econometrica*, 48, 1507–1520.

Laffont, J.-J. and D. Martimort (2000), "Mechanism Design with Collusion and Correlation," *Econometrica*, 48, 309–342.

Laffont, J.-J. and M. Meleu (1996), "Mutual Supervision and Collusion," mimeo, IDEI.

Laffont, J.-J. and J.-C. Rochet (1997), "Collusion in Organizations," *Scandinavian Journal of Economics*, 99(4), 485–495.

Laffont, J.-J. and J. Tirole (1993), *A Theory of Incentives in Procurement and Regulation*. Cambridge, MA: MIT Press.

Laffont, J.-J. and J. Tirole (2000), *Competition in Telecommunications*. Cambridge, MA: MIT Press.

Lambson, V. E. (1987), "Optimal Penal Codes in Price-Setting Supergames with Capacity Constraints," *Review of Economic Studies*, 54, 385–397.

Lambson, V. E. (1994), "Some Results on Optimal Penal Codes in Asymmetric Bertrand Supergames," *Journal of Economic Theory*, 62, 444–468.

Lambson, V. E. (1995), "Optimal Penal Codes in Nearly Symmetric Bertrand Supergames with Capacity Constraints," *Journal of Mathematical Economics*, 24(1), 1–22.

Lenoir, N. (1991), "Optimal Structure of Regulatory Agencies Facing the Threat of Political Influence," Master's Thesis in the Science of Transportation, Massachusetts Institute of Technology.

Malik, A. and R. Schwab (1991), "The Economics of Tax Amnesty," *Journal of Public Economics*, 46, 29–49.

Martimort, D. (1999), "The Life Cycle of Regulatory Agencies: Dynamic Capture and Transaction Costs," *Review of Economic Studies*, 66, 929–947.

Maskin, E. (1977), "Nash Implementation and Welfare Optimality," mimeo, Massachusetts Institute of Technology; (1999), *Review of Economic Studies*, 66(1), 23–38.

Mason, C. F., O. R. Phillips, and C. Nowell (1992), "Duopoly Behavior in Asymmetric Markets: An Experimental Evaluation," *Review of Economics and Statistics*, 74, 662–670.

McCubbins, M. and T. Schwartz (1984), "Congressional Oversight Overlooked: Police Patrols vs Fire Alarms," *American Journal of Political Science*, 28, 165–179.

McCutcheon, B. (1997), "Do Meetings in Smoke-Filled Rooms Facilitate Collusion?" *Journal of Political Economy*, 105(3), 330–350.

Moore, J. (1992), "Implementation, Contracts, and Renegotiation in Environments with Complete Information," in *Advances in Economic Theory, Sixth World Congress*, Vol. 1, (ed. by J.-J. Laffont), New York: Cambridge University Press, 182–282.

Motta, M. and M. Polo (2000), "Leniency Programs and Cartel Prosecution," mimeo, available at http://www.iue.it/Personal/Motta/.

Neven, D., R. Nuttall, and P. Seabright (1993), *Merger in Daylight*. London: CEPR.

Pénard, T. (1997), "Choix de Capacités et Comportements Stratégiques: Une Approche par les Jeux Répétés," *Annales d'Economie et de Statistique*, 46, 203–224.

Perry, M. and R. Porter (1985), "Oligopoly and the Incentive for Horizontal Merger," *American Economic Review*, 75, 219–227.

Polinski, A. M. and S. Shavell (2000), "The Economic Theory of Public Enforcement of Law," *Journal of Economic Literature*, 38, 45–76.

Porter, R. (1983), "A Study of Cartel Stability: The Joint Executive Committee, 1880–1886," *Bell Journal of Economics*, 14, 301–314.

Roberts, K. (1985), "Cartel Behavior and Adverse Selection," *Journal of Industrial Economics*, 33, 401–413.

Salant, S., S. Switzer, and R. Reynolds (1983), "Losses from Horizontal Merger: The Effects of an Exogenous Change in Industry Structure on Cournot–Nash Equilibrium," *Quarterly Journal of Economics*, 98, 185–199.

Selten, R. (1973), "A Simple Model of Imperfect Competition where Four Are Few and Six Are Many," *International Journal of Game Theory*, 2, 141–201.

Selten, R. (1984), "Are Cartel Laws Bad for Business?," in *Operations Research and Economic Theory*, (ed. by H. Hauptmann, W. Krelle, and K. C. Mosler), Berlin: Springer-Verlag.

Smith, A. (1776), *The Wealth of Nations*. New York: The Modern Library.

Souam, S. (1997), Instruments et Mécanismes des Politiques de la Concurrence: Les Incitations Comme Fondement du Contrôle des Comportements et des Structures de Marché, Ph.D. Thesis, University Paris.

Spagnolo, G. (2000a), "Optimal Leniency Programs," mimeo, Stockholm School of Economics.

Spagnolo, G. (2000b), "Self-Defeating Antitrust Laws: How Leniency Programs Solve Bertand's Paradox and Enforce Collusion in Auctions," mimeo, Stockholm School of Economics.

Sutton, J. (1991), *Sunk Cost and Market Structure*. Cambridge, MA: MIT Press.

Sutton, J. (1998), *Technology and Market Structure*. Cambridge, MA: MIT Press.

Symeonidis, G. (2000), "Price Competition and Market Structure: The Impact of Cartel Policy on Concentration in the UK," *Journal of Industrial Economics*, 48, 1–26.

Tirole, J. (1988), *The Theory of Industrial Organization*. Cambridge, MA: MIT Press.

Tirole, J. (1992), "Collusion and the Theory of Organizations," in *Advances in Economic Theory*, Vol. 2, (ed. by J. J. Laffont), New York: Cambridge University Press, pp. 151–206.

Tokar, S. (2000), "Whistleblowing and Corporate Crime," mimeo, European University Institute.

U.S. Department of Justice (1993), "Corporate Leniency Policy," Antitrust Division, available at http://www.usdoj.gov/atr/public/guidelines/lencorp.htm; also see (1994), "Individual Leniency Policy," available at http://www.usdoj.gov/atr/public/guidelines/lenind.htm.

Van Huyck, J., R. Battalio, and R. Beil (1990), "Tacit Coordination Games, Strategic Uncertainty, and Coordination Failure," *American Economic Review*, 80(1), 234–248.

Vanek, J. (1970), *The General Theory of Labor-Managed Market Economies*, Ithaca, NY: Cornell University Press.

Ward, B. (1958), "The Firm in Illyria: Market Syndicalism," *American Economic Review*, 48, 566–588.

CHAPTER 4

Identification and Estimation of Cost Functions Using Observed Bid Data

An Application to Electricity Markets

Frank A. Wolak

1. INTRODUCTION

This paper presents several techniques for recovering cost function estimates for electricity generation from a model of optimal bidding behavior in a competitive electricity market. These procedures are applied to actual data from the Australian National Electricity Market (NEM1) to recover cost function estimates for a specific market participant. I find close agreement between the cost functions recovered from these procedures and those obtained from engineering estimates. The techniques developed in this paper for recovering cost function estimates are not limited to markets for electricity generation. They can be used to recover cost function estimates for a participant in any bid-based centralized market.

There are number of uses for the procedures developed in this paper. The primary use is to measure the extent of market power possessed by a market participant using only bid information and market-clearing prices and quantities. A major research effort in empirical industrial organization is the measurement of market power. Bresnahan (1989) summarizes much of this research, although there has been an explosion of recent research on this general topic. The techniques presented in this paper are a logical extension of the techniques described by Bresnahan (1989) to bid-based markets.

A major challenge for designers of competitive electricity markets is to devise market rules that limit the ability of generation unit owners to exercise market power. Market power is the ability of a firm owning generation assets to raise the market price by its bidding behavior and to profit from this price increase. Until the recent trend toward industry restructuring, electricity was supplied by vertically integrated geographic monopolies regulated by state public utilities commissions in the United States or by goverment-owned national or state monopolies in other countries around the world. All of the industry characteristics that resulted in these two market structures make wholesale markets for electricity generation ripe for the exercise of market power. Electricity is extremely costly to store, there are binding short-run capacity constraints on its production, and demand must equal supply throughout the electricity grid at every moment in time. In addition, because of the manner in which electricity was sold to final customers during the former vertically integrated regime, the retail demand for electricity is very price inelastic on hour-ahead, day-ahead, and even month-ahead time horizons. These features of the electricity production process and the insensitivity of retail demand to wholesale price fluctuations allow small defects in market design to enhance significantly the ability of generation unit owners to exercise market power.

For this same reason, seemingly innocuous changes in market rules can produce a large impact on market outcomes. Consequently, market design is an extremely important aspect of the ongoing industry restructuring process. The optimal market design problem can be thought of as a single-principal (the market designer), multiple-agent (many electricity generation unit owners and wholesale energy purchasers) problem. Although the market design problem fits into the general class of common agency problems, given the complexity of even an extremely simple transmission network, solving for the optimal market design is an immensely complex task. The set of feasible mechanisms for compensating generators for the energy and generation reserves they supply and charging wholesale consumers for the energy they demand is enormous. Consequently, for a given market structure there are many feasible market designs, but the optimal market design is unknown.

Fortunately, all markets that currently exist in the United States have in place a process whereby the market monitoring unit within the Independent System Operator (ISO), the entity that operates the wholesale market and transmission grid, studies all aspects of market performance in order to detect design flaws that degrade market performance and enhance the ability of firms to exercise market power.[1] The next step in this market design process is to devise and implement market rule changes that eliminate these design flaws and move closer to finding the optimal set of market rules for that market structure. Although economic theory plays a major role in this process, there are very few empirical methods with a firm foundation in economic theory for analyzing the vast volumes of bid and market outcomes data available to these market monitoring units. This paper develops these sorts of tools and illustrates their use in the market monitoring and design process.

The specific application I consider is the estimation of forward market energy positions from spot market bid functions. In virtually all competitive wholesale electricity markets generators and loads engage in forward financial or hedge contracts, which allow them to fix the price for a specified amount of energy delivered and consumed in real time.[2] As noted in Wolak (2000), even with

[1] For an across-country discussion of the institutions and performance of competitive electricity markets, see Wolak (1999).

[2] Hedge contracts are typically signed between a generating company and an electricity retailer. They are purely financial obligations that guarantee the price at which a fixed quantity of electricity will be sold at a mutually agreed-on time in the future to the purchaser of the forward contract. If the relevant spot market price exceeds the contract price, then the contract seller pays to the buyer the difference between these two prices times the contract quantity. If the market price is less than the contract price, the buyer pays the absolute value of this same price difference times the contract quantity to the seller.

knowledge of a firm's bidding behavior in a competitive electricity market, it is difficult, if not impossible, to determine if the firm is able to exercise market power without knowing the generation unit owner's forward contract position. For a specific bid function and marginal cost function, there is a portfolio of forward financial contracts that can rationalize that bid function as expected profit maximizing. Wolak (2000) also demonstrates the enormous influence a generation unit owner's financial contract position has on his or her incentive to bid to attempt to increase the market-clearing price in the spot electricity market. Consequently, a technique for estimating a market participant's hedge contract or forward market position from bids submitted to a spot market can allow a market monitor to determine more precisely when the generation unit owner is likely to possess significant market power.

The remainder of the paper proceeds as follows. Section 2 relates this research to the general literature in empirical industrial organization on measuring market power using data on market-clearing prices and quantities. This section discusses the gain in econometric identification of the underlying cost function that results from using bids in addition to market-clearing prices and quantities in the estimation process. Section 3 presents a model of optimal bidding behavior with hedge contracts for a generic competitive electricity market. This section defines a *best-response bidding strategy* as the set of daily bid prices and quantities that maximize expected daily variable profits given the strategies of other firms participating in the market. The section also defines the *best-response price* as the market-clearing price that maximizes the realized profits of the firm given the bids actually submitted by its competitors, the realized value of the stochastic shock to the price-setting process. Both of these concepts are used to derive estimates of the cost function for a bidder in a competitive electricity market using actual bid information, the firm's hedge contract position, and actual market outcomes.

Section 4 presents my estimation methodology based on the best-response price concept. Section 5 presents this methodology based on the best-response bidding strategy. Section 6 then describes the essential features of the Australian National Electricity Market and the data set used in my empirical analysis. Section 7 presents the results of this empirical analysis. Section 8 describes how these techniques might be used in the market design process and discusses directions for future research.

2. IDENTIFYING MARGINAL COST FUNCTIONS FROM BIDS AND MARKET PRICES AND QUANTITIES

Beginning with Rosse (1970), empirical industrial organization (IO) economists have devised estimation procedures to recover cost functions from data on market-clearing prices and quantities. Rosse used a sample of monopoly local newspapers and the assumption of profit maximization to estimate the underlying marginal cost function of the monopolists. Porter (1983) employed a related approach in his study of price wars in the U.S. railroad industry during the 1880s.

He assumes a firm-level, homogeneous product, quantity-setting conjectural variation oligopoly equilibrium. He aggregates the firm-level first-order conditions to produce an industrywide supply function, which he jointly estimates along with an industry-level demand function. Bresnahan (1981, 1987) quantifies the extent of market power possessed by each vehicle model in the U.S. automobile industry using a discrete choice, differentiated products model of individual demand with vertical product differentiation in the unobserved product quality dimension. Aggregating these discrete purchase decisions across U.S. households yields an aggregate demand system for all automobile models. Bresnahan assumes Nash–Bertrand competition among the automobile makers facing this aggregate demand system to estimate the implied marginal cost of producing automobiles of each quality level. More recently, Berry (1994), Berry, Levinsohn, and Pakes (1995), and Goldberg (1995) have extended the techniques pioneered by Bresnahan to discrete choice oligopoly models with horizontal product differentiation.

The basic idea of all the techniques just described can be illustrated using the following example, which follows from the intuition given in Rosse (1970). Let $P(q, W, \theta, \varepsilon)$ denote the inverse demand function facing a monopolist and $C(q, Z, \theta, \eta)$ its total cost function. The variables W and Z are demand and cost function shifters, respectively. Here θ is the vector of parameters to be estimated, and ε and η are unobserved, to the econometrician, stochastic shocks. These shocks are assumed to be observable to the monopolist. The profit function of the monopolist is

$$\pi(q) = P(q, W, \theta, \varepsilon)q - C(q, Z, \theta, \eta). \tag{2.1}$$

The first-order condition for profit maximization is

$$\pi'(q) = P'(q, W, \theta, \varepsilon)q + P(q, W, \theta, \varepsilon) - C'(q, Z, \theta, \eta) = 0. \tag{2.2}$$

The researcher is assumed to have only market-clearing price and quantity data and the values of the demand and supply shifters, W and Z, for a cross section of monopolists selling the same homogeneous product. The econometrician does not have information on production costs for any of the firms. The researcher could also have a time series of observations on the same information for one or a small number of monopolists over time. This lack of cost data is the standard case faced by empirical researchers studying unregulated industries, such as those for automobiles, airlines, and personal computers, which are a few of the industries in which these techniques have been applied. Industry associations or government regulators usually publicly disclose information on market-clearing prices and quantities, but they give little information on production costs.

For the econometrician to make any statement about the extent of market power exercised in this market, she or he must have an estimate of the marginal cost function, $C'(q, Z, \theta, \eta)$. This estimate is constructed in the following manner. The econometrician first specifies a functional form for the inverse demand function. Suppose she or he selects $P(q, W, \theta, \varepsilon) = \mathrm{a} + \mathrm{b}q + \mathrm{c}W + \varepsilon$, where a, b, and c are elements of θ. The parameters of a, b, and c must be estimated by standard instrumental variables techniques to account for the fact that observed q and unobserved ε are correlated. This correlation occurs because the observed market-clearing quantity is determined by solving the first-order condition for profit maximization given in (2.2). This implies that q^E, the equilibrium quantity, is a function of η and ε and the demand and supply shifters, W and Z, so that $q^E = f(W, Z, \eta, \varepsilon)$. The market-clearing price is then determined by substituting q^E into the inverse demand function.

Given these estimates for a, b, and c, the econometrician can then solve for the value of $C'(q, Z, \theta, \eta)$ implied by the first-order conditions for profit maximization given in (2.2), using the observed market-clearing prices and quantities. Rearranging (2.2) for the assumed parametric inverse demand function yields

$$C'(q^E, Z, \theta, \eta) = P'(q^E, W, \theta, \varepsilon)q + P(q^E, W, \theta, \varepsilon) = \mathrm{b}q^E + p^E. \tag{2.3}$$

For each value of p^E and q^E, the market-clearing prices and quantities, compute an estimate of the marginal cost, $C'(q^E, Z, \theta, \eta)$, using the right-hand side of (2.3) and an estimate of the demand parameter b. This marginal cost estimate can then be used to compute an estimate of the amount of market power possessed by the firm in each market, by computing the Lerner index:

$$L = [p^E - C'(q^E, Z, \theta, \eta)]/p^E. \tag{2.4}$$

The assumption of firm-level profit maximization implies that estimates of only the parameters of the demand function are needed to compute an estimate of the Lerner index of market power.

Researchers often select a functional form for $C'(q^E, Z, \theta, \eta)$ and use the implied marginal costs derived from (2.3) to estimate the elements of θ contained in the cost function. An alternative approach, beginning with Rosse (1970), estimates the parameters of the inverse demand and cost function jointly using the assumption of profit maximization to identify the marginal cost function from observed market-clearing prices and quantities.

The intuition embodied in this example is used in all of the papers described thus far. Porter (1983) estimates the aggregate demand function facing the oligopoly that he studies. He makes assumptions on the functional form of costs for each individual firm and the nature of the strategic interaction among firms – cartel or perfect competition – to deliver an aggregate supply function for the industry under each of these two behavioral assumptions. Then he jointly estimates these aggregate supply and demand equations as a switching regression model, using the assumption of profit maximization to identify parameters of the underlying individual cost functions from time series data on market-clearing prices and quantities.

Bresnahan (1987) specifies a discrete-choice demand structure in which each individual decides whether to purchase an automobile, and if so, which model.

He aggregates this discrete-choice demand structure across all consumers to derive a system of aggregate demand equations. Using various assumptions about the nature of strategic interaction – specifically, Nash–Bertrand competition or collusion among automobile producers – he estimates the parameters of this aggregate demand system along with the parameters of the marginal cost functions implied by the first-order conditions associated with profit maximization. Bresnahan (1981) allows for a richer stochastic specification in the aggregate demand system, but follows the same basic procedure to recover estimates of the marginal cost function.

Berry et al. (1995) allow for a multinomial logit discrete-choice demand structure at the consumer level and assume unobservable (to the econometrician) stochastic consumer-level marginal utilities of product attributes. These marginal utilities are assumed to be independent and nonidentically normally distributed across product attributes and independent and identically distributed across consumers. Integrating individual-level purchase decisions with respect to these normal distributions yields a product-level aggregate demand system for automobiles. The authors assume that the conditional indirect utility functions for each consumer contain the same vector of unobservable (to the econometrician) product characteristics, and that these product characteristics are uncorrelated with all observable product characteristics. This stochastic structure induces correlation between equilibrium prices and the vector of unobserved random product characteristics in the aggregate demand system. Berry et al. propose and implement an instrumental variables estimation technique that exploits this lack of correlation between observed and unobserved product characteristics to estimate the demand system jointly with the marginal cost function under the assumption of Nash–Bertrand competition among automobile producers. In contrast, Bresnahan (1981, 1987) relies on maximum likelihood techniques.

Goldberg (1995) uses individual household-level data to estimate a general discrete-choice model for automobile purchases at the household level. She then uses weights giving the representativeness of each of these households in the population of U.S. households to produce a system of aggregate demand functions for automobiles based on the choice probabilities implied by her model of household-level automobile demand. Using the assumption of Nash–Bertrand competition among automobile producers, she then computes implied marginal cost estimates similar to those given in (2.3), which she uses to estimate a marginal cost function for each automobile model.

The most important conclusion to draw from this line of research is that all marginal cost estimates are the direct result of the combination of the assumed functional form for the aggregate demand for the products under consideration and the assumed model of competition among firms. Similar to the example given here, the first-order conditions for profit maximization and the demand function for each product determine the implied marginal cost for that product. Consequently, a major focus of this research has been on increasing the flexibility and credibility of the aggregate demand system used. However, because supply and demand functions are the econometrician's creation for describing the observed joint distribution of market-clearing prices and quantities across markets, for any finite data set, a functional form for the demand curves faced by the oligopolists or the monopoly must be assumed in order to estimate any demand function. Although it may be possible to apply the nonparametric and semiparametric identification and estimation strategies described in Blundell and Powell (2003) and in Florens (2003) to this economic environment, all existing work on modeling oligopoly equilibrium has relied on functional form restrictions and models of firm-level profit-maximizing behavior to identify the underlying demand and cost functions.

Rosse (1970) and Porter (1983) explicitly make this functional form assumption for aggregate demand. Bresnahan (1981, 1987) and Berry et al. (1995) assume a functional form for the probabilities that determine individual purchase decisions. They derive the aggregate demand system actually estimated by summing these individual choice probabilities across consumers. Goldberg (1995) specifies a household-level choice model, which she estimates using household-level data. The aggregate demand functions entering into her oligopoly model are an appropriately weighted sum of these estimated household-level demand systems across all U.S. households.

3. MODELS OF BEST-RESPONSE BIDDING AND BEST-RESPONSE PRICING

This section shows how the techniques described herein can be extended to estimate underlying marginal cost functions using data on bids and market-clearing prices and quantities from competitive electricity markets. Specifically, I demonstrate how the availability of bids allows the econometrician to identify the underlying firm-level cost function purely through an assumption about firm behavior. A functional form assumption for aggregate demand is no longer necessary. I consider two models of optimizing behavior by the firm that recover an estimate of the firm's marginal cost function. The first model makes the unrealistic but simplifying assumption that the firm is able to choose the market-clearing price that maximizes its profits given the bids submitted by its competitors. The second model is more realistic, but entails a significantly greater computation burden. It imposes all of the constraints implied by the market rules on the bids used by the firm to set the market-clearing price. The firm is assumed to bid according to the rules of the competitive electricity market to maximize its expected profits. The second approach explicitly imposes the reality that the only way the firm is able to influence the market-clearing price is through the bids it submits.

A first step in describing both of these methodologies is a description of the payoff functions and strategy space for participants in a generic competitive electricity market. Specifically, I describe the structure of bids and how these bids are translated into the payoffs that a market participant receives for supplying energy to the wholesale electricity market.

A competitive electricity market is an extremely complicated noncooperative game with a very high-dimensional strategy space. A firm owning a single generating set (genset) competing in a market with half-hourly prices must, at a minimum, decide how to set the daily bid price for the unit and the quantity bid for forty-eight half-hours during the day.[3] In all existing electricity markets, firms have much more flexibility in how they bid their generating facilities. For instance, in NEM1, firms are allowed to bid daily prices and half-hourly quantities for ten bid increments per genset. For a single genset, this amounts to a 490-dimensional strategy space (ten prices and 480 half-hourly quantities). Bid prices can range from −9,999.99 $AU to 5,000.00 $AU, which is the maximum possible market price. Each of the quantity increments must be greater than or equal to zero and their sum less than or equal to the capacity of the genset. Most of the participants in this market own multiple gensets, so the dimension of the strategy space for these firms is even larger. The England and Wales electricity market imposes similar constraints on the bid functions submitted by market participants. Each genset is allowed to bid three daily price increments and 144 half-hourly quantity increments. Genset owners also submit start-up and no-load costs as part of the day-ahead bidding process. Bidders in the California ISO's real-time electricity market bid eleven price and quantity increments, both of which can vary on an hourly basis.

To compute the profit function associated with any set of bids the firm might submit, I must have an accurate model of the process that translates the bids that generators submit into the actual market prices they are paid for the electricity for all possible bids submitted by them and their competitors and all possible market demand realizations. The construction of a model of the price-setting process in NEM1 that is able to replicate actual market prices with reasonable accuracy is a necessary first step in the process of estimating cost functions from generator bidding behavior and market outcomes. Wolak (2000) devotes significant attention to demonstrating that the model of the price-setting process used here accurately reflects the actual price-setting process in NEM1.

In preparation for the empirical portion of the paper, I describe the two procedures for cost function estimation for NEM1 in Australia, although the modifications necessary to apply these methods to other competitive electricity markets and other bid-based markets are straightforward. In NEM1, each day of the market, d, is divided into the half-hour load periods i beginning with 4:00 A.M. to 4:30 A.M. and ending with 3:30 A.M. to 4:00 A.M. the following day. Let Firm A denote the generator whose bidding strategy is being computed. Define

Q_{id}, Total market demand in load period i of day d;

$\mathrm{SO}_{id}(p)$, Amount of capacity bid by all other firms besides Firm A into the market in load period i of day d at price p;

[3] Electricity-generating plants are usually divided into multiple gensets or units. For example, a 2-GW plant will usually be divided into four 500-MW gensets.

$\mathrm{DR}_{id}(p) = Q_{id} - \mathrm{SO}_{id}(p)$, Residual demand faced by Firm A in load period i of day d, specifying the demand faced by Firm A at price p;

QC_{id}, Contract quantity for load period i of day d for Firm A;

PC_{id}, Quantity-weighted average (over all hedge contracts signed for that load period and day) contract price for load period i of day d for Firm A;

$\pi_{id}(p)$, Variable profits to Firm A at price p, in load period i of day d;

MC, Marginal cost of producing a megawatt hour by Firm A; and

$\mathrm{SA}_{id}(p)$, Bid function of Firm A for load period i of day d giving the amount it is willing to supply as a function of the price p.

For ease of exposition, I assume that MC, the firm's marginal cost, does not depend on the level of output it produces. For the general case of recovering marginal cost function estimates, I relax this assumption.

The market-clearing price p is determined by solving for the smallest price such that the equation $\mathrm{SA}_{id}(p) = \mathrm{DR}_{id}(p)$ holds. The magnitudes QC_{id} and PC_{id} are usually set far in advance of the actual day-ahead bidding process. Generators sign hedge contracts with electricity suppliers or large consumers for a pattern of prices throughout the day, week, and month, for an entire year or for a number of years. There is some short-term activity in the hedge contract market for electricity purchasers requiring price certainty for a larger or smaller than planned quantity of electricity at some point during the year.

In terms of the aforementioned notation, I can define the variable profits[4] (profits excluding fixed costs) earned by Firm A for load period i during day d at price p as

$$\pi_{id}(p) = \mathrm{DR}_{id}(p)(p - \mathrm{MC}) - (p - \mathrm{PC}_{id})\mathrm{QC}_{id}. \tag{3.1}$$

The first term is the variable profits from selling electricity in the spot market. The second term captures the payoffs to the generator from buying and selling hedge contracts. Assuming $\mathrm{QC}_{id} > 0$ (the generator is a net seller of hedge contracts), if $p > \mathrm{PC}_{id}$, the second term is the total payments made to purchasers of hedge contracts during that half-hour by Firm A. If $p < \mathrm{PC}_{id}$, the second term is the total payments made by purchasers of hedge contracts to Firm A. Once the market-clearing price is determined for the period, Equation (3.1) can be used to compute the profits for load period i in day d.

Financial hedge contracts impose no requirement on the generator to deliver actual electricity. These contracts are merely a commitment between the seller

[4] For the remainder of the paper, I use variable profits and profits interchangeably, with the understanding that I am always referring to variable profits.

(usually a generator) and the purchaser (usually a large load or load-serving entity) to make the payment flows described herein contingent on the value of the spot market-clearing price relative to the contract price. However, as discussed in detail in Wolak (2000), a generator that has sold a significant quantity of financial hedge contracts will find it optimal to bid more aggressively (to sell a larger quantity of energy in the spot market) than one that has sold little or no hedge contracts. This point can be illustrated by computing the first-order conditions for maximizing (3.1) with respect to p:

$$\pi'_{id}(p) = \mathrm{DR}'_{id}(p)(p - \mathrm{MC}) - (\mathrm{DR}_{id}(p) - \mathrm{QC}_{id}) = 0. \tag{3.2}$$

Because the residual demand curve is downward sloping and the firm can produce only a nonnegative quantity of electricity ($\mathrm{DR}_{id}(p) \geq 0$), the price that solves (3.2) for $\mathrm{QC}_{id} > 0$ is smaller than the price that solves (3.2) for $\mathrm{QC}_{id} = 0$. This result implies that, for the same values of MC and $\mathrm{DR}_{id}(p)$, the firm finds it profit maximizing to produce a larger amount of energy for $\mathrm{QC}_{id} > 0$ than it does for $\mathrm{QC}_{id} = 0$. Figure 1 from Wolak (2000) gives a graphical presentation of this logic. Another implication of the first-order condition (3.2) is that the contract price, PC_{id}, has no effect on the firm's profit-maximizing market-clearing price or output quantity. The level of the contract price simply determines the magnitude of the transfers that flow between the buyer and seller of the hedge contract. These same incentives to participate aggressively in the spot electricity market are also valid for a firm that has a contract guaranteeing physical delivery of QC_{id} units of electricity at price PC_{id} during hour i of day d.

The expression for Firm A's profits given in (3.1) illustrates two very important aspects of competitive electricity markets. First, unless a firm is able to move the market-clearing price by its bidding strategy, its profits are independent of its bidding strategy for a given hedge contract quantity and price. Given the market-clearing price, all of the terms in (3.1), the firm's actual variable profit function for load period i in day d, depend on factors unrelated to the bids it submits into the electricity market. Second, the difference between Equation (3.1) and the usual oligopoly model profit function is that the residual demand function $\mathrm{DR}_{id}(p)$ faced by Firm A is ex post directly observable given the bids submitted by all other market participants besides Firm A. As shown herein, the residual demand curve faced by Firm A at each price, p, is simply the aggregate demand function less the aggregate bid curve of all market participants besides Firm A, $\mathrm{DR}_{id}(p) = Q_{id} - SO_{id}(p)$.

In the standard oligopoly context, the residual demand faced by each market participant is not directly observable, because the aggregate demand function is not observable ex post. For example, in the Cournot duopoly model, the residual demand curve faced by one firm is simply the market demand, $D(p)$, less the quantity made available by that firm's competitor: $\mathrm{DR}(p) = D(p) - q_c$, where q_c is the quantity made available by the firm's competitor. Different from the case of a competitive electricity market, $D(p)$ is not directly observable, so that an estimate of $\mathrm{DR}(p)$ cannot be constructed without first econometrically estimating $D(p)$, the market aggregate demand function. In the case of a bid-based market such as electricity, even if load-serving entities could submit price-responsive demands, as is currently possible in most competitive electricity markets, the residual demand curve facing any competitor in these markets can be directly computed using all of the bids submitted by all other market participants.

Because this residual demand function can be constructed by the econometrician using bid data, there is no need to make a functional form assumption for the demand curve the bidder faces in order to compute its implied marginal cost for any level of output. Given a model for the price-setting process in this market and a behavioral model for bidders, implied marginal costs can be constructed for each observed level of output by Firm A.

The ex post observability of each generator's residual demand function has important implications for designing a competitive electricity market. Because the price elasticity of the residual demand curve faced by a bidder determines the extent of market power that it is able to exercise, the goal of the market design process is to face all bidders with a perfectly price-elastic residual demand function. Under these circumstances, no generator possesses market power. However, the residual demand curve faced by one market participant depends on the bids submitted by all other market participants. Therefore, aggressive bidding (very price-elastic bid supply functions) by a firm's competitors will leave it with a very elastic residual demand. This will cause the firm to bid very aggressively. This aggressive bidding will leave its competitors with elastic residual demand curves, which will cause them to bid more aggressively. This sequence of self-reinforcing aggressive bidding also works in the opposite direction to reinforce less price-elastic bidding. Specifically, if a firm bids a steep supply curve, that increases the incentive for its competitors to bid steep supply curves, because they now face more inelastic residual demand curves. Consequently, a very important aspect of the market design process is putting in place very strong incentives for aggressive spot market bidding.

Active participation by wholesale demanders in the forward electricity market is crucial to providing strong incentives for aggressive spot market bidding by generation unit owners. If a firm that owns significant generating capacity does not have a large fraction of this capacity tied up in forward contracts, then given the extreme inelasticity of the demand for electricity in any hour, this firm will find it profit maximizing to bid substantially in excess of its variable costs into the spot market during any hour that it knows some of its capacity is needed to meet total demand. Forward market commitments for a significant fraction of its capacity make this strategy less attractive because the firm earns the spot price only on any spot market sales in excess of its forward market commitments, rather than on all of its sales. In addition, because the number of firms that can compete to supply forward contracts far in advance of the delivery date is significantly greater than the number of firms that compete to supply electricity on a month-ahead, day-ahead, or hour-ahead basis, the price of electricity for delivery during these high-demand hours purchased far in advance is significantly cheaper than electricity purchased in short-term markets.

In particular, at time horizons greater than two years in advance of delivery, both existing and new entrants can compete to supply electricity. In contrast, a few hours before delivery, only the large generating units that are operating are able to compete to deliver electricity. Significant quantities of forward contracts guarantee that enough of these large generating units will be operating a few hours before delivery to ensure a workably competitive spot market for electricity. Wolak (2002) discusses these issues and the essential role of retail competition in developing an active forward electricity market.

I now introduce notation necessary to present the two procedures for recovering marginal cost estimates from bid data. Suppose that there are stochastic demand shocks to the price-setting process each period, and that Firm A knows the distribution of these shocks. This uncertainty could be due to the fact that Firm A does not exactly know the form of SO(p) – this function has a stochastic component to it – or it does not know the market demand used in the price-setting process when it submits its bids – Q is known only up to an additive error. Because I am not solving for an equilibrium bidding strategy, I do not need to be specific about the sources of uncertainty in the residual demand that Firm A faces. Regardless of the source of this uncertainty, Firm A will attempt to maximize profits conditional on the value of this uncertainty if the firm can observe it. If Firm A cannot observe this uncertainty, it will then choose its bids to maximize expected profits given an assumed distribution for this uncertainty. The two procedures for recovering the firm's underlying cost function from bid data differ in terms of their assumptions about whether the firm is able to achieve prices that maximize profits given the realization of this uncertainty or achieve only market prices that maximize expected profits taken with respect to the distribution of this uncertainty.

Let ε_i equal this shock to Firm A's residual demand function in load period i ($i = 1, \ldots, 48$). Rewrite Firm A's residual demand in load period i, accounting for this demand shock as $\mathrm{DR}_i(p, \varepsilon_i)$. Define $\Theta = (p_{11}, \ldots, p_{JK}, q_{1,11}, \ldots, q_{11,JK}, q_{2,11}, \ldots, q_{2,JK}, \ldots, q_{48,11}, \ldots, q_{48,JK})$ as the vector of daily bid prices and quantities submitted by Firm A. There are K increments for each of the J gensets owned by firm A. The rules of the NEM1 market require that a single price, p_{kj}, is set for each of the $k = 1, \ldots, K$ bid increments for each of the $j = 1, \ldots, J$ gensets owned by Firm A for the entire day. However, the quantity q_{ikj} made available to produce electricity in load period i from each of the $k = 1, \ldots, K$ bid increments for the $j = 1, \ldots, J$ gensets owned by Firm A can vary across $i = 1, \ldots, 48$ load periods throughout the day. In NEM1, the value of K is 10, so the dimension of Θ is $10J + 48 \times 10J$. Firm A owns a number of gensets so the dimension of Θ is more than several thousand. Let $\mathrm{SA}_i(p,\Theta)$ equal Firm A's bid function in load period i as parameterized by Θ. Note that by the rules of the market, bid increments are dispatched based on the order of their bid prices, from lowest to highest. This means that $\mathrm{SA}_i(p, \Theta)$ must be nondecreasing in p.

Figure 4.1 gives an example of two bid functions for different half-hours of the same day that are consistent with the rules of the Australian electricity

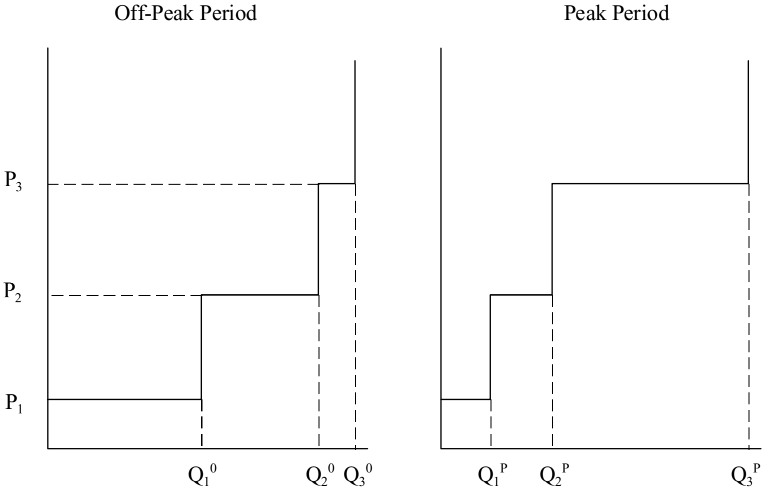


Figure 4.1. Sample bid functions for Australian electricity market.

market for the case of three bid price increments. Note that the only difference between the two bid curves is the quantity of energy that is made available at each price level during the two half-hours. Both the peak and off-peak period bid curves have the same price bids, as is required by NEM1 market rules, but the peak period bid curve assigns a large quantity of the capacity from the genset to the highest-price bid increment because the generator is reasonably confident that its highest-price bid will set the market-clearing price during this period. However, during the off-peak period that generator reduces the quantity that it bids at the highest price in order to be ensured that a significant fraction of its capacity will be sold at or above the intermediate bid price.

Let $p_i(\varepsilon_i, \Theta)$ denote the market-clearing price for load period i given the residual demand shock realization, ε_i, and daily bid vector Θ. It is defined as the solution in p to the equation $\mathrm{DR}_i(p, \varepsilon_i) = \mathrm{SA}_i(p, \Theta)$. Let $f(\varepsilon)$ denote the probability density function of $\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{48})'$. Define

$$E(\Pi(\Theta)) = \int_0^\infty \ldots \int_0^\infty \sum_{i=1}^{48} [\mathrm{DR}_i(p_i(\varepsilon_i, \Theta), \varepsilon_i)(p_i(\varepsilon_i, \Theta) - \mathrm{MC}) \\ - (p_i(\varepsilon_i, \Theta) - \mathrm{PC}_i)\,\mathrm{QC}_i] f(\varepsilon) d\,\varepsilon_1 \cdots d\varepsilon_{48} \qquad (3.3)$$

as the expected profits to Firm A for the daily bid vector Θ.

Firm A's best-reply bidding strategy is the solution to the following optimization problem:

$$\max_{\Theta} E(\Pi(\Theta)) \text{ subject to } b_U \geq R\Theta \geq b_L\,. \qquad (3.4)$$

Define Θ^* as the expected profit-maximizing value of Θ. Besides the extremely large dimension of Θ, there are several other reasons to expect this problem to be difficult to solve. First, in general, the realization of each residual demand function faced by Firm A is a nondecreasing, discontinuous step function, because the aggregate supply curve of all participants besides Firm A is a nondecreasing step function. Second, to compute the value of the objective function requires integrating with respect to a forty-eight-dimensional random vector ε. Most important, the dimension of Θ for Firm A is greater than 2,000. The linear inequality constraints represented by the matrix R and vectors of upper and lower bounds b_U and b_L imply that none of the q_{ik} can be negative and the sum of the q_{ik} relevant to a given genset cannot be greater than the capacity of the genset and that the prices for each bid increment cannot be smaller than −9,999.99 $AU or larger than 5,000.00 $AU. Wolak (2001a) computes this optimal bidding strategy for one market participant in the Australian electricity market and compares actual market outcomes with those that would exist under this optimal bidding strategy for a sample of days in NEM1.

At this point it is useful to compare the optimal bidding strategy problem given in (3.4) to the problem of computing an optimal supply function with demand uncertainty discussed in Klemperer and Meyer (1989) and applied to the electricity supply industry in England and Wales by Green and Newbery (1992). Rewrite Equation (3.1) with the residual demand function for load period i that includes the shock for period i as

$$\pi_{id}(p, \varepsilon_i) = \mathrm{DR}_{id}(p, \varepsilon_i)(p - \mathrm{MC}) - (p - \mathrm{PC}_{id})\mathrm{QC}_{id}. \tag{3.5}$$

Solving for the value of p that maximizes (3.5) yields $p_i{*}(\varepsilon_i)$, which is the profit-maximizing market-clearing price given that Firm A's competitors bid to yield the residual demand curve, $\mathrm{DR}_{id}(p, \varepsilon_i)$, with demand shock realization, ε_i, for the hedge contract position, QC_{id} and PC_{id}. This optimal price also depends on QC_{id}, PC_{id}, and MC. I write it as $p_i{*}(\varepsilon_i)$ to emphasize that it is Firm A's profit-maximizing price given the realization of ε_i. Because this price maximizes the ex post realized profits of Firm A, for the remainder of the paper I will refer to it as the *best-response price* for the residual demand curve $\mathrm{DR}_{id}(p, \varepsilon_i)$ with demand shock realization ε_i for the hedge contract position QC_{id} and PC_{id}. Substituting this value of p into the residual demand curve yields $\mathrm{DR}_{id}(p_i{*}(\varepsilon_i), \varepsilon_i)$. This price and quantity combination yields Firm A the maximum profit that it can earn given the bidding behavior of its competitors and the demand shock realization, ε_i.

Klemperer and Meyer (1989) impose sufficient restrictions on the underlying economic environment – the demand function, cost functions, and distribution of demand shocks – so that tracing out the price–quantity pairs $(p_i{*}(\varepsilon_i)$, $\mathrm{DR}_{id}(p_i{*}(\varepsilon_i), \varepsilon_i))$ for all values of ε_i yields a continuous, strictly increasing equilibrium supply curve, $\mathrm{SA}_i(p)$, for their duopolists. For each demand shock realization, their supply curve yields the best-response price for each duopolist given the bidding strategies of its competitor. Because each realization of ε_i in the Klemperer and Meyer model is associated with a unique price–quantity pair, the symmetric equilibrium duopoly supply function in the Klemperer and Meyer model does not depend on the distribution of ε_i. For this same reason, the Klemperer and Meyer framework can allow ε_i to be only one dimensional.

However, the NEM1 market rules explicitly prohibit firms from submitting continuous, strictly increasing bid functions. They must submit step functions, where bid price remains constant for all forty-eight half-hours of the day, but the length of each price increment can change on a half-hourly basis. This market rule constrains the ability of generation unit owners to submit supply bids that set the ex post profit-maximizing price for all possible realizations of the ex post residual demand function they face. Therefore, the expected profit-maximizing step function bid function, $\mathrm{SA}_i(p, \Theta)$, depends on the distribution of ε_i. For this reason, our best-reply bidding framework can allow for residual demand uncertainty, ε_i, that is multidimensional.

Because the market rules and market structure in NEM1 constrain the feasible set of price and quantity pairs that Firm A can bid in a given load period, it may be unable to achieve $p_i{}^*(\varepsilon_i)$ for all realizations of ε_i using its allowed bidding strategy. As noted herein, the allowed bidding strategy constrains Firm A to bid ten bid increments per genset, but, more importantly, the prices of these ten bid increments must be the same for all forty-eight load periods throughout the day. This may severely limit the ability of Firm A to achieve $p_i{}^*(\varepsilon_i)$. To the extent that it does, our best-response pricing procedure will yield unreliable estimates of the firm's underlying cost functions.

Best-response prices must yield the highest profits, followed by best-response bidding, because the former is based on the realization of ε_i as shown in (3.5), whereas the latter depends on the distribution of ε as shown in (3.3). The expected value of the generator's actual profits can only be less than or equal to the expected value of the best-response bidding profits. Recall that, by definition, the best-response price, $p_i{}^*(\varepsilon_i)$, yields the maximum profits possible given the bidding strategies of Firm A's competitors and the realized value of the residual demand shock, ε_i. The best-response bidding strategy that solves (3.3) for the expected profit-maximizing vector of allowable daily bid prices and quantities, Θ^*, yields the highest level of expected profits for Firm A within the set of allowable bidding strategies. Therefore, by definition, this bidding strategy should lead to average profits that are greater than or equal to Firm A's average profits from its current bidding strategy for the same set of competitors' bids and own hedge contract positions. The extent to which profits from a best-response bidding strategy lie below the maximum possible obtainable from best-response prices is not addressed here. Wolak (2001a) shows that a significant fraction of the difference between the actual variables profits earned by a firm in the Australian electricity market and the profits that it would earn from best-reply prices is due to the fact that the market rules constrain the ability of the firm to achieve $p_i{}^*(\varepsilon_i)$ for every realization of ε_i using a bidding strategy that respects the NEM1 market rules. In addition, given the high-dimensional strategy space available to Firm A, Wolak (2001a) also shows that a nonnegligible portion of the difference between the best-response pricing variable profits and variable profits under Firm A's current bidding strategy can be attributed to the use of bidding strategies that are not best response in the sense of not exactly solving the optimization problem (3.4).

Before both cost function estimation procedures are described in detail, it is useful to compare their properties. The best-response pricing approach has the advantage of computational simplicity and is broadly consistent with the approach used in the empirical IO literature, which uses a parametric model for demand and the assumption of profit-maximizing behavior to recover a cost function estimate. Similar to the empirical IO approach, this approach yields estimated marginal cost values for each observed level of output in the sample. The validity of the best-response pricing approach relies on the assumption that the firm is somehow able to achieve $p_i{*}(\varepsilon_i)$ for every realization of ε_i. As discussed herein, this is unlikely to be strictly valid for NEM1. In contrast, the best-response bidding strategy approach respects all of the rules governing bidding behavior and market price determination in NEM1 and relies only on the assumption of bidding to maximize expected profits to recover cost function estimates. Because it imposes all of the restrictions on bidding behavior implied by the market rules, this approach is able to recover genset-level cost function estimates. If the assumptions necessary for the validity of the best-response pricing approach hold, then both approaches will yield valid cost function estimates, but the best-response bidding approach should yield more precise cost function estimates.

4. RECOVERING COST FUNCTION ESTIMATES FROM BEST-RESPONSE PRICES

This section describes a procedure for recovering marginal cost function estimates based on my model of best-response pricing. This procedure can also be used to recover estimates of a generator's forward hedge contract position. Recall that my assumption of best-response pricing does not impose any of the restrictions implied by the market rules on the behavior of the firm under consideration. This procedure assumes that the firm is able to observe the market demand and the bids submitted by all other market participants. It then constructs the realized value of its residual demand function implied by the market demand and these bids and then selects the profit-maximizing price associated with this residual demand given the firm's hedge contract position and marginal cost function. Because of its computational simplicity, this approach should be a useful diagnostic tool in recovering an estimate of a firm's marginal cost function or in diagnosing the extent of market power a firm might possess when the assumptions required for its validity are approximately true.

Let $C(q)$ denote the total variable cost associated with producing output level q. Rewrite the period-level profit function for Firm A in terms of this general variable cost function as

$$\pi(p) = \mathrm{DR}(p, \varepsilon)p - C(\mathrm{DR}(p, \varepsilon)) - (p - \mathrm{PC})\mathrm{QC}. \tag{4.1}$$

To compute the best-reply price associated with this realization of the residual demand function, $\mathrm{DR}(p, \varepsilon)$, differentiate (4.1) with respect to p and set the result equal to zero:

$$\pi'(p) = \mathrm{DR}'(p, \varepsilon)(p - C'(\mathrm{DR}(p, \varepsilon))) + (\mathrm{DR}(p, \varepsilon) - \mathrm{QC}) = 0. \tag{4.2}$$

This first-order condition can be used to compute an estimate of the marginal cost at the observed market-clearing price, p^E, as

$$C'(\mathrm{DR}(p^E, \varepsilon)) = p^E - (\mathrm{QC} - \mathrm{DR}(p^E, \varepsilon))/\mathrm{DR}'(p^E, \varepsilon). \tag{4.3}$$

$\mathrm{DR}(p^E, \varepsilon)$ can be directly computed by using the actual market demand and bid functions submitted by all other market participants besides Firm A. The market-clearing price, p^E, is directly observed. I also assume that QC is observed. Computing $\mathrm{DR}'(p^E, \varepsilon)$ is the only complication associated with applying (4.3) to obtain an estimate of the marginal cost of Firm A at $\mathrm{DR}(p^E, \varepsilon)$.

For most competitive electricity markets, bidders submit step functions rather than piecewise linear functions. Consequently, strictly speaking, $\mathrm{DR}'(p^E, \varepsilon)$ is everywhere equal to zero.[5] However, because of the large number of bid increments permitted for each generating facility in the Australian market – ten per generating unit – and the close to 100 generating units in the Australian electricity market, the number of steps in the residual demand curve facing any market participant is very large. In addition, because of the competition among generators to supply additional energy from their units, there are usually a large number of small steps in the neighborhood of the market-clearing price. Nevertheless, some smoothness assumption on $\mathrm{DR}(p, \varepsilon)$ is still needed to compute a value for $\mathrm{DR}'(p^E, \varepsilon)$ to use in Equation (4.3).

I experimented with a variety of techniques for computing $\mathrm{DR}'(p^E, \varepsilon)$ and found that the results obtained are largely invariant to the techniques used. One technique approximates $\mathrm{DR}'(p^E, \varepsilon)$ by $(\mathrm{DR}(p^E + \delta, \varepsilon) - \mathrm{DR}(p^E, \varepsilon))/\delta$, for values of δ ranging from ten Australian cents to one Australian dollar. Another technique approximates the residual demand function by

$$\mathrm{DR}(p, \varepsilon) = Q_d(\varepsilon) - \mathrm{SO}_h(p, \varepsilon), \tag{4.4}$$

where the aggregate bid supply function of all other market participants besides Firm A is equal to

$$\mathrm{SO}_h(p, \varepsilon) = \sum_{n=1}^{N} \sum_{k=1}^{10} qo_{nk}\, \Phi((p - po_{nk})/h). \tag{4.5}$$

[5] For the now-defunct California Power Exchange (PX), bidders submitted piecewise linear bid functions starting at point (0, 0) in price-quantity space and ending at (2500, x) for any positive value of x. There were no limits on the number of these bid functions that any market participant could submit for a given hour. Therefore, the residual demand function facing any PX market participant was a piecewise linear function. Consequently, except at the points where two linear functions join, $\mathrm{DR}'(p^E, \varepsilon)$ is a well-defined concept.

Here qo_{nk} is the kth bid increment of genset n and po_{nk} is the bid price for increment k of genset n, where N is the total number of gensets in the market excluding those owned by Firm A. The function $\Phi(t)$ is the standard normal cumulative distribution function and h is a user-selected smoothing parameter. This parameter smooths the corners on the aggregate supply bid function of all other market participants besides Firm A. Smaller values of h introduce less smoothing at the cost of a value for $\mathrm{DR}'(p^E, \varepsilon)$ that may be at one of the smoothed corners. This second technique was adopted because it is very easy to adjust the degree of smoothing in the resulting residual demand function. Using this technique results in

$$\mathrm{DR}'_h(p, \varepsilon) = -\frac{1}{h}\sum_{n=1}^{N}\sum_{k=1}^{10} qo_{nk}\, \varphi((p - po_{nk})/h), \tag{4.6}$$

where $\varphi(t)$ is the standard normal density function. Using this method to compute $\mathrm{DR}'(p^E, \varepsilon)$, I can compute $C'(\mathrm{DR}(p^E, \varepsilon))$ by using Equation (4.3) for each market-clearing price.

There are variety of procedures to estimate the function $C'(q)$ given the $C'(q)$ and $q = \mathrm{DR}(p^E, \varepsilon)$ pairs implied by (4.3) applied to a sample of market-clearing prices and generator bids. In the empirical portion of the paper, I present a scatter plot of these $(C'(q), q)$ pairs and one estimate of $C'(q)$.

The first-order condition for best-reply pricing can also be used to compute an estimate of the value of QC for that half-hour for an assumed value for $C'(q)$ at that level of output. Rewriting (4.2) yields

$$\mathrm{QC} = (p^E - C'(\mathrm{DR}(p^E, \varepsilon)))\mathrm{DR}'(p^E, \varepsilon) + \mathrm{DR}(p^E, \varepsilon). \tag{4.7}$$

Different from the case of estimating the generator's marginal cost function, I expect QC to vary on a half-hourly basis both within and across days. Nevertheless, there are deterministic patterns in QC within the day and across days of the week. In the empirical portion of the paper, I quantify the extent to which the half-hourly implied values of QC follow the same deterministic patterns within the day and across days of the week as the actual values of QC.

In concluding this section, I find it important to emphasize that, strictly speaking, this procedure for estimating Firm A's marginal cost is valid only if the firm is somehow to able to obtain best-reply prices for all realizations of ε_i. As shown in Wolak (2000, 2001a), this is not possible because the market rules constrain the ability of expected profit-maximizing generators to set best-reply prices for all realizations from the joint distribution of forty-eight residual demand functions that Firm A faces each day. Nevertheless, as Section 7 shows, the deviation of actual prices from best-reply prices for Firm A is not so great as to make these calculations uninformative about Firm A's marginal cost function or its half-hourly hedge contract holdings. Given that these calculations are relatively straightforward to perform, I believe that they can be very useful diagnostic tools for computing marginal cost function estimates or forward contract position estimates that can be used in market power monitoring and analysis.

5. RECOVERING COST FUNCTION ESTIMATES FROM BEST-RESPONSE BIDDING

This section uses the assumption of best-response bidding, or, equivalently, bidding to maximize expected profits subject to the market rules on feasible bid functions, to recover estimates of genset-level marginal cost functions for Firm A. Imposing all of the bidding rules on the methodology used to recover the firm's marginal cost function will produce more accurate estimates than the methodology outlined in Section 4, even if the assumptions required for the validity of this simple approach hold. However, the procedure described here involves significantly more computational effort and econometric complexity. Specifically, I derive a generalized method of moments (GMM) estimation technique to recover genset-level cost functions for all of the units bid into the market by Firm A.

Deriving this estimation procedure requires additional notation. Define

$\mathrm{SA}_{ij}(p, \Theta)$, The amount bid by genset j at price p during load period i;

$\mathrm{C}_j(q, \beta_j)$, The variable cost of producing output q from genset j;

β_j, The vector of parameters of the cost function for genset j; and

$\mathrm{SA}_i(p, \Theta) = \sum_{j=1}^{J} \mathrm{SA}_{ij}(p, \Theta)$, The total amount bid by Firm A at price p during load period i.

In terms of this notation, write the realized variable profit for Firm A during day d as

$$\begin{aligned}\Pi_d(\Theta, \varepsilon) = \sum_{i=1}^{48} \Big[& \mathrm{DR}_i(p_i(\varepsilon_i, \Theta), \varepsilon_i)\, p_i(\varepsilon_i, \Theta) \\ & - \sum_{j=1}^{J} C_j(\mathrm{SA}_{ij}(p_i(\varepsilon_i, \Theta), \Theta), \beta_j) \\ & - (p_i(\varepsilon_i, \Theta) - \mathrm{PC}_i)\,\mathrm{QC}_i \Big],\end{aligned}$$

where ε is the vector of realizations of ε_i for $i = 1, \ldots, 48$. As discussed herein, $p_i(\varepsilon_i, \Theta)$, the market-clearing price for load period i given the residual demand shock realization, ε_i, and daily bid vector Θ, is the solution in p to the equation $\mathrm{DR}_i(p, \varepsilon_i) = \mathrm{SA}_i(p, \Theta)$. As a way to economize on notation, in the development that follows I abbreviate $p_i(\varepsilon_i, \Theta)$ as p_i. The best-reply bidding strategy maximizes the expected value of $\Pi_d(\Theta, \varepsilon)$ with respect to Θ, subject to the constraints that all bid quantity increments, q_{ikj}, must be greater than or equal to zero for all load periods, i, bid increments, k, and gensets, j, and that for each genset the sum of bid quantity increments during each load period is less than the capacity, CAP_j, of genset j. As discussed earlier, there are also upper and lower bounds on the daily bid prices. However, Firm A's price bids for all bid increments, k, and gensets, j, and days, d, during my sample period are strictly below the upper bound and strictly above the lower bound.

This result allows me to use the first-order conditions for daily expected profit maximization with respect to Firm A's choice of the daily bid price increments to derive a GMM estimator for the genset-level cost function parameters. For all days, d, the moment restrictions implied by these first-order conditions are

$$E_\varepsilon\left(\frac{\partial\, \Pi_d(\Theta_d, \varepsilon)}{\partial\, p_{km}}\right) = 0 \tag{5.1}$$

for all gensets, m, and bid increments, k. I index Θ by d to denote the fact that there are different values of Θ for each day during the sample period. Equation (5.1) defines the $J \times K$ moment restrictions that I will use to estimate the parameters of the genset-level cost functions. The sample analog of this moment restriction is as follows:

$$\begin{aligned}\frac{\partial\, \Pi_d(\Theta_d, \varepsilon)}{\partial\, p_{km}} = \sum_{i=1}^{48}\Bigg[\Bigg(&\mathrm{DR}_i'(p_i(\varepsilon_i, \Theta), \varepsilon_i)\, p_i(\varepsilon_i, \Theta)\\ &+ (\mathrm{DR}_i(p_i(\varepsilon_i, \Theta), \varepsilon_i) - \mathrm{QC}_i)\\ &- \sum_{j=1}^{J} C_j'(\mathrm{SA}_{ij}(p_i(\varepsilon_i, \Theta)), \beta_j)\left(\frac{\partial\, \mathrm{SA}_{ij}}{\partial\, p_i}\right)\Bigg)\frac{\partial\, p_i}{\partial\, p_{km}}\\ &- \sum_{j=1}^{J} C_j'(\mathrm{SA}_{ij}(p_i(\varepsilon_i, \Theta)), \beta_j)\frac{\partial\, \mathrm{SA}_{ij}}{\partial\, p_{km}}\Bigg],\end{aligned} \tag{5.2}$$

where p_i is shorthand for the market-clearing price in load period i. Let $\ell_d(\beta)$ denote the $J \times K$ dimensional vector of partial derivatives given in (5.2), where β is the vector composed of β_j for $j = 1, \ldots, J$. Assuming that the functional form for $C_j(q, \beta_j)$ is correct, the first-order conditions for expected profit maximization with respect to daily bid prices imply that $E(\ell_d(\beta^0)) = 0$, where β^0 is the true value of β. Consequently, solving for the value of b that minimizes

$$\left[\frac{1}{D}\sum_{d=1}^{D}\ell_d(b)\right]'\left[\frac{1}{D}\sum_{d=1}^{D}\ell_d(b)\right] \tag{5.3}$$

will yield a consistent estimate of β. Let $b(I)$ denote this consistent estimate of β, where I denotes the fact that the identity matrix is used as the GMM weighting matrix. I can construct a consistent estimate of the optimal GMM weighting matrix using this consistent estimate of β as follows:

$$V_D(b(I)) = \frac{1}{D}\sum_{d=1}^{D}\ell_d(b(I))\,\ell_d(b(I))'. \tag{5.4}$$

The optimal GMM estimator finds the value of b that minimizes

$$\left[\frac{1}{D}\sum_{d=1}^{D}\ell_d(b)\right]' V_D(b(I))^{-1}\left[\frac{1}{D}\sum_{d=1}^{D}\ell_d(b)\right]. \tag{5.5}$$

Let $b\,(O)$ denote this estimator, where O denotes the fact this estimator is based on a consistent estimate of the optimal weighting matrix.

Operationalizing this estimation procedure requires computing values for the partial derivative of $\mathrm{SA}_{ij}(p, \Theta)$ with respect to p and p_{km} and the partial derivative of $p_i(\varepsilon_i, \Theta)$ with respect to p_{kj}. I use the same smoothing technique used in the previous section to compute the derivative of the residual demand function with respect to the market price to compute these partial derivatives. Define $\mathrm{SA}^h_{ij}(p, \Theta)$ as

$$\mathrm{SA}^h_{ij}(p, \Theta) = \sum_{k=1}^{10} q_{ikj}\ \Phi((p - p_{kj})/h), \tag{5.6}$$

which implies

$$\mathrm{SA}^h_{i}(p, \Theta) = \sum_{j=1}^{J} \sum_{k=1}^{10} q_{ikj}\ \Phi((p - p_{kj})/h). \tag{5.7}$$

This definition of $\mathrm{SA}_{ij}(p, \Theta)$ yields the following two partial derivatives:

$$\begin{aligned} \frac{\partial\,\mathrm{SA}_{ij}}{\partial p} &= \frac{1}{h}\sum_{k=1}^{10} q_{ikj}\ \phi((p - p_{kj})/h) \quad \mathit{and} \\ \frac{\partial\,\mathrm{SA}_{ij}}{\partial\, p_{kj}} &= -\frac{1}{h}\, q_{ikj}\ \phi((p - p_{kj})/h). \end{aligned} \tag{5.8}$$

The final partial derivative required to compute the sample analog of (5.1) can be computed by applying the implicit function theorem to the equation $\mathrm{DR}_i(p, \varepsilon_i) = \mathrm{SA}_i(p, \Theta)$. This yields the expression

$$\frac{\partial\, p_i(\varepsilon_i, \Theta))}{\partial\, p_{kj}} = \frac{\dfrac{\partial\,\mathrm{SA}_i(p_i(\varepsilon_i, \Theta), \Theta)}{\partial\, p_{kj}}}{\mathrm{DR}'_i(p_i(\varepsilon_i, \Theta), \varepsilon_i) - \mathrm{SA}'_i(p_i(\varepsilon_i, \Theta), \Theta)}, \tag{5.9}$$

where the derivative of the residual demand curve with respect to price used in this expression is given in Equation (4.6) and the other partial derivatives are given in (5.8). Given data on market-clearing prices and the bids for all market participants, I can compute all of the inputs into Equation (5.2). I only need to choose a value for h, the smoothing parameter that enters the smoothed residual demand function and the smoothed bid functions of Firm A. Once this smoothing parameter has been selected, the magnitudes given in (5.8) and (5.9) remain constant for the entire estimation procedure.

The final step necessary to implement this estimation technique is choosing the functional form for the marginal cost function for each genset. Firm A owns two power plants. One power plant has four identical gensets that the firm operates during the sample period. I refer to this facility as Plant 1. The gensets at Plant 1 have a maximum capacity of 660 MW and a lower operating limit of 200 MW. The other power plant has three identical gensets that the firm operates during the sample period. I refer to this facility as Plant 2. The gensets at Plant 2 have a maximum capacity of 500 MW and a lower operating limit of 180 MW. Because it is physically impossible for a genset to supply energy safely at a rate below its lowest operating limit, I specify a functional form for marginal cost to take this into account. Consequently, I assume the following parametric functional forms for the two unit-level marginal cost functions:

$$C_1'(q, \beta_1) = \beta_{10} + \beta_{11}(q - 200) + \beta_{12}(q - 200)^2, \tag{5.10}$$

$$C_2'(q, \beta_2) = \beta_{20} + \beta_{21}(q - 180) + \beta_{22}(q - 180)^2. \tag{5.11}$$

These functional forms are substituted into (5.2) to construct the sample moment restrictions necessary to construct the objective function I minimize to estimate

$$\beta = (\beta_{10}, \beta_{11}, \beta_{12}, \beta_{20}, \beta_{21}, \beta_{22})'.$$

Recall that for each genset the value of q entering (5.10) and (5.11) in the estimation procedure is the actual level of output produced by that unit during the half-hour period under consideration. I now turn to summarizing the relevant features of the NEM1 market in Australia necessary to understand the empirical work.

6. OVERVIEW OF NEM1

The Victoria Power Exchange (VPX) is the longest-running wholesale electricity market in Australia. It was established under the Electricity Industry (Amendment) Act of 1994 and formally began operation on July 1, 1994. The New South Wales (NSW) State Electricity Market (SEM) began operation May 10, 1996. NEM1 is the competitive electricity market established jointly by NSW and Victoria on May 4, 1997. It introduced unrestricted competition for generation dispatch across the two states; that is, the cheapest available generation, after allowing for transmission losses and constraints, is called on regardless of where it is located, and all wholesale energy is traded through the integrated market. The spot price in each state is determined with electricity flows in and between the state markets based on competitive bids or offers received in both markets.

The formation of NEM1 started the harmonization of the rules governing the operation of the two markets in Victoria and NSW. The market structures of the two electricity supply industries in Victoria and NSW are similar in terms of the relative sizes of the generation firms and the mix of generation capacity by fuel type, although the NSW industry is a little less than twice the size (as measured by installed capacity) of the Victoria industry and the largest three generators in NSW control a larger fraction of the total generation capacity in their state than the three largest generators in Victoria control in their state.

6.1. Market Structure in NEM1

Restructuring and privatization of the State Electricity Commission of Victoria (SECV) in 1994 took place at the power station level.[6] Each power station was formed into a separate entity to be sold. All former SECV generation capacity is now privately owned. The new owners are from within Australia and abroad. Currently there are eight generating companies competing in the VPX. The NSW–SEM has four generators competing to supply power. All generating assets are still owned by the NSW government. There are seven corporatized state-owned electricity distribution and supply companies serving NSW and the Australian Capital Territory (ACT). The eventual goal is to privatize both the generation and supply companies.

In both Victoria and NSW, there is an accounting separation within the distribution companies between their electricity distribution business and their electricity supply business. All other retailers have open and nondiscriminatory access to any of the other distribution companies wires. In NSW, the high-voltage transmission grid remains in government hands. In Victoria, the high-voltage transmission grid was initially owned by the government and was called PowerNet Victoria. It was subsequently sold to the New Jersey-based U.S. company, GPU, and renamed GPU-PowerNet. In NSW it is called TransGrid. Both the state markets operating under NEM1 – SEM in NSW and VPX in Victoria – were state-owned corporatized entities separate from the bulk transmission entities.

During 1997, the year of our sample, peak demand in Victoria was approximately 7.2 GW. The maximum amount of generating capacity that could be supplied to the market was approximately 9.5 GW. Because of this small peak demand, and despite the divestiture of generation to the station level, three of the largest baseload generators had sufficient generating capacity to supply at least 20 percent of this peak demand. More than 80 percent of the generating plant is coal fired, although some of this capacity does have fuel-switching capabilities. The remaining generating capacity is shared equally between gas turbines and hydroelectric power.

During 1997, the NSW market had a peak demand of approximately 10.7 GW and the maximum amount of generating capacity that could be supplied to the market was approximately 14 GW. There were two large generation companies, each of which controlled coal-fired capacity sufficient to supply more than 40 percent of NSW peak demand. The remaining large generators had hydroelectric, gas turbine, and coal-fired plants. The Victoria peak demand tends to occur during the summer month of January, whereas peak demand in NSW tends to occur in the winter month of July.

The full capability of the transmission link between the two states is nominally 1,100 MW from Victoria to NSW, and 1,500 MW in the opposite direction,

[6] Wolak (1999) provides a more detailed discussion of the operating history of the VPX and compares its market structure, market rules, and performance to the markets in England and Wales, Norway and Sweden, and New Zealand.

although this varies considerably, depending on temperature and systems conditions. If there are no constraints on the transfer between markets, then both states see the same market price at the common reference node. If a constraint limits the transfer, then prices in both markets diverge, with the importing market having a higher price than the exporting market.

6.2. Market Rules in NEM1

With a few minor exceptions, NEM1 standardized the price-setting process across the two markets. Generators are able to bid their units into the pool in ten price increments that cannot be changed for the entire trading day – the twenty-four-hour period beginning at 4:00 A.M. and ending at 4:00 A.M. the next day. The ten quantity increments for each genset can be changed on a half-hourly basis. Demanders can also submit their willingness to reduce their demand on a half-hourly basis as a function of price according to these same rules. Nevertheless, there is very little demand-side participation in the pool. A few pumped storage facilities and iron smelter facilities demand-side bid, but these sources total less than 500 MW of capacity across the two markets. All electricity is traded through the pool at the market price, and all generators are paid the market price for their energy.

7. RECOVERING IMPLIED MARGINAL COST FUNCTIONS AND HEDGE CONTRACT QUANTITIES

This section presents the results of applying the procedures described in Sections 4 and 5 to actual market outcomes. This requires collecting data on generator bids and market outcomes for a time period in which I also have information on the half-hourly values of QC, the quantity of the firm's forward contract obligations. I was able to obtain this information for a market participant in the Australian market for the period from May 15, 1997 to August 24, 1997. As discussed earlier, a major source of potential error in this analysis is the possibility that I have not adequately modeled the actual price-setting process in the Australian electricity market. Wolak (2000, Section 5) compares different models of the half-hourly price-setting procedure to determine which one does the best job of replicating observed half-hourly prices. This analysis found that the process I use in this paper – setting the half-hourly price equal to the price necessary to elicit sufficient supply from the aggregate half-hourly bid supply curve to meet the half-hourly market demand – replicates actual prices with sufficient accuracy.

I first compute implied marginal cost estimates using bid data submitted by Firm A's competitors, actual market prices, and total market demand. To give some idea of the range of residual demand curves faced by Firm A within the same day, Figures 4.2 and 4.3 plot the actual ex post residual demand curve faced by a firm in a representative off-peak demand period and on-peak demand period for July 28, 1997. These curves have been smoothed using

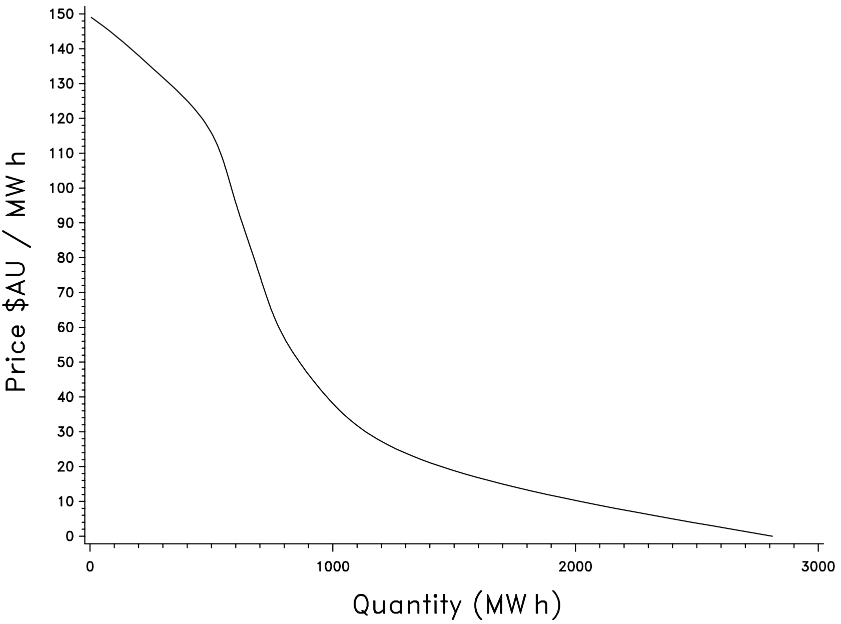


Figure 4.2. Residual demand curve for July 28, 1997 low demand.

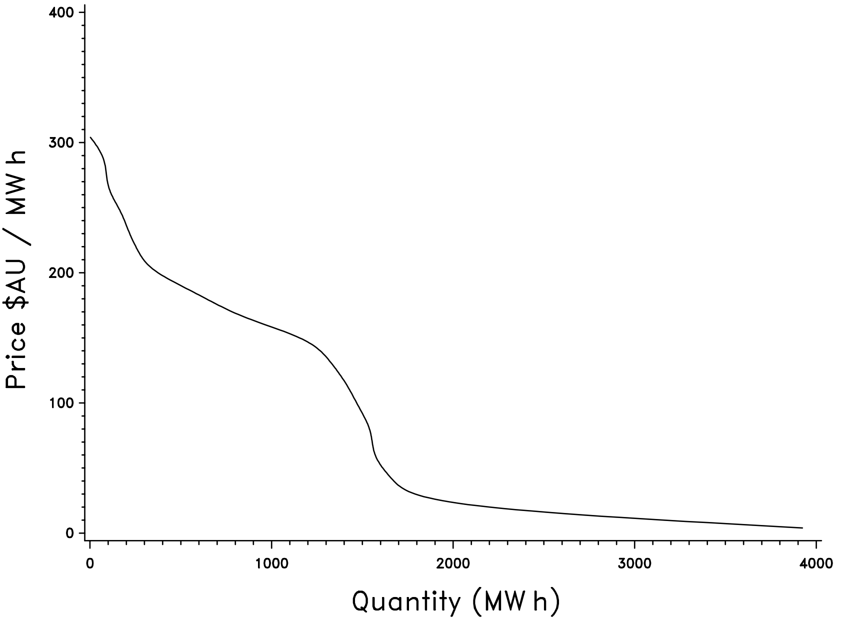


Figure 4.3. Residual demand curve for July 28, 1997 high demand.

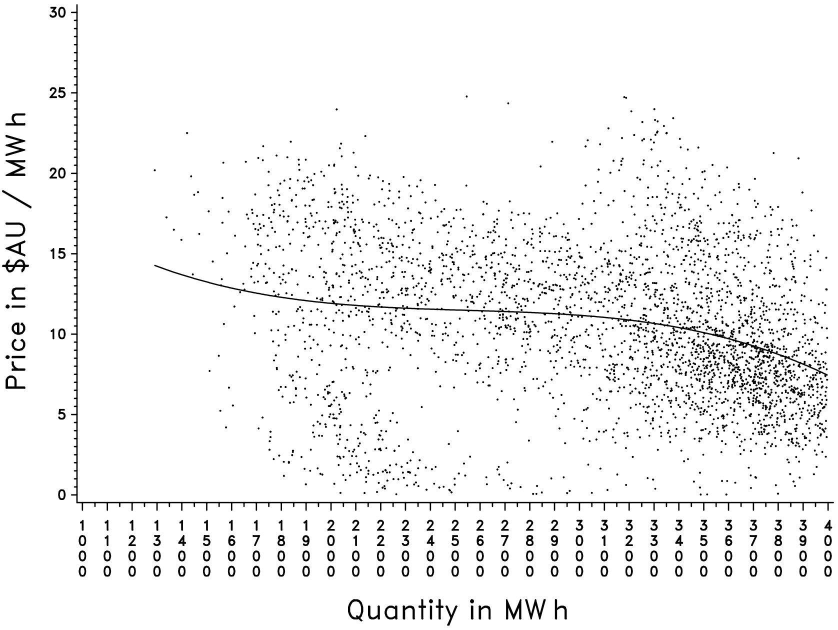


Figure 4.4. Implied marginal cost.

the expression for the residual demand curve given in (4.4) and (4.5), using a value of $h = 1$ $AU. These curves confirm the standard intuition that Firm A possesses greater opportunities to exercise its market power during high market demand periods as opposed to low market demand periods. At every price level, Firm A faces a significantly higher residual demand during the high-demand load period than in the low-demand load period. I use the value of h employed to plot these figures for all of the results reported in this section. Repeating these results for values of $h = 10$ $AU and 0.10 $AU did not substantially alter any of the conclusions reported in the paragraphs that follow.

Figure 4.4 is a plot of the marginal cost and associated output demanded pairs, $(C'(\mathrm{DR}(p^E, \varepsilon)), \mathrm{DR}(p^E, \varepsilon))$, for all of the half-hourly market-clearing prices, p^E. The figure also plots the predicted values from the following cubic regression of the implied marginal cost, $C'(q)$, on q, the associated implied output of Firm A:

$$C'(q) = a + bq + cq^2 + dq^3 + \eta.$$

Table 4.1 gives the results of this implied marginal cost function regression. Although there is a considerable amount of noise in the sample of implied marginal cost and output pairs, the regression results are broadly consistent with the magnitudes of marginal costs implied by the heat rates and fuel prices of the facilities owned by Firm A. In particular, in discussions with the management of Firm A, they confirmed that their best estimates of their own marginal costs fluctuate between 15 $AU/MWh and 10 $AU/MWh.

Table 4.1. *Implied marginal cost regression: C(q) = a + bq + cq^2 + dq^3*

Parameter	Estimate	Standard Error	t-Stat
a	31	4.43	7.00
b	-2.21×10^{-2}	4.61×10^{-3}	−4.78
c	8.47×10^{-6}	1.54×10^{-6}	5.49
d	-1.11×10^{-9}	1.61×10^{-10}	−6.65

I now examine the accuracy of my procedure for estimating the half-hourly values of QC. This process requires selecting values for the marginal cost at any given output level. Consistent with the usual circumstances one is likely to face in market monitoring, I assume that the form of the generator's cost function is unknown. Therefore, I perform this procedure by assuming a rough estimate of the firm's marginal cost function. I assume a constant value for the marginal cost for all levels of output. More sophisticated marginal cost functions can be assumed, but the results with constant marginal cost should be the most informative about usefulness of this technique in market monitoring, because more accurate information on generation unit heat rates and output rates are generally unavailable. The two values of MC, the fixed level of marginal cost used in the procedure to recover estimates of QC for each half-hour, bound the fitted marginal cost curve given in Figure 4.4. These values are 10 $AU/MWh and 15 $AU/MWh. Figures 4.5 and 4.6 plot the implied half-hourly values of QC and the associated actual half-hourly values of QC for MC equal to 10 $AU/MWh and 15 $AU/MW h, respectively. Both of the figures also have a graph of the line QC(implied) = QC(actual) to illustrate visually how closely my procedure tracks the desired relationship between these two magnitudes. Both values of the marginal cost show a clear positive correlation between QC(implied) and QC(actual), although the consistency with the desired relationship seems greatest for an MC equal to 10 $AU/MWh.

The values of QC(actual) vary on a half-hourly basis within the day and across days. However, there are still systematic patterns to these changes within the day and across days. On average, QC(actual) is higher on weekdays than weekends and higher during the peak hours of the day than the off-peak hours of the day. Consequently, another way to determine the usefulness of my procedure is to see if it captures the systematic variation in the values of QC(actual) within the day and across days of the week. To do this, estimate the following regression for both QC(actual) and QC(implied) for all load periods, $i = 1, \ldots, 48$, and days, $d = 1, \ldots, D$:

$$\mathrm{QC}(J)_{id} = \alpha + \sum_{m=1}^{3} \rho_m \,\mathrm{DMN}(m)_{id} + \sum_{p=1}^{6} \gamma_{p\mathrm{DWKD}}(p)_{id} + \sum_{r=1}^{47} \psi_r \,\mathrm{DPD}(r)_{i,d} + \nu_{id}, \tag{7.1}$$

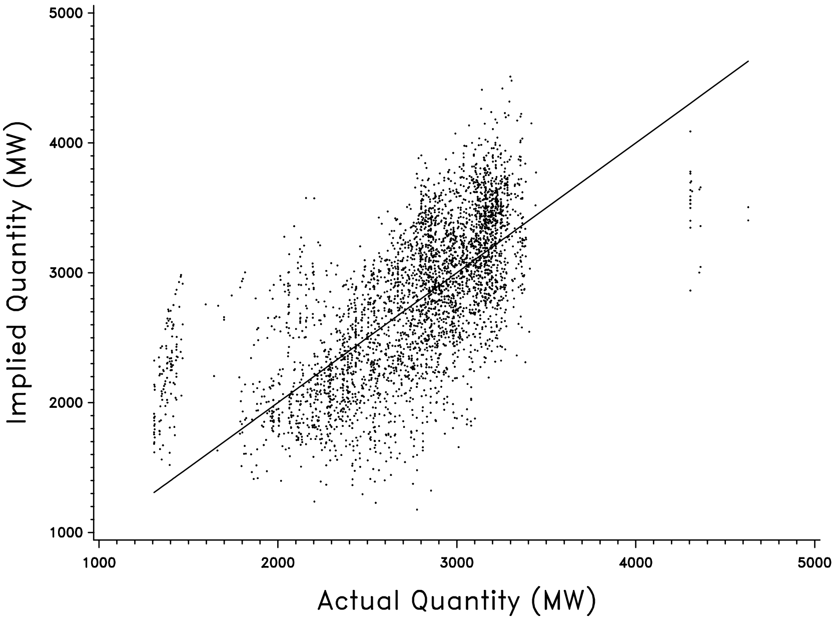


Figure 4.5. (MC = 10 $AU/MW h) Implied vs. actual contract quantities.

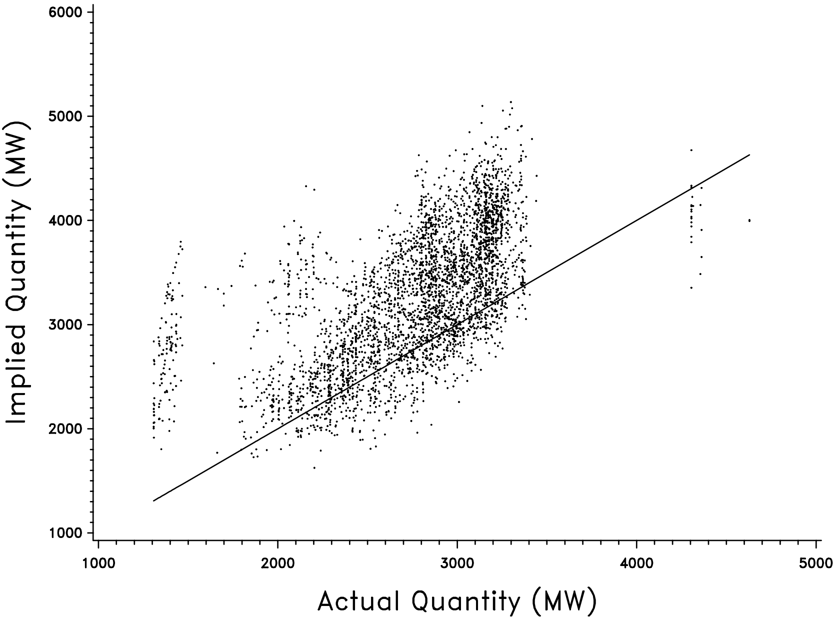


Figure 4.6. (MC = 15 $AU/MW h) Implied vs. actual contract quantities.

where $\mathrm{DMN}(m)_{id}$ is a dummy variable equal to one when day d is in month m and zero otherwise, $\mathrm{DWKD}(p)_{id}$ is a dummy variable equal to one when day d is on day-of-the week p and zero otherwise, and $\mathrm{DPD}(r)_{id}$ is dummy variable equal to one when load period i is in load period-within-the-day r and zero otherwise. I compute estimates of the ρ_m, γ_p, and ψ_r for both QC(actual) and QC(implied), by estimating (7.1) by ordinary least squares. Table 4.2 reports the results of this estimation for QC(implied) with MC set equal to 10 $AU/MW h. Figures 4.7 and 4.8 plot estimated values of γ_p $(p = 1, \ldots, 6)$ and ψ_r $(r = 1, \ldots, 48)$ for QC(implied) with MC equal to 10 $AU/MWh and for QC(actual). The first value of γ_p is associated with Sunday and the excluded day of the week is Saturday. The first value of ψ_r is the half-hour beginning at 4:00 A.M. and ending at 4:30 A.M. and the excluded load period is the one beginning at 3:30 A.M. and ending at 4:00 A.M. the following day. These figures show a remarkable level of agreement between the deterministic part of the within-day and across-day variation in QC(implied) and QC(actual). These results provide strong evidence that even applying my procedure with this crude assumed marginal cost function yields a considerable amount of information about the level and pattern of forward contracting that a firm has undertaken.

At this point is it important to note that a generation unit owner's forward contract position is generally unknown to the market monitor. However, the analysis given here demonstrates that, with the use of the assumption of expected profit maximization with data on actual bidding behavior, something that is readily available to the market monitor, accurate estimates of the hourly levels of forward contract obligations can be obtained. Consequently, even this very rough procedure, which relies on best-response pricing, can be a very powerful tool for determining those instances when a market participant is likely to attempt to exercise market power in a spot electricity market. As shown in Wolak (2000), a generation unit owner's forward contract position is a very important determinant of the amount of market power that it is able to exercise in a spot electricity market.

I now examine the properties of my procedure for recovering genset-level marginal cost functions implied by best-reply bidding. As discussed in Section 5, Firm A has two types of identical gensets. Consequently, I estimate two genset-level marginal cost functions, applying the GMM estimation technique outlined in that section. I compute estimates of these unit-level marginal cost functions using both the identity matrix and a consistent estimate of the optimal weighting matrix. For all of the estimations reported, I assume $h = 1$ $AU, although the results did not change appreciably for values of h ranging from 0.10 $AU to 50 $AU.

Table 4.3 reports the results of estimating $C_1'(q, \beta_1)$ and $C_2'(q, \beta_2)$, using the identity matrix as the weighting matrix and a consistent estimate of the optimal weighting matrix. Wolak (2001b) proves the consistency of these parameter estimates under the assumption that h tends to zero as the number of observations, D, tends to infinity. This paper also derives an expression for the variance

Table 4.2. *Contract quantity regression for QC(implied) for (MC = $10)*

Variable	Estimate	Standard Error	t-Stat
Constant	1882.37	59.70	31.53
DWKD1	−126.34	25.02	−5.05
DWKD2	215.72	25.95	8.31
DWKD3	282.29	25.64	11.01
DWKD4	330.84	25.54	12.95
DWKD5	391.34	25.01	15.65
DWKD6	397.24	25.45	15.61
DMN1	−12.46	40.29	−0.31
DMN2	34.90	39.19	0.89
DMN3	398.21	39.10	10.19
DPD1	−165.64	67.18	−2.47
DPD2	−147.10	67.42	−2.18
DPD3	−26.59	65.32	−0.41
DPD4	−54.94	65.11	−0.84
DPD5	113.07	65.90	1.72
DPD6	315.23	66.53	4.74
DPD7	602.74	65.90	9.15
DPD8	581.09	66.56	8.73
DPD9	654.61	67.07	9.76
DPD10	674.01	66.62	10.12
DPD11	743.62	66.84	11.13
DPD12	736.64	67.07	10.98
DPD13	777.06	66.18	11.74
DPD14	822.59	66.18	12.43
DPD15	898.80	66.18	13.58
DPD16	877.76	66.16	13.27
DPD17	820.91	66.39	12.37
DPD18	800.76	66.15	12.10
DPD19	757.67	66.15	11.45
DPD20	714.33	66.15	10.80
DPD21	706.38	65.93	10.71
DPD22	648.04	65.72	9.86
DPD23	663.86	65.93	10.07
DPD24	692.77	65.93	10.51
DPD25	741.97	66.14	11.22
DPD26	876.84	65.95	13.30
DPD27	866.17	65.91	13.14
DPD28	793.67	67.74	11.72
DPD29	739.79	67.68	10.93
DPD30	624.01	68.45	9.12
DPD31	695.82	66.32	10.49
DPD32	770.62	66.36	11.61
DPD33	879.34	66.82	13.16
DPD34	858.81	66.38	12.94
DPD35	848.25	67.27	12.61
DPD36	623.35	66.17	9.42
DPD37	739.35	67.07	11.02
DPD38	522.93	66.41	7.87
DPD39	462.97	67.04	6.91
DPD40	432.56	66.62	6.49
DPD41	421.45	67.06	6.28
DPD42	300.60	67.06	4.48
DPD43	145.39	67.06	2.17
DPD44	138.65	66.60	2.08
DPD45	64.44	66.60	0.97
DPD46	87.50	66.13	1.32
DPD47	55.02	65.68	0.84

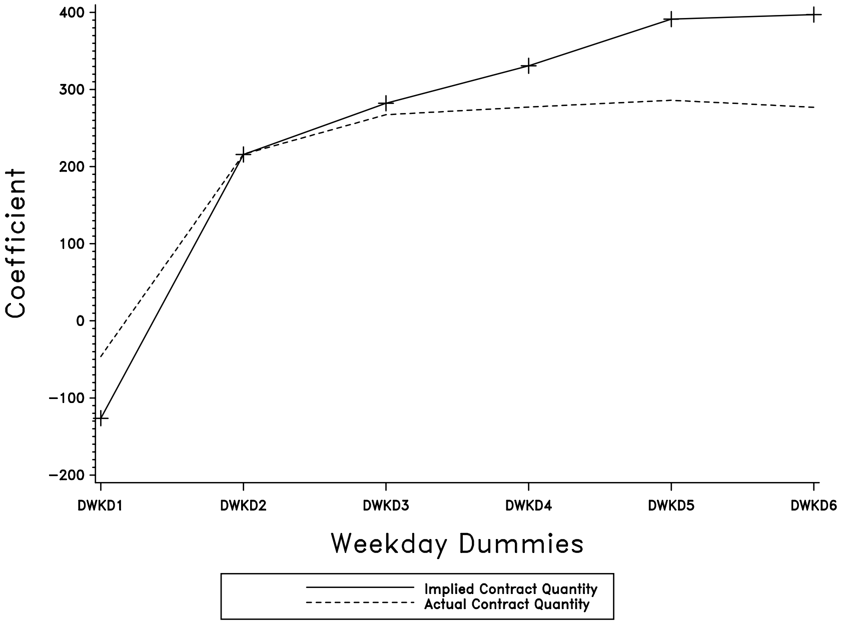


Figure 4.7. Contract quantity regression (MC = $AU10).

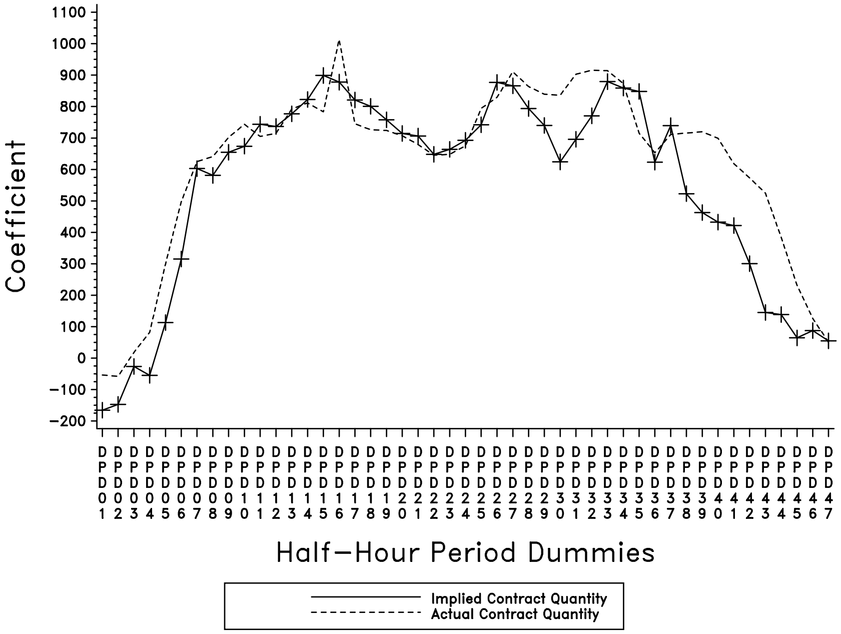


Figure 4.8. Contract quantity regression (MC = $AU10).

Table 4.3. *Genset-level marginal cost functions*

	Plant 1		Plant 2	
	Identity	Optimal	Identity	Optimal
β_{0k}	10.1	9.32	4.36	12.14
SE(β_{0k})	(1.23)	(1.14)	(1.53)	(0.74)
β_{1k}	−0.002	0.00103	−0.000448	0.0017
SE(β_{1k})	(0.006)	(0.000087)	(0.0041)	(0.000784)
β_{2k}	0.00000669	0.0000917	0.00031	0.0000686
SE(β_{2k})	(0.00001)	(0.00001)	(0.000085)	(0.00001)

Note: SE(β) = estimated standard error of the coefficient estimate, using the asymptotic covariance matrix given in Hansen (1982).

of the asymptotic normal distribution of these parameter estimates under the same assumptions.

The coefficient estimates are fairly precisely estimated across the four columns of results. As expected, the GMM estimates using a consistent estimate of the optimal weighting matrix appear to be more precisely estimated. The optimized value of the objective function from the GMM estimation with the consistent estimate of the optimal weighting matrix can be used to test the overidentifying restrictions implied by best-reply bidding. To estimate the six parameters of $C_1'(q, \beta_1)$ and $C_2'(q, \beta_2)$, I use seventy moment restrictions – ten bid increments for seven gensets. From the results of Hansen (1982), the optimized value of the objective function is asymptotically distributed as a chi-squared random variable with 64 degrees of freedom – the number of moment restrictions less the number of parameters estimated – under the null hypothesis that all of the moment restrictions imposed to estimate the parameters are valid. The optimized value of the objective function using a consistent estimate of the optimal weighting matrix is 75.40, which is less than the 0.05 critical value from a chi-squared random variable with 64 degrees of freedom. This implies that the null hypothesis of the validity of the moment restrictions given in (5.1) cannot be rejected by the actual bid data. This hypothesis test implies that given the parametric genset-unit cost functions in Equations (5.10) and (5.11), the overidentifying moment restrictions implied by the assumption of expected profit-maximizing behavior by Firm A cannot be rejected.

Figures 4.9 and 4.10 plot the estimated genset-level marginal cost functions for Plant 1 and Plant 2 along with pointwise 95 percent confidence intervals for the case of the consistent estimate of the optimal weighting matrix estimation results. Using the identity matrix as the GMM weighting matrix did not yield significantly different results. The confidence intervals indicate that the marginal cost curves are fairly precisely estimated. The results are broadly consistent with the results for the case of best-reply pricing. However, considerably more insight about the structure of Firm A's costs can be drawn from these results. Specifically, these results indicate the Plant 1 gensets are, for the same

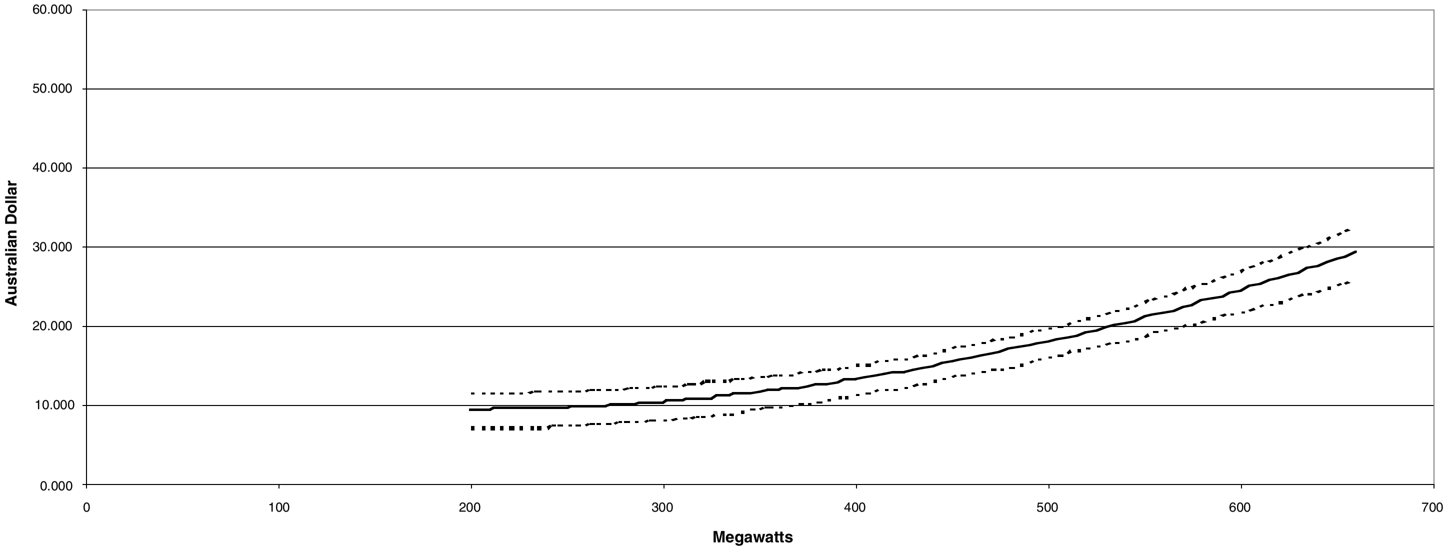


Figure 4.9. Marginal costs for Plant 1 of Firm A with optimal weighting matrix.

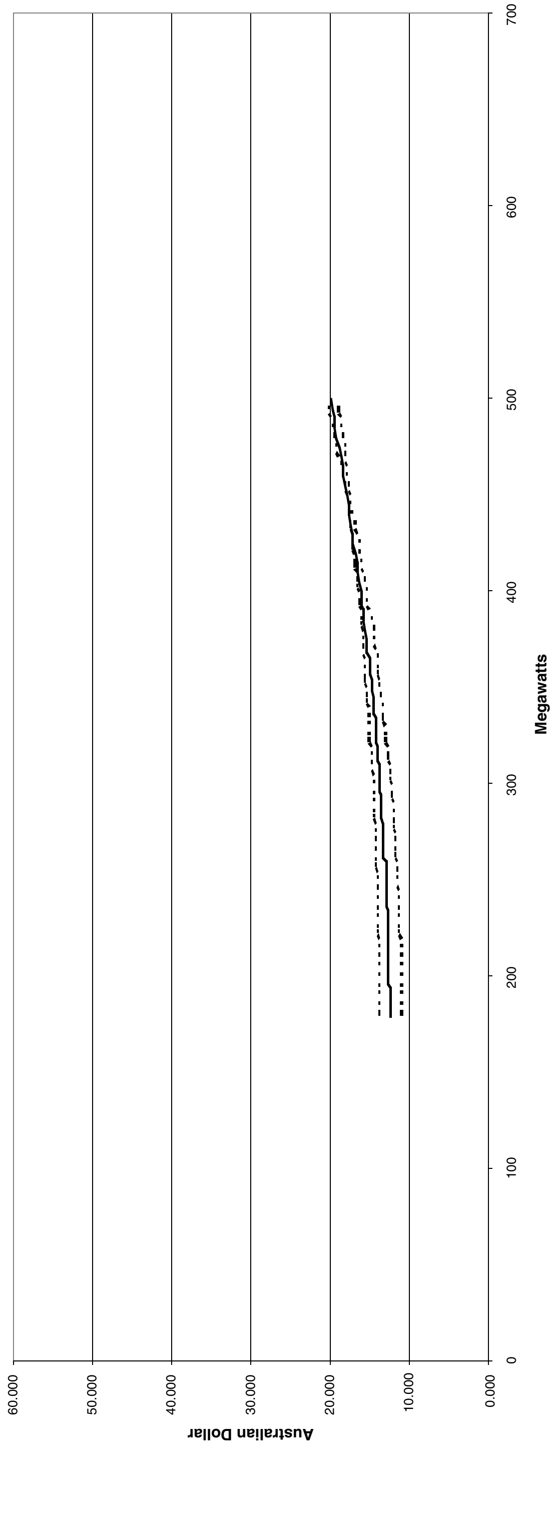


Figure 4.10. Marginal costs for Plant 2 of Firm A with optimal weighting matrix.

output levels, lower cost than Plant 2 gensets. This result was also confirmed by discussions with plant operators at Firm A and the fact that Plant 2 gensets are used less intensively by Firm A than are Plant 1 gensets.

The other result to emerge from this analysis is the increasing, convex marginal cost curves for all cases except the identity weighting matrix and the Plant 1 genset. One potential explanation for this result comes from discussions with market participants in wholesale electricity markets around the world. They argue that genset owners behave as if their marginal cost curves look like those in Figures 4.8 and 4.9 because they are hedging against the risk of unit outages when they have sold a significant amount of forward contracts. Because of the enormous financial risk associated with losing a genset in real time combined with the inability to quickly bring up another unit in time to meet this contingency, generation unit owners apply a large and increasing opportunity cost to the last one-third to one-quarter of the capacity of each genset. That way they will leave sufficient unloaded capacity on all of their units in the hours leading up to the daily peak so that they can be assured of meeting their forward financial commitments for the day even if one of their units is forced out.

This desire to use other units as physical hedges against the likelihood of a forced outage seems to be a very plausible explanation for the form of the marginal cost functions I recover, in light of the following facts about Firm A. First, during this time period, Firm A sold forward a large fraction of its expected output, and in some periods even more than its expected output. Second, all of Firm A's units are large coal-fired units, which can take close to 24 hours to start up and bring to their minimum operating level. Both of these facts argue in favor of Firm A's operating its units as if there were increasing marginal costs at an increasing rate as output approached the capacity of the unit.

8. IMPLICATIONS FOR MARKET MONITORING AND DIRECTIONS FOR FUTURE RESEARCH

There are a variety of uses for these results in market monitoring. Perhaps the most obvious is in constructing an estimate of the magnitude of variable profits earned by the firm over some time period. A major topic of debate among policymakers and market participants is the extent to which price spikes are needed for generation unit owners to earn an adequate return on the capital invested in each generating facility. This debate is particularly contentious with respect to units that supply energy only during peak periods. The methodology presented in this paper can be used to inform this debate.

Using these estimated marginal cost functions and actual market outcomes, one can compute an estimate of the magnitude of variable profits a generating unit earns over any time horizon. This information can then be used to determine whether the variable profit level earned on an annual basis from this unit is sufficient to adequately compensate the owner for the risk taken. This calculation should be performed on a unit-by-unit basis to determine the extent to which some units earn higher returns than other units. By comparing these variable profit levels to the annual fixed cost and capital costs of the unit, one can make a determination of the long-term profitability of each unit. Borenstein, Bushnell, and Wolak (2002) present a methodology for computing marketwide measures of the extent market power exercised in a wholesale electricity market. They apply their procedure to the California electricity market over the period from June 1998 to October 2000. These sorts of results should provide useful input into the regulatory decision-making process on the appropriate magnitude of allowable price spikes and the necessity of bid caps in competitive electricity markets. Determining the answers to these questions is particularly important in light of the events in all wholesale electricity markets throughout the United States during the summers of 1999 and 2000.

The framework outlined here can be extended in a number of directions. One extension involves using these methods in multisettlement electricity markets such as the California electricity supply industry. Here market participants make day-ahead commitments to supply or demand a fixed quantity of electricity and then purchase or sell any imbalance energy necessary to meet their actual supply and demand obligations in the ISO's real-time energy market. In this case the generator's profits from supplying electricity for the day are the result of selling into two sequential electricity markets. Consequently, one way to model this process is to assume best-reply pricing for the firm in both markets (and that the firm knows that it will attain best-reply prices in the real-time market when bidding into the PX market) and derive the implied marginal cost for the generator. Preliminary results from applying this procedure to California ISO and PX data are encouraging. Extending this procedure to the case of best-reply bidding in both markets is significantly more challenging.

A second direction for extensions is to specify Firm A's cost function as a multigenset cost function such as $C(q_1, \ldots, q_7, \beta)$. Assuming this functional form, I can examine the extent to which complementarities exist in the operation of different units. The marginal cost function for a given genset could also be generalized to allow dependence across periods within the day in the marginal cost of producing in a given hour, so that the variable cost for genset k might take the form $C_k(q_{1k}, \ldots, q_{48,k}, \beta_k)$, to quantify the impact of ramping constraints and other factors that prevent units from quickly moving from one output level to another.

ACKNOWLEDGMENTS

I thank Jeremy Fox, Jun Ishii, and Marshall Yan for outstanding research assistance. Bob Wilson and Lars Peter Hansen provided very helpful comments on previous drafts. Partial financial support for this research was provided by the National Science Foundation.

References

Berry, S. T. (1994), "Estimating Discrete-Choice Models of Product Differentiation," *RAND Journal of Economics*, 25(2), 242–262.

Berry, S. T., J. L. Levinsohn, and A. Pakes (1995), "Automobile Prices in Market Equilibrium," *Econometrica*, 63(4), 841–890.

Bresnahan, T. (1981), "Departures from Marginal-Cost Pricing in the American Automobile Industry," *Journal of Econometrics*, 17, 201–227.

Bresnahan, T. (1987), "Competition and Collusion in the American Automobile Market: The 1955 Price War," *Journal of Industrial Economics*, June, 45(4), 457–482.

Blundell, R. and J. L. Powell (2003), "Endogeneity in Nonparametric and Semiparametric Regression Models," in *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress*, Vol. 11 (ed. by Mathias Dewatripont, Lars Peter Hansen, and Stephen J. Turnovsky), Cambridge: Cambridge University Press.

Borenstein, S., J. Bushnell, and F. A. Wolak (2002), "Measuring Inefficiencies in California's Re-structured Electricity Market," forthcoming, *American Economic Review*, available from http://www. stanford.edu/∼wolak.

Florens, J.-P. (2003), "Inverse Problems and Structural Econometrics: The Example of Instrumental Variables," in *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress*, Vol. 11 (ed. by Mathias Dewatripont, Lars Peter Hansen, and Stephen J. Turnovsky), Cambridge: Cambridge University Press.

Goldberg, P. K. (1995), "Product Differentiation & Oligopoly in International Markets: The Case of the U.S. Automobile Industry," *Econometrica*, 63(4), 891–952.

Green, R. J. and D. M. Newbery (1992), "Competition in the British Electricity Spot Market," *Journal of Political Economy*, 100(5), 929–953.

Hansen, L. P. (1982), "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50, 1029–1054.

Klemperer, P. D. and M. A. Meyer (1989), "Supply Function Equilibria in Oligopoly under Uncertainty," *Econometrica*, 57(6), 1243–1277.

Porter, R. H. (1983), "A Study of Cartel Stability: The Joint Executive Committee, 1880–1886," *Bell Journal of Economics*, Autumn, 14(2), 301–314.

Rosse, J. N. (1970), "Estimating Cost Function Parameters without Using Cost Data: Illustrated Methodology," *Econometrica*, 38(2), 256–275.

Wolak, F. A. (1999), "Market Design and Price Behavior in Restructured Electricity Markets: An International Comparison," in *Competition Policy in the Asia Pacific Region*, *EASE Vol. 8* (ed. by T. Ito and A. Krueger), Chicago: University of Chicago Press.

Wolak, F. A. (2000), "An Empirical Analysis of the Impact of Hedge Contracts on Bidding Behavior in a Competitive Electricity Market," *International Economic Journal*, 14(2), 1–40.

Wolak, F. A. (2001a), "A Model of Optimal Bidding Behavior in a Competitive Electricity Market," January, available from http://www.stanford.edu/∼wolak.

Wolak, F. A. (2001b), "Identification, Estimation and Inference of Cost Functions in Multi-Unit Auction Models with an Application to Competitive Electricity Markets," February, available from http://www.stanford.edu/∼wolak.

Wolak, F. A. (2002), "Designing a Competitive Wholesale Market that Benefits Consumers," available from http://www.stanford.edu/∼wolak.

CHAPTER 5

Liquidity, Default, and Crashes

Endogenous Contracts in General Equilibrium

John Geanakoplos

1. LIQUIDITY CRISES

In 1994 and again in 1998, fixed income markets, especially derivatives and mortgage derivatives, suffered terrible liquidity crises, which at the time seemed to threaten the stability of the whole financial system. Though we shall see that economists have had trouble precisely defining liquidity, the general features of the liquidity crises can be succinctly described. In both episodes one saw the following:

1. There was a price crash in defaultable assets, especially for the riskiest assets, but *without* a commensurate increase in subsequent defaults.
2. These effects spilled over many markets, such as high-risk corporate bonds and mortgages, even though the risks of default are probably not correlated between the markets.
3. There was a huge income loss for the most adventurous buyers (e.g., hedge funds purchasing derivatives).
4. There was an increase in the spread between more "liquid" and less "liquid" securities (like off-the-run Treasuries and on-the-run Treasuries), even though the assets had the same probability of default. Thus default spreads and liquidity spreads both increased.
5. The margin requirements on borrowing were raised.
6. Borrowing decreased.

Another crucial observation is that the crises did not seem to be driven by changes in the riskless interest rate. In 1994, Treasury interest rates were rising before the crisis, whereas in 1998 they were falling. Moreover, when the margin requirements on borrowing were raised, the interest rate charged remained virtually the same.

The thesis of this paper is that a liquidity crisis begins when bad news about assets raises their probability of default, which then redistributes wealth away from their natural buyers. But the crisis reaches its climax only when the margin requirements of borrowers using the assets as collateral are tightened.

The major part of my argument shows how the equilibrium forces of supply and demand can endogenously determine margin requirements. Next I show that most kinds of bad shocks loosen margin requirements. But shocks that indicate default are not only more likely, but also quicker, to lead to tighter margin requirements.

In Section 2, I explain how the possibility of default paradoxically makes the asset structure of an economy endogenous. In Sections 3 and 4, I describe the general equilibrium model of collateral equilibrium, including a precise description of endogenous margins. In Section 5, I explain how wealth redistributions and changes in margin requirements exacerbate asset price volatility. In Section 6, I present a concrete two-period example of a liquidity crisis. The result is not satisfactory because the margin requirements move in the wrong directions after a bad shock. However, the example is an important building block to the more elaborate and more satisfying three-period example presented in Section 7, in which volatility is increased by margin feedbacks and wealth redistributions. Sections 8 and 9 extend the analysis to many assets, permitting a rigorous explanation of liquidity spreads and spillovers. Section 10 reviews other possible explanations of liquidity crises. Finally, Section 11 suggests a formal definition of liquidity.

2. DEFAULT AND ENDOGENOUS CONTRACTS

Standard general equilibrium theory is unable to answer the question: Which contracts are traded in equilibrium? I argue that introducing default into general equilibrium makes room for a competitive theory of endogenous contracts, and that in such a model, liquidity and liquidity crises can be explained.

Let $\mathcal{C}$ be the set of marketed contracts, and let $\mathcal{C}^*$ be the set of contracts that are actively traded by at least one agent in equilibrium. A contract in $\mathcal{C}\backslash\mathcal{C}^*$ is priced by the market, but untraded. If there are far fewer promises in $\mathcal{C}^*$, then we can say that the forces of supply and demand select the set of actively traded promises.

When there is the possibility of default, promises must be augmented by contract provisions that give the seller the incentive to deliver what he or she promised. These generally take one of two forms – punishment or collateral. It would seem to be far more daunting a task for competitive equilibrium theory to explain the terms of the loan contracts, as well as their promises and prices. Given a fixed promise, there are many attendant terms, such as how much collateral has to be put up, what the penalty for default should be, what the maximum allowable sales is, and so on. It would seem that instead of one equation matching supply and demand and one endogenous price, as in conventional general equilibrium theory, there is now a whole host of new endogenous variables representing contract terms, but the same single market-clearing equation for each promise. Equilibrium looks to be underdetermined.

The answer to the puzzle is to let each specification of contract terms $c \in C$ define another market, and therefore another market-clearing price. The contract terms themselves are not endogenous variables like prices, which get set by equilibrium at one determinate value. Instead, they are parameters that help to define the different markets; but equilibrium can set their values just as well. Equilibrium will choose determinate levels of trade q_c in each market $c \in C$. And if, for example, $q_c = 0$ for all $c \neq c^*$, then we can say that the forces of supply and demand have determined the contract terms c^*. This possibility is often obscured by the economist's preoccupation with price.

The public, and unfortunately the Federal Reserve, also share the economists' preoccupation with price. Every day the newspapers print the interest rates, and the Federal Reserve monitors them closely and systematically. But it might happen that the contract terms c^* attending most new loans dramatically change, while interest rates stay put. (This would imply that the prices for loans at the old terms had also dramatically shifted, but the newspapers do not print the prices of loans that are hardly transacted.) A change in c^* may be a more important harbinger of a liquidity crisis than a change in interest rates.

Scarce collateral provides a straightforward and compelling explanation for endogenous limits on contract trade. Simply put, the quantity of desired promises exceeds the value of the available collateral, and so the forces of supply and demand (operating through margin requirements) will ration the volume of trade. The rationing does not reduce the volume of trade in each contract proportionately, but instead it chokes off all trade in most contracts. As real conditions and expectations change, the margin requirements will need to change in order to maintain equilibrium. These margin changes will in turn have real effects, necessitating further adjustments in margins, and occasionally creating an *equilibrium cascade* into crisis.

The mechanisms by which scarce collateral and punishment ration contracts are similar. Both make the marginal utility of buying less than the marginal disutility of selling. With a positive probability of actual default, the buyer of a promise receives less than the seller delivers. For example, if a seller partially defaults and serves time in jail as punishment, she or he delivers both goods and jail time, whereas the buyer of the promise receives only the goods. Similarly, a provision of the contract might be that the seller is forced to put up collateral, that is, to buy and hold some durable good that the seller otherwise might not want, or to hold cash reserves that she or he would otherwise spend. The seller of the promise delivers goods to the buyer of the promise, but the seller also delivers the disutility of making an inconvenient transaction with a third party. The marginal utility of buying a promise may thus be less than the marginal disutility of selling the promise.

When the marginal utility–disutility gap is big for each agent, there is a real possibility that there will be an overlap containing a price greater than every agent's marginal utility of buying the contract, and less than every agent's marginal disutility of selling the contract, at which the contract will not be traded at all. In standard general equilibrium theory, this almost never happens to a nonredundant asset because every agent's marginal utility of buying is equal to his or her marginal disutility of selling. Standard general equilibrium theory cannot explain which assets are traded because it does not leave room for assets that are *not* traded. General equilibrium with default does.[1]

Together with Dubey and Shubik (Dubey, Geanakoplos, and Shubik, 2001) and with Dubey (Dubey and Geanakoplos, 2001a, 2001b), I built a theory of endogenous punishment and endogenous quantity constraints on sales of promises. In earlier work (Geanakoplos, 1997; Geanakoplos and Zame, 1988). I constructed a model of endogenous collateral levels. In this paper I build on this latter work, reinterpreting collateral levels in terms of liquidity and explaining how shifts in equilibrium collateral levels (margin requirements) can cause equilibrium crises.

3. DEFAULT AND COLLATERAL

The difficulty with promises is that they require some mechanism to make sure they are kept. This can take the form of penalties, administered by the courts, or collateral. As I mentioned at the outset, more and more often collateral has displaced penalties. In this paper I deal exclusively with collateral, by supposing that there is no penalty, legal or reputational, to defaulting. Of course, even collateral requires the courts to make sure the collateral changes hands in case of default.

The simplest kind of collateral is pawn-shop collateral – valuable goods such as watches or jewelry left with third parties (warehoused) for safekeeping. Financial markets have advanced as the number of goods that could function as collateral has increased, from watches and jewelry to stocks and bonds. A further advance occurred when lenders (instead of warehouses) held collateral, such as paintings, that afforded them utility. This required a more sophisticated court system, because the lender had to be obliged to return the collateral if the promise was kept. The biggest advance, however, was in allowing borrowers themselves to continue to hold the collateral. This enabled houses, and later cars, to be used as collateral, which again is possible only because of a finely tuned court system that can enforce the confiscation of collateral.

More recently, the complexity of collateral has taken several more giant steps forward. Pyramiding occurs when an agent A puts up collateral for his promise to B, and then B in turn uses A's promise to her, and hence in effect the same collateral, for a promise she makes to C, who in turn reuses the same collateral for a promise he makes to D. Mortgage passthrough securities offer a classic example of pyramiding. Pyramiding naturally gives rise to chain reactions, as a default by Mr. A ripples through, often all the way to D.

Still more complex is tranching, which arises when the same collateral backs several promises to different lenders. Needless to say, the various lenders will

[1] Moreover, it is not necessarily the default, nor even the probability of default, but the potential for default that puts the wedge between buying and selling utilities. Even if it is known that the default will not occur, given the contract provisions, these provisions may be so onerous as to choke off trade in the contract.

be concerned about whether their debts are adequately covered. Tranching usually involves a legal trust that is assigned the duty of dividing up the collateral among the different claims according to some contractual formula. Again, collateralized mortgage obligations offer a classic example of tranching.

Every one of these innovations is designed to increase or to stretch the available collateral to cover as many promises as possible. We shall see later that active default is another way of stretching the available collateral.

For the formal analysis in this paper, I avoid pyramiding and tranching. All collateral will, by assumption, be physical commodities. Collateral must be put up at the moment the promise is sold, even if the delivery is not scheduled for much later. Agents are not allowed to pledge their future endowment as collateral, because that would raise questions in the minds of lenders about whether the borrowers actually will have the endowments they pledged, and therefore it would once again destroy the anonymity of markets.

3.1. Contracts with Collateral

Let there be two periods, S states of nature, and L goods. To each contract j we must formally associate a promise $A_j \in R_+^{SL}$, and levels of collateral. Any good can potentially serve as collateral, and there is no reason why the single promise A_j cannot be backed by a collection of goods. The bundle of goods that is required to be warehoused for contract j is denoted $C_j^W \in R_+^L$, the vector of goods that the lender is allowed to hold is denoted $C_j^L \in R_+^L$, and the vector of goods the borrower is obliged to hold is denoted $C_j^B \in R_+^L$. A contract j is defined by the promise it makes *and* the collateral backing it, $(A_j, C_j^W, C_j^L, C_j^B)$. It is quite possible that there will be contracts that make the same promises $A_j = A_{j'}$, but trade at different prices because their collateral levels are different: $(C_j^W, C_j^L, C_j^B) \neq (C_{j'}^W, C_{j'}^L, C_{j'}^B)$. Similarly, the two contracts might require exactly the same collaterals, but trade at different prices because their promises are different.

The price of contract j is denoted by π_j. A borrower sells contract j, in effect borrowing π_j, in return for which he or she promises to make deliveries according to A_j.

3.2. Production

Collateral is useful only to the extent that it is still worth something when the default occurs. Durability is a special case of production, so we introduce production into our model, and allow all goods to be durable, to varying degrees.

For ease of notation we shall suppose that production is of the fixed coefficient, constant returns to scale variety. One unit of commodity ℓ becomes a vector of commodities next period. A house may become a house that is one year older, wine may become a wine that is one year older, grapes may become wine one year later, and so on. In these examples, one good became a different good the next period, but there is no reason not to permit one good to become several goods. By linearity, we can talk more succinctly about the transformation of a vector of goods $x \in \mathbb{R}_+^L$ into goods $f_s(x) \in \mathbb{R}_+^L$ for each $s \in S$.

The transformation of a commodity depends, of course, on how it is used. We suppose a bundle of goods $x \in \mathbb{R}_+^L$ is transformed into a vector $f_s^0(x) \in R_+^L$ in each state s if it is used for consumption (e.g., living in a house, or using a light bulb). If it is warehoused, then we assume that it becomes a vector $f_s^W(x) \in R_+^L$ in each state s. Likewise, if it is held as collateral by the lender, it becomes a vector $f_s^L(x) \in R_+^L$ in each state s, whereas if it is held by the borrower it becomes the vector $f_s^B(x) \in R_+^L$ in each state s. The linear functions f^0, f^W, f^L, and f^B summarize these different durabilities.

Observe that we have allowed for differential durability depending on the use to which the commodity is put. However, we have not allowed the durability to be affected by the identity of the user. In this way the anonymity of markets is maintained, and our modeling problem becomes easier.

Given the collateral requirements (C_j^W, C_j^L, C_j^B) for each contract j, the security they provide in each state s is

$$p_s \cdot \left[f_s^W\left(C_j^W\right) + f_s^L\left(C_j^L\right) + f_s^B\left(C_j^B\right)\right].$$

The collateral is owned by the borrower but may be confiscated by the lender (actually by the courts on behalf of the lender) if the borrower does not make his or her promised deliveries. Because we have assumed that the borrower has nothing to lose but his or her collateral from walking away from his or her promise, it follows that the actual delivery by every agent h on asset j in state s will be

$$D_{sj}^h = \min\left\{p_s \cdot A_s^j,\, p_s \cdot \left[f_s^W\left(C_j^W\right) + f_s^L\left(C_j^L\right) + f_s^B\left(C_j^B\right)\right]\right\}.$$

4. COLLATERAL EQUILIBRIUM

We are now ready to put together the various elements of our model. An economy E is defined by a vector

$$E = \left((u^h, e^h)_{h\in H}, \left(A^j, C_j^W, C_j^L, C_j^B\right)_{j\in J}, \left(f^0, f^W, f^L, f^B\right)\right)$$

of agent utilities $u^h : R_+^{(1+S)L} \to R$, and endowments $e^h \in R_{++}^{(1+S)L}$, asset promises and collateral levels, and the durability of goods kept by consumers, warehouses, lenders, and borrowers, respectively. We assume that the utilities u^h are continuous, concave, and weakly monotonic.

In keeping with the standard methodological approach of general equilibrium and perfect competition, we suppose that *in equilibrium* agents take the prices (p, π) of commodities and assets as given.

Our price-taking hypothesis has the implication that agents have perfect conditional foresight, in that they anticipate at time 0 what the prices p_s will be, depending on which state s prevails at time 1. Because they know the collateral that has been put up, and they know the production technology, they also understand in each state how much each asset will actually pay.

It might seem therefore that we could simply replace each asset promise A_j with an actual delivery vector, and thereby bypass the complications of collateral. However, this is not possible, because whether an asset defaults or not in state s depends on whether the promise or the collateral is worth more. Because both are vectors, this cannot be known in advance until the prices $p_s \in R^L_+$ have been determined in equilibrium.

4.1. The Budget Set

Given the prices (p, π), each agent h decides what commodities to consume, x^h_0, and what commodities, x^h_W, to save in a warehouse. The agent also decides what contract purchases θ and what contract sales φ he or she will make at time 0. Note that for every promise φ_j that the agent makes, he or she must put up the corresponding collateral $(C^W_j, C^L_j, C^B_j)\varphi_j$. The value of all his or her net trades at time 0 must be less than or equal to zero; that is, the agent cannot purchase anything without raising the money by selling something else (initial endowments of money are taken to be zero).

After the state of nature is realized in period 1, the agent must again decide on his or her net purchases of goods $(x_s - c^h_s - f^0_s(x_0) - f^W_s(x_W))$. Recall that the goods x_0 whose services the agent consumed at time 0 may be durable, and still available, in the form $f^0(x_0)$, for consumption at time 1 in each state s. These net expenditures on goods can be financed out of sales of the collateral that the agent put up in period 0, and from the receipts from contracts j that he or she purchased at time 0, less the deliveries the agent makes on the contracts he or she sold at time 0. Putting all these transactions together, and noting again that the agent cannot buy anything without also selling something else of at least equal value, we derive the budget set for agent h:

$$B^h(p, \pi) = \big\{(x_0, x_W, (x_s)_{s\in S}, \theta, \varphi) \in R^L_+ \times \mathbb{R}^L_+ \times R^{SL}_+ \times R^J_+ \times R^J_+ :$$

$$p_0\big(x_0 + x_W - e^h_0\big) + \pi(\theta - \varphi) + p_0 \sum_{j\in J} \big(C^W_j + C^L_j + C^B_j\big)\varphi_j \leq 0$$

and, for all $s \in S$,

$$\begin{aligned} & p_s\big(x_s - e^h_s - f^0_s(x_0) - f^W_s(x_W)\big) \\ & \quad \leq \sum_{j\in J} \varphi_j p_s \cdot \big[f^W_s\big(C^W_j\big) + f^L_s\big(C^L_j\big) + f^B_s\big(C^B_j\big)\big] \\ & \qquad + \sum_{j\in J} (\theta_j - \varphi_j) \min\big\{p_s \cdot A^j_s,\ p_s \cdot \big[f^W_s\big(C^W_j\big) + f^L_s\big(C^L_j\big) + f^B_s\big(C^B_j\big)\big]\big\}. \end{aligned}$$

4.2. Equilibrium

The economy $E = ((u^h, e^h)_{h\in H}, (A_j, C_j^W, C_j^L, C_j^B)_{j\in J}, (f^0, f^W, f^L, f^B))$ is in equilibrium at macro prices and individual choices $((p, \pi), (x^h, \theta^h, \varphi^h)_{h\in H})$ if supply equals demand in all the goods markets and asset markets, and if, given the prices, the designated individual choices are optimal, that is, if

$$\sum_{h\in H}\left(x_0^h + x_W^h - e_0^h + \sum_{j\in J}\left(C_j^W + C_j^L + C_j^B\right)\varphi_j^h\right) = 0,$$

$$\text{and, for all } s \geq 1, \qquad (4.1)$$

$$\sum_{h\in H}\left(x_s^h - e_s^h - f_s^0\left(x_0^h\right) - f_s^W\left(x_W^h\right)\right) - \sum_{j\in J}\sum_{h\in H}\varphi_j^h\left[f_s^W\left(C_j^W\right) + f_s^L\left(C_j^L\right) + f_s^B\left(C_j^B\right)\right] = 0, \qquad (4.1')$$

$$\sum_{h\in H}(\theta^h - \varphi^h) = 0, \qquad (4.2)$$

$$(x^h, \theta^h, \varphi^h) \in B^h(p, \pi), \qquad (4.3)$$

$$(x, \theta, \varphi) \in B^h(p, \pi) \Rightarrow u^h\left(x_0 + \sum_{j\in J}\left[C_j^B\varphi_j + C_j^L\theta_j\right], \bar{x}\right) \leq u^h\left(x_0^h + \sum_{j\in J}\left[C_j^B\varphi_j^h + C_j^L\theta_j^h\right], \bar{x}^h\right). \qquad (4.4)$$

We write $x^h = (x_0^h, \bar{x}^h)$, so consumption at time 0 is $x_0^h + \sum_{j\in J}[C_j^B\varphi_j^h + C_j^L\theta^h]$, and consumption at time 1 is $\bar{x}^h = (x_1^h, \ldots, x_S^h)$.

4.3. The Orderly Function of Markets

The agents we have described must anticipate not only what the prices will be in each state of nature, and not only what the contracts promise in each state of nature, but also what they will actually deliver in each state of nature. The hypothesis of agent rationality is therefore slightly more stringent in this model than in the conventional models of intertemporal perfect competition. Nevertheless, equilibrium always exists in this model (under the assumptions made so far), yet in the standard model of general equilibrium with incomplete asset markets, equilibrium may not exist. The following theorem is taken from Geanakoplos and Zame (1998).

Theorem 4.1. *Under the assumptions on endowments and utilities already specified, equilibrium must exist, no matter what the structure of contracts and collateral.*

In standard general equilibrium theory, everybody keeps every promise, so agents are allowed to sell as much of a promise as they like, provided they are sure to get the funds somewhere to deliver on their promise. For example, an agent could sell a huge amount of one promise and use the proceeds to buy another promise that would deliver enough for him or her to cover the promise the agent sold. It is this potential for unbounded trades that sometimes compromises the existence of equilibrium.

When promises are kept only insofar as they are collateralized, this unboundedness problem disappears. If a contract contains no provision for collateral whatsoever, then of course everybody will rightly anticipate that it will deliver nothing, and its equilibrium price will be zero. Indeed, the economy would function exactly the same way if it were not available at all. For assets with some nonzero collateral, agents will not be able to sell arbitrarily large quantities, because the required collateral is a physical good in finite supply. As agents try to sell more of the promise, their demand for the physical collateral will eventually drive its price up above the sales price of the promise, so that on net the asset sellers will have to pay for the privilege of selling. Their sales will be limited by their budget, guaranteeing the existence of equilibrium.

4.4. Endogenous Contracts

One of the major shortcomings of the standard general equilibrium model is that it leaves unexplained which contracts are traded. Generically, all the contracts exogenously allowed into the model will be traded. When default can be avoided only by collateral, the situation is different and much more interesting.

The crucial idea is that without the need for collateral, the marginal utility $\mu_j^h(B)$ to an agent h of buying the first unit of a contract j is almost exactly the same as the marginal utility loss $\mu_j^h(S)$ in selling the first unit of the contract; we can call both μ_j^h. Only by an incredible stroke of luck will it turn out that $\mu_j^h = \mu_j^{h'}$ for different agents h and h', and hence contract j will almost surely be traded in a GEI equilibrium. When collateral must be provided by the seller, the disutility of making a promise goes up, sometimes by as much as the consumption forgone by buying the collateral. If the required collateral is borrower held, and if it is something that agent h planned to hold anyway, then there is no extra utility loss from selling the first unit of contract j. But if agent h did not plan to hold the collateral for consumption, or if all that this agent intended to hold as consumption has already been allocated as collateral for other promises, then the loss in utility from selling even the first unit of contract j would be larger than the marginal utility from buying the first unit of contract j, $\mu_j^h(S) > \mu_j^h(B)$. It might well transpire that

$$\min_{h\in H} \mu_j^h(S) > \pi_j > \max_{h\in H} \mu_j^h(B),$$

and hence that contract j does not trade at all in equilibrium.

This situation can be most clearly seen when the value of the Arrow–Debreu promises in some state exceeds the salvage value of all the durable goods carried over into that state. It is then physically impossible to collateralize every socially useful promise up to the point that every delivery is guaranteed without exception. The market system, through its equilibrating mechanism, must find a way to ration the quantity of promises. This rationing is achieved by a scarcity of collateral. The resulting wedge between the buying marginal utility of each contract and the selling marginal utility of the contract not only serves to limit the quantity of each promise, but more dramatically, it chokes off most promises altogether, so that the subset of contracts that are actually traded is endogenous and potentially much smaller than the set of available contracts.

The endogeneity of contracts applies to promises as well as to collateral levels (see Geanakoplos, 1997). However, in this paper, we shall be interested only in the latter. Let $\mathcal{C} = \{(C^W, C^L, C^B) \in Z_+^L \times Z_+^L \times Z_+^L : C_\ell^i \leq 10^{100}\}$ be a finite set of (virtually) all potential collateral levels. Fix a promise $a \in \mathbb{R}_+^{SL}$. Consider the set $J(a) = \mathcal{C}$ of all possible contracts with promise a and collateral levels $c \in \mathcal{C}$. In equilibrium, all of these many contracts will be priced, but only a very few of them will actually be traded. The rest will not be observable in the marketplace, and therefore the appearance will be given of many missing markets. The untraded contracts will lie dormant not because their promises are irrelevant to spreading risk efficiently, but because the scarce collateral does not permit more trade.

For each $\ell \in L$, consider the set $\mathcal{C}_\ell = \{(C^W, C^L, C^B) \in \mathcal{C} : C_k^W = C_k^L = C_k^B = 0 \text{ if } k \neq \ell\}$ of potential collaterals that use only the commodity ℓ. Consider the set $J_\ell(a) = \mathcal{C}_\ell$ of all possible contracts with promise a and collateral levels $c \in \mathcal{C}_\ell$. In equilibrium, all of these many contracts will be priced, but often only one of them will be *actively* traded. Thus houses are always used as borrower held collateral, and the present value of the promises is usually 80 percent of the value of the house. Mortgage derivatives are always lender held collateral, and the present value of the promises is a number that varies from time to time (90 percent of the value of the collateral in 1997, and 50 percent in 1998).

4.5. Margins and Liquidity

Let contract j be given by the vector $(A_j, C_j^W, C_j^B, C_j^L)$. Define $p(C_j) = p_0 \cdot [C_j^W + C_j^B + C_j^L]$. In equilibrium, we will always have $\pi_j \leq p(C_j)$, because by assumption the payoff from the contract will never exceed the value of the collateral.

The margin on a contract j in equilibrium is defined as

$$m_j = \frac{p(C_j) - \pi_j}{p(C_j)}.$$

The margin m_j will be positive for essentially three reasons. First, the collateral may provide utility before the promises come due, boosting the price of the collateral above the price of the promise. Second, there may be a mismatch between future collateral values and the promises, so that in some states, the collateral is worth more than the promises. Third, to the extent the mismatch is variable, risk-averse lenders might prefer higher margins m_j to higher interest rates (i.e., to bigger A_j).

We shall see that sometimes we can associate with each collateral a single promise that is actively traded. In that case, we can think of the active margin requirement as pertaining to the collateral. Each durable good ℓ might then have an active margin requirement m_ℓ. As we said, houses generally are bought with 20 percent cash and the rest is borrowed. We shall see in later sections how this active margin requirement is determined endogenously in equilibrium.

Provisionally we shall think of the active equilibrium margin requirements as the liquidity of the system (the higher the margin, the lower the liquidity). Section 11 gives a more careful definition of liquidity.

4.5.1. Assets and Contracts

Each actively traded contract defines a margin requirement. We can think of this margin symmetrically as the margin pertaining to the promise of the contract, or as the margin pertaining to the collateral. Our focus so far has been on the promise, but gradually we shall shift our attention to the collateral. When the collateral is a single durable good, we shall often call it an *asset*. The active margin requirement has a big effect on asset prices, as we shall see.

4.6. Collateral and Default

It would be interesting to catalog the rules by which the market implicitly chooses one promise over another, or one level of collateral over another. This issue is more fully developed in Geanakoplos (1997) and in Geanakoplos and Zame (1998), but let us note some things here. The active margin requirement determines how much default there is in the economy. Collateral is scarce. The easiest way of economizing on collateral is by allowing default in some states of nature. If one vector of collaterals guarantees full delivery in every state of nature, there is no point in trading the same promise collateralized by greater levels of collateral. Finally, if a vector of promises is very different from the vector of its collateral values across the states of nature, then the contract is not well drawn. In some states there will be too much collateral, and in others not enough. One might suspect that such a contract would also not be traded. The general principle is that the market chooses contracts that are as efficient as possible, given the prices. This is made precise in the next section.

4.7. Constrained Efficiency

It is to be expected that an increase in available collateral, either through an improvement in the legal system (e.g., borrower held collateral), or through the increased durability of goods, will be welfare improving. But could it lower welfare in a pathological case? More subtly, we might wonder whether government intervention could improve the functioning of financial markets given a fixed level of available collateral. After all, the unavailability of collateral might create a wedge that prevents agents from trading the promises in J that would lead to a Pareto-improving sharing of future risks. If the government transferred wealth to those agents unable to afford collateral, or subsidized some market to make it easier to get collateral, could the general welfare be improved? What if the government prohibited trade in contracts with low collateral levels? The answer, surprisingly, is no, government intervention cannot be Pareto improving, at least under some important restrictions. See the theorem from Geanakoplos and Zame (1998) that follows.

Constrained Efficiency Theorem. *Each collateral equilibrium is Pareto efficient among the allocations that (1) are feasible and (2) given whatever period 0 decisions are assigned, respect each agent's budget set at every state s at time 1 at the old equilibrium prices, and (3) entail that agents will deliver no more on their contract promises than they have to, namely the minimum of the promise and the value of the collateral put up at time 0, given the original prices.*

In particular, no matter how the government redistributes income in period 0, and taxes and subsidizes various markets at time 0, if it allows markets to clear on their own at time 1, then we can be sure that if the time 1 market-clearing relative prices are the same as they were at the old equilibrium, then the new allocation cannot Pareto dominate the old equilibrium allocation. This will be illustrated in our examples.

In contrast, this theorem does *not* hold if relative prices do change. By intervening to prohibit high leverage (selling contracts with low levels of collateral), the government can reduce the volatility of future prices. If the volatility is sufficiently disruptive, leveraging limits can actually be Pareto improving (see Geanakoplos and Kubler, 1999, for an example).

5. VOLATILITY

5.1. Natural Buyers, the Marginal Buyer, and the Distribution of Wealth

In any general economic equilibrium, the price of a good depends on the utilities of the agents and the distribution of wealth. If the agents who are fondest of the good are also relatively wealthy, the good's price will be particularly high. Any redistribution of wealth away from these "natural buyers" toward agents who like the good less will tend to lower the price of the good.

To a large extent, the value of durable goods depends on the expectations, and, when markets are incomplete, on the risk aversion of potential investors, as well as on the intrinsic utility of the good. These multiple determinants of value make it quite likely that there will be wide divergences in the valuations different agents put on durable goods.

For example, farms in 1929 could be thought of as an investment, available to farmers and bankers, but to farmers there is a superior intrinsic value that made it sensible for them to own them and use them at the same time. Because the farmers did not have enough money to buy farms outright, they typically borrowed money and used their farms as collateral. Similarly, mortgage derivatives in the 1990s were worth much more to investors who had the technology and understanding to hedge them than they were to the average investor.

Since the 1929 stock market crash, it has been widely argued that low margin requirements can increase the volatility of stock prices. The argument is usually of the following kind: When there is bad news about the stocks, margins are called and the agents who borrowed using the stocks as collateral are forced to put them on the market, which lowers their prices still further.

The trouble with this argument is that it does not quite go far enough. In general equilibrium theory, every commodity (and thus every asset) is for sale at every moment. Hence the crucial step in which the borrowers are forced to put the collateral up for sale has by itself no bite. Nevertheless, the argument is exactly on the right track.

We argue that indeed using houses or stocks, or mortgage derivatives, as collateral for loans (i.e., allowing them to be bought on margin) makes their prices more volatile. The reason is that those agents with the most optimistic view of the assets' future values, or simply the highest marginal utility for their services, will be enabled by buying on margin to hold a larger fraction of them than they could have afforded otherwise.

The initial price of those assets will be much higher than if they could not be used as collateral for two reasons: Every agent can afford to pay more for them by promising future wealth, and second, the marginal buyer will tend to be somebody with a higher marginal utility for the asset than would otherwise be the case.

As a result of the margin purchases, the investment by the optimistic agents is greatly leveraged. When the asset rises in value, these agents do exceedingly well, and when the asset falls in price, these agents do exceedingly badly. Thus on bad news the stock price falls for two reasons: The news itself causes everyone to value it less, and this lower valuation causes a redistribution of wealth away from the optimists and toward the pessimists who did not buy on margin. The marginal buyer of the stock is therefore likely to be someone less optimistic than would have been the case had the stock not been purchased on margin, and the income redistribution not been so severe. Thus the fall in price is likely to be more severe than if the stock could not have been purchased on margin.[2]

[2] Instead of imagining that the shock and income redistribution causes the assets to become partly owned by less enthusiastic buyers, which we called the marginal buyer effect, we could imagine instead that the original buyers themselves became less enthusiastic as their diminished wealth (and inevitable diminished consumption) lowered the asset's marginal utility relative to the marginal utility of consumption.

5.2. Volatility and Incomplete Markets

The analysis here depends on the (endogenous) incompleteness of risk markets. When risk markets are incomplete, trade in assets and contracts might make the distribution of wealth *more* volatile, especially across low-probability events. If asset prices are very sensitive to the distribution of wealth, this can lead to occasional, large changes in asset prices.

When risk markets are complete, trade in contracts will tend to make the distribution of wealth fairly constant across states, eliminating the wild swings possible with incomplete markets.

Scarce collateral endogenously limits the contract trade, forcing incomplete risk markets even when any contract could be written (but delivery not enforced).

5.3. Volatility II: Asset Values and Margin Requirements

Even without any change in the distribution of income, changes in margin requirements can dramatically affect asset values. Durable assets can provide dividends for years into the future, vastly exceeding the quantity of consumption goods in the present. However, if the buyers of the assets cannot spend by borrowing against their future wealth, the price of the assets might remain quite low simply because the "liquidity constrained" buyers cannot call on enough financial resources. A toughening of margin requirements can thus cause asset prices to tumble.

5.4. Why Margin Requirements Get Tougher

The thesis of this paper is that liquidity crises reach their climax with a stiffening of margin requirements following bad news. But many kinds of bad news should have the effect of loosening margin requirements. Of course, regulators can suddenly enforce higher margins. If returns become more volatile, or lenders become more risk averse, margin requirements will likely stiffen. Furthermore, if worries about adverse selection or moral hazard increase, margin requirements will toughen. All these factors probably contributed to the crises of 1994 and 1998. But here we are seeking reasons stemming purely from the logic of collateral equilibrium that could explain why adverse news might lead to tighter margins, and thus a multiplier effect on prices.

One possibility is that when lenders lose money to some defaulting borrowers, their decreased wealth makes them more risk averse. They might then demand higher margins, and this may lead to a fall in asset prices. However, in the mortgage derivative crisis of 1998, none of the lenders lost any money, including the creditors of Long-Term Capital.

One important question is whether a fall in asset prices themselves will lead to higher margins. The answer depends on what caused the fall in price. Very often, bad news for asset prices will lead to a reduction in margin requirements.

For example, if asset prices decline because of an income shock to the natural buyers, lenders may demand less onerous margins because they feel asset prices have less far to fall.

If asset values follow a geometric random walk, then after an adverse shock, prices may be lower, but the standard deviation of outcomes is also scaled down, so the margin requirement (which is a ratio) may very well hold constant.

A productivity shock that raises the probability of (the same) bad outcome will tend to lower asset prices, but also to *ease* margins. For example, suppose that some asset Y could produce 1 with probability b or $R < 1$ with probability $1 - b$. Suppose the contracts that are backed by Y involve promises that are multiples of the payoffs (1, 1) across the two states. If the natural buyers were risk neutral and believed in b, the asset would sell for $p_Y = b1 + (1 - b)R$, provided that these buyers had access to enough cash. If the lenders were infinitely risk averse, it is not unreasonable to guess that they would lend at most R against one unit of Y as collateral. The margin requirement would then be

$$\begin{aligned} m &= \frac{p_Y - R}{p_Y} = \frac{b1 + (1-b)R - R}{b1 + (1-b)R} = \frac{b(1-R)}{b(1-R) + R} \\ &= \frac{1}{1 + \{R/[b(1-R)]\}}. \end{aligned}$$

It is clear that if the probability of a bad event increases, b goes down, and *m decreases*, if $0 < R < 1$. The reason is that the drop Δ in p_Y causes a percentage drop Δ/p_Y in the price of Y, but a bigger percentage drop, $\Delta/(p_Y - R)$, in the required down payment.

In contrast, in the same situation, a productivity shock that lowers R, keeping b fixed, dramatically *raises* the margin requirement, as can be seen in the formula just given.

Bad news about assets typically does not take the form that, *if* default occurs, the recovery will be less. Typically the bad news suggests that default is more likely, but not worse. So how can bad news create tighter margins? By indicating that the default, if it comes, will come sooner! We shall see that the combination of *more likely* and *sooner* can lead to higher margins (even though *more likely* by itself often leads to lower margins).

We must rigorously investigate how the margin is set. In the last paragraph, we described utilities for which it seemed plausible that the margin would be set high enough to eliminate default. Suppose instead that the natural buyers of Y are risk neutral as before, but that they also get a utility boost simply from holding Y. Suppose the lenders are also risk neutral, and agree with the buyers that the probability of the good state is b. Then it can easily be shown that the relevant loan will promise 1 in both states, but because of default it will deliver what Y delivers, namely 1 or R (thereby defaulting by $1 - R$ in the bad state).

The next sections present a more elaborate example, worked out in detail, to see how equilibrium determines a unique collateral level.

6. ENDOGENOUS COLLATERAL WITH HETEROGENOUS BELIEFS: A SIMPLE EXAMPLE

Let us begin with the same example in which there are two goods ($L = 2$), X and Y, in each state $s = 0, 1, 2$. X is a storable consumption good, like tobacco, and Y is an investment good (say a tobacco plant) that delivers 1 unit of X when things go well in state $s = 1$, and a smaller amount $R < 1$ in state $s = 2$. Y is reminiscent of a defaultable bond or a securitized mortgage, for which there are normal payments in state $s = 1$ and default with recovery R in state $s = 2$.

In Section 5 we assumed infinitely risk-averse lenders and trivially deduced that active margin requirements would rule out default. Next we assumed completely risk-neutral borrowing and lending and trivially deduced that there would be active default. We turn now to a more subtle situation in which there are optimists who think that state 1 is very likely and pessimists who do not. The price of Y (in terms of X) at time 0 will naturally be somewhere between 1 and R, reflecting the average opinion about the probability of the good state. At that price, the optimists would like to buy Y from the pessimists, but they do not have the cash. They would gladly borrow the money, but they must put up Y as collateral for their loans. There will be a menu of loans, some with low required collateral (low margin), but high interest rates, and other contracts with low interest rates but high margin requirements. Will only one contract be traded in equilibrium, thus determining both the interest rate and the margin requirement? If so, will it be the socially efficient contract? Let us be precise.

Let each agent $h \in H \subset [0, 1]$ assign probability h to $s = 1$ and probability $1 - h$ to $s = 2$ (see Figure 5.1). Agents with h near 1 are optimists; agents with h near 0 are pessimists. (The heterogeneity in beliefs may be regarded as a reduced-form version of a more complicated model in which low-h agents are more risk averse, or have relatively bigger endowments in state 1.) Suppose that each unit of X gives 1 unit of consumption utility in each state and that Y gives no utility of consumption:

$$u^h(x_0, y_0, x_1, y_1, x_2, y_2) = x_0 + hx_1 + (1 - h)x_2.$$

Suppose that each agent h has an endowment of e units of good X and 1 unit

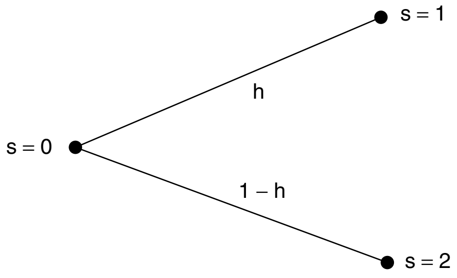


Figure 5.1.

of good Y in state $s = 0$ and nothing otherwise:

$$e^h = \left(e^h_{0x}, e^h_{0y}, e^h_{1x}, e^h_{1y}, e^h_{2x}, e^h_{2y}\right) = (e, 1, 0, 0, 0, 0).$$

Suppose that X is perfectly durable if warehoused and extinguished if consumed (like tobacco). Suppose that 1 unit of Y gives 1 unit of X in state $s = 1$ and $R < 1$ units of X in $s = 2$.

We can write this formally as

$$\begin{aligned} &f^0_s((x, y)) = f^L_s((x, y)) = f^B_s((x, y)) = (0, 0), \quad s = 1, 2,\\ &f^W_s((x, y)) = (x + y, 0), \quad s = 1,\\ &f^W_s((x, y)) = (x + Ry, 0), \quad s = 2.\end{aligned}$$

We suppose that every contract j promises 1 unit of X in each state $s = 1, 2$:

$$A^j_s = (1, 0), \quad s = 1, 2, \quad j \in J.$$

The collateral required by contract j is j units of good Y in a warehouse:

$$\begin{aligned} &C^L_j = C^B_j = (0, 0), \quad j \in J,\\ &C^W_j = (0, j), \quad j \in J.\end{aligned}$$

Buying 1 unit of Y on margin via contract j in state 0 means selling $1/j$ units of contract j for π_j/j, then paying $p_{0Y} - \pi_j/j$ cash margin plus the borrowed π_j/j for the 1 unit of Y.

For convenience, we take a continuum of agents $H = [0, a]$ and assets $J = [0, 10^{100}]$. (The definition of equilibrium must then be modified in the obvious way, replacing the sum $\sum_h$ by the integral $\int dh$ and restricting each agent to trade a finite number of contracts.) The parameter a will control the number of optimists. We proceed to compute equilibrium.

The first (and perhaps most important) property of equilibrium is indeed that only one contract will be traded. In fact, it is the contract with $j^* = 1/R$, guaranteeing that full delivery is just barely made in state $s = 2$ (and made with ease in $s = 1$). Let us temporarily take this claim on faith and construct the equilibrium, verifying the claim at the end.

We choose X as numeraire, fixing $p_{sX} = 1\ \forall s = 0, 1, 2$. Clearly, $p_{1Y} = p_{2Y} = 0$.

Some agent $b \in (0, a)$ will be indifferent to buying or selling Y at time 0. Because of the linear utilities, we guess that agents $h > b$ will buy all they can afford of Y (after selling all their X and borrowing to the max), and agents $h < b$ will sell all they have of Y, lend (buy contract j^*), and consume X. Because there is no default, and no impatience (discounting), the price $\pi_{j^*} = 1$, and the interest rate is zero. The total money spent on purchases of Y will be the X endowments of agents $h \in (b, a]$, totalling $e(a - b)$, plus the money they can borrow, which is R on each unit of Y they own, plus R on each unit of Y they buy. Total net sales of Y are the b units of agents $h \in [0, b)$, giving a price in equilibrium of

$$p_{0Y} = \frac{e(a-b) + (a-0)R}{b}. \tag{6.1}$$

A buyer on margin of Y must put down $p_{0Y} - R = [e(a-b) + aR)]/b - R = [(a-b)(e+R)]/b$ of his or her own money, getting a payoff of $1 - R$ in state 1 and zero in state 2. Because $h = b$ is indifferent to buying on margin, $[(a-b)/b](e+R) = b(1-R)$, or $b^2(1-R) + b(e+R) - a(e+R) = 0$, or

$$b = \frac{-(e+R) + \sqrt{(e+R)^2 + 4a(e+R)(1-R)}}{2(1-R)}. \tag{6.2}$$

Notice that agent b must also be indifferent to buying Y directly from cash, without borrowing, so

$$p_{0Y} = b1 + (1-b)R. \tag{6.3}$$

The price of Y is given by the marginal utilities of the *marginal buyer b*.

It follows that buying Y on margin via contract j^* costs on net $p_{0Y} - R = b(1-R)$, and pays $1 - R$ in state 1 and zero in state 2.

Thus for $h > b$, $x_0^h = 0$, $y_0^h = 0$, $C_Y^W \varphi_{j^*}^h = 1 + b/(a-b) = a/(a-b)$, $\varphi_{j^*}^h = R[a/(a-b)]$, $x_1^h = (1-R)[a/(a-b)]$, $x_2^h = 0$, and all other choice variables equal zero. For $h < b, x_0^h = e + [(a-b)/b]e, y_0^h = 0, \theta_{j^*}^h = R(a/b)$, $x_1^h = R(a/b) = x_2^h$, and all other choice variables equal zero. One can easily check that supply equals demand, and that each agent is balancing his or her budget, using the definition of p_{0Y}.

To finish the description of the equilibrium, we must describe all the other prices, and show that the agent actions are optimal. In particular, we must check that no agent wants to buy or sell (lend or borrow) any contract j with collateral level $C_j \neq C_{j^*}$. This is surprising, because optimists are very eager to buy Y, and one might imagine that they would be willing to pay a high interest rate (i.e., get a low π_j) to borrow via contract j with a lower collateral level. However, we shall see that the equilibrium interest rate $(1/\pi_j - 1)$ will be so high that the optimists will choose not to borrow at collateral levels $j \neq j^*$. We must also check that, at such a high interest rate, nobody wants to lend.

We already said that for collateral level $C_{j^*} = (0, j^*) = (0, 1/R)$, $\pi_{j^*} = 1$. In general, we set

$$\pi_j = b \min\{1, j\} + (1-b) \min\{1, jR\} \tag{6.4}$$

equal to the marginal utility of agent b. For $j > j^*$, collateral levels are wasteful, because then the collateral more than covers the loan. Thus $\pi_j = 1$ for all $j > j^*$. Nobody has any reason to lend (buy) via contract $j > j^*$, because he or she gets the same price and return as with contract j^*. Similarly, nobody would sell (borrow via) $j > j^*$, because the price is the same on j as j^*, and the collateral terms are more onerous.

We now turn to contracts $j < j^*$. These contracts involve default, but they demand higher interest (lower price for the same promise). In effect, they pay less in state 2 but more in state 1 than asset j^*. This is bad for optimistic borrowers $h > b$ and also bad for pessimistic lenders $h < b$, because these contracts deliver more in the event borrowers think will happen and lenders think will not happen. If anything, cautious optimists with h barely bigger than b might want to lend via contract j. But lending requires money, and they would rather spend all their free liquidity on Y_0. We now show rigorously that there will be no trade in contracts $j < j^*$.

A buyer of contract j receives $D_1^j = \min\{1, j\}$ in state 1 and $D_2^j = \min\{1, jR\} = jR < D_1^j$ in state 2. A seller of contract j must also buy the collateral consisting of j units of Y. On net in state s, he or she receives $-D_s^j + jf_{s1}(0, 1)$. In state 1 this is $-\min\{1, j\} + j1 \geq 0$, and in state 2 this is $-\min\{1, jR\} + jR = 0$. Notice that both the buyer and seller of contract j get a payoff at least as high in state 1 as in state 2. All prices are determined linearly by taking expectations with respect to $(b, 1 - b)$. Agents $h < b$ will therefore regard each payoff as too expensive, or at best, as break even. To see that agents $h > b$ do not wish to trade either side of contracts $j \neq j^*$, observe that their budget set is included in $B \equiv \{(x_0, x_1, x_2) : x_0 + bx_1 + (1 - b)x_2 = e + b1 + (1 - b)R\}$. Every asset and contract trades at a price equal to its contingent X payoffs, valued at price $(1, b, 1 - b)$. The collateral requirements make trades more difficult, reducing the real budget set strictly inside B. In B, agents $h > b$ clearly would take $x_1 = [e + b + (1 - b)R]/b$, $x_0 = x_2 = 0$; that is, they would spend all their wealth in state 1. But, as we saw, that is exactly what they are able to do via margin borrowing on contract j^*. Therefore, they have no incentive to trade any other contract $j \neq j^*$.

Table 5.1 gives are equilibria for various values of the exogenous parameters (R, a, e).

Consider the case where $a = e = 1$ and $R = 0.2$. The marginal buyer is $b \approx 0.69$, and the price of the asset $p_{0Y} \approx 0.69(1) + 0.31(0.02) = 0.75$.

Table 5.1.

R	0	0.1	0.2	0	0.2	0	0.2
a	1	1	1	0.75	0.75	1	1
e	1	1	1	1	1	0.75	0.75
b	0.618034	0.652091	0.686141	0.5	0.549038	0.568729	0.647233
p_{0Y}	0.618034	0.686882	0.748913	0.5	0.63923	0.568729	0.717786
m	1	0.854415	0.732946	1	0.687124	1	0.721366
x_{0H}	0	0	0	0	0	0	0
x_{1H}	2.618034	2.586882	2.548913	3	2.985641	2.318729	2.267786
x_{2H}	0	0	0	0	0	0	0
x_{0L}	1.618034	1.533529	1.457427	1.5	1.366025	1.318729	1.158779
x_{1L}	0	0.153353	0.291485	0	0.273205	0	0.309008
x_{2L}	0	0.153353	0.291485	0	0.273205	0	0.309008

6.1. The Marginal Buyer

A striking property of the example is that the prices of the asset Y and all the contracts $j \in J$ are set by the marginal utilities of a particular marginal buyer $b \in H$.

6.2. Endogenous Margin Requirement

We saw that equilibrium endogenously sets the maximum loan backed by 1 unit of Y at R, thus ruling out default. The margin requirement is then

$$m \equiv \frac{p_{0Y} - (1/j^*)\pi_{j^*}}{p_{0Y}} = 1 - \frac{R}{p_{0Y}}; \quad (1-m) = \frac{R}{p_{0Y}}, \tag{6.5}$$

where $p_{0Y} = b1 + (1-b)R$. For $(R, a, e) = (0.2, 1, 1)$, the margin requirement is $m \approx (0.75 - 0.2)/0.75 \approx 0.73$.

6.3. Margin Feedback Effects

In Table 5.1 we see that a decrease in R leads to a decline in p_{0Y}. This is natural, because with lower R, Y has lower expected payoff. The interesting point is that p_{0Y} falls by *more* than the expected output of Y, calculated with respect to the probabilities of the marginal buyer b in the old equilibrium. For example, when $(R, a, e) = (0.2, 1, 1)$, $p_{0Y} = b1 + (1-b)0.2 \approx 0.69(1) + 0.31(0.2) \approx 0.75$. When R falls to 0.1, expected output at the old b falls to $0.69(1) + (0.31)(0.1) \approx 0.72$. But actual p_{0Y} falls to 0.69, as can be seen in Table 5.1. Thus p_{0Y} falls by twice as much as would be expected from its drop in expected output.

Of course, the reason for this large fall in p_{0Y} is that b falls. As R falls, formula (6.2) shows that b must fall.[3] By (6.3), p_{0Y} falls for two reasons. It falls because with lower R, the expected payoff from Y falls, computed with respect to the old probability b. But p_{0Y} falls again because the new marginal buyer is less optimistic, $\tilde{b} < b$, and with $\tilde{b}$ replacing b, p_{0Y} would fall even with the old R.

The reason for the drop in b is the sharp increase in margin requirements. With Y much less expensive, one would expect b to rise, because presumably a smaller crowd of optimists could now afford to buy up all the Y. The only possible explanation is that the equilibrium margin requirements have gone way up, which we can confirm analytically. Using the margin requirement m in (6.5), we can write $R = (1-m)p_{0Y}$. Plugging that into the right-hand side of (6.1), we get

$$p_{0Y} = \frac{e(a-b)}{b - a(1-m)} = \frac{1}{-1 + [am/(a-b)]}. \tag{6.6}$$

[3] To see this analytically, consider the equation $f(b, R) = b^2(1-R) + b(e+R) - a(e+R)$, and recall that $b(R)$ is defined so that $f(b(R), R) = 0$. Clearly $\partial f/\partial b > 0$, and $\partial f/\partial R = -b^2 + b - a < 0$. If R falls, b must fall to restore $f(b(R), R) = 0$.

Because b and p_{0Y} fall when R falls, it follows from (6.6) that m must rise as R falls. Indeed, at $(R, a, e) = (0.1, 1, 1)$, the margin requirement rises to $m \approx 0.85$.

Thus a fall in R has a direct effect on p_{0Y}, because it lowers expected output, but it also has an indirect effect on p_{0Y} by raising margin requirements. And the indirect effect can be as large as the direct effect.

To put the matter in different words, an asymmetrically perceived decrease in the productivity and safety of the asset Y leads to an even greater fall in its price, because it also makes it harder to borrow, and markets become less liquid.

By contrast, consider the effect of a decrease in liquid wealth e. This also reduces the value of Y. A drop in $(R, a, e) = (0.2, 1, 1)$ to $(0.2, 1, 0.75)$ causes p_{0Y} to drop from 0.75 to 0.72, and b to drop from 0.69 to 0.65. However, the drop in liquidity is partly ameliorated by a decrease in margin requirements, from $m = 0.73$ to $m = 0.72$.

Similarly, a fall in the number of optimistic buyers a naturally leads to a drop in p_{0Y} and in b. As (R, a, e) falls from $(0.2, 1, 1)$ to $(0.2, 0.75, 1)$, p_{0Y} falls from 0.75 to 0.64. However, m also falls from 0.73 to 0.69, partly damping what would have been a worse fall in p_{0Y}.

Thus we see that, on one hand, certain kinds of shocks tend to reduce asset prices, but in a damped way because they also lower margin requirements. On the other hand, we shall see that shocks that reduce value less for buyers than for sellers lower price by more than they lower expected value to the original marginal buyer, because they also tend to raise the margin requirement, making a less optimistic buyer the marginal buyer, giving a second reason for prices to fall.

6.4. Endogenous Default

We saw in the example that the equilibrium margin requirements were set so that there would be no default, but that is not necessarily the case. Consider a variant of the last example in which there are three states, with payoffs of Y and agent-dependent probabilities given as shown in Figure 5.2. Note that all agents agree on the probability of $s = 3$. It is easy to check that for any $\tilde{R} < R$, in equilibrium, *only* asset $j^* = 1/R$ will be traded, exactly as before. If $\tilde{R} < R$, then there will be defaults in state 3. Rather than adjusting the collateral level to maintain zero default, equilibrium will adjust the price of all the loans to compensate lenders for the higher expected loss from default. In the new equilibrium, the price of Y and every contract j again is calculated according to the probabilities of the new marginal trader $\tilde{b}$:

$$\tilde{p}_{0Y} = \frac{\tilde{b}}{1+\varepsilon} 1 + \frac{1-\tilde{b}}{1+\varepsilon} R + \frac{\varepsilon}{1+\varepsilon} \tilde{R},$$

$$\tilde{\pi}_j = \frac{\tilde{b}}{1+\varepsilon} \min\{1, j\} + \frac{1-\tilde{b}}{1+\varepsilon} \min\{1, jR\} + \frac{\varepsilon}{1+\varepsilon} \min\{1, j\tilde{R}\}.$$

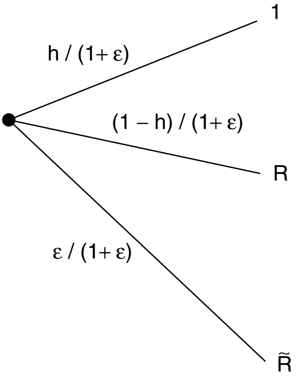


Figure 5.2.

If the new equilibrium with $\varepsilon > 0$ had the same marginal buyer as before, when $\varepsilon = 0$, then the new price $\tilde{p}_{0Y}$ would be less than the old p_{0Y} by the expected loss in output $[\varepsilon/(1+\varepsilon)](p_{0Y} - \tilde{R})$. The fall in $R\pi_{j^*}$ would, however, only be $[1/(1+\varepsilon)](R - \tilde{R})$, which is smaller. Hence agents would need less cash after borrowing to buy Y_0. This drives the price of Y_0 up, or equivalently it drives the marginal buyer $\tilde{b} > b$. (A countervailing force is that agents $h > b$ can borrow less on the Y they begin by owning. For $\tilde{R}$ near R, this is a less important effect.) Thus the equilibrium price $\tilde{p}_{Y0}$ falls *less* than the expected drop in output at the old b. Far from a feedback, news of a potential default, if universally agreed upon in probability, lowers asset prices by less than the direct effect.

This can be verified by noting that the economy with parameters $(a, e, R, \varepsilon, \tilde{R})$ has the same equilibrium marginal buyer as the old economy with parameters $(a, \tilde{e}, R)$, where $\tilde{e} = e + b^2(1 - R)/(a - b) - (R - \tilde{R})$.

6.5. Efficiency Versus Constrained Efficiency

Collateral equilibrium is clearly not Pareto efficient. In our example, agents $h < b$ end up consuming $R(a/b)$ units in states $s = 1$ and $s = 2$. In particular, agent $h = 0$, who attaches probability zero to $s = 1$, consumes $R(a/b) > 0$ in state 1, if $R > 0$. It would be better if this agent could sell some of his or her $s = 1$ consumption to agent $h = b - \varepsilon$ in exchange for some $s = 2$ consumption.

When $R = 0$, agents $h > b$ consume nothing in state 2, and agents $h < b$ consume nothing in state 1, but still the collateral equilibrium is inefficient, because agents $h < b$ consume $1 + [(a - b)/b]$ units at time 0. Again agents $h = 0$ and $h = b - \varepsilon$ could *both* be made better off if they could trade some x_0 for some x_2.

We now compute the Arrow–Debreu prices $(1, b^*, 1 - b^*)$. They must induce all agents $h \in (b^*, a]$ to consume only in state 1, and all agents $h \in [0, b^*)$ to consume only in state 2, for some b^*. Because aggregate output in state 1 is $ea + a$, and in state 2 it is $ea + aR$, we conclude that $[I(a - b^*)]/b^* = ea + a$

and $Ib^*/(1-b^*) = ea + aR$, where $I = b^*(e+1) + (1-b^*)(e+R)$ is the wealth at prices b^* of every agent $h \in [0, a]$. It follows from some algebra that

$$b^* = \frac{-(1+a)(e+R) + \sqrt{(1+a)^2(e+R)^2 + 4a(e+R)(1-R)}}{2(1-R)} < b. \tag{6.7}$$

We see, therefore, that in collateral equilibrium it is possible for asset prices to be much higher than in Arrow–Debreu equilibrium. This of course is confirmed by simulation. When $R = 0$ and $e = a = 1$, then $b^* = p^*_{0Y} = 0.414 < 0.62 = p_{0Y}$. When $R = 0.2$ and $e = a = 1$, $p^*_{0Y} = 0.43 < 0.75 = p_{0Y}$ and so on.

Because collateral equilibrium gives rise to asset prices that are too high ($p_{0Y} > p^*_{0Y}$), one is tempted to think that government intervention to impose high margin requirements would be beneficial. It is particularly tempting when there are defaults, as in the variant of the example considered in Section 6.4. But by the constrained efficiency theorem in Geanakoplos and Zame, no allocation achievable with the collateral enforcement mechanism for delivery could do better, because relative prices $p_{sY}/p_{sX} = 0$ at every equilibrium, for $s = 1, 2$.

7. CRASHES

We turn now to a dynamic context in which we can find feedback from wealth redistribution and margin changes at the same time. Imagine a multiperiod model, with the agent-specific probabilities and payoffs from asset Y indicated in Figure 5.3. It is now convenient to label agent h's opinion of the probability of up by $g(h)$.

The tree roughly corresponds to the possibility of default getting closer, as well as more probable. An asset Y can pay off 1 or default and pay off $R < 1$. Each period, there is either good news or bad news, independently drawn. The asset Y defaults only if there are two bad signals. After the first bad signal, the probability of default rises, and the horizon over which there may be a default

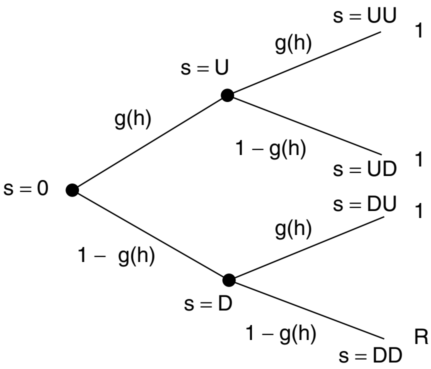


Figure 5.3.

shortens. (The rest of the tree, e.g., where there is one good signal and two bad signals, is compressed for simplicity into the simple three-stage tree shown here.)

Take the case where $g(h) = 1 - (1 - h)^2$. At time 0, agent h attaches probability $(1 - h)^4$ of eventual default in asset Y. If this agent gets bad news, $s = D$, then his or her probability rises to $(1 - h)^2$ in the next period. For an optimist with h near 1, this may be hardly any change at all.

Each node or state s in the tree is defined by its history of Us and Ds. The node sD means the node where the move D occurred after the history s. This is similarly true for sU.

Again let there be two goods X and Y in each state, where X is like cigarettes, and Y is like a tobacco plant that will produce only in the last period. Y produces one cigarette, unless two independent events go bad, in which case it produces only $R < 1$.

Endowments are 1 unit of X and Y at $s = 0$, and zero otherwise:

$$e^h_{0X} = e^h_{0Y} = 1, \quad e^h_{sX} = e^h_{sY} = 0 \; \forall s \neq 0, \quad \forall h \in H.$$

As before, X is durable but extinguishable by production, and Y is durable until its final output of X is produced. Let $f_s(z_1, z_2)$ denote the output in state s from the inputs (z_1, z_2) of X and Y in the unique state s^* preceding s. Then for $s \neq 0$, consumption destroys the good:

$$f^0_s(z_1, z_2) = f^B_s(z_1, z_2) = f^L_s(z_1, z_2) = 0.$$

Warehousing, in contrast, produces

$$\begin{aligned} f^W_U(z_1, z_2) &= f^W_D(z_1, z_2) = (z_1, z_2), \\ f^W_{UU}(z_1, z_2) &= f^W_{UD}(z_1, z_2) = f^W_{DU}(z_1, z_2) = (z_1 + z_2, 0), \\ f^W_{DD}(z_1, z_2) &= (z_1 + Rz_2, 0). \end{aligned}$$

Utility as before is given by the expected consumption of x:

$$\begin{aligned} U^h(x, x_W, y) = x_0 + g(h)x_U + (1 - g(h))x_D + g^2(h)x_{UU} \\ + g(h)(1 - g(h))[x_{UD} + x_{DU}] + (1 - g(h))^2 x_{DD}. \end{aligned}$$

We assume now that $H = [0, \alpha]$.

In each state s^* there is a contract $j \in J$ that promises 1 unit of X in each successive state s and requires j units of Y as collateral at time s^*. We write

$$\begin{aligned} A_{sj} &= (1, 0) \; \forall s \neq 0, \\ C_{s^*j} &= (0, j) \; \forall s \neq 0. \end{aligned}$$

Prices, as before, are given by p_{sX}, p_{sY} $\forall s$, and π_{sj} for all states s and all $j \in J$. It is understood that $e^h_{0^*} = 0$ and that $\pi_{sj} = 0$ for the terminal states $s \in \{UU, UD, DU, DD\}$.

The budget set for each agent h is given by exactly the same equations as before, but for every state s separately. It is understood that the output $f_s(z_1, z_2)$ belongs to the owner of the input (z_1, z_2) at state s^*.

Let us now compute equilibrium. It is clear that in state U we will have $p_{UY} = 1 = \pi_{Uj^*}$ where $j^* = 1$. The price p_{DY} will depend on what happens at $s = 0$, and on the resulting redistribution of wealth at $s = D$. Let us guess again that all contract trade at $s = D$ takes place via the contract j_D, where $j_D = 1/R$, and that all contract trade at $s = 0$ takes place via the contract j_0, where $j_0 = 1/p_{DY}$.

Following this guess we further suppose that at time 0, all agents $h \in (a, \alpha]$ borrow to the max and buy up all the Y. In the likely state $s = U$, they get rich. In the (for them) unlikely state D, they lose everything. The rest of the agents $h \in [0, a)$ sell Y and lend at $s = 0$. Thus in state $s = D$, agents $h \in [0, a)$ begin with endowments α/a of both X and Y. Of these, agents $h \in (b, a)$ will borrow to the max to buy Y in state D, and agents $h \in [0, b)$ will sell Y and lend to them, as in our first example.

If our guess is right, then the price p_{DY} will crash far below p_{0Y} for three reasons. First, every agent believes that D is bad news, and so by each agent's reckoning, the expected output of Y is lower. Second, the optimistic agents at $s = 0$ leverage, by borrowing to the hilt, and so they suffer a huge wealth setback at $s = D$, creating a feedback on prices p_{DY}, as we saw in the last section. (The elimination of the top echelon of optimists reduces the price at $s = D$.) Third, the margin requirement increases.

Computing equilibrium is similar to the simple example from Section 6, but with one wrinkle.

Agent a is the marginal buyer at $s = 0$, but at state $s = D$ this agent is much more optimistic than the marginal buyer b. Therefore he or she anticipates that \$1 is worth much more than \$1 worth of consumption of x_D. Indeed, it is worth $g(a)/g(b)$ times as much. The reason is exactly as we saw in Section 6. Agent a can buy y_D on the margin, paying $g(b)\delta$ at time D to get δ in state DU, which gives expected utility $g(a)\delta$. It follows that he or she should not consume x_0, but rather save it, then consume it if $s = U$, but if $s = D$ use the X to buy Y on margin. The marginal utility to a of x_0 is therefore $g(a)1 + (1 - g(a))[g(a)/g(b)]$.

The marginal utility to agent h from buying Y at $s = 0$ and holding it to the end is

$$MU_Y^h \equiv [1 - (1 - g(h))^2]1 + (1 - g(h))^2 R.$$

Thus we must have

$$\frac{[1 - (1 - g(a))^2]1 + (1 - g(a))^2 R}{p_{0Y}} = g(a)1 + (1 - g(a))\frac{g(a)}{g(b)}1. \tag{7.1}$$

Agents $h \in (a, \alpha]$ will buy Y on margin, spending in total $(\alpha - a) + \alpha p_{DY}$, and agents $h \in [0, a)$ will sell Y, giving

$$p_{0Y} = \frac{(\alpha - a) + \alpha p_{DY}}{a}. \tag{7.2}$$

Table 5.2.

R	0	0.1	0.2
b	0.5012	0.55055	0.60022
$1 - g(b)$	0.24880144	0.202005	0.159824
$g(b)$	0.75119856	0.797995	0.840176
p_{DY}	0.75119856	0.818195	0.872141
a	0.877700718	0.910007	0.936414
$1 - a$	0.122299282	0.89993	0.063586
p_{0Y}	0.995211492	0.998002	0.999267
$1 - g(a)$	0.014957114	0.008099	0.004043
$g(a)$	0.985042886	0.991901	0.995957
$(1 - g(a))^2$	0.000223715	6.56×10^{-5}	1.63×10^{-5}
$g(a)/g(b)$	1.31129496	1.242992	1.185415
$E_a Y/p_{0Y}$	1.004656074	1.001968	1.00075
$m0$	0.245134929	0.180146	0.127194
mD	1	0.87778	0.770679

However, as before,

$$p_{DY} = g(b)1 + (1 - g(b))R, \tag{7.3}$$

and

$$p_{DY} = \frac{(a - b) + aR}{b}. \tag{7.4}$$

Combining (7.3) and (7.4) gives

$$a = \frac{b[(1 + g(b)) + (1 - g(b))R]}{1 + R}. \tag{7.5}$$

These five equations can be solved simultaneously by means of a simple algorithm. Choose b arbitrarily. From the last equation compute a. Then compute p_{DY} and then p_{0Y}. Finally, check that Equation (7.1) holds. If not, iterate.

Table 5.2 describes the equilibrium for three different values of R, given $g(h) \equiv 1 - (1 - h)^2$, and $\alpha = 1$.

7.1. What Caused the Crash? Feedback

Consider the case $R = 0.2$. In state D, the asset price p_{DY} crashes, falling from a price of $p_{0Y} = 0.9993$ to a price $p_{DY} = 0.8721$. Three factors explain this change. First, the probability of default increased from $(1 - h)^4$ to $(1 - h)^2$ for each agent h. For the marginal buyer $a = 0.9364$, this represents an increase from virtually zero to 0.0040, still a negligible number. The drop in expected output from Y is thus about 0.003, which itself is negligible, compared to the drop in price of $(0.9993 - 0.8721) = 0.1272$.

Second, the drop in value of the price destroyed the wealth of the most optimistic buyers, effectively eliminating the purchasing power of every agent

Table 5.3.

α	0.9364136
R	0.2
b	0.563
$1 - g(b)$	0.190969
$g(b)$	0.809031
p_{DY}	0.8472248
a	0.866656302
$1 - a$	0.133343698
p_{0Y}	0.995908206
$1 - g(a)$	0.017780542
$g(a)$	0.982219458
$(1 - g(a))^2$	0.000316148
$g(a)/g(b)$	1.214069001
$E_a Y/p_{0Y}$	1.003806263
$m0$	0.149335291
mD	0.763935144

$h > a = 0.9364$. We can see what effect the disappearance of these agents would have *ceteris paribus*, by recomputing equilibrium in the three-period model with $\alpha = 0.9364$. The result is listed in Table 5.3.

We see that there is almost no effect on equilibrium prices from eliminating the 7 percent of the most optimistic buyers. *Ceteris paribus*, p_{0Y} drops from 0.9993 to 0.9959.

Third, the time of default gets closer, and the margin requirement jumps from 12.7 percent to 77 percent. We can compute the effect this change would have itself by returning to our two-period model, but with $g(h) = 1 - (1 - h)^4$, which is the probability each agent h attaches to no-default in the three-period model. The result is listed in Table 5.4.

We see again that the effect of changing the margin requirement from 12.7 percent to 79.6 percent (as well as bringing the possibility of default nearer) reduces price p_{0Y} from 0.9993 to 0.9807, again close to negligible.

The conclusion I draw is that the price crash in the example is not due to any one factor, but is due to the reinforcement each brings to the others.

Table 5.4.

R	0.2
a	1
e	1
b	0.60585
$1 - g(b)$	0.024134934
$g(b)$	0.975865066
p_{0Y}	0.980692052
m	0.796062383

7.2. Why Did the Margin Increase?

The margin requirement on Y increased because the potential crash grew nearer. An implication of drawing nearer is that the rate of information flow increased, yet agents continued to disagree in their forecasts. The pieces of information D and DD are completely symmetric, as the probability of bad news is $1 - g(h)$ in both cases, and the two events are independent. However, from D, the significance of the information to be revealed at the next step is huge. It resolves whether Y is worth 1 or R, whereas at $s = 0$, the next step will resolve whether Y is worth 1 or p_{DY}. When put another way, the variance of the price of Y one period after $s = D$ is much higher than the variance of Y one period after $s = 0$. Because agents continued to disagree about the probability of future news, the higher volatility must result in higher margins.

7.3. Liquidity and Differences of Opinion

The size of the crash depends on how far b is from a, and on how fast $g(h)$ changes as h changes. With b near a, $g(b)$ is near $g(a)$ and b's valuation of Y is not much different from a's. However, as b moves away from a, this difference accelerates, given the functional form $g(h) = 1 - (1 - h)^2$. Had we made $g(h)$ a constant, so there were no differences of opinion, there would have been no crash.

With $g(h)$ a constant, there is a deep reservoir of potential buyers of the asset at the same price. With $1 - g(h)$ very convex, this pool erodes at an accelerating pace, so that twice the bad news does more than twice the damage. Hence the power of multiple factors in the crash, when each alone makes little difference.

This appears to give us a different perspective on liquidity, closer to one of the conventional definitions. In that definition, liquidity is defined as the sensitivity of the reaction function of the price when an agent tries to sell more. It would appear from the foregoing that we might describe a market as illiquid and vulnerable to crashes if changes in the supply of the asset dramatically affect its price.

This definition does not, however, capture what is going on. *Doubling* the supply of the asset Y (which is equivalent to reducing every agent's endowment of X by 50 percent) would change equilibrium p_{0Y} from 0.9993 to 0.9920, a negligible change (see Table 5.5). It is interesting that after the doubling, the economy becomes much more vulnerable to the shock D, because then the price drops from 0.9920 to 0.7746. We will give a more appropriate definition of liquidity in Section 11.

7.4. Profits After the Crash and Cautious Speculators

Before we leave the crash example, it is instructive to reconsider why it is difficult to imagine a crash in a rational expectations world. One would think that if the crash is foreseen, then nobody would want to hold the asset before

Table 5.5.

α	1
R	0.2
b	0.46915
$1 - g(b)$	0.281801723
$g(b)$	0.718198278
p_{DY}	0.774558622
a	0.854227396
$1 - a$	0.145772604
p_{0Y}	0.992060109
$1 - g(a)$	0.021249652
$g(a)$	0.978759348
$(1 - g(a))^2$	0.000451548
$g(a)/g(b)$	1.362785708
E_aY/p_{0Y}	1.00770907
m_0	0.219188192
m_D	0.741788427

the crash. Or better, that investors would hold their capital, waiting to buy after the crash. After the crash, optimistic investors could make a far greater return than they could before the crash. Investor $a = 0.9364$ can see that he or she could make an expected return of 18 percent $(g(a)/g(b))$ above the riskless rate starting at $s = D$. Why don't investors wait to invest until after the crash (thereby eliminating the crash)?

In fact, a group of investors do wait. At $s = 0$, investor $h = a$ calculates the expected output of Y per dollar at 1.00075. Unleveraged, this investor anticipates a 0.075 percent return on his or her money, above the riskless rate, from investing in Y. This investor is risk neutral, yet he or she holds off investing in Y. Why? Because the investor foresees that if he or she keeps the money in liquid X, he or she can earn an 18 percent return $(g(a)/g(b))$ on the money above the riskless rate, after leverage, if state D should occur. There is a whole group of agents $h \in (\underline{a}, a)$ who regard Y_0 as a profitable investment, but who choose instead to sell it in order to stay liquid in X in anticipation of the crash. The probability of the crash is so low, however, that not many investors bother to prepare themselves this way, and so the crash still occurs.

8. THE LIQUIDITY SPREAD

Consider two assets that are physically identical, but suppose that only the first can be used as collateral. Will their prices be the same? To some extent this situation prevails with on-the-run and off-the-run Treasuries. The percentage of off-the-run Treasuries that are used as collateral is much smaller than the on-the-run Treasuries, and they sell for a lower price.

We can see in our simple example why this should be so. Suppose a fraction f of each agent's Y is painted blue, and can be used as collateral, while the

Table 5.6.

f	0.4	0.5	.06
α	1	1	1
e	1	1	1
R	0.2	0.2	0.2
a	0.873007	0.841774	0.810657
b	0.627968	0.636558	0.645501
p	0.702374	0.709246	0.7164
p^*	0.746014	0.746839	0.747544

remaining fraction $(1 - f)$ is painted red and cannot. What will their equilibrium prices be? If the price p^* of blue is the same as the price p of red, then all h above the marginal buyer b will spend all their money on blue (because they strictly prefer Y to X, and leveraging is the way to get as much Y as possible). All the agents $h < b$ will sell Y (since they strictly prefer X to Y). Thus there will be no buyers for red Y, and markets will fail to clear. It follows that $p^* > p$. A moment's thought shows that in equilibrium, households $h \in [0, \alpha]$ will split into three pieces. The most optimistic $h \in (a, \alpha]$ will leverage and buy blue Y. Agent a will be indifferent to buying blue on margin at the high price, and red at the low price. Agents $h \in (b, a)$ will buy only the red Y, selling their blue, and X. Agents $h \in [0, b)$ will sell all their Y. Agent b is indifferent between buying red Y and holding X.

More precisely, we can find equilibrium by solving the following equations:

$$1b + (1 - b)R = p, \tag{8.1}$$

$$\frac{e(a - b) + p^* f(a - b)}{(1 - f)(\alpha - (a - b))} = p, \tag{8.2}$$

$$\frac{e(\alpha - a) + p(1 - f)(\alpha - a) + f\alpha R}{fa} = p^*, \tag{8.3}$$

$$\frac{a(1 - R)}{p^* - R} = \frac{a1 + (1 - a)R}{p}. \tag{8.4}$$

Equation (8.1) asserts that agent b is indifferent between red Y and X. Equation (8.2) says that agents $h \in (b, a)$ take all their cash, plus the money they get selling off their blue Y, and spend it all on red Y. Everyone else sells their red Y. Equation (8.3) says that agents $h \in (a, \alpha]$ take all their cash, plus all the money they get selling their red Y plus all the money they can borrow in the blue Y, and use it to buy all the blue Y that is sold by agents $h \in [0, a)$. Finally, Equation (8.4) ensures that for agent a, the marginal utility of \$1 in blue Y is equal to the marginal utility of \$1 in red Y.

Table 5.6 gives equilibrium for various values of f, fixing $\alpha = 1$, $R = 0.2$, and $e = 1$.

The equilibrium equations sharpen our intuition about why the prices of blue Y and red Y differ, despite the fact that they are perfect substitutes. The buyers of blue Y and red Y can be disjoint sets. Y bought on the margin gives extreme payoffs $(1 - R, 0)$ that are not collinear with the payoffs $(1, R)$ from buying Y with cash.

One can see from the Table 5.6 that, as f declines, the total value of Y falls, the spread between red and blue Y increases, and both blue Y and red Y fall in value. The fact that the total value of Y falls is obvious. Y is harder to purchase if its liquidity is lower.

The fact that blue Y is more valuable than its perfect substitute, red Y, just because it can be used as collateral, is of extreme importance, as is the principle that this spread gets wider as the general liquidity in the economy falls. This liquidity spread widening is one of the hallmarks of a liquidity crisis. In our example, spread widening is inevitable because the supply of blue Y went down and the supply of red Y went up. The only curiosity is that the price of blue Y went down. This is an accidental artifact of our parameters, coming from the fact that as p declines the liquid wealth of the superoptimists $h \in (a, \alpha]$, who are sellers of red Y, declines, thereby reducing their purchasing power for blue Y.

A subtler proposition is that when one asset Y becomes less liquid, say because margin requirements are raised on it, then the spread between liquid and less liquid assets that are unrelated to Y also tends to increase. We consider such questions in the next section.

9. SPILLOVERS

Since the collapse of Long-Term Capital Management in 1998, it has become clear that many assets are much more correlated in times of (liquidity) crisis than they are otherwise. Our simple example of Section 8 can be extended to show some reasons why.

Consider the situation in which there are two assets Y and Z, and suppose that the margin requirement on Y is increased, say because R falls. Why should we expect the price of Z to fall?

At least three reasons come to mind. First, the same optimistic buyers might hold Y and Z. A negative shock to their wealth, or to their liquidity, will reduce their demand for all normal goods. Second, a decline in their liquidity will give them the incentive to shift into more liquid assets; if Z has relatively high margin requirements, and there is another comparable asset Z' with easier margin requirements, they will demand less Z. Finally, the equilibrium margin requirement may rise on Z, as a result of decreased recovery R on Y.

9.1. Correlated Output

At first glance it would seem that, if two assets had very similar returns, then they would be close substitutes. If R fell for Y, impairing its value, we might expect investors to switch to Z, possibly raising its value. However, this substitution

Table 5.7.

α	1	1
f	0.5	0.5
e	1	1
R	0.3	0.2
a	0.85873	0.841774
b	0.646957	0.636558
p	0.752871	0.709246
p^*	0.802224	0.746839

effect can easily be swamped by an income effect. If Y and Z are closely correlated, it is likely that optimists about Y are also optimistic about Z. The fall in R causes an income shock to Y buyers, which impairs their ability to buy Z.

When R falls, we saw that the price of Y falls for two reasons: First, because the expected output goes down, and second because the new marginal buyer is a more pessimistic fellow. If Y and Z are very correlated, then a more pessimistic buyer for Y will be more pessimistic about Z, and so the price of Z should fall as well.

We can see this in the example from the last section. Holding the fraction of blue Y fixed at 0.5, and lowering R on both blue Y and Z = red Y, reduces the price of both by more than expected output decreases, as can be seen from Table 5.7.

When R falls from 0.3 to 0.2, both prices p and p^* fall by more than the expected output of Y and Z, calculated with respect to either the possibilities $(a, 1 - a)$ or $(b, 1 - b)$. The gap between p^* and p narrows from 0.050 to 0.037.

In the example there is no substitution effect. Agents either prefer to buy expensive Y on the margin, or they prefer to buy cheaper Z. A change in the margin requirement simply reduces the amount of Y that can be bought on margin, but it does not by itself induce an agent to switch. If we had three states and a more complicated example, we could have had agents holding both Y and Z and then adjusting the proportions of each. Then the gap might have narrowed more.

A similar example in which Y and Z are correlated but not identical is the following. Let Y pay 1 or R, as usual. Let Z pay 1 or 0. It is easy to see that the equilibrium is the same as it would be with one asset $W = Y + Z$. Lowering R for W will reduce p_W and make the marginal buyer b more pessimistic, but that lowers the price of both Y and Z.

9.2. Independent Outputs and Correlated Opinions

It is perfectly possible for each agent h to think that the returns from Y and Z are independent, yet for optimists about Y to be optimistic about Z. For example, we could imagine four states of nature giving payoffs from Y and Z as follows: (1, 1), (1, R), (R, 1), and (R, R). Each household h might regard his or her probabilities as $(h^2, h(1-h), (1-h)h, (1-h)^2)$, respectively. Thus everybody might agree that defaults by the Russian government and American homeowners are independent. Yet many hedge funds might have been optimistic about both, and thus simultaneously invested in Russian debt and mortgages.

In our example, every agent is risk neutral, so equilibrium is exactly the same for the independent case as for the perfectly correlated case just given. As in the example of Subsection 9.1, a decrease in R for Russian debt will lower American mortgage prices.

9.3. Cross-Collateralization and the Margin Requirement

Many lenders cross-collateralize their loans. Thus if the same promise (say of \$1) is taken out by a borrower using C_1 as collateral, and another promise is made by the same borrower using C_2 as collateral, then the lender is paid in each state s

$$\min\{2, f_s^W(C_1) + f_s^W(C_2)\},$$

where $f_s^W(\,)$ is the value of the collateral in state s.

Consider the situation in the example in Subsection 9.2 in which assets Y and Z had independent payoffs. The total value of $Y + Z$ in the four states would then be $(2, 1+R, 1-R, 2R)$. If lenders could count on borrowers' taking out an equal amount of Y-backed loans as Z-backed loans, then they might loan $1 + R$ for each collateral of $Y + Z$ (charging a higher interest rate to compensate for the chance of default). But the margin requirement is then only $[2p - (1 + R)]/2p = 1 - (1 + R)/2p$, which is less than the margin requirement for Z alone, $(p - R)/p = 1 - R/p$. Thus cross-collateralization often leads to more generous loan terms.

If Y disappears, say because Russian debt collapsed, then lenders will be lending against only Z collateral, and thus margin requirements may rise on mortgages.

9.4. Rational Expectations and Liquidity Risk

We have assumed in our examples that agents may differ in their probability assessment of exogenous events (U or D or UU), but that they all completely understand the endogenous implications of each event. In reality, of course, agents do not have identical opinions about endogenous variables. In particular, there are probably wide disparities in the probability assessments of a liquidity crisis. An optimist about liquidity crises would then be optimistic about all kinds of assets that crash in liquidity crises. He or she might therefore be led to hold all of them. However, if enough liquidity optimists do this, then they create precisely the conditions we have been describing that lead to spillovers in a liquidity crisis.

10. TWO MORE CAUSES OF LIQUIDITY CRISES

There are other explanations of liquidity crises that the examples given here suggest but that are not pursued. The first is that when lenders cross-collateralize, but leave it to the borrower to choose the proportions of collateral, there is a moral hazard problem. Desperate hedge funds facing collapse might be tempted to gamble, thus holding a less hedged portfolio, for example, not balancing Y with Z. Anticipating this, lenders might raise margin requirements, thus causing the collapse they feared.

Second, I took the possibility of default (the state in which output is $R < 1$) to be exogenous, and I looked for endogenous liquidity crashes. In reality, there is a long chain of interlocking loans and the probability of a cascade of defaults is endogenous, and also an effect of liquidity, rather than just a cause.

11. A DEFINITION OF LIQUIDITY AND LIQUID WEALTH

Liquidity is an elusive concept in economics. Sometimes it is used to refer to the volume of trade in a particular market; sometimes it means the average time needed to sell; sometimes it means the bid–ask spread in the market; sometimes it means the price function relating the change in price to the change in quantity orders; and sometimes it refers to the spread between two assets with the same promises (such as the spread between on-the-run and off-the-run Treasuries).

Some of these definitions seem to require a noncompetitive view of the world, as they presume that trades are not instantly transacted at one price. Yet some of the other definitions apply in competitive markets. It is evident that economists do not all have the same notion in mind when they speak of liquidity. However, every one of these standard definitions of liquidity is applied in each market separately.

By contrast, the examples of collateral equilibrium discussed earlier suggest a new definition of the "liquidity of the system" that depends on the interactions of agents between markets.

For simplicity, let us suppose that the production of commodities whose services are being consumed does not depend on whether they are borrower held, or lender held, or owner held: $f^0 = f^B = f^L$. Then by a small change in notation, we can more simply write the budget set as

$$
\begin{aligned}
B^h(p,\pi) \; = \Big\{ &(x,\theta,\varphi) \in R^L_+ \times \mathbb{R}^L_+ \times R^{SL}_+ \times R^J_+ \times R^J_+ : \\
& p_0\big(x_0 + x_W - e^h_0\big) + \pi(\theta - \varphi) \le 0, \\
\sum_{j\in J} \varphi_j C^W_j \le x_W, & \sum_{j\in J} \varphi_j C^B_j + \sum_{j\in J} \theta_j C^L_j \le x_0, \\
& p_s\big(x_s - e^h_s - f^0_s(x_0) - f^W_s(x_W)\big) \le \sum_{j\in J} (\theta_j - \varphi_j) D_{sj}, \\
D_{sj} = \min \big\{ &p_s \cdot A^j_s,\ p_s \cdot f^W_s\big(C^W_j\big) + p_s \cdot f^0_s\big(C^B_j + C^L_j\big)\big\}.
\end{aligned}
$$

Now the utilities u^h depend solely on $(x_0, x_1, ..., x_S)$, because x_0 now includes the goods put up as collateral and held by the borrowers or lenders.

This budget set is identical to the standard general equilibrium budget set with incomplete contracts (GEI), except for two changes. The contract payoffs D_{sj} are endogenous, instead of exogenous, and contract sales φ are restricted by the collateral requirement, which shows up in the $2L$ constraints in the third line of the budget set.

I define the liquidity of the system by how closely the collateral budget set comes to attaining the GEI budget set. One crude way of measuring this is by taking the maximum expenditure that can be made on x_0, x_W, and θ in period 0 without violating any of the budget constraints. We might call this the liquid wealth of the agent at time 0.

Liquidity can suddenly deteriorate if the collateral levels increase (i.e., if the contract at which trade is actually taking place shifts to one with the same promise but a higher collateral level).

References

Aiyagari, S. R. and M. Gertler (1995), "Overreaction of Asset Prices in General Equilibrium," mimeo, New York University.

Allen, F. and D. Gale (1994), *Financial Innovation and Risk Sharing*. Cambridge, MA: MIT Press.

Baron, D. P. (1976), "Default and the Modigliani–Miller Theorem: A Synthesis," *American Economic Review*, 66, 204–212.

Dubey, P., J. Geanakoplos, and M. Shubik (2001), "Default and Punishment in General Equilibrium," Discussion Paper 1304, Cowles Foundation.

Dubey, P. and J. Geanakoplos (2001a), "Signalling and Default: Rothschild–Stiglitz Reconsidered," Discussion Paper 1305, Cowles Foundation.

Dubey, P. and J. Geanakoplos (2001b), "Insurance Contracts Designed by Competitive Pooling," Discussion Paper 1315, Cowles Foundation.

Geanakoplos, J. (1997), "Promises, Promises," in *The Economy as an Evolving Complex System, II*, (ed. by W. B. Arthur, S. Durlauf, and D. Lane), Reading, MA: Addison–Wesley, 285–320.

Geanakoplos, J. and F. Kubler (1999), "Currencies, Crises, and Collateral," mimeo, Yale University.

Geanakoplos, J. and W. Zame (1998), "Default, Collateral, and Crashes," mimeo, Yale University.

Hellwig, M. F. (1981), "Bankruptcy, Limited Liability and the Modigliani–Miller Theorem," *American Economic Review*, 71, 155–170.

Hellwig, M. F. (1981), "A Model of Borrowing and Lending with Bankruptcy," *Econometrica*, 45, 1879–1905.

Kiyotaki, N. and J. Moore. (1997), "Credit Cycles," *Journal of Political Economy*, 105(2), 211–248.

Modigliani, F. and M. H. Miller (1958), "The Cost of Capital, Corporation Finance, and the Theory of Investment," *American Economic Review*, 48, 261–297.

Smith, V. L. (1972), “Default Risk, Scale and the Homemade Leverage Theorem,” *American Economic Review*, 62, 66–76.

Stiglitz, J. E. (1974), “The Irrelevance of Corporate Financial Policy,” *American Economic Review*, 64, 851–866.

Stiglitz, J. and A. Weiss, (1981), “Credit Rationing in Markets with Imperfect Information,” *American Economic Review*, 72, 393–410.

Zame, W. (1993), ”Efficiency and the Role of Default when Security Markets are Incomplete,” *American Economic Review*, 83, 1142–1164.

CHAPTER 6

Trading Volume

Andrew W. Lo and Jiang Wang

1. INTRODUCTION

One of the most fundamental notions of economics is the determination of prices through the interaction of supply and demand. The remarkable amount of information contained in equilibrium prices has been the subject of countless studies, both theoretical and empirical, and with respect to financial securities, several distinct literatures devoted solely to prices have developed.[1] Indeed, one of the most well-developed and most highly cited strands of modern economics is the *asset-pricing* literature.

However, the intersection of supply and demand determines not only equilibrium prices but also equilibrium quantities, yet quantities have received far less attention, especially in the asset-pricing literature (is there a parallel *asset-quantities* literature?). One explanation for this asymmetry is the fact that, for most individual investors, financial markets have traditionally been considered close to perfectly competitive, so that the size of a typical investment has little impact on prices. For such scale-free investment opportunities, quantities are largely irrelevant and returns become the basic objects of study, not prices. But for large investors such as institutional investors, the securities markets are not perfectly competitive, at least not in the short run. Moreover, when investors possess private information – about price movements, their own trading intentions, and other market factors – perfect competition is even less likely to hold.

For example, if a large pension fund were to liquidate a substantial position in one security, that security's price would drop precipitously if the liquidation were attempted through a single sell-order, yielding a significant loss in the value of the security to be sold. Instead, such a liquidation would typically

[1] For example, the *Journal of Economic Literature* classification system includes categories such as Market Structure and Pricing (D4), Price Level, Inflation, and Deflation (E31), Determination of Interest Rates and Term Structure of Interest Rates (E43), Foreign Exchange (F31), Asset Pricing (G12), and Contingent and Futures Pricing (G13).

be accomplished over several days, with a professional trader managing the liquidation process by breaking up the entire order into smaller pieces, each executed at opportune moments so as to minimize the trading costs and the overall impact of the sale on the market price of the security.[2] This suggests that there is information to be garnered from quantities as well as prices; a 50,000-share trade has different implications than a 5,000-share trade, and the *sequence* of trading volume contains information as well. The fact that the demand curves of even the most liquid financial securities are downward sloping for institutional investors, and that information is often revealed through the price-discovery process, implies that quantities are as fundamental as prices and equally worthy of investigation. Even in the absence of market imperfections, quantities reveal important information about the underlying risks of the securities and their prices. After all, such risks constitute an important motive for trading in financial securities.

In this paper, we hope to provide some balance to the asset-pricing literature by developing quantity implications of a dynamic general equilibrium model of asset markets under uncertainty, and investigating those implications empirically. Through theoretical and empirical analysis, we seek to understand the motives for trade, the process by which trades are consummated, the interaction between prices and volume, and the roles that risk preferences and market frictions play in determining trading activity as well as price dynamics. In Section 2, we propose a continuous-time model of asset prices and trading activity in which a linear factor structure emerges not only for asset returns, but also for trading activity. We examine these implications empirically in Sections 3 and 4, using recently available trading volume data for individual U.S. equity securities from 1962 to 1996, and we find that substantial cross-sectional variation in trading volume can be explained by economic factors such as trading costs. In Section 5, we examine additional implications of the model for a dynamic volume-return relation, both theoretically and empirically. We focus on trading costs explicitly in Section 6 by incorporating them into our theoretical model, and we show that even small fixed costs can induce the type of trading activity we observe in existing financial markets. In Section 7, we turn our attention to technical analysis, a controversial method of forecasting future price movements based on geometric patterns in past prices and volume. Despite its somewhat dubious academic standing, technical analysis has always emphasized the importance of volume in determining the impact of price movements over time, and the fact that many professional traders are still devoted to its practice today suggests that a deeper investigation is warranted. We conclude in Section 8 with several suggestions for future research directions for incorporating trading volume into a more complete understanding of the economics of financial markets.

[2] See Chan and Lakonishok (1995) for further discussion of the price impact of institutional trades.

2. A DYNAMIC EQUILIBRIUM MODEL

Here we develop a simple equilibrium model of asset trading and pricing in a dynamic setting. We restrict our attention to the case in which investors have homogeneous information about the economy; the extension to the case of heterogeneous information is discussed briefly at the end of this section. Because our motivation for developing this model is to derive qualitative implications for the behavior of volume and returns, we keep the model as parsimonious as possible.

2.1. The Economy

We consider an economy defined on a continuous time horizon $[0, \infty)$. There is a single, perishable good, which is also used as the numeraire. The underlying uncertainty of the economy is characterized by an n-dimensional Brownian motion $B = \{B_t : t \geq 0\}$, defined on its filtered probability space $(\Omega, \mathcal{F}, F, P)$. The filtration $F = \{\mathcal{F}_t : t \geq 0\}$ represents the information revealed by B over time.

There are J risky assets in the economy, which we call stocks. Each stock pays a stream of dividends over time. Let D_{jt} denote the cumulative dividend of stock j paid up to time t and P_{jt} its exdividend price. We assume that

$$D_{jt} = \mu_{Dj} t + \sigma_{Dj} B_t \quad (j = 1, \ldots, J), \tag{2.1}$$

where $\mu_{Dj} > 0$. Without loss of generality, we assume that the total number of shares outstanding is unity for all stocks.

In addition to the stock, there is also a risk-free bond that yields a constant, positive interest rate r.

There are I investors in the economy, each initially endowed with equal shares of the stocks and zero bonds, and with a stream of nonfinancial income (e.g., income from labor or nontraded assets). Let N_t^i denote investor i's cumulative nonfinancial income up to t. We assume that

$$N_t^i = \int_0^t \left(X_s^i + Y_s^i + Z_s\right) \sigma_N dB_s \quad (i = 1, \ldots, I), \tag{2.2}$$

where

$$X_t^i = \int_0^t \sigma_X^i dB_s, \tag{2.3a}$$

$$Y_t^i = \int_0^t -\alpha_Y Y_s^i ds + \sigma_Y^i dB_s, \tag{2.3b}$$

$$Z_t = \int_0^t -\alpha_Z Z_s ds + \sigma_Z dB_s, \tag{2.3c}$$

where $\alpha_Y > 0$ and $\alpha_Z > 0$. We also assume that for all $1 \leq k, l \leq I$, there is perfect symmetry between (X_t^k, Y_t^k) and (X_t^l, Y_t^l) in the joint distribution of the state variables ($\{D_t, X_t^i, Y_t^i, Z_t; i = 1, \ldots, I\}$), and

$$\sum_{i=1}^{I} X_t^i = \sum_{i=1}^{I} Y_t^i = 0 \quad \forall\, t \geq 0. \tag{2.4}$$

From (2.2), investor i's nonfinancial income from t to $t + dt$ is $(X_t^i + Y_t^i + Z_t)\sigma_N dB_t$. Thus, $\sigma_N dB_t$ defines the risk in investors' nonfinancial income and $(X_t^i + Y_t^i + Z_t)$ gives investor i's risk exposure. Because Z_t is common to all investors, it measures the aggregate exposure to the risk in nonfinancial income. In contrast, $(X_t^i + Y_t^i)$ measures the idiosyncratic exposure of investor i. From (2.3), it is clear that X_t^i measures the permanent component of the idiosyncratic exposure, and Y_t^i measures the transitory component.[3]

Each investor chooses his or her consumption and stock-trading policies to maximize his or her expected utility over lifetime consumption. Let C denote the investor's set of consumption policies and S the set of stock-trading policies, respectively. The consumption and stock-trading polices are F-adapted processes that satisfy certain technical conditions (see, e.g., Huang and Pages, 1990). Under a particular consumption-trading policy within the policy set, investor i's financial wealth (the market value of his or her stock and bond holdings), denoted by W_t^i, satisfies

$$W_t^i = \int_0^t \left(rW_s^i - c_s^i\right) ds + \sum_{j=1}^{J} S_{js}^i (dD_{js} + dP_{js} - rP_{js}ds) + dN_t^i. \tag{2.5}$$

We denote the set of consumption-trading policies that satisfy the budget constraint (2.5) by Φ.

Investors are assumed to choose the consumption-trading policy within the set Φ to maximize their expected lifetime utility of the following form:

$$E_0\left[-\int_0^{\infty} e^{-\rho s - \gamma c_s^i} ds\right] \quad (\rho > 0 \text{ and } \gamma > 0), \tag{2.6}$$

subject to the terminal wealth condition,

$$\lim_{t\to\infty} E\left[-e^{-\rho t - r\gamma W_t^i}\right] = 0, \tag{2.7}$$

[3] Here, we have omitted any nonrisky component in investors' nonfinancial income. Also, by (2.3), we have assumed that the steady-state distribution of the investors' exposure has a mean of zero. Relaxing these assumptions has no impact on our results. We adopt them merely for parsimony in exposition. For example, under the assumption of constant absolute risk aversion for the investors, which we make later, there is no income effect in their demand for the stocks. Thus, the level of their mean nonfinancial income plays no role in their stock trading. Also, a nonzero mean in their idiosyncratic exposure to the nonfinancial risk may change their average stock demand, but not their trading, which is generated by changes in their exposures. See Subsection 2.2 for more discussion on this point.

where γ is their risk aversion coefficient, ρ is their time-discount coefficient, and $i = 1, \dots, I$.[4] In addition, we require $4\gamma^2\sigma_N^2\sigma_Z^2 < 1$ to ensure that the model is well behaved.

2.2. Discussion, Notation, and Simplifications

The economy defined herein has several special features. For tractability, we assume that stock dividends are normally distributed and investors have constant absolute risk aversion. These assumptions have restrictive implications. For example, investors' stock demand becomes independent of their wealth, which makes the model tractable but less realistic. As a result, investors have no need to change their stock holdings as their wealth changes, which is an otherwise natural motive to trade.

To provide motivation for trading, we endow investors with nonfinancial income that is positively correlated with stock dividends. The investors' desire to manage their total risk exposure gives rise to their need to trade the stock. For example, when investors' exposure to nonfinancial risks increases, they would like to reduce their exposure to financial risks from their stock holdings because the two risks are positively correlated. In other words, they sell stock from their portfolios. Each investor's nonfinancial risk is determined by two factors: his or her exposure to the risk, which is determined by $(X_t^i + Y_t^i + Z_t)$, and the risk itself, which is given by $\sigma_N B_t$. Moreover, $(X_t^i + Y_t^i + Z_t)$ gives the idiosyncratic exposure and Z_t the aggregate exposure. As their exposures change, they adjust their stock holdings. In particular, through trading, they mutually ensure the idiosyncratic part of their nonfinancial risk.

Another feature of our model is that the interest rate on the bond is assumed to be constant; hence the bond market is required to clear. This is a modeling choice we have made to simplify our analysis and to focus on the stock market. As will become clear later, changes in the interest rate are not important for the issues we are concerned about in this paper. Also, from an empirical perspective, given the frequency we are interested in (daily or weekly), changes in interest rates are usually small.[5]

For convenience, we introduce some notational conventions. Given m scalar elements, $e_1, \dots, e_m$, let $(e_1, \dots, e_m)$ denote the row vector and $(e_1; \dots; e_m)$ the column vector formed from the m elements. Let M' denote the transpose of a matrix M. Thus, $D_t \equiv (D_{1t}; \dots; D_{Jt})$ is the column vector of cumulative stock dividends, $\mu_D \equiv (\mu_{D1}; \dots; \mu_{DJ})$ is its expected growth rate, and $D_t = \mu_D t + \sigma_D B_t$, where $\sigma_D \equiv (\sigma_{D1}; \dots; \sigma_{DJ})$.

[4] The terminal wealth condition is imposed on the strategies to prevent investors from running Ponzi schemes.

[5] Endogenizing the interest by clearing the bond market introduces additional risks (the interest rate risk). This creates additional risk-sharing motives. Also, bonds with longer maturities (if they are traded) provide additional vehicles for risk sharing.

A portfolio of stocks is given by the number of shares of each stock it contains. Let S_j be the number of shares of stock j in a portfolio, where $j = 1, \ldots, J$. Then, $S \equiv (S_1; \ldots; S_j; \ldots; S_J)$ defines the portfolio. A portfolio of particular importance is the market portfolio, denoted by S^M, which is given by

$$S^M = \iota, \tag{2.8}$$

where ι is a column vector of 1s with rank J.

For any two processes F_t and G_t, let $\langle F_t, G_t \rangle$ denote their cross-variation process and $\sigma_{F\sigma} \equiv d\langle F_t, G_t \rangle / dt$ denote their instantaneous cross-variation.

For expositional clarity, we make the following simplifying assumptions on the correlation (cross-variation) among the state variables:

$$\sigma_X^{i\,\prime}\sigma_D = \sigma_Y^{i\,\prime}\sigma_D = 0 \quad \text{and} \quad \sigma_N = \iota'\sigma_D. \tag{2.9}$$

The first condition states that the idiosyncratic components of investors' exposure to nonfinancial risk are uncorrelated with stock dividends. The second condition states that nonfinancial risk itself is perfectly correlated with the aggregate risk in stock dividends. We also assume that $n = J + 2I + 1$ and σ_{DD} has full rank to avoid redundancies.

2.3. The Equilibrium

We now define and derive the equilibrium of the economy just described. Let $P_t \equiv (P_{1t}; \ldots; P_{Jt})$ be the vector of (exdividend) stock prices and $S_t^i \equiv (S_{1t}^i; \ldots; S_{Jt}^i)$ be the vector of investor i's stock holdings.

Definition 2.1. *An equilibrium is given by a price process* $\{P_t : t \geq 0\}$ *and the investors' stock holdings* $\{S_t^i : t \geq 0\} \in \Phi$, $i = 1, \ldots, I$, *such that:*

1. S_t^i *solves investor i's optimization problem*:

$$\text{Max} \;\; E_0\left[-\int_t^\infty e^{-\rho t - \gamma c_t^i} dt\right] \tag{2.10}$$

$$\text{s. t.} \;\; W_t^i = W_0^i + \int_0^t \left(rW_s^i - c_s^i\right) ds + S_s^{i\,\prime}(dD_s + dP_s - rP_s ds) + dN_t^i \tag{2.11}$$

$$\lim_{t\to\infty} E\left[e^{-\rho t - r\gamma W_t^i}\right] = 0.$$

2. *The stock market clears*:

$$\sum_{i=1}^{i} S_t^i = \iota \quad \forall\, t \geq 0. \tag{2.12}$$

This definition of equilibrium is standard, except that the clearing of the bond market is not imposed, as already discussed.

For expositional convenience, we define Q_t to be the vector of cumulative excess dollar returns on the stocks:

$$Q_t = \int_0^t (dD_s + dP_s - rP_t ds)\,. \tag{2.13}$$

For the remainder of this section, returns on the stocks always refer to their excess dollar returns (2.13). The solution to the equilibrium is summarized in the following theorem (see the Appendix for its derivation):

Theorem 2.1. *The economy just defined has a linear equilibrium in which*

$$P_t = \frac{1}{r}\mu_D - a - bZ_t, \tag{2.14}$$

and

$$S_t^i = I^{-1}\iota + \left(X_t^i + Y_t^i\right) S^I + \left(h_{IIX} X_t^i + h_{IIY} Y_t^i\right) S^{II}, \tag{2.15}$$

where

$$a = \bar{\gamma}\sigma_{QQ}\iota + \lambda_a \sigma_{QZ}, \quad b = \frac{r\gamma}{r+\alpha_Z}\sigma_{QN} + \lambda_b \sigma_{QZ},$$

$\bar{\gamma} = \gamma/I$, λ_a, λ_b, h_{IIX}, h_{IIY} *are constants given in the Appendix,* ι *is the market portfolio, and*

$$S^I \equiv (\sigma_{QQ})^{-1}\sigma_{QN}, \quad S^{II} \equiv (\sigma_{QQ})^{-1}\sigma_{QZ}$$

are two hedging portfolios.

The equilibrium has the following properties. First, stock prices equal the expected value of future dividends discounted at the risk-free rate (μ_D/r) minus a risk discount $(a + bZ_t)$, which is an affine function of Z_t, the aggregate exposure to the nonfinancial risk. Stock prices are independent of the other state variables. Second, four-fund separation holds for the investors' portfolio choices; that is, all investors hold combinations of the same four portfolios: the market portfolio, two hedging portfolios, and the riskless bond.[6] The first hedging portfolio S^I is used to hedge nonfinancial risks and the second hedging portfolio S^{II} is used to hedge changes in market conditions, which are driven by changes in Z_t. Third, stock returns are not *iid* over time; in particular, we have

$$dQ_t = [ra + (r + \alpha_Z)bZ_t]\,dt + \sigma_Q dB_t, \tag{2.16}$$

[6] As a matter of convention, we use the phrase "$(K+1)$-fund separation" to describe the situation in which investors hold combinations of the same $K+1$ funds – the riskless bond and K stock funds. Therefore, four-fund separation implies three stock funds.

where $\sigma_Q = \sigma_D - b\sigma_Z$. Thus, expected stock returns are affine functions of Z_t, which follows an AR(1) process.

To develop further intuition for the model and the nature of the equilibrium, consider the special case when $Z_t = 0$:

Corollary 2.1. *When $Z_t = 0 \; \forall \; t \geq 0$, the equilibrium stock price is*

$$P_t = \frac{1}{r}\mu_D - \bar{\gamma}\sigma_{DD}\iota,$$

and the investors' stock holdings are

$$S_t^i = \left(I^{-1} - X_t^i - Y_t^i\right)\iota\,.$$

In this case, the stock prices are constant over time and all investors hold a fraction of the market portfolio. Moreover, stock returns are given by

$$dQ_t = (r\bar{\gamma}\sigma_{DD}\iota)\,dt + \sigma_D dB_t,$$

and the return on the market portfolio is

$$dQ_{Mt} = \iota' dQ_t = r\bar{\gamma}(\iota'\sigma_{DD}\iota)\,dt + \iota'\sigma_D dB_t.$$

Let $\sigma_M^2 = \iota'\sigma_{DD}\iota$ denote the squared volatility of the return on the market portfolio. We can define

$$\beta_M \equiv \left(1/\sigma_M^2\right) d\langle Q_t, Q_{Mt}\rangle/dt = \left(1/\sigma_M^2\right)\sigma_{DD}\iota$$

to be the vector of the stocks' βs with respect to the market portfolio. Then the expected returns (per unit of time) of the stocks are given by

$$\bar{Q} \equiv E\left[dQ_t\right]/dt = r\bar{\gamma}\sigma_{DD}\iota = \beta_M \bar{Q}_M,$$

where $\bar{Q}_M = r\bar{\gamma}\sigma_M^2$ is the expected return on the market portfolio. This is the well-known pricing relation of the Sharpe–Lintner Capital Asset Pricing Model (CAPM).

Thus, when $Z_t = 0 \; \forall \; t \geq 0$, each investor holds only a fraction of the market portfolio. In other words, two-fund separation holds. When the investors adjust their stock investments, as their exposure to the nonfinancial risk changes, they trade only in the market portfolio. Furthermore, stock returns are IID over time and the CAPM holds.

In the more general case when the aggregate exposure to the nonfinancial risk Z_t is changing over time, the situation is more complicated. First, in addition to the market portfolio, the investors invest in two other portfolios to hedge their nonfinancial risk and changes in market conditions, respectively. Second, stock returns are no long *iid* over time; in general, they are predictable. Third, the CAPM no longer holds. In addition to contemporaneous market risk (the risk with respect to the market portfolio), there is also the risk of changing market conditions. For convenience, we refer to these two risks as the static risk and the dynamic risk, respectively. Different stocks have different exposures to the dynamic risk as well as the static risk. The expected returns on the stocks now depend on their exposures to these two different risks.

2.4. Implications for Trading and Returns

In the remainder of this section, we explore in more detail the implications of our model for the behavior of trading and returns.

2.4.1. Trading Activity

As (2.15) shows, the investors' portfolio choices satisfy four-fund separation, where the three stock funds are the market portfolio, ι; the hedging portfolio, S^I (which allows investors to hedge their current nonfinancial risk); and the hedging portfolio, S^{II} (which allows investors to hedge against changes in market conditions, driven by changes in the aggregate exposure to nonfinancial risk Z_t). The fact that investors hold and trade in only a few funds has strong implications about the behavior of trading activities in the market, especially their cross-sectional characteristics.

As an illustration, consider the special case when $Z_t = 0$ for all $t \geq 0$. In this case, investors trade only in the market portfolio, implying that when an investor reduces his or her stock investments, he or she sells stocks in proportion to their weights in the market portfolio. Under our normalization, this investor sells an equal number of shares of each stock. Consequently, the turnover ratio must be numerically identical across all stocks; that is, turnover exhibits an exact one-factor structure.

In the more general case of changing market conditions, investors trade in three stock portfolios, which, in general, can give rise to complex patterns in the turnover across stocks. However, when the trading in the market portfolio dominates the trading in the other two portfolios, turnover exhibits an approximate three-factor structure. There is a dominating factor – representing the trading in the market portfolio – and two minor factors, representing the trading in the other two portfolios. Furthermore, when the trading in the market portfolio dominates and the approximate three-factor structure holds for the cross section of turnover, the loadings on the two minor factors are proportional to the share weights of the two hedging portfolios. This provides a way to empirically identify the hedging portfolio from the data on the turnover of individual stocks. In Sections 3 and 4, we discuss the empirical implications for volume in more detail and present some supporting empirical evidence for the cross-sectional behavior of volume.

2.4.2. Stock Returns

The identification of the hedging portfolio allows us to further explore the predictions of the model for the behavior of returns. For example, because the stock returns are changing over time as Z_t changes, the second hedging portfolio $S^{II} = (\sigma_{QQ})^{-1}\sigma_{QZ}$ allows investors to hedge the risk of changing expected returns. As Merton (1971) has shown, among all the portfolios, the return on the hedging portfolio S^{II} has the highest correlation with changes in expected returns. In other words, it best predicts future stock returns. Moreover, the return on S^{II} serves as a proxy for the dynamic risk, whereas the return on the market portfolio serves as a proxy for the static risk. Consequently, the returns on these two portfolios give the two risk factors. The stocks' expected returns depend only on their loadings on these two portfolio returns. We obtain a two-factor pricing model, which extends the static CAPM when returns are *iid* into an intertemporal CAPM when returns are changing over time.

In contrast to the approach of Fama and French (1992), which seeks purely empirically based factor models, the approach here is to start with a structural model that specifies the risk factors, identify the risk factors empirically by using the predictions of the model on trading activities, and further test its pricing relations. We explore this approach in more detail in Lo and Wang (2000b).

2.4.3. Volume-Return Relations

The previous discussion mainly focused on the behavior of trading activity and the behavior of returns separately. We now consider the joint behavior of return and trading activities. To fix ideas, we divide the investors into two groups: those for whom the correlation between their idiosyncratic exposure to the nonfinancial risk (X_t^i and Y_t^i) and the aggregate exposure (Z_t) is positive and those for whom the correlation is negative. For now, we call them group A and group B.

When the investors' exposure to nonfinancial risk changes, their stock demand also shifts. In equilibrium, they trade with each other to revise their stock holdings and the stock prices adjust to reflect the change in demand. Thus, an immediate implication of our model is that the absolute changes in prices are positively correlated with contemporaneous trading activities, which has been documented empirically (for a survey of earlier empirical literature, see Karpoff, 1987).

Another implication of our model involves the dynamic relation between return and volume: current returns and volume can predict future returns. Campbell, Grossman, and Wang (1993) and Wang (1994) have examined this relation (see also Antoniewicz, 1993; LeBaron, 1992; Llorente et al., 2000). The intuition for such a dynamic volume-return relation is as follows. When the investors' exposure to nonfinancial risk changes, they wish to trade. Stock prices must adjust to attract other investors to take the other side. However, these price changes are unrelated to the stocks' future dividends. Thus, they give rise to changes in expected future returns. For example, when group A investors' exposure to nonfinancial risk increases, they want to sell the stocks in their portfolios. To attract group B investors to buy these stocks, stock prices have to decrease, yielding a negative return in the current period. Because there is no change in expectations about future dividends, the decrease in current stock prices implies an increase in expected future returns. In fact, it is this increase in expected returns that induces group B investors to increase their stock holdings. Thus, low current returns accompanied by high volume portends high future returns. In the opposite situation, when group B investors' exposure to nonfinancial risk decreases, we have high current returns (group A investors want to buy and drives up the prices) accompanied by high volume, which predicts lower future returns (with unchanged expectation of future dividends but higher prices). Hence, our model leads to the following dynamic volume-return relation: Returns accompanied by high volume are more likely to exhibit reversals in the future. In Section 5, we discuss the empirical analysis of this relation.

2.4.4. Merton's ICAPM

Our model bears an important relation to the intertemporal CAPM (ICAPM) framework of Merton (1973). On the basis of a class of assumed price processes, Merton has characterized the equilibrium conditions – including mutual-fund separation – for the investors' portfolios and the risk factors in determining expected returns. However, except for one special case when the static CAPM is obtained, Merton does not provide any specification of the primitives of the economy that can support the assumed price processes in equilibrium. Several authors, such as Ohlson and Rosenberg (1976), have argued that it is difficult to construct an economy capable of supporting Merton's price processes. Our model provides a concrete example in which equilibrium prices indeed fit Merton's specification.[7] Obviously, restrictive assumptions are required to achieve this specification.

3. THE DATA

One of the most exciting aspects of the recent literature on trading volume is the close correspondence between theory and empirical analysis, thanks to newly available daily volume data for individual U.S. securities from the Center for Research in Security Prices (CRSP). In this section, we describe some of the basic characteristics of this data set as a prelude to the more formal econometric and calibration analyses of Sections 4–7. We begin in Subsection 3.1 by reviewing the various measures of volume, and in light of the implications of the model in Section 2, we argue that turnover is the most natural measure of trading activity. In Subsection 3.2, we describe the construction of our turnover database, which is an extract of the CRSP volume data, and we report some basic summary statistics for turnover indexes in Subsection 3.3.

[7] Wu and Zhou (2001) considers a model that has many features similar to ours.

3.1. Volume Measures

To analyze the behavior of volume empirically, one must have an appropriate definition of volume. In the existing literature, many different volume measures have been used, including share volume, dollar volume, share turnover ratio, dollar turnover ratio, number of trades, and so on. Certainly, the choice of volume measure should depend on what is supposed to be measured, that is, what information we would like the measure to convey. A good measure of volume should contain useful information about underlying economic conditions, especially why investors trade and how they trade. Thus, the appropriate measure of volume is intimately tied to the particular model and motives for trading.

3.1.1. A Numerical Example

The theoretical model we developed in Section 2 suggests that the turnover ratio provides a good measure, in the sense that it clearly reflects the underlying economic regularities in trading activities. To illustrate this point, we consider a simple numerical example in the special case of our model when investors trade only in the market portfolio (i.e., when $Z_t = 0$ for all $t \geq 0$).

Suppose there are only two stocks, 1 and 2. For concreteness, assume that stock 1 has ten shares outstanding and is priced at \$100 per share, yielding a market value of \$1,000, and stock 2 has thirty shares outstanding and is priced at \$50 per share, yielding a market value of \$1,500. Let N_{1t} and N_{2t} denote the numbers of shares outstanding for the two stocks, respectively (for expositional convenience, we do not normalize the number of shares outstanding to one in this example). We have $N_{1t} = 10$, $N_{2t} = 30$, $P_{1t} = 100$, and $P_{2t} = 50$. In addition, suppose there are only two investors in this market – investors 1 and 2 – and two-fund separation holds for their stock investments so that both investors hold different amounts of the market portfolio. Specifically, let investor 1 hold one share of stock 1 and three shares of stock 2, and let investor 2 hold nine shares of stock 1 and twenty-seven shares of stock 2. Thus, $S^1_{1t-1} = 1$, $S^1_{2t-1} = 3$, $S^2_{1t-1} = 9$, and $S^2_{2t-1} = 27$. In this way, all shares are held and both investors hold a (fraction of) the *market* portfolio (ten shares of stock 1 and thirty shares of stock 2).

Now suppose that investor 2 liquidates \$750 of his or her portfolio – three shares of stock 1 and nine shares of stock 2 – and investor 1 is willing to purchase exactly this amount from investor 2 at the prevailing market prices.[8] After completing the transaction, investor 1 owns four shares of stock 1 and twelve shares of stock 2, and investor 2 owns six shares of stock 1 and eighteen

[8] In our model, in absence of aggregate exposure to the nonfinancial risk, the trading needs of the two investors are exactly the opposite of each other. Thus, trade occurs without any impact on prices.

Table 6.1. *Volume measures for a two-asset two-investor numerical example*

Volume Measure	*A*	*B*	Aggregate
Number of trades	1	1	2
Shares traded	3	9	12
Dollars traded	\$300	\$450	\$750
Share turnover	0.3	0.3	0.3
Dollar turnover	0.3	0.3	0.3
Relative dollar turnover	0.4	0.6	1.0
Share-weighted turnover	—	—	0.3
Equal-weighted turnover	—	—	0.3
Value-weighted turnover	—	—	0.3

Note: The example assumes that two-fund separation holds.

shares of stock 2. In other words, $S^1_{1t} = 4$, $S^1_{2t} = 12$, $S^2_{1t} = 6$, and $S^2_{2t} = 18$. What kind of trading activity does this transaction imply?

For individual stocks, we can construct the following measures of trading activity:

- Number of trades per period,
- Share volume, $V_{jt} \equiv \frac{1}{2}\sum_{i=1}^{2}\left|S^i_{jt} - S^i_{jt-1}\right|$,
- Dollar volume, $P_{jt}V_{jt}$,
- Relative dollar volume, $P_{jt}V_{jt}/\sum_j P_{jt}V_{jt}$,
- Share turnover, $\tau_{jt} \equiv V_{jt}/N_{jt}$, and
- Dollar turnover, $\nu_{jt} \equiv (P_{jt}V_{jt})/(P_{jt}N_{jt}) = \tau_{jt}$,

where $j = 1, 2$.[9] To measure aggregate trading activity, we can define similar measures:

- Number of trades per period,
- Total number of shares traded, $V_{1t} + V_{2t}$,
- Dollar volume, $P_{1t}V_{1t} + P_{2t}V_{2t}$,
- Share-weighted turnover, $\tau^{\mathrm{SW}}_t \equiv \omega^{\mathrm{SW}}_1\tau_{1t} + \omega^{\mathrm{SW}}_2\tau_{2t}$, where $\omega^{\mathrm{SW}}_j \equiv N_j/(N_1 + N_2)$ and $j = 1, 2$,
- Equal-weighted turnover, $\tau^{\mathrm{EW}}_t \equiv \frac{1}{2}(\tau_{1t} + \tau_{2t})$, and
- Value-weighted turnover, $\tau^{\mathrm{VW}}_t \equiv \omega^{\mathrm{VW}}_{1t}\tau_{1t} + \omega^{\mathrm{VW}}_{2t}\tau_{2t}$, where $\omega^{\mathrm{VW}}_j \equiv P_{jt}N_j/(P_{1t}N_1 + P_{2t}N_2)$ and $j = 1, 2$.

Table 6.1 reports the values that these various measures of trading activity take on for the hypothetical transaction between investors 1 and 2. Though these values vary considerably – two trades, twelve shares traded, \$750 traded – one regularity does emerge: The turnover measures are all identical. This is no

[9] Although the definition of dollar turnover may seem redundant because it is equivalent to share turnover, it will become more relevant in the portfolio case that follows.

coincidence, but is an implication of two-fund separation from our equilibrium model. If all investors hold the same relative proportions of stocks at all times, then it can be shown that trading activity, as measured by turnover, must be identical across all stocks. Although the other measures of volume do capture important aspects of trading activity, if the focus is on the relation between volume and equilibrium models of asset markets, such as the ICAPM we developed in Section 2, turnover yields the sharpest empirical implications and is the most natural measure. For this reason, we use turnover in our empirical analysis.

3.1.2. *Defining Individual and Portfolio Turnover*

For each individual stock j, its share volume at time t is defined by

$$V_{jt} = \frac{1}{2} \sum_{i=1}^{I} \left| S_{jt}^{i} - S_{jt-1}^{i} \right|. \tag{3.1}$$

Its turnover is defined by

$$\tau_{jt} \quad \equiv \quad \frac{V_{jt}}{N_j}, \tag{3.2}$$

where N_j is the total number of shares outstanding of stock j.[10]

For the specific purpose of investigating the volume implications of our model, we also introduce a measure of portfolio trading activity, defined as follows: For any portfolio p defined by the vector of shares held $S_t^p = (S_{1t}^p; \ldots; S_{Jt}^p)$ with nonnegative holdings in all stocks, that is, $S_{jt}^p \geq 0$ for all j, and strictly positive market value, that is, $S_t^{p\prime} P_t > 0$, let $\omega_{jt}^p \equiv S_{jt}^p P_{jt} / (S_t^{p\prime} P_t)$ be the fraction invested in stock j, with $j = 1, \ldots, J$. Then its turnover is defined to be

$$\tau_t^p \equiv \sum_{j=1}^{J} \omega_{jt}^p \tau_{jt} \ . \tag{3.3}$$

Under this definition, the turnovers of value-weighted and equal-weighted indexes are well-defined

$$\tau_t^{\mathrm{VW}} \equiv \sum_{j=1}^{J} \omega_{jt}^{\mathrm{VW}} \tau_{jt}, \qquad \tau_t^{\mathrm{EW}} \equiv \frac{1}{J} \sum_{j=1}^{J} \tau_{jt}, \tag{3.4}$$

respectively, where $\omega_{jt}^{\mathrm{VW}} \equiv N_j P_{jt} / (\sum_j N_j P_{jt})$, for $j = 1, \ldots, J$.

Although (3.3) gives a reasonable definition of portfolio turnover within the context of our model, care must be exercised in interpreting it and generalizing it into a different context. Although τ_t^{VW} and τ_t^{EW} are relevant to the volume implications of our model, they should be viewed in a more general context

[10] Although we define the turnover ratio by using the total number of shares traded, it is obvious that using the total dollar volume normalized by the total market value gives the same result.

only as particular weighted averages of individual turnover, not necessarily as the turnover of any specific trading strategy. In particular, our definition for portfolio turnover cannot be applied too broadly. For example, when short sales are allowed, some portfolio weights can be negative and (3.3) can be quite misleading because the turnover of short positions will offset the turnover of long positions. In general, the appropriate turnover measure for portfolio trading crucially depends on why these portfolios are traded and how they are traded. See Lo and Wang (2000a) for further discussion.

3.1.3. Time Aggregation

Given our choice of turnover as a measure of volume for individual securities, the most natural method of handling time aggregation is to sum turnover across dates to obtain time-aggregated turnover. Formally, if the turnover for stock j at time t is given by τ_{jt}, the turnover between $t-1$ and $t+q$, for any $q \geq 0$, is given by

$$\tau_{jt}(q) \equiv \tau_{jt} + \tau_{jt+1} + \cdots + \tau_{jt+q}. \tag{3.5}$$

3.2. MiniCRSP Volume Data

Having defined our measure of trading activity as turnover, we use the CRSP Daily Master File to construct *weekly* turnover series for individual NYSE and AMEX securities from July 1962 to December 1996 (1,800 weeks), using the time-aggregation method discussed in Subsection 3.1.[11] We choose a weekly horizon as the best compromise between maximizing sample size and minimizing the day-to-day volume and return fluctuations that have less direct economic relevance. Because our focus is the implications of portfolio theory for volume behavior, we confine our attention to ordinary common shares on the NYSE and AMEX (CRSP sharecodes 10 and 11 only), omitting ADRs, SBIs, REITs, closed-end funds, and other such exotica whose turnover may be difficult to interpret in the usual sense.[12] We also omit NASDAQ stocks altogether, because

[11] To facilitate research on turnover and to allow others to easily replicate our analysis, we have produced daily and weekly "MiniCRSP" data set extracts composed of returns, turnover, and other data items for each individual stock in the CRSP Daily Master file, stored in a format that minimizes storage space and access times. We have also prepared a set of access routines to read our extracted data sets via either sequential or random access methods on almost any hardware platform, as well as a user's guide to MiniCRSP (see Lim et al., 1998). More detailed information about MiniCRSP can be found at the website http://lfe.mit.edu/volume/.

[12] The bulk of NYSE and AMEX securities are ordinary common shares; hence limiting our sample to securities with sharecodes 10 and 11 is not especially restrictive. For example, on January 2, 1980, the entire NYSE/AMEX universe contained 2,307 securities with sharecode 10, thirty securities with sharecode 11, and fifty-five securities with sharecodes other than 10 and 11. Ordinary common shares also account for the bulk of the market capitalization of the NYSE and AMEX (excluding ADRs, of course).

the differences between NASDAQ and the NYSE/AMEX (market structure, market capitalization, etc.) have important implications for the measurement and behavior of volume (see, e.g., Atkins and Dyl, 1997), and this should be investigated separately.

Throughout our empirical analysis, we report turnover and returns in units of percent per week; they are *not* annualized.

Finally, in addition to the exchange and sharecode selection criteria imposed, we also discard thirty-seven securities from our sample because of a particular type of data error in the CRSP volume entries.[13]

3.3. Turnover Indexes

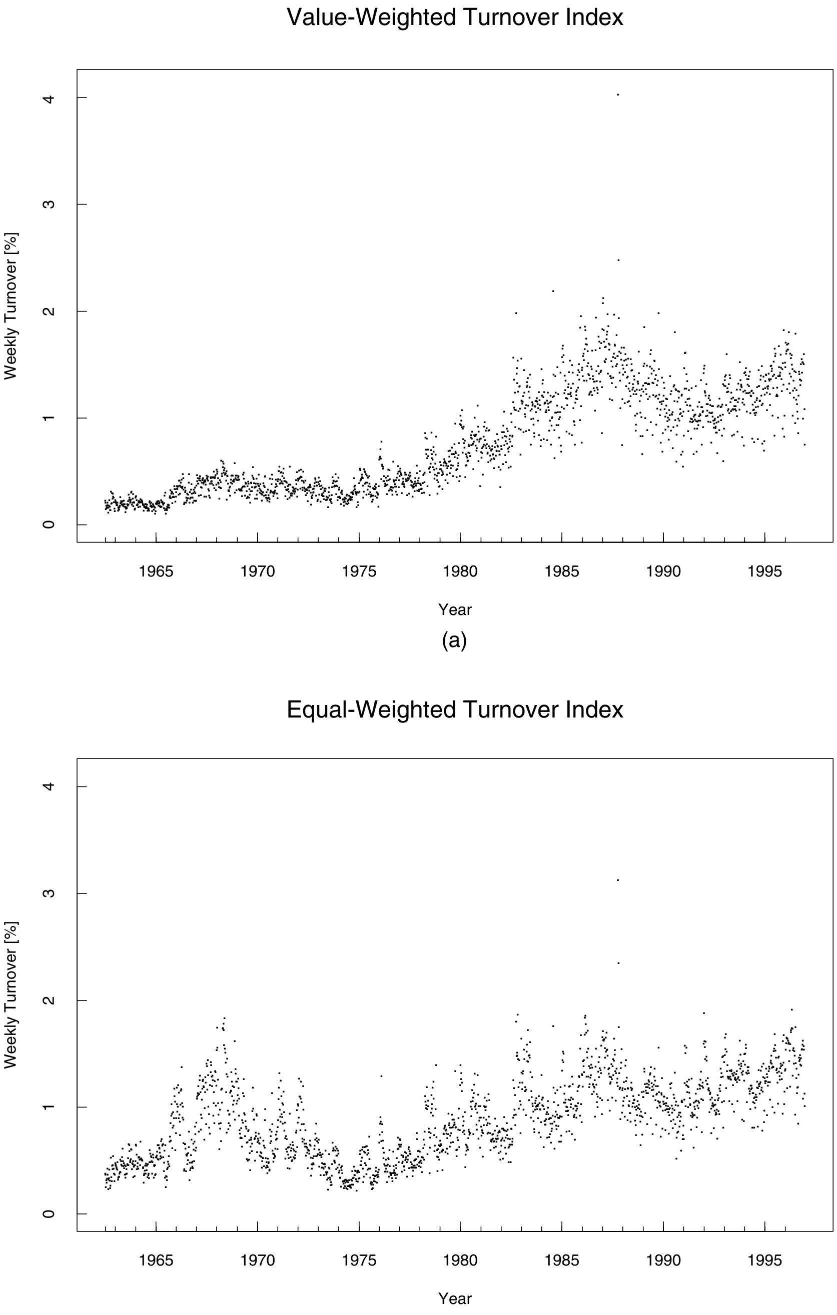
Although it is difficult to develop simple intuition for the behavior of the entire time-series–cross-section volume data set (a data set containing between 1,700 and 2,200 individual securities per week over a sample period of 1,800 weeks), some gross characteristics of volume can be observed from value-weighted and equal-weighted turnover indexes.[14] These characteristics are presented in Figures 6.1 and 6.2, and in Table 6.2.

Figure 6.1(a) shows that the value-weighted turnover has increased dramatically since the mid-1960s, growing from less than 0.20 percent to over 1 percent per week. The volatility of value-weighted turnover also increases over this period. However, equal-weighted turnover behaves somewhat differently: Figure 6.1(b) shows that it reaches a peak of nearly 2 percent in 1968, and then declines until the 1980s, when it returns to a similar level (and goes well beyond it during October 1987). These differences between the value- and equal-weighted indexes suggest that smaller-capitalization companies can have high turnover.

Table 6.3 reports various summary statistics for the two indexes over the 1962–1996 sample period as well as over five-year subperiods. Over the

[13] Briefly, the NYSE and AMEX typically report volume in round lots of 100 shares – "45" represents 4,500 shares – but, on occasion, volume is reported in shares and this is indicated by a "Z" flag attached to the particular observation. This Z status is relatively infrequent, usually valid for at least a quarter, and may change over the life of the security. In some instances, we have discovered daily share volume increasing by a factor of 100, only to decrease by a factor of 100 at a later date. Although such dramatic shifts in volume are not altogether impossible, a more plausible explanation – one that we have verified by hand in a few cases – is that the Z flag was inadvertently omitted when in fact the Z status was in force. See Lim et al. (1998) for further details.

[14] These indexes are constructed from weekly individual security turnover, where the value-weighted index is reweighted each week. Value-weighted and equal-weighted return indexes are also constructed in a similar fashion. Note that these return indexes do not correspond exactly to the time-aggregated CRSP value-weighted and equal-weighted return indexes, because we have restricted our universe of securities to ordinary common shares. However, some simple statistical comparisons show that our return indexes and the CRSP return indexes have very similar time-series properties.



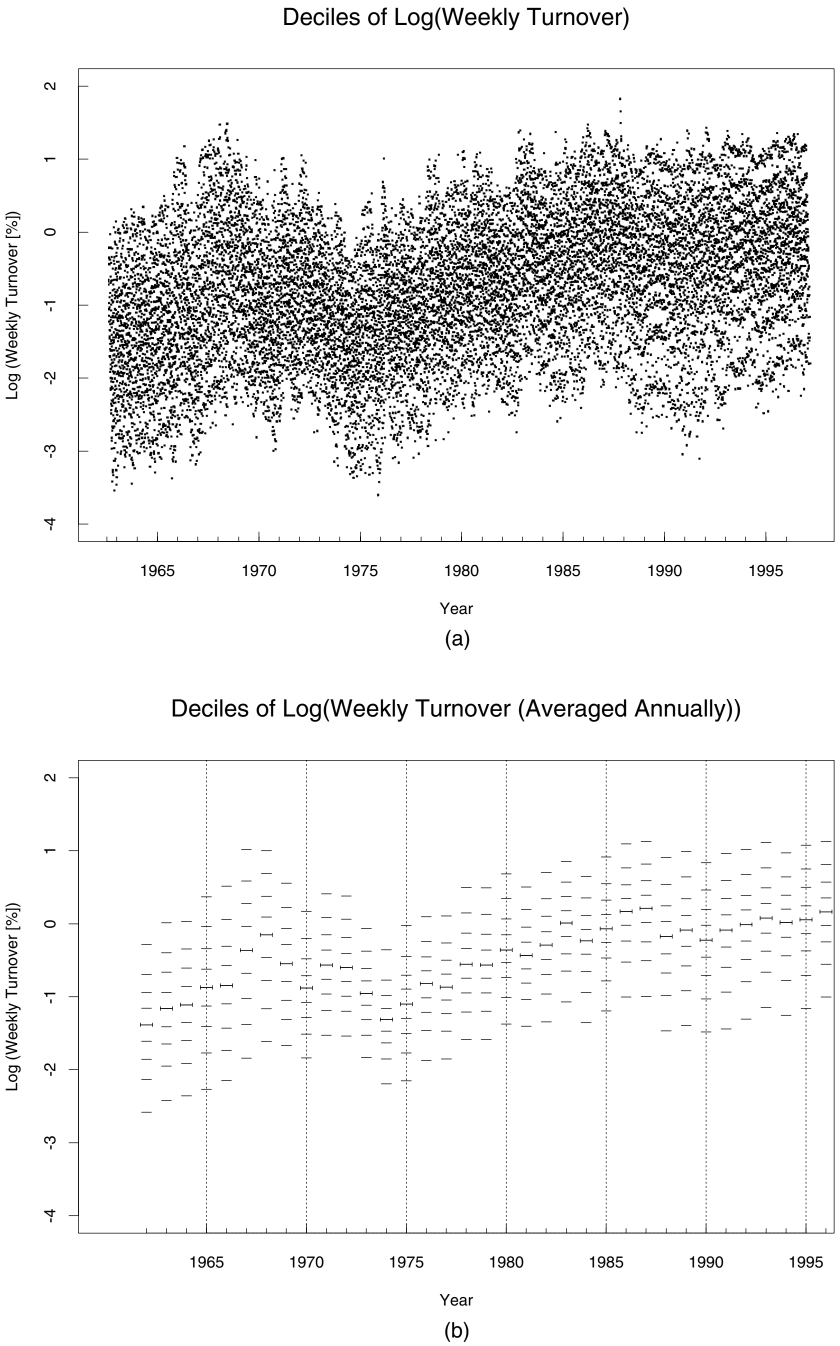
Value-Weighted Turnover Index

(a)

Equal-Weighted Turnover Index

(b)

Figure 6.1. Weekly value-weighted and equal-weighted turnover indexes, 1962–1996.



Deciles of Log(Weekly Turnover)

(a)

Deciles of Log(Weekly Turnover (Averaged Annually))

(b)

Figure 6.2. The cross-section of the logarithm of weekly turnover, 1962–1996.

Table 6.2. *Summary statistics for weekly value- and equal-weighted turnover and return indexes of NYSE and AMEX ordinary common shares*

Statistic	τ^{VW}	τ^{EW}	R^{VW}	R^{EW}
Mean	0.78	0.91	0.23	0.32
SD	0.48	0.37	1.96	2.21
Skewness	0.66	0.38	−0.41	−0.46
Kurtosis	0.21	−0.09	3.66	6.64
Percentiles				
Min.	0.13	0.24	−15.64	−18.64
5%	0.22	0.37	−3.03	−3.44
10%	0.26	0.44	−2.14	−2.26
25%	0.37	0.59	−0.94	−0.80
50%	0.64	0.91	0.33	0.49
75%	1.19	1.20	1.44	1.53
90%	1.44	1.41	2.37	2.61
95%	1.57	1.55	3.31	3.42
Max.	4.06	3.16	8.81	13.68

Statistic	τ^{VW}	τ^{EW}	R^{VW}	R^{EW}
	1962–1966 (234 weeks)			
Mean	0.25	0.57	0.23	0.30
SD	0.07	0.21	1.29	1.54
Skewness	1.02	1.47	−0.35	−0.76
Kurtosis	0.80	2.04	1.02	2.50
	1967–1971 (261 weeks)			
Mean	0.40	0.93	0.18	0.32
SD	0.08	0.32	1.89	2.62
Skewness	0.17	0.57	0.42	0.40
Kurtosis	−0.42	−0.26	1.52	2.19
	1972–1976 (261 weeks)			
Mean	0.37	0.52	0.10	0.19
SD	0.10	0.20	2.39	2.78
Skewness	0.93	1.44	−0.13	0.41
Kurtosis	1.57	2.59	0.35	1.12

Autocorrelations				
ρ_1	91.25	86.73	5.39	25.63
ρ_2	88.59	81.89	−0.21	10.92
ρ_3	87.62	79.30	3.27	9.34
ρ_4	87.44	78.07	−2.03	4.94
ρ_5	87.03	76.47	−2.18	1.11
ρ_6	86.17	74.14	1.70	4.07
ρ_7	87.22	74.16	5.13	1.69
ρ_8	86.57	72.95	−7.15	−5.78
ρ_9	85.92	71.06	2.22	2.54
ρ_{10}	84.63	68.59	−2.34	−2.44
Box–Pierce Q_{10}	13,723.0	10,525.0	23.0	175.1
	(0.000)	(0.000)	(0.010)	(0.000)

	1977–1981 (261 weeks)			
Mean	0.62	0.77	0.21	0.44
SD	0.18	0.22	1.97	2.08
Skewness	0.29	0.62	−0.33	−1.01
Kurtosis	−0.58	−0.05	0.31	1.72
	1982–1986 (261 weeks)			
Mean	1.20	1.11	0.37	0.39
SD	0.30	0.29	2.01	1.93
Skewness	0.28	0.45	0.42	0.32
Kurtosis	0.14	−0.28	1.33	1.19
	1987–1991 (261 weeks)			
Mean	1.29	1.15	0.29	0.24
SD	0.35	0.27	2.43	2.62
Skewness	2.20	2.15	−1.51	−2.06
Kurtosis	14.88	12.81	7.85	16.44
	1992–1996 (261 weeks)			
Mean	1.25	1.31	0.27	0.37
SD	0.23	0.22	1.37	1.41
Skewness	−0.06	−0.05	−0.38	−0.48
Kurtosis	−0.21	−0.24	1.00	1.30

Note: Table shows CRSP sharecodes 10 and 11, excluding 37 stocks containing Z errors in reported volume, for July 1962 to December 1996 (1,800 weeks) and subperiods. Turnover and returns are measured in percent per week and p values for Box–Pierce statistics are reported in parentheses.

Table 6.3. *Cross-sectional regressions of median weekly turnover of NYSE and AMEX ordinary common shares*

c	$\hat{\alpha}_{r,j}$	$\hat{\beta}_{r,j}$	$\hat{\sigma}_{\epsilon,r,j}$	v_j	p_j	d_j	SP500$_j$	$\hat{\gamma}_{r,j}(1)$	R^2 (%)
1962–1966 (234 weeks, 2,073 stocks)									
0.742	0.059	0.354	0.043	−0.064	0.150	0.071	0.048	0.004	41.8
(0.108)	(0.019)	(0.014)	(0.006)	(0.006)	(0.014)	(0.081)	(0.018)	(0.001)	
−0.306	0.068	0.344	0.053	—	0.070	0.130	−0.006	0.006	38.8
(0.034)	(0.020)	(0.015)	(0.006)		(0.012)	(0.083)	(0.018)	(0.001)	
0.378	0.111	0.401	0.013	−0.028	—	0.119	0.048	0.005	38.7
(0.105)	(0.019)	(0.014)	(0.005)	(0.005)		(0.083)	(0.019)	(0.001)	
1967–1971 (261 weeks, 2,292 stocks)									
0.289	0.134	0.448	0.095	−0.062	0.249	0.027	0.028	0.006	44.7
(0.181)	(0.024)	(0.023)	(0.009)	(0.010)	(0.023)	(0.235)	(0.025)	(0.002)	
−0.797	0.152	0.434	0.112	—	0.173	0.117	−0.026	0.007	43.7
(0.066)	(0.024)	(0.023)	(0.009)		(0.020)	(0.237)	(0.024)	(0.002)	
−0.172	0.209	0.507	0.057	−0.009	—	−0.108	0.023	0.011	41.9
(0.180)	(0.023)	(0.023)	(0.009)	(0.009)		(0.241)	(0.026)	(0.002)	
1972–1976 (261 weeks, 2,084 stocks)									
0.437	0.102	0.345	0.027	−0.041	0.171	−0.031	0.031	0.001	38.0
(0.092)	(0.015)	(0.013)	(0.003)	(0.005)	(0.012)	(0.079)	(0.015)	(0.001)	
−0.249	0.111	0.320	0.032	—	0.114	−0.058	−0.007	0.002	36.5
(0.027)	(0.015)	(0.013)	(0.003)		(0.009)	(0.080)	(0.014)	(0.001)	
−0.188	0.141	0.367	0.008	0.008	—	−0.072	0.020	0.003	32.7
(0.085)	(0.015)	(0.014)	(0.003)	(0.004)		(0.082)	(0.015)	(0.001)	
1977–1981 (261 weeks, 2,352 stocks)									
−0.315	−0.059	0.508	0.057	−0.001	0.139	0.015	0.013	0.005	44.2
(0.127)	(0.020)	(0.018)	(0.006)	(0.007)	(0.017)	(0.069)	(0.019)	(0.002)	
−0.344	−0.058	0.508	0.057	—	0.137	0.015	0.011	0.005	44.2
(0.035)	(0.019)	(0.017)	(0.005)		(0.013)	(0.069)	(0.018)	(0.002)	
−0.810	−0.008	0.534	0.040	0.037	—	−0.001	−0.001	0.009	42.6
(0.114)	(0.019)	(0.018)	(0.005)	(0.006)		(0.070)	(0.020)	(0.002)	
1982–1986 (261 weeks, 2,644 stocks)									
−1.385	0.051	0.543	0.062	0.071	0.085	−0.223	0.091	0.006	31.6
(0.180)	(0.025)	(0.027)	(0.007)	(0.010)	(0.023)	(0.081)	(0.031)	(0.001)	
−0.193	0.018	0.583	0.057	—	0.170	−0.182	0.187	0.005	30.4
(0.051)	(0.024)	(0.027)	(0.007)		(0.020)	(0.081)	(0.028)	(0.001)	
−1.602	0.080	0.562	0.048	0.091	—	−0.217	0.085	0.006	31.3
(0.170)	(0.023)	(0.027)	(0.005)	(0.009)		(0.081)	(0.031)	(0.001)	
1987–1991 (261 weeks, 2,471 stocks)									
−1.662	0.155	0.791	0.038	0.078	0.066	−0.138	0.131	0.003	31.9
(0.223)	(0.027)	(0.034)	(0.005)	(0.013)	(0.024)	(0.097)	(0.041)	(0.001)	
−0.313	0.153	0.831	0.035	—	0.158	−0.128	0.252	0.003	30.9
(0.052)	(0.027)	(0.033)	(0.005)		(0.019)	(0.098)	(0.036)	(0.001)	
−1.968	0.171	0.795	0.031	0.100	—	−0.122	0.119	0.003	31.7
(0.195)	(0.026)	(0.034)	(0.005)	(0.010)		(0.097)	(0.041)	(0.001)	
1992–1996 (261 weeks, 2,520 stocks)									
−1.004	−0.087	0.689	0.077	0.040	0.262	−0.644	0.029	0.000	29.6
(0.278)	(0.034)	(0.033)	(0.007)	(0.016)	(0.033)	(0.164)	(0.049)	(0.001)	
−0.310	−0.095	0.708	0.076	—	0.314	−0.641	0.087	−0.001	29.4
(0.061)	(0.034)	(0.032)	(0.007)		(0.026)	(0.164)	(0.043)	(0.001)	
−2.025	−0.025	0.711	0.046	0.115	—	−0.590	−0.005	0.000	27.8
(0.249)	(0.034)	(0.033)	(0.006)	(0.012)		(0.166)	(0.049)	(0.001)	

Note: Table shows CRSP sharecodes 10 and 11, excluding 37 stocks containing Z errors in reported volume, for subperiods of the sample period from July 1962 to December 1996. The explanatory variables are as follows: $\hat{\alpha}_{r,j}$, $\hat{\beta}_{r,j}$, and $\hat{\sigma}_{\epsilon,r,j}$ (the intercept, slope, and residual, respectively, from the time-series regression of an individual security's return on the market return); v_j (natural logarithm of market capitalization); p_j (natural logarithm of price); d_j (dividend yield); SP500$_j$ (S&P 500 indicator variable); and $\hat{\gamma}_{r,j}(1)$ (first-order return autocovariance).

entire sample, the average weekly turnover for the value-weighted and equal-weighted indexes is 0.78 percent and 0.91 percent, respectively. The standard deviation of weekly turnover for these two indexes is 0.48 percent and 0.37 percent, respectively, yielding a coefficient of variation of 0.62 for the value-weighted turnover index and 0.41 for the equal-weighted turnover index. In contrast, the coefficients of variation for the value-weighted and equal-weighted *returns* indexes are 8.52 and 6.91, respectively. Turnover is not nearly so variable as returns, relative to their means.

Table 6.3 also illustrates the nature of the secular trend in turnover through the five-year subperiod statistics. Average weekly value-weighted and equal-weighted turnover is 0.25 percent and 0.57 percent, respectively, in the first subperiod (1962–1966); this grows to 1.25 percent and 1.31 percent, respectively, by the last subperiod (1992–1996). At the beginning of the sample, equal-weighted turnover is three to four times more volatile than value-weighted turnover (0.21 percent vs. 0.07 percent in 1962–1966, and 0.32 percent vs. 0.08 percent in 1967–1971), but by the end of the sample their volatilities are comparable (0.22 percent vs. 0.23 percent in 1992–1996).

The subperiod containing the October 1987 crash exhibits a few anomalous properties: excess skewness and kurtosis for both returns and turnover, average value-weighted turnover slightly higher than average equal-weighted turnover, and slightly higher volatility for value-weighted turnover. These anomalies are consistent with the extreme outliers associated with the 1987 crash (see Figure 6.1).

Table 6.3 also reports the percentiles of the empirical distributions of turnover and returns that document the skewness in turnover that Figure 6.1 hints at, as well as the first ten autocorrelations of turnover and returns and the corresponding Box–Pierce Q-statistics. Unlike returns, turnover is highly persistent, with autocorrelations that start at 91.25 percent and 86.73 percent for the value-weighted and equal-weighted turnover indexes, respectively, decaying very slowly to 84.63 percent and 68.59 percent, respectively, at lag 10. This slow decay suggests some kind of nonstationarity in turnover – perhaps a stochastic trend or *unit root* (see, e.g., Hamilton, 1994). For these reasons, many empirical studies of volume use some form of detrending to induce stationarity. This usually involves either taking first differences or estimating the trend and subtracting it from the raw data. However, Lo and Wang (2000a) show that detrending methods can alter the time-series properties of the data in significant ways, and that without further economic structure for the source of the trend, there is no canonical method for eliminating the trend (see Lo and Wang, 2000a, Table 4, for a more detailed analysis of detrending). Therefore, we shall continue to use raw turnover rather than its first difference or any other detrended turnover series in much of our empirical analysis (the sole exception is the eigenvalue decomposition of the first differences of turnover in Table 6.5). As a way to address the problem of the apparent time trend and other nonstationarities in raw turnover, the empirical analysis of Subsection 4.2 is conducted within five-year subperiods only.

4. CROSS-SECTIONAL CHARACTERISTICS OF VOLUME

The theoretical model of Section 2 leads to sharp predictions about the cross-sectional behavior of trading activity. In this section, we examine the empirical support for these predictions. The empirical results presented here are from Lo and Wang (2000a), in which the starting point was mutual-fund separation, and our dynamic equilibrium model of Section 2 provides the theoretical foundations for such a starting point.

4.1. Theoretical Implications for Volume

The dynamic equilibrium model of Section 2 leads to $(K+1)$-fund separation for the investors' stock investments. In the special case with *iid* stock returns, we have two-fund separation. In the general case with time-varying stock returns, we have four-fund separation. In this case, expected stock returns are driven by a one-dimensional state variable Z_t. However, our model can easily be generalized to allow Z_t to be multidimensional, in which case more than three funds would emerge as the separating stock funds. Thus, in the following discussion, we leave K unspecified, except that it is a small number when compared with the total number of stocks, J.

Let $S_t^k = (S_1^k; \ldots; S_J^k)$, $k = 1, \ldots, K$, denote the K separating stock funds, where the separating funds are expressed in terms of the number of shares of their component stocks. Each investor's stock holdings can be expressed in terms of his or her holdings of the K separating funds:

$$S_t^i = \sum_{k=1}^{K} h_{kt}^i S^k, \quad i = 1, \ldots, I. \tag{4.1}$$

It should be emphasized that, from our theoretical model, the separating stock funds are constant over time, which leads to much simpler behavior in volume. Because in equilibrium, $\sum_{i=1}^{I} S_{i,t} = S^M$ for all t, we have

$$\sum_{k=1}^{K} \left(\sum_{i=1}^{I} h_{kt}^i \right) S^k = S^M.$$

Thus, without loss of generality, we can assume that the market portfolio S^M is one of the separating stock funds, which we label as the first fund. The remaining stock funds are *hedging* portfolios (see Merton, 1973).[15]

[15] In addition, we can assume that all the separating stock funds are mutually orthogonal, that is, $S^{k\prime} S^{k'} = 0$, $k = 1, \ldots, K$, $k' = 1, \ldots, K$, and $k \neq k'$. In particular, $S^{M\prime} S_k = \sum_{j=1}^{J} S_j^k = 0$, and $k = 2, \ldots, K$; hence, the total number of shares in each of the hedging portfolios sum to zero under our normalization. For this particular choice of the separating funds, h_{kt}^i has the simple interpretation that it is the projection coefficient of S_t^i on S^k. Moreover, $\sum_{i=1}^{I} h_{1t}^i = 1$ and $\sum_{i=1}^{I} h_{kt}^i = 0$, $k = 2, \ldots, K$.

From (4.1), investor i's holding in stock j is $S^i_{jt} = \sum_{k=1}^K h^i_{kt} S^k_j$. Therefore, the turnover of stock j at time t is

$$\tau_{jt} = \frac{1}{2}\sum_{i=1}^{I} \left| S^i_{jt} - S^i_{jt-1} \right| = \frac{1}{2}\sum_{i=1}^{I} \left| \sum_{k=1}^{K} \left(h^i_{kt} S^k_j - h^i_{kt-1} S^k_j \right) \right|, \qquad j = 1, \ldots, J. \quad (4.2)$$

To simplify notation, we define $\tilde{h}^i_{kt} \equiv h^i_{kt} - h^i_{kt-1}$ as the change in investor i's holding of fund k from $t-1$ to t.

We now impose the assumption that the amount of trading in the hedging portfolios is small (relative to the trading in the market portfolio) for all investors:

Assumption 4.1. *For $k = 1, \ldots, K$ and $i = 1, \ldots, I$, $|\tilde{h}^i_{1t}| < H < \infty$ and $|\tilde{h}^i_{kt}| \le \lambda H < \infty$ for $1 < k \le K$, where $0 < \lambda \ll 1$, and $\tilde{h}^i_{1t}, \tilde{h}^i_{2t}, \ldots, \tilde{h}^i_{Jt}$ have a continuous joint probability density.*

We then have the following proposition (Lo and Wang, 2000a):

Proposition 4.1. *Under Assumption 4.1, the turnover of stock j at time t can be approximated by*

$$\tau_{jt} \approx \frac{1}{2}\sum_{i=1}^{I} |\tilde{h}^i_{1t}| + \frac{1}{2}\sum_{k=2}^{K} \left[\sum_{i=1}^{I} \operatorname{sgn}\left(\tilde{h}^i_{1t}\right) \tilde{h}^i_{kt} \right] S^k_j, \quad j = 1, \ldots, J, \quad (4.3)$$

and the nth absolute moment of the approximation error is $o(\lambda^n)$.

Now define the following "factors":

$$F_{1t} \equiv \frac{1}{2}\sum_{i=1}^{I} |\tilde{h}^i_{1t}|,$$

$$F_{kt} \equiv \frac{1}{2}\sum_{i=1}^{I} \operatorname{sgn}\left(\tilde{h}^i_{1t}\right) \tilde{h}^i_{kt}, \quad k = 2, \ldots, K.$$

Then the turnover of each stock j can be represented by an approximate K-factor model

$$\tau_{jt} = F_{1t} + \sum_{k=2}^{K} S^k_j F_{kt} + o(\lambda), \quad j = 1, \ldots, J. \quad (4.4)$$

Equation (4.4) summarizes the implication of our model on the cross-sectional behavior of volume, which we now examine empirically.

4.2. The Cross Section of Turnover

To develop a sense for cross-sectional differences in turnover over the sample period, we turn our attention from turnover indexes to the turnover of individual securities. Because turnover is, by definition, an asymmetric measure of trading activity – it cannot be negative – its empirical distribution is naturally skewed. Taking natural logarithms may provide a clearer visual representation of its behavior; hence we plot in Figure 6.2(a) the weekly deciles for the cross section of the logarithm of weekly turnover for each of the 1,800 weeks in the sample period. Figure 6.2(b) simplifies this by plotting the deciles of the cross section of *average* log turnover, averaged within each year.

Figure 6.2(b) shows that the median log turnover, shown by the horizontal bars with vertical sides, has a positive drift over time, but the cross-sectional dispersion is relatively stable. This suggests that the cross-sectional distribution of log turnover is similar over time up to a location parameter, and implies a potentially useful reduced-form description of the cross-sectional distribution of turnover: an identically distributed random variable multiplied by a time-varying scale factor.

An important aspect of individual turnover data that is not immediately obvious from Figure 6.2 is the frequency of turnover outliers among securities and over time. To develop a sense for the magnitude of such outliers, we plot in Figure 6.3 the turnover and returns of a stock in the 1992–1996 subperiod that exhibited large turnover outliers: UnionFed Financial Corporation. Over the course of just a few weeks in the second half of 1993, UnionFed's weekly turnover jumped to a level of 250 percent. This was likely the result of significant news regarding the company's prospects; on June 15, 1993, the *Los Angeles Times* reported the fact that UnionFed, a California savings and loan, "has become 'critically undercapitalized' and subject to being seized within 90 days by federal regulators." In such cases, turnover outliers are not surprising, yet it is interesting that the returns of UnionFed during the same period do not exhibit the same extreme behavior.

Figure 6.4 displays the time series of turnover and return for four randomly selected stocks during the same 1992–1996 subperiod, and although the outliers are considerably smaller here, nevertheless there are still some rather extreme turnover values in their time series. For example, for most of 1996, Culligan Water Technologies Inc. exhibited weekly turnover in the 2–3 percent range, but toward the end of the year, there is one week in which the turnover jumped to 13 percent. Other similar patterns in Figure 6.4 seem to suggest that, for many stocks, there are short bursts of intense trading activity lasting only a few weeks, but with turnover far in excess of the "typical" values during the rest of the time. This characteristic of individual turnover is prevalent in the entire database, and must be taken into account in any empirical analysis of trading activity.

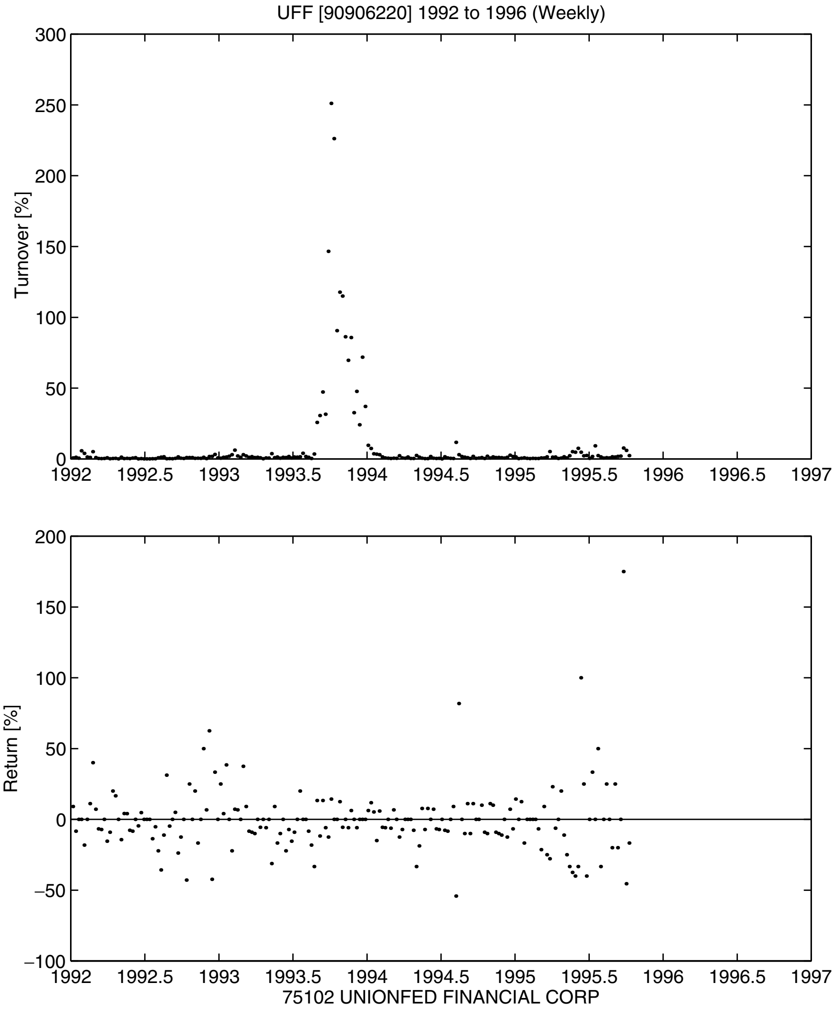


Figure 6.3. Turnover outliers and returns for UnionFed Financial Corp. in the 1992–1996 subperiod.

4.2.1. Cross-Sectional Regressions

The volume implications of our theoretical model provide a natural direction for empirical analysis: look for linear factor structure in the turnover cross section. If two-fund separation holds, turnover should be identical across all stocks, that is, a one-factor linear model where all stocks have identical factor loadings. If $(K + 1)$-fund separation holds, turnover should satisfy a K-factor linear model. We examine these hypotheses in this section.

It is clear from Figure 6.2 that turnover varies considerably in the cross section; hence two-fund separation may be rejected out of hand. However, the

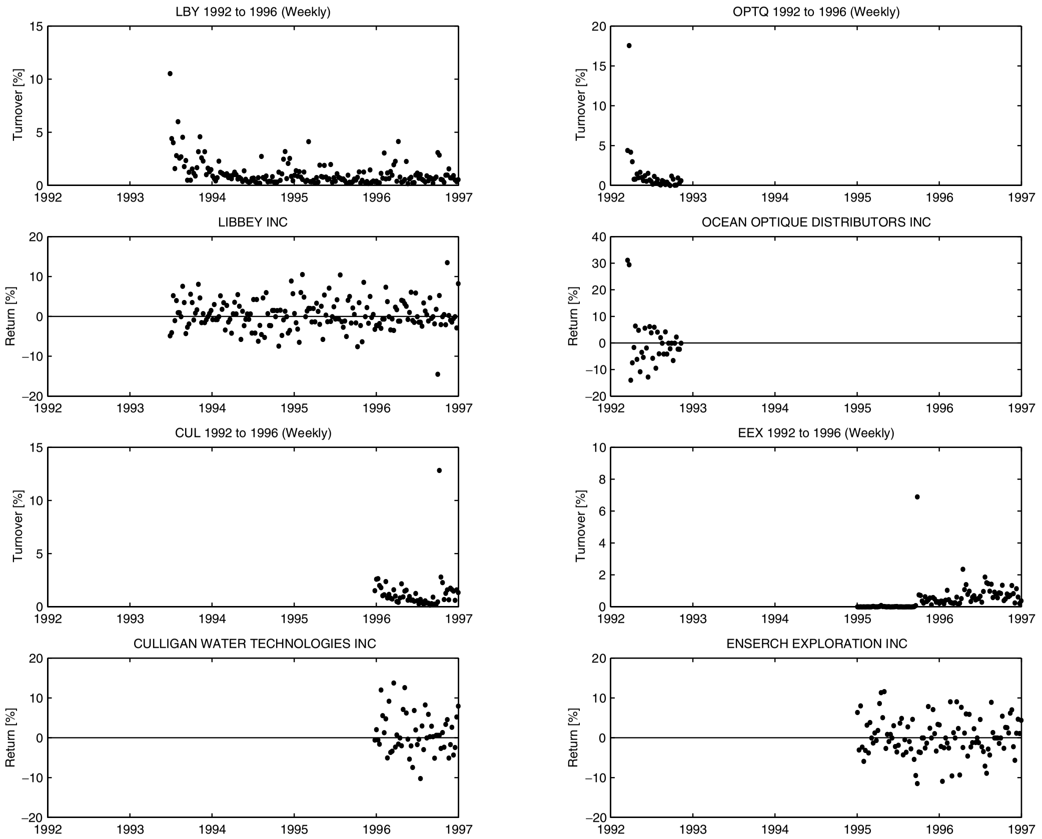


Figure 6.4. Turnover and returns of four randomly selected stocks, 1992–1996.

turnover implications of two-fund separation might be *approximately* correct in the sense that the cross-sectional variation in turnover may be “idiosyncratic” white noise, for example, cross-sectionally uncorrelated and without common factors. We shall test this and the more general $(K + 1)$-fund separation hypothesis, but, before doing so, we first consider a less formal, more exploratory analysis of the cross-sectional variation in turnover. In particular, we wish to examine the explanatory power of several economically motivated variables such as expected return, volatility, and trading costs in explaining the cross section of turnover.

To do this, we estimate cross-sectional regressions over five-year subperiods in which the dependent variable is the median turnover $\tilde{\tau}_j$ of stock j. We use median turnover instead of mean turnover to minimize the influence of outliers, which can be substantial in this data set (see Figures 6.3 and 6.4 and the corresponding discussion mentioned previously).[16] The explanatory variables are the following stock-specific characteristics:[17]

$\hat{\alpha}_{r,j}$, Intercept coefficient from the time-series regression of stock j’s return on the value-weighted market return;

$\hat{\beta}_{r,j}$, Slope coefficient from the time-series regression of stock j’s return on the value-weighted market return;

$\hat{\sigma}_{\epsilon,r,j}$, Residual standard deviation of the time-series regression of stock j’s return on the value-weighted market return;

v_j, Average of natural logarithm of stock j’s market capitalization;

p_j, Average of natural logarithm of stock j’s price;

d_j, Average of dividend yield of stock j; where dividend yield in week t is defined by $d_{jt} = \max\{0,\ \log[(1 + R_{jt})V_{jt-1}/V_{jt}]\}$ and V_{jt} is j’s market capitalization in week t;

SP500_j, Indicator variable for membership in the S&P 500 Index; and

$\hat{\gamma}_{r,j}(1)$, First-order autocovariance of returns.

The inclusion of these regressors in our cross-sectional analysis is loosely motivated by various intuitive “theories” that have appeared in the volume literature.

The motivation for the first three regressors comes partly from linear asset-pricing models such as the CAPM and asset-pricing theory; they capture excess expected return ($\hat{\alpha}_{r,j}$), systematic risk ($\hat{\beta}_{r,j}$), and residual risk ($\hat{\sigma}_{\epsilon,r,j}$), respectively. To the extent that expected excess return ($\hat{\alpha}_{r,j}$) may contain a premium associated with liquidity (see, e.g., Amihud and Mendelson, 1986a, 1986b; and Hu, 1997) and heterogeneous information (see, e.g., He and Wang, 1995 and Wang, 1994), it should also give rise to cross-sectional differences in turnover.

[16] Also, within each five-year period, we exclude all stocks that are missing turnover data for more than two-thirds of the subsample.

[17] We use median turnover instead of mean turnover to minimize the influence of outliers, which can be substantial in this data set (see Figures 6.3 and 6.4 and the corresponding discussion). Also, within each five-year period, we exclude all stocks that are missing turnover data for more than two-thirds of the subsample.

Although a higher premium from lower liquidity should be inversely related to turnover, a higher premium from heterogeneous information can lead to either higher or lower turnover, depending on the nature of information heterogeneity. The two risk measures of an asset, $\hat{\beta}_{r,j}$ and $\hat{\sigma}_{\epsilon,r,j}$, also measure the volatility in its returns that is associated with systematic risk and residual risk, respectively. Given that realized returns often generate portfolio-rebalancing needs, the volatility of returns should be positively related to turnover.

The motivation for log market capitalization (v_j) and log price (p_t) is twofold. On the theoretical side, the role of market capitalization in explaining volume is related to Merton's (1987) model of capital market equilibrium in which investors hold only the assets they are familiar with. This implies that larger-capitalization companies tend to have more diverse ownership, which can lead to more active trading. The motivation for log price is related to trading costs. Given that part of trading costs comes from the bid–ask spread, which takes on discrete values in dollar terms, the actual costs in percentage terms are inversely related to price levels. This suggests that volume should be positively related to prices.

On the empirical side, there is an extensive literature documenting the significance of log market capitalization and log price in explaining the cross-sectional variation of expected returns. See, for example, Banz (1981), Black (1976), Brown, Van Harlow, and Tinic (1993), Marsh and Merton (1987), and Reinganum (1992). If size and price are genuine factors driving expected returns, they should drive turnover as well (see Lo and Wang, 2000b, for a more formal derivation and empirical analysis of this intuition).

Dividend yield (d_j) is motivated by its (empirical) ties to expected returns, but also by *dividend-capture* trades – the practice of purchasing stock just before its exdividend date and then selling it shortly thereafter.[18] Often induced by differential taxation of dividends vs. capital gains, dividend-capture trading has been linked to short-term increases in trading activity. See, for example, Karpoff and Walking (1988, 1990), Lakonishok and Smidt (1986), Lakonishok and Vermaelen (1986), Lynch-Koski (1996), Michaely (1991), Michaely and Murgia (1995), Michaely and Vila, (1995, 1996), Michaely, Vila, and Wang (1996), and Stickel (1991). Stocks with higher dividend yields should induce more dividend-capture trading activity, and this may be reflected in higher median turnover.

The effects of membership in the Standard & Poors (S&P) 500 have been documented in many studies, for example, those by Dhillon and Johnson (1991), Goetzmann and Garry (1986), Harris and Gurel (1986), Jacques (1988), Jain (1987), Lamoureux and Wansley (1987), Pruitt and Wei (1989), Shleifer (1986), Tkac (1996), and Woolridge and Ghosh (1986). In particular, Harris and Gurel

[18] Our definition of d_j is meant to capture net corporate distributions or outflows (recall that returns R_{jt} are inclusive of all dividends and other distributions). The purpose of the nonnegativity restriction is to ensure that inflows, for example, new equity issues, are not treated as negative dividends.

(1986) document increases in volume just after inclusion in the S&P 500, and Tkac (1996) uses an S&P 500 indicator variable to explain the cross-sectional dispersion of relative turnover (relative dollar volume divided by relative market capitalization). The obvious motivation for this variable is the growth of indexation by institutional investors, and by the related practice of *index arbitrage*, in which disparities between the index futures price and the spot prices of the component securities are exploited by taking the appropriate positions in the futures and spot markets. For these reasons, stocks in the S&P 500 index should have higher turnover than others. Indexation began its rise in popularity with the advent of the mutual-fund industry in the early 1980s, and index arbitrage first became feasible in 1982 with the introduction of the Chicago Mercantile Exchange's S&P 500 futures contracts. Therefore, the effects of S&P 500 membership on turnover should be more dramatic in the later subperiods. Another motivation for S&P 500 membership is its effect on the publicity of member companies, which leads to more diverse ownership and more trading activity in the context of Merton (1987).

The last variable, the first-order return autocovariance, $\hat{\gamma}_{r,j}(1)$, serves as a proxy for trading costs, as in Roll's (1984) model of the "effective" bid–ask spread. In that model, Roll shows that, in the absence of information-based trades, prices' bouncing between bid and ask prices implies the following approximate relation between the spread and the first-order return autocovariance:

$$\frac{s_{r,j}^2}{4} \approx -\text{cov}[R_{jt}, R_{jt-1}] \equiv -\gamma_{r,j}(1), \tag{4.5}$$

where $s_{r,j} \equiv s_j/\sqrt{P_{aj}P_{bj}}$ is the percentage effective bid–ask spread of stock j as a percentage of the geometric average of the bid and ask prices P_{bj} and P_{aj}, respectively, and s_j is the dollar bid–ask spread.

Rather than solve for $s_{r,j}$, we choose instead to include $\hat{\gamma}_{r,j}(1)$ as a regressor to sidestep the problem of a positive sample first-order autocovariance, which yields a complex number for the effective bid–ask spread. Of course, using $\hat{\gamma}_{r,j}(1)$ does not eliminate this problem, which is a symptom of a specification error, but rather is a convenient heuristic that allows us to estimate the regression equation. (Complex observations for even one regressor can yield complex parameter estimates for all the other regressors as well!) This heuristic is not unlike Roll's method for dealing with positive autocovariances; however, it is more direct.[19]

Under the trading-cost interpretation for $\hat{\gamma}_{r,j}(1)$, we should expect a positive coefficient in our cross-sectional turnover regression; a large negative value for $\hat{\gamma}_{r,j}(1)$ implies a large bid–ask spread, which should be associated with lower turnover. Alternatively, Roll (1984) interprets a positive value for $\hat{\gamma}_{r,j}(1)$ as a negative bid–ask spread, and thus turnover should be higher for such stocks.

[19] In a parenthetical statement in footnote *a* of Table I, Roll (1984) writes "the sign of the covariance was preserved after taking the square root."

These eight regressors yield the following regression equation to be estimated:

$$\tilde{\tau}_j = \gamma_0 + \gamma_1\hat{\alpha}_{r,j} + \gamma_2\hat{\beta}_{r,j} + \gamma_3\hat{\sigma}_{\epsilon,r,j} + \gamma_4 v_j + \gamma_5 p_j \\ + \gamma_6 d_j + \gamma_7 \mathrm{SP500}_j + \gamma_8\hat{\gamma}_{r,j}(1) + \epsilon_j. \tag{4.6}$$

Table 6.3 contains the estimates of the cross-sectional regression model (4.6). We estimated three regression models for each subperiod: one with all eight variables and a constant term included, one excluding log market capitalization, and one excluding log price. Because the log price and log market capitalization regressors are so highly correlated (see Lim et al., 1998), regressions with only one or the other included were estimated to gauge the effects of multicollinearity. The exclusion of either variable does not affect the qualitative features of the regression – no significant coefficients changed sign other than the constant term – though the quantitative features were affected to a small degree. For example, in the first subperiod, v_j has a negative coefficient (-0.064) and p_j has a positive coefficient (0.150), both significant at the 5 percent level. When v_j is omitted, the coefficient of p_j is still positive but smaller (0.070), and when p_j is omitted, the coefficient of v_j is still negative and also smaller in absolute magnitude (-0.028); in both of these cases, the coefficients retain their significance.

The fact that size has a negative impact on turnover while price has a positive impact is an artifact of the earlier subperiods. This can be seen heuristically in the time-series plots of Figure 6.1; compare the value-weighted and equal-weighted turnover indexes during the first two or three subperiods. Smaller-capitalization stocks seem to have higher turnover than larger-capitalization stocks.

This begins to change in the 1977–1981 subperiod: The size coefficient is negative but not significant, and when price is excluded, the size coefficient changes sign and becomes significant. In the subperiods after 1977–1981, both size and price enter positively. One explanation of this change is the growth of the mutual-fund industry and other large institutional investors in the early 1980s. As portfolio managers manage larger asset bases, it becomes more difficult to invest in smaller-capitalization companies because of liquidity and corporate-control issues. Therefore, the natural economies of scale in investment management coupled with the increasing concentration of investment capital make small stocks less actively traded than large stocks. Of course, this effect should have implications for the equilibrium return of small stocks vs. large stocks, and we investigate such implications in ongoing research.

The first-order return autocovariance has a positive coefficient in all subperiods except the second regression of the last subperiod (in which the coefficient is negative but insignificant), and these coefficients are significant at the 5 percent level in all subperiods except 1972–1976 and 1992–1996. This is consistent with the trading-cost interpretation of $\hat{\gamma}_{r,j}(1)$: A large negative return autocovariance implies a large effective bid–ask spread, which, in turn, should imply lower turnover.

Membership in the S&P 500 also has a positive impact on turnover in all subperiods as expected, and the magnitude of the coefficient increases dramatically in the 1982–1986 subperiod – from 0.013 in the previous period to 0.091 – also as expected given the growing importance of indexation and index arbitrage during this period, and the introduction of S&P 500 futures contracts in April of 1982. Surprisingly, in the 1992–1996 subperiod, the S&P 500 coefficient declines to 0.029, perhaps because of the interactions between this indicator variable and size and price (all three variables are highly positively correlated with each other; see Lim et al., 1998, for further details). When size is omitted, S&P 500 membership becomes more important, yet when price is omitted, size becomes more important and S&P 500 membership becomes irrelevant. These findings are roughly consistent with those in Tkac (1996).[20]

Both systematic and idiosyncratic risk, $\hat{\beta}_{r,j}$ and $\hat{\sigma}_{\epsilon,r,j}$, have a positive and significant impact on turnover in all subperiods. However, the impact of excess expected returns $\hat{\alpha}_{r,j}$ on turnover is erratic: negative and significant in the 1977–1981 and 1992–1996 subperiods, and positive and significant in the others.

The dividend-yield regressor is insignificant in all subperiods but two: 1982–1986 and 1992–1996. In these two subperiods, the coefficient is negative, which contradicts the notion that dividend-capture trading affects turnover.

In summary, the cross-sectional variation of turnover does seem related to several stock-specific characteristics such as risk, size, price, trading costs, and S&P 500 membership. The explanatory power of these cross-sectional regressions, as measured by R^2, ranges from 29.6 percent (1992–1996) to 44.7 percent (1967–1971), rivaling the R^2s of typical cross-sectional return regressions. With sample sizes ranging from 2,073 (1962–1966) to 2,644 (1982–1986) stocks, these R^2s provide some measure of confidence that cross-sectional variations in median turnover are not purely random but do bear some relation to economic factors.

4.2.2. *Tests of* $(K+1)$*-Fund Separation*

Because two-fund and $(K+1)$-fund separation imply an approximately linear factor structure for turnover, we can investigate these two possibilities by using principal components analysis to decompose the covariance matrix of turnover (see Muirhead, 1982, for an exposition of principal components analysis). If turnover is driven by a linear K-factor model, the first K principal components should explain most of the time-series variation in turnover. More formally, if

$$\tau_{jt} = \alpha_j + \delta_1 F_{1t} + \cdots + \delta_K F_{Kt} + \epsilon_{jt}, \tag{4.7}$$

[20] In particular, she finds that S&P 500 membership becomes much less significant after controlling for the effects of size and institutional ownership. Of course, her analysis is not directly comparable with ours because she uses a different dependent variable (monthly relative dollar volume divided by relative market capitalization) in her cross-sectional regressions, and she considers only a small sample of the very largest NYSE/AMEX stocks (809) over the four-year period from 1988 to 1991.

where $E[\epsilon_{jt}\epsilon_{j't}] = 0$ for any $j \neq j'$, then the covariance matrix Σ of the vector $\tau_t \equiv (\tau_{1t}; \ldots; \tau_{Jt})$ can be expressed as

$$\text{var}[\tau_t] \equiv \Sigma = H\Theta H', \tag{4.8}$$

$$\Theta = \begin{bmatrix} \theta_1 & 0 & \cdots & 0 \\ 0 & \theta_2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & 0 & \theta_N \end{bmatrix}, \tag{4.9}$$

where Θ contains the eigenvalues of Σ along its diagonal, and H is the matrix of corresponding eigenvectors. Because Σ is a covariance matrix, it is positive semidefinite and thus all the eigenvalues are nonnegative. When normalized to sum to one, each eigenvalue can be interpreted as the fraction of the total variance of turnover attributable to the corresponding principal component. If (4.7) holds, it can be shown that as the size N of the cross section increases without bound, exactly K normalized eigenvalues of Σ approach positive finite limits, and the remaining $N - K$ eigenvalues approach zero (see, e.g., Chamberlain, 1983 and Chamberlain and Rothschild, 1983). Therefore, the plausibility of (4.7), and the value of K, can be gauged by examining the magnitudes of the eigenvalues of Σ.

The only obstacle is the fact that the covariance matrix Σ must be estimated; hence we encounter the well-known problem that the standard estimator

$$\widehat{\Sigma} \equiv \frac{1}{T}\sum_{t=1}^{T}(\tau_t - \bar{\tau})(\tau_t - \bar{\tau})'$$

is singular if the number of securities J in the cross section is larger than the number of time-series observations T.[21] Because J is typically much larger than T – for a five-year subperiod T is generally 261 weeks, and J is typically well over 2,000 – we must limit our attention to a smaller subset of stocks. We do this by following the common practice of forming a small number of portfolios (see Campbell, Lo, and MacKinlay, 1997, Chapter 5), sorted by turnover beta to maximize the dispersion of turnover beta among the portfolios.[22] In particular,

[21] Singularity by itself does not pose any problems for the computation of eigenvalues – this follows from the singular-value decomposition theorem – but it does have implications for the statistical properties of estimated eigenvalues. In some preliminary Monte Carlo experiments, we have found that the eigenvalues of a singular estimator of a positive-definite covariance matrix can be severely biased. We thank Bob Korajczyk and Bruce Lehmann for bringing some of these issues to our attention and plan to investigate them more thoroughly in ongoing research.

[22] Our desire to maximize the dispersion of turnover beta is motivated by the same logic used in Black, Jensen, and Scholes (1972): A more dispersed sample provides a more powerful test of a cross-sectional relationship driven by the sorting characteristic. This motivation should not be taken literally in our context because the theoretical implications of Section 2 need not imply a prominent role for turnover beta (indeed, in the case of two-fund separation, there is no cross-sectional variation in turnover betas). However, given the factor structure implied by $(K + 1)$-fund separation (see Subsection 4.1), sorting by turnover betas seems appropriate.

within each five-year subperiod, we form ten turnover-beta-sorted portfolios by using betas estimated from the previous five-year subperiod, estimate the covariance matrix $\widehat{\Sigma}$ by using 261 time-series observations, and perform a principal-components decomposition on $\widehat{\Sigma}$. For purposes of comparison and interpretation, we perform a parallel analysis for returns, using ten return-beta-sorted portfolios. The results are reported in Table 6.4.

Table 6.4 contains the principal-components decomposition for portfolios sorted on out-of-sample betas, where the betas are estimated in two ways: relative to value-weighted indexes (τ^{VW} and R^{VW}) and equal-weighted indexes (τ^{EW} and R^{EW}).[23] The first principal component typically explains between 70 percent and 85 percent of the variation in turnover, and the first two principal components explain almost all of the variation. For example, the upper-left subpanel of Table 6.4 shows that in the second five-year subperiod (1967–1971), 85.1 percent of the variation in the turnover of turnover-beta-sorted portfolios (using turnover betas relative to the value-weighted turnover index) is captured by the first principal component, and 93.6 percent is captured by the first two principal components. Although using betas computed with value-weighted instead of equal-weighted indexes generally yields smaller eigenvalues for the first principal component (and therefore larger values for the remaining principal components) for both turnover and returns, the differences are typically not large.

The importance of the second principal component grows steadily through time for the value-weighted case, reaching a peak of 15.6 percent in the last subperiod, and the first two principal components account for 87.3 percent of the variation in turnover in the last subperiod. This is roughly comparable with the return portfolios sorted on value-weighted return betas; the first principal component is by far the most important, and the importance of the second principal component is most pronounced in the last subperiod. However, the lower-left subpanel of Table 6.4 shows that for turnover portfolios sorted by betas computed against equal-weighted indexes, the second principal component explains approximately the same variation in turnover, varying between 6.0 percent and 10.4 percent across the six subperiods.

Of course, one possible explanation for the dominance of the first principal component is the existence of a time trend in turnover. Despite the fact that we have limited our analysis to five-year subperiods, within each subperiod there is a certain drift in turnover; might this account for the first principal component? To investigate this conjecture, we perform eigenvalue decompositions for the covariance matrices of the *first differences* of turnover for the ten turnover portfolios.

These results are reported in Table 6.5 and are consistent with those in Table 6.4: the first principal component is still the most important, explaining

[23] In particular, the portfolios in a given period are formed by ranking on betas estimated in the immediately preceding subperiod; for example, the 1992–1996 portfolios were created by sorting on betas estimated in the 1987–1991 subperiod. Hence the first subperiod in Table 6.4 begins in 1967, not 1962.

Table 6.4. *Eigenvalues $\hat{\theta}_i$, $i = 1, \ldots, 10$, of the covariance matrix of ten out-of-sample beta-sorted portfolios of weekly turnover and returns of NYSE and AMEX ordinary common shares*

θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8	θ_9	θ_{10}	Period	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8	θ_9	θ_{10}
Turnover-Beta-Sorted Turnover Portfolios (τ^{VW})											Return-Beta-Sorted Return Portfolios (R^{VW})									
85.1	8.5	3.6	1.4	0.8	0.3	0.2	0.1	0.0	0.0	1967–1971	85.7	5.9	2.0	1.4	1.4	1.1	0.8	0.7	0.5	0.4
(7.5)	(0.7)	(0.3)	(0.1)	(0.1)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)		(7.5)	(0.5)	(0.2)	(0.1)	(0.1)	(0.1)	(0.1)	(0.1)	(0.0)	(0.0)
82.8	7.3	4.9	2.0	1.4	0.8	0.5	0.2	0.1	0.1	1972–1976	90.0	3.8	1.8	1.0	0.9	0.7	0.6	0.6	0.4	0.3
(7.3)	(0.6)	(0.4)	(0.2)	(0.1)	(0.1)	(0.0)	(0.0)	(0.0)	(0.0)		(7.9)	(0.3)	(0.2)	(0.1)	(0.1)	(0.1)	(0.1)	(0.0)	(0.0)	(0.0)
83.6	8.6	2.3	2.0	1.2	0.8	0.6	0.4	0.4	0.1	1977–1981	85.4	4.8	4.3	1.4	1.3	0.9	0.6	0.5	0.4	0.3
(7.3)	(0.8)	(0.2)	(0.2)	(0.1)	(0.1)	(0.1)	(0.0)	(0.0)	(0.0)		(7.5)	(0.4)	(0.4)	(0.1)	(0.1)	(0.1)	(0.1)	(0.0)	(0.0)	(0.0)
78.9	7.9	3.6	2.9	2.4	1.4	1.3	0.8	0.5	0.4	1982–1986	86.6	6.1	2.4	1.6	1.0	0.6	0.5	0.5	0.4	0.3
(6.9)	(0.7)	(0.3)	(0.3)	(0.2)	(0.1)	(0.1)	(0.1)	(0.0)	(0.0)		(7.6)	(0.5)	(0.2)	(0.1)	(0.1)	(0.1)	(0.0)	(0.0)	(0.0)	(0.0)
80.1	6.2	5.2	2.4	1.6	1.3	1.0	1.0	0.8	0.5	1987–1991	91.6	2.9	1.7	1.1	0.7	0.6	0.6	0.4	0.3	0.2
(7.0)	(0.5)	(0.5)	(0.2)	(0.1)	(0.1)	(0.1)	(0.1)	(0.1)	(0.0)		(8.0)	(0.3)	(0.1)	(0.1)	(0.1)	(0.1)	(0.0)	(0.0)	(0.0)	(0.0)
71.7	15.6	4.5	2.9	1.8	1.2	0.9	0.8	0.5	0.3	1992–1996	72.4	11.6	4.4	3.5	2.2	1.8	1.5	1.1	0.8	0.6
(6.3)	(1.4)	(0.4)	(0.3)	(0.2)	(0.1)	(0.1)	(0.1)	(0.0)	(0.0)		(6.3)	(1.0)	(0.4)	(0.3)	(0.2)	(0.2)	(0.1)	(0.1)	(0.1)	(0.1)
Turnover-Beta-Sorted Turnover Portfolios (τ^{EW})											Return-Beta-Sorted Return Portfolios (R^{EW})									
86.8	7.5	3.0	1.3	0.6	0.5	0.2	0.1	0.1	0.0	1967–1971	87.8	4.3	2.2	1.5	1.0	0.9	0.8	0.5	0.5	0.5
(7.6)	(0.7)	(0.3)	(0.1)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)		(7.7)	(0.4)	(0.2)	(0.1)	(0.1)	(0.1)	(0.1)	(0.0)	(0.0)	(0.0)
82.8	6.0	5.4	2.9	1.2	1.0	0.4	0.2	0.1	0.0	1972–1976	91.6	4.1	0.9	0.8	0.6	0.5	0.4	0.4	0.3	0.3
(7.3)	(0.5)	(0.5)	(0.3)	(0.1)	(0.1)	(0.0)	(0.0)	(0.0)	(0.0)		(8.0)	(0.4)	(0.1)	(0.1)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)
79.1	8.5	5.4	2.8	1.4	1.0	0.7	0.6	0.3		1977–1981	91.5	3.9	1.4	0.8	0.6	0.5	0.4	0.3	0.3	0.3
(6.9)	(0.7)	(0.5)	(0.2)	(0.1)	(0.1)	(0.1)	(0.0)	(0.0)	(0.0)		(8.0)	(0.3)	(0.1)	(0.1)	(0.1)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)
78.0	10.4	3.1	2.3	2.0	1.3	1.3	0.8	0.6	0.4	1982–1986	88.9	4.4	2.3	1.3	0.7	0.7	0.6	0.5	0.4	0.4
(6.8)	(0.9)	(0.3)	(0.2)	(0.2)	(0.1)	(0.1)	(0.1)	(0.1)	(0.0)		(7.8)	(0.4)	(0.2)	(0.1)	(0.1)	(0.1)	(0.0)	(0.0)	(0.0)	(0.0)
82.5	4.8	3.2	2.4	2.0	1.4	1.3	0.9	0.9	0.6	1987–1991	92.7	3.0	1.2	0.7	0.7	0.4	0.4	0.4	0.3	0.2
(7.2)	(0.4)	(0.3)	(0.2)	(0.2)	(0.1)	(0.1)	(0.1)	(0.1)	(0.1)		(8.1)	(0.3)	(0.1)	(0.1)	(0.1)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)
79.0	8.5	4.9	2.6	1.5	1.1	0.9	0.6	0.5	0.4	1992–1996	76.8	10.4	3.9	2.7	1.9	1.1	1.0	0.9	0.7	0.6
(6.9)	(0.7)	(0.4)	(0.2)	(0.1)	(0.1)	(0.1)	(0.1)	(0.0)	(0.0)		(6.7)	(0.9)	(0.3)	(0.2)	(0.2)	(0.1)	(0.1)	(0.1)	(0.1)	(0.1)

Note: Table shows CRSP sharecodes 10 and 11, excluding 37 stocks containing Z errors in reported volume – in percentages (where the eigenvalues are normalized to sum to 100 percent) – for subperiods of the sample period from July 1962 to December 1996. Turnover portfolios are sorted by out-of-sample turnover betas and return portfolios are sorted by out-of-sample return betas, where τ^{VW} and R^{VW} indicate that the betas are computed relative to value-weighted indexes, and τ^{EW} and R^{EW} indicate that they are computed relative to equal-weighted indexes. Standard errors for the normalized eigenvalues are given in parentheses and are calculated under the assumption of *iid* normality.

Table 6.5. *Eigenvalues* $\hat{\theta}_i$, $i = 1, \ldots, 10$*, of the covariance matrix of the first-differences of the weekly turnover of ten out-of-sample beta-sorted portfolios of NYSE and AMEX ordinary common shares*

Period	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$	$\hat{\theta}_5$	$\hat{\theta}_6$	$\hat{\theta}_7$	$\hat{\theta}_8$	$\hat{\theta}_9$	$\hat{\theta}_{10}$
Out-of-Sample Turnover-Beta-Sorted Turnover-Differences Portfolios (τ^{VW})										
1967–1971	82.6	7.1	5.1	2.0	1.6	0.8	0.5	0.1	0.1	0.1
	(7.2)	(0.6)	(0.5)	(0.2)	(0.1)	(0.1)	(0.0)	(0.0)	(0.0)	(0.0)
1972–1976	81.2	6.8	4.7	2.8	2.0	1.0	0.9	0.4	0.2	0.1
	(7.1)	(0.6)	(0.4)	(0.2)	(0.2)	(0.1)	(0.1)	(0.0)	(0.0)	(0.0)
1977–1981	85.2	4.5	2.9	2.6	1.6	1.2	0.8	0.5	0.5	0.2
	(7.5)	(0.4)	(0.3)	(0.2)	(0.1)	(0.1)	(0.1)	(0.0)	(0.0)	(0.0)
1982–1986	81.3	5.1	3.5	2.7	2.2	1.7	1.3	0.9	0.7	0.6
	(7.1)	(0.4)	(0.3)	(0.2)	(0.2)	(0.2)	(0.1)	(0.1)	(0.1)	(0.1)
1987–1991	73.1	10.9	4.1	3.0	2.2	1.7	1.6	1.4	1.1	0.9
	(6.4)	(1.0)	(0.4)	(0.3)	(0.2)	(0.2)	(0.1)	(0.1)	(0.1)	(0.1)
1992–1996	78.4	8.6	4.0	2.8	2.1	1.2	1.0	0.9	0.6	0.4
	(6.9)	(0.8)	(0.4)	(0.2)	(0.2)	(0.1)	(0.1)	(0.1)	(0.0)	(0.0)
Out-of-Sample Turnover-Beta-Sorted Turnover-Differences Portfolios (τ^{EW})										
1967–1971	82.2	8.0	4.5	2.3	1.4	0.7	0.4	0.3	0.1	0.0
	(7.2)	(0.7)	(0.4)	(0.2)	(0.1)	(0.1)	(0.0)	(0.0)	(0.0)	(0.0)
1972–1976	79.3	7.5	4.8	4.0	1.9	1.3	0.6	0.4	0.2	0.1
	(7.0)	(0.7)	(0.4)	(0.4)	(0.2)	(0.1)	(0.1)	(0.0)	(0.0)	(0.0)
1977–1981	80.3	5.3	4.8	3.8	2.0	1.4	1.2	0.7	0.5	0.2
	(7.0)	(0.5)	(0.4)	(0.3)	(0.2)	(0.1)	(0.1)	(0.1)	(0.0)	(0.0)
1982–1986	82.6	5.0	3.0	2.6	2.0	1.7	1.1	0.9	0.7	0.4
	(7.3)	(0.4)	(0.3)	(0.2)	(0.2)	(0.1)	(0.1)	(0.1)	(0.1)	(0.0)
1987–1991	77.2	5.5	4.3	2.7	2.5	2.3	1.8	1.6	1.2	1.0
	(6.8)	(0.5)	(0.4)	(0.2)	(0.2)	(0.2)	(0.2)	(0.1)	(0.1)	(0.1)
1992–1996	80.4	6.4	4.6	2.6	1.7	1.4	1.1	0.7	0.5	0.4
	(7.1)	(0.6)	(0.4)	(0.2)	(0.1)	(0.1)	(0.1)	(0.1)	(0.0)	(0.0)

Note: Table shows CRSP sharecodes 10 and 11, excluding 37 stocks containing Z errors in reported volume – in percentages (where the eigenvalues are normalized to sum to 100 percent – for subperiods of the sample period from July 1962 to December 1996. Turnover betas are calculated in two ways: with respect to a value-weighted turnover index (τ^{VW}) and an equal-weighted turnover index (τ^{EW}). Standard errors for the normalized eigenvalues are given in parentheses and are calculated under the assumption of *iid* normality.

between 60 percent and 88 percent of the variation in the first differences of turnover. The second principal component is typically responsible for another 5 percent to 20 percent. And in one case – in-sample sorting on betas relative to the equal-weighted index during 1987–1991 – the third principal component accounts for an additional 10 percent. These figures suggest that the trend in turnover is unlikely to be the source of the dominant first principal component.

In summary, the results of Tables 6.4 and 6.5 indicate that a one-factor model for turnover is a reasonable approximation, at least in the case of turnover-beta-sorted portfolios, and that a two-factor model captures well over 90 percent of the time-series variation in turnover. This lends some support to the practice of estimating "abnormal" volume by using an event-study style "market model," such as that used by Bamber (1986), Jain and Joh (1988), Lakonishok and Smidt (1986), Morse (1980), Richardson, Sefcik, and Thompson (1986), Stickel and Verrecchia (1994), and Tkac (1996).

As compelling as these empirical results seem to be, several qualifications should be kept in mind. First, we have provided little statistical inference for our principal-components decomposition. In particular, the asymptotic standard errors reported in Tables 6.4 and 6.5 were computed under the assumption of *iid* Gaussian data, hardly appropriate for weekly U.S. stock returns and even less convincing for turnover (see Muirhead, 1982, Chapter 9 for further details). In particular, Monte Carlo simulations should be conducted to check the robustness of these standard errors under a variety of data-generating processes.

5. DYNAMIC VOLUME-RETURN RELATION

In this section, we examine the empirical support for the implications of our model for the joint behavior of returns and volume. We first use a simple version of the model to derive its specific implications for a dynamic volume-return relation. We then present the empirical evidence based on the work of Campbell, Grossman, and Wang (CGW, 1993) and Llorente, Michaely, Saar, and Wang (LMSW, 2001).

5.1. Theoretical Implications for a Volume-Return Relation

Our model leads to a set of predictions about the joint behavior of volume and returns, especially how they are related intertemporally. A specific relation we want to examine is how current volume and returns can forecast future returns. In the context of our model, this relation can be formally expressed by the following conditional expectation: $E\left[\tilde{Q}_{t+1} | \tilde{Q}_t, \tau_t\right]$, where $\tilde{Q}_t \equiv Q_t - Q_{t-1}$ is the (excess dollar) return on the stocks from $t-1$ to t and τ_t is the vector of turnover at t.

To simplify our analysis, we consider the special case of a single stock ($J = 1$) and two investors ($I = 2$).[24] Furthermore, we let $X_t^1 = Z_t = -X_t^2$, $Y_t^1 = Y_t^2 = 0$. In this case, the return on the stock and the investors' stock holdings can be expressed as follows:

$$\tilde{Q}_{t+1} = [ra + (r + \alpha_Z)bZ_t] + \epsilon_{Qt+1}, \tag{5.1a}$$

$$S_t^i = \left(\frac{1}{2} - X_t^i\right) \quad (i = 1, 2), \tag{5.1b}$$

where ϵ_{Qt+1} is normally distributed and uncorrelated with Z_t.[25] The turnover

[24] In our model, even with multiple stocks, when $\sigma_D' \sigma_Z = 0$ and σ_Z^2 is small, $S^I \approx S^M$, and we obtain approximate two-fund separation. Effectively, the model reduces to the one-stock case we consider here, where the market portfolio plays the role of the single stock.

[25] From Theorem 2.1, $(Q_t; Z_t)$ is a Gaussian process. Furthermore, $E[\tilde{Q}_{t+1} | Z_t] = ra + (r + \alpha_{s}z)bZ_t$ gives the expectation of $\tilde{Q}_{t+1}$ conditional on Z_t and U_{t+1} is the residual, which is also normally distributed and independent of Z_t.

of the stock is then

$$\tau_t \equiv \frac{1}{2}\left(|X_t^1 - X_{t-1}^1| + |X_t^2 - X_{t-1}^2|\right) = |Z_t - Z_{t-1}|. \tag{5.2}$$

It should be emphasized that in our continuous-time model, in absence of any frictions, there is no natural time scale and the share volume over any finite time interval is not well defined. However, a natural time scale does emerge and the share volume is well defined once transactions costs are introduced; we return to this issue in Section 6. For the moment, we assume that trading occurs at certain time intervals (at least in expectation), and volume and return are measured over this interval, which is taken to be the unit of time.

We can now compute the expected return on the stock conditional on the current return and volume. The result is summarized in the following proposition (its proof can be found in Wang, 1994):

Proposition 5.2. *From (5.1) and (5.2), we have*

$$E\big[\tilde{Q}_{t+1}|\tilde{Q}_t, \tau_t\big] = \theta_0 + \theta_1\tilde{Q}_t + \theta_2\tau_t\tilde{Q}_t + \text{higher-order terms in } \tilde{Q}_t \text{ and } \tau_t, \tag{5.3}$$

and $\theta_2 \leq 0$.

In other words, returns accompanied by high volume are more likely to exhibit reversals. CGW and LMSW have explicitly tested this dynamic relation between volume and returns in the form of (5.3). We discuss their empirical findings in Subsection 5.3.

5.2. The Impact of Asymmetric Information

An important aspect of financial markets that our dynamic equilibrium model of Section 2 neglected is the possibility of asymmetric information, which is critical to an understanding of dynamic volume-return relations.

Without information asymmetry, returns accompanied with higher volume are more likely to reverse themselves, as Proposition 5.2 states. The intuition behind this result is simple: Suppose (a subset of) investors sell stocks to adjust their risk profile in response to exogenous shocks to their risk exposure or preferences (e.g., their exposure to the nonfinancial risk); stock prices must decrease to attract other investors to take the other side of the trade. Consequently, we have a negative current return accompanied by high volume. Because the expectation of future stock dividends has not changed, the decrease in current prices corresponds to an increase in expected future returns.

In the presence of information asymmetry, especially when some investors have superior information about the stocks relative to other investors, the dynamic volume-return relation can be different. Suppose, for example, better informed investors reduce their stock positions in response to negative private information about future dividends. Current prices have to decrease to attract other investors to buy. However, prices will not drop by the amount that fully reflects the negative information because the market, in general, is not informationally efficient (in our model, because of incompleteness). As a result, they drop further later as more of the negative information gets impounded into prices (through additional trading or public revelation). In this case, we have negative current returns accompanied by high volume, followed by lower returns later.

Of course, investors may still trade for noninformational reasons (e.g., hedging their nonfinancial risk), in which case the opposite is true, as discussed before. The net effect, that is, the expected future return conditioned on current return and volume, depends on the relative importance of asymmetric information. Wang (1994) has shown that, in the presence of severe information asymmetry, it is possible that θ_2 becomes negative. Under the simplifying assumption that private information is short lived and investors are myopic, LMSW further show that θ_2 decreases monotonically with the degree of information asymmetry. The differences in the degree of information asymmetry can give rise to differences in the dynamic volume-return relation across stocks.

5.3. Empirical Evidence

Many authors have explored the dynamic volume-return relation (5.3). For example, CGW and LeBaron (1992) provided supporting evidence based on market indices and aggregate volume measures. Antoniewicz (1993) and Stickel and Verrecchia (1994) examine the same relation by using pooled returns and volume of individual stocks and find conflicting evidence. Wang (1994) develops a model to incorporate the effect of information asymmetry on volume-return relations, which provides a framework to explain the difference between market indices and individual stocks. LMSW sharpened the results in Wang (1994) and empirically examined the cross-sectional variations in the volume-return relation predicted by the model. Their results reconcile the previous empirical evidence on the volume-return relation.

We now discuss the empirical results presented in CGW and LMSW in testing (5.3) for market indices and individual stocks, respectively. In both studies, turnover for the market and individual stocks are used as a measure of trading volume. In particular, when the existing literature is followed, a detrended log turnover is used[26]:

$$\tilde{\tau}_t = \log(\tau_t + \epsilon) - \frac{1}{N} \sum_{s=-N}^{-1} \log(\tau_{t+s} + \epsilon). \tag{5.4}$$

Here, a moving average of past turnover is used for detrending. The average

[26] As discussed in Section 3, detrending can affect the time-series properties of the data. The detrending scheme in CGW and LMSW is merely to make the results comparable to theirs or to report their results.

is taken over N days, where N is set at 250 in CGW and 200 in LMSW. For individual stocks, daily trading volume is often zero, in which case a small constant ϵ, which is set at 0.00000255 in LMSW, is added to the turnover before logarithms are taken.[27] For market indexes, daily volume is always positive and ε is set to zero.

Both CGW and LMSW estimate the dynamic volume-return relation of the following form:

$$R_{t+1} = \theta_0 + \theta_1 R_t + \theta_2 \tilde{\tau}_t R_t + \epsilon_{t+1} \tag{5.5}$$

for market indexes and individual stocks, respectively, where R_t is the rate of return on an index or an individual stock. LMSW further examine the cross-sectional variation in the coefficient θ_2. Note that there is a slight gap between the variables proposed in the theoretical model and those used in the empirical part. In particular, the model considers excess dollar returns per share and turnover, whereas the empirical analysis considers returns per dollar and detrended log turnover. The difference between the theoretical and corresponding empirical variables is mainly a matter of normalization. At the daily frequency that we focus on, the relation among these variables should not be very sensitive to the normalizations used here.

CGW used the value-weighted return on stocks traded on the NYSE and AMEX, as provided by CRSP, from July 1962 to December 1988. For the turnover measure of the market, they used the total number of shares traded for all the stocks (in the sample), normalized by the total number of shares outstanding. As shown in Lo and Wang (2000a), this turnover measure is equivalent to the weighted average of individual turnover, where the weight for each stock is proportional to its number of shares outstanding. To be more compatible with our current analysis, we repeat the CGW procedure by using our data set, which is more complete (including the additional period from January 1989 to December 1996), and by using the value-weighted turnover index as a measure of market turnover. The results are reported in Table 6.6. We find that for market indexes, θ_2 is negative for the whole sample period and for each of the subperiods, confirming the prediction of the model.

LMSW examine (5.5) at the level of individual stocks and its cross-sectional differences. From the CRSP data set, they have chosen the sample to be stocks traded on the NYSE and AMEX between January 1, 1983 and December 31, 1992.[28] Stocks with more than twenty consecutive days of missing values or

[27] The value of the constant is chosen to maximize the normality of the distribution of daily trading volume. See Richardson et al. (1986), Cready and Ramanan (1991), and Ajinkya and Jain (1989) for an explanation.

[28] The main reason for their choosing this ten-year period it that it overlaps with the TAQ data on bid–ask spreads of individual stocks, which they use as a potential measure of information asymmetry among investors. In the published version of LMSW, they use a more recent but shorter sample, which is January 1, 1993 to December 31, 1998.

Table 6.6. *Impact of volume on autocorrelation of market returns*

Sample Period	$\hat{\theta}_2$ (SE)
1962–1966	−0.466 (0.117)
1967–1971	−0.425 (0.106)
1972–1976	−0.413 (0.106)
1977–1981	−0.334 (0.115)
1982–1986	−0.061 (0.102)
1987–1991	0.098 (0.065)
1992–1996	−0.383 (0.141)
1962–1996	−0.161 (0.030)

Note: This is as measured by the estimated parameter $\hat{\theta}_2$ from the regression: $R_{t+1}^{\mathrm{VW}} = \theta_0 + (\sum_{i=1}^{5} \theta_{1i} D_{it} + \theta_2 \tilde{\tau}_t^{\mathrm{VW}}) R_t^{\mathrm{VW}} + \epsilon_{t+1}$, where R_t^{VW} is the daily CRSP value-weighted return index for NYSE and AMEX stocks, $\tilde{\tau}_t^{\mathrm{VW}}$ is the daily value-weighted turnover index, and D_{it} is an indicator variable for the ith day of the week, $i = 1, \ldots, 5$.

more than twenty consecutive days without a single trade are excluded from the sample. The characteristics of the sample are given in Table 6.7, which reports the summary statistics of the average market capitalization, average daily share volume, average daily turnover, and average price for the whole sample and for the five quintiles.

Their results on (5.5) are reproduced in Table 6.8. It shows that, for stocks in the largest three quintiles, the average of their estimates for θ_2 is positive, as our model predicts. This is consistent with the results of CGW and LeBaron (1992) based on market indexes or large stocks. However, for the two smallest quintiles, the average of the θ_2 estimate is negative. As LMSW suggest, if market capitalization can be used as a negative proxy for the degree of information asymmetry about a stock, the decrease in the value of θ_2 as market capitalization decreases is consistent with the above theoretical discussion. The negative θ_2 estimates for many small stocks are consistent with the negative θ_2 Antoniewicz (1993) reported based on pooled individual returns and volume.

What we conclude from the discussion in this section is that our theoretical model leads to important implications about the joint behavior of volume and

Table 6.7. *Summary statistics for the 1,108 firms used in LMSW.*

Sample Quintiles	AvgCap ($\times\$10^6$)	AvgTrd ($\times 100$)	AvgTurn (%)	AvgPrc ($)
Mean (Q1)	26.53	73	0.162	9.46
Median	23.01	42	0.139	7.94
SD	16.54	93	0.121	6.63
Obs.	(222)	(222)	(222)	(222)
Mean (Q2)	119.97	241	0.216	17.39
Median	115.12	145	0.202	16.62
SD	40.18	306	0.126	8.05
Obs.	(222)	(222)	(222)	(222)
Mean (Q3)	401.96	477	0.233	26.33
Median	384.57	311	0.194	25.16
SD	141.47	481	0.139	12.32
Obs.	(222)	(222)	(222)	(222)
Mean (Q4)	1210.95	1181	0.269	33.46
Median	1150.49	953	0.258	31.90
SD	378.44	930	0.128	14.51
Obs.	(221)	(221)	(221)	(221)
Mean (Q5)	6553.68	3426	0.280	49.38
Median	4020.11	2739	0.259	44.65
SD	8032.98	2623	0.133	28.35
Obs.	(221)	(221)	(221)	(221)
Mean (sample)	1658.61	1077	0.232	27.18
Median	383.10	366	0.212	24.76
SD	4359.62	1768	0.136	21.03
Obs.	(1108)	(1108)	(1108)	(1108)

Note: To be included in the sample, a firm had to have a stock traded on the NYSE or AMEX for the entire sample period (1983–1992) with return and volume information in the CRSP database. AvgCap is the average end-of-year market capitalization; AvgTurn is the average daily turnover; AvgTrd is the average number of shares traded daily; and AvgPrc is the average stock price over the sample period.

returns. In particular, its implication on the dynamic volume-return relation is generally supported by the empirical findings. However, the model ignores many other factors in the market, such as the existence of private information, frictions, and a rich set of trading motives. These factors can be important in developing a more complete understanding of the behavior of volume and its relation with returns. For the specific dynamic volume-return relations examined by LMSW, for example, the existence of information asymmetry is crucial. In this sense, we should view our model as merely a starting point from which we can build a more complete model. Our discussion in the next two sections, which deals with transactions costs and trading behavior based on technical analysis, is one approach to pursuing this broader agenda.

Table 6.8. *Impact of volume on the autocorrelation of stock returns*

Quintile	θ_0 (# < 0)	θ_1 (# < 0)	$-\theta_2$ (# < 0)	t_0 (|#| > 1.64)	t_1 (|#| > 1.64)	$-t_2$ (|#| > 1.64)	R^2 (%)	AvgCap ($\times\$10^6$)
Q1	0.000935	−0.104318	0.032001	0.899	−4.557	2.712	2.766	26.53
($n = 222$)	(29)	(171)	(32)	(32)	(172)	(154)		
Q2	0.000448	0.003522	0.015588	0.994	−0.178	0.995	1.313	119.97
($n = 222$)	(36)	(101)	(85)	(55)	(155)	(107)		
Q3	0.000593	0.041883	−0.005034	1.582	1.599	−0.217	0.879	401.96
($n = 222$)	(13)	(63)	(131)	(100)	(160)	(84)		
Q4	0.000602	0.055762	−0.016220	1.683	2.255	−0.677	0.800	1210.95
($n = 221$)	(7)	(44)	(140)	(109)	(157)	(88)		
Q5	0.000696	0.048459	−0.019830	1.956	1.968	−0.700	0.581	6553.68
($n = 221$)	(1)	(42)	(137)	(144)	(137)	(84)		

Note: This is by average market capitalization quintiles (where the average is computed over the entire sample period for each stock). For each stock, we measure the impact of volume on the autocorrelation of stock returns by the parameter θ_2 from the regression: $R_{jt+1} = \theta_0 + \theta_1 R_{jt} + \theta_2 \tilde{\tau}_{jt} R_{jt} + \epsilon_{jt+1}$, where R_{jt} is the daily return of stock j and $\tilde{\tau}_{jt}$ is its daily detrended log turnover. We report the mean value of each parameter for five size quintiles. The number of negative parameters as well as the number of statistically significant (at the 10 percent level) parameters are also noted, and t statistics are reported in parentheses (t_k denotes the t statistic for θ_k, where $k = 0, 1, 2$).

6. TRADING VOLUME AND TRANSACTIONS COSTS

The theoretical model of Section 2 assumes that investors can trade in the market continuously at no cost. Consequently, the volume of trade is unbounded over any finite time interval. In our empirical analysis, we have avoided this issue by implicitly assuming that investors trade at a finite frequency. Trading volume over the corresponding trading interval is determined by the desired changes in the investors' stock positions over the interval, which is finite. This shortcut is intuitive. In practice, transactions costs would prevent any investor from trading continuously. Investors trade discretely only over time. However, the actual trading interval does depend on the nature of the costs as well as investors' motives to trade.

In this section, we address this issue directly within the framework of our theoretical model by incorporating transactions costs explicitly into our analysis. Our objective is twofold. The narrow objective is to provide a theoretical justification for the shortcut in our empirical analysis. The broader objective is to provide a benchmark for understanding the level of volume. It has often been argued that the level of volume we see in the market is too high to be justifiable from a rational asset-pricing perspective, in which investors trade to share risk and to smooth consumption over time (see, e.g., Ross, 1989). Yet, in the absence of transactions costs, most dynamic equilibrium models imply that trading volume is infinite when the information flow to the market is continuous (i.e., a diffusion). Thus, the level of volume crucially depends on the nature and the magnitude of transactions costs. By considering the impact of transactions costs on the investors' trading behavior as well as the stock prices in our theoretical model, we hope to better understand the level of trading volume.

Our discussion in this section is based on the work of Lo, Mamaysky, and Wang (2000a; LMW1, hereafter). Only the relevant results are presented here, and the readers are referred to LMW1 for details.

6.1. Equilibrium Under Fixed Transactions Costs

We start by introducing fixed transactions costs into our model. It is obvious that fixed costs makes continuous trading prohibitively costly. Consequently, investors do not adjust their stock positions continuously as new shocks hit the economy. Instead, as market conditions and their own situations change, investors will wait until their positions are far from the optimal ones under the new conditions and then trade discretely toward the optimal positions. As a result, they trade only infrequently, but by discrete amounts when they do trade. How frequently and how much they trade depend on the magnitude of the cost and the nature of their trading needs.

For tractability, we again consider a simple case of our model with only one stock and two investors. We also assume that $Z_t = 0$ for all $t \geq 0$. In this case there is no aggregate exposure to nonfinancial risk and the stock returns are *iid* over time. It should be pointed out that, in our model, it is the difference in the investors' exposures to the nonfinancial risk that gives rise to the trading between them. The aggregate exposure does not give rise to trading needs because it affects all investors in the same way. In addition, we let $Y_t^i = 0$ for all $t \geq 0$ $(i = 1, 2)$ and $\alpha_X = 0$.[29]

In addition, we assume that investors have to pay a fixed cost each time they trade the stock. In particular, for each stock transaction, the two sides have to pay a total fixed cost of κ, which is exogenously specified and independent of the amount transacted. This cost is allocated between the buyer and seller as follows. For a trade δ, the transactions cost is given by

$$\kappa(\delta) = \begin{cases} \kappa^+ & \text{for } \delta > 0 \\ 0 & \text{for } \delta = 0, \\ \kappa^- & \text{for } \delta < 0 \end{cases} \tag{6.1}$$

where δ is the signed volume (positive for purchases and negative for sales), κ^+ is the cost for purchases, κ^- is the cost for sales, and the sum $\kappa^+ + \kappa^- = \kappa$. The allocation of the fixed cost, given by κ^+ and κ^-, is determined endogenously in equilibrium.

Under fixed transactions costs, investors trade only infrequently. We limit their stock-trading policy set to impulse policies defined as follows:

Definition 6.2. *Let* $\mathbf{N}_+ \equiv \{1, 2, \ldots\}$. *An impulse trading policy* $\{(\tau_k, \delta_k) : k \in \mathbf{N}_+\}$ *is a sequence of trading times* τ_k *and trade amounts* δ_k*, where (1)* $0 \leq \tau_k \leq \tau_{k+1}$ *a.s.* $\forall\ k \in \mathbf{N}_+$*, (2)* τ_k *is a stopping time of* F*, and (3)* δ_k *is progressively measurable with respect to* F_{τ_k}.

Following an impulse-trading policy, investor i's stock holding at time t is S_t^i, given by

$$S_t^i = S_{0^-}^i + \sum_{\{k:\, \tau_k^i \leq t\}} \delta_k^i, \tag{6.2}$$

where $S_{0^-}^i$ is this investor's initial endowment of stock shares, which is assumed to be $\bar{S}$.

We denote the set of impulse-trading policies by S_Δ and the set of consumption-trading policies by Φ_Δ. For the remainder of this section, we restrict the investor's policies to the set Φ_Δ.

We also need to revise the notion of equilibrium in order to accommodate the situation of infrequent trading:

[29] There is no loss of generality by setting $Y_t^i = 0$ because X_t^i captures the same effect. Setting $\alpha_X = 0$ implies that shocks to investors' exposure to nonfinancial risk are permanent, not transitory, which simplifies the analysis (see LMW1 for more details).

Definition 6.3. *In the presence of fixed transactions costs, an equilibrium in the stock market is defined by (a) a price process $P = \{P_t : t \geq 0\}$ progressively measurable with respect to F, (b) an allocation of the transaction cost (κ^+, κ^-) as defined in (6.1), (c) agents' consumption-trading policies $(c^i, S^i) \in \Phi_\Delta$, where $i = 1, 2$, such that (i) each agent's consumption-trading policy solves his or her optimization problem as defined in (2.10) and (ii) the stock market clears:*

$$\forall\, k \in \mathbf{N}_+ : \quad \tau_k^1 = \tau_k^2, \tag{6.3a}$$

$$\delta_k^1 = -\delta_t^2. \tag{6.3b}$$

The solution to the equilibrium involves two standard steps: First to solve for the optimal consumption-trading policy for the two investors, given a stock price process and an allocation of the fixed transaction cost, and next to find the stock price process and cost allocation such that the stock market clears. However, in the presence of transactions costs, the market-clearing condition consists of two parts: investors' trading times always match, which is (6.3), and their desired trade amounts also match, which is (6.3). In other words, "double-coincidence of wants" must always be guaranteed in equilibrium, which is a very stringent condition when investors trade only occasionally.

Before solving it, we make several comments about the equilibrium (assuming it exists). We have skipped technical details in supporting some of the comments and refer the readers to LMW1 for formal derivations.

First, in the absence of transaction costs, our model reduces to (a special version of) the model considered in Section 2. Investors trade continuously in the stock market to hedge their nonfinancial risk. Because their nonfinancial risks offset each other, the investors can eliminate their nonfinancial risk through trading. Consequently, the equilibrium price remains constant over time, independent of the idiosyncratic nonfinancial risk as characterized by $X_t^1 = -X_t^2$. In particular, the equilibrium price has the following form:

$$P_t = \frac{\mu_D}{r} - a \quad \forall\, t \geq 0, \tag{6.4}$$

where $a \equiv \bar{a} = \gamma\sigma_D^2\bar{S}$ gives the risk discount on the price of the stock to compensate for its risk. The investors' optimal stock holding is linear in their exposure to the nonfinancial risk:

$$S_t^i = \bar{S} - X_t^i, \tag{6.5}$$

where $\bar{\theta}$ is a number of stock shares per capita.

Second, in the presence of transaction costs, investors trade only infrequently. However, whenever they trade, we expect them to reach optimal risk sharing. This implies, as in the case of no transactions costs, that the equilibrium price at all trades should be the same, independent of the idiosyncratic nontraded risk X_t^i $(i = 1, 2)$. Thus, we consider the candidate stock price processes of the form of (6.4) even in the presence of transaction costs.[30] The discount a now reflects the price adjustment of the stock for both its risk and illiquidity.

Third, as the stock price stays constant (in the conjectured equilibrium, which is verified later), changes in investors' stock demand are purely driven by changes in their exposure to the nonfinancial risk. Given the correlation between the nonfinancial risk and the stock risk, which becomes perfect when $Z_t = 0$ as assumed here, each investor can completely hedge his or her nonfinancial risk by trading the stock. The net risk an investor actually bears is the difference between his or her stock position and his or her exposure to the nonfinancial risk. In other words, what the investor would like to control is really $z_t^i = S_t^i - X_t^i$, which determines his or her net risk exposure. Thus, z_t^i becomes the effective state variable (in addition to his or her wealth) that drives investor i's optimal policy and determines his or her value function.

Investor i can control z_t^i by adjusting his or her stock holding S_t^i, from time to time. In particular, his or her optimal trading policy falls into the class of bang-bang policies:[31] The investor maintains z_t^i within a certain band, characterized by three parameters, (z_l, z_m, z_u). When z_t^i hits the upper bound of the band z_u, investor i sells $\delta^- \equiv z_u - z_m$ shares of the stock to move z_t^i down to z_m. When z_t^i hits the lower bound of the band z_l, investor i buys $\delta^+ \equiv z_m - z_l$ shares of the stock to move z_t^i up to z_m. Thus, trading occurs at the first passage time when z_t^i hits z_l or z_m, and the trade amount is δ^+ or δ^-, correspondingly. The optimal policy is then further determined by the optimal choice of (z_l, z_m, z_u), which depends on the stock price (i.e., a) and the allocation of transaction costs.

Fourth, given the nature of the investors' trading policies, an equilibrium can be achieved by choosing a and $\kappa^{\pm}$, such that (1) $\delta^+ = \delta^-$ (i.e., $z_u - z_m = z_m - z_l$), and (2) $z_m = \bar{S}$. The first condition guarantees a match of trading timing and amount for both investors (noting that $z_t^1 = z_t^2$). The second condition guarantees that the two investors always hold all the shares of the stock at the prevailing prices.

In light of this discussion, the solution to the optimal trading policy becomes solving for z_m, δ^+, and δ^- given a and $\kappa^{\pm}$, and the solution to the equilibrium reduces to finding the right a and $\kappa^{\pm}$ such that $\delta^+ = \delta^-$ and $z_m = \bar{S}$.

For an arbitrary fixed cost, only numerical solutions to the equilibrium are available (see LMW1). However, when the fixed cost is small, an approximate analytical solution to the equilibrium can be obtained. In particular, we seek the solution to each agent's optimal trading policy, the equilibrium stock price, and cost allocation and stock price that can be approximated by powers of $\epsilon \equiv \kappa^{\alpha}$,

[30] Given the perfect symmetry between the two agents, the economy is invariant under the following transformation: $X_t^1 \to -X_t^1$. This implies that the price must be an even function of X_t. A constant is the simplest even function.

[31] See LMW1 for more discussion on the optimal policy and references therein.

where α is a positive constant. Especially, $\kappa^{\pm}$ takes the following form:

$$\kappa^{\pm} = \kappa \left(\tfrac{1}{2} \pm \sum_{n=1}^{\infty} k^{(n)} \epsilon^{n} \right). \tag{6.6}$$

The following theorem summarizes the equilibrium (see LMW1 for the proof):

Theorem 6.2. *Let $\epsilon \equiv \kappa^{1/4}$. For (a) small κ and $\kappa^{\pm}$ in the form of (6.6), and (b) the value function analytic for small z and ϵ, the investors' optimal trading policy is given by*

$$\delta^{\pm} = \phi\kappa^{1/4} + o(\kappa^{1/2}), \tag{6.7a}$$

$$z_m = \bar{S}. \tag{6.7b}$$

In equilibrium, the stock price and allocation of transaction costs are given by

$$a = \bar{a}\left(1 + \frac{1}{6} r\gamma^2\sigma_D^2\phi^2\kappa^{1/2}\right) + o(\kappa^{1/2}), \tag{6.8a}$$

$$\kappa^{\pm} = \kappa\left[\frac{1}{2} \pm \frac{2}{15} r\gamma a\phi\kappa^{1/4} + o(\kappa^{1/4})\right], \tag{6.8b}$$

where

$$\phi = \left(\frac{6\sigma_z^2}{r\gamma\sigma_D^2}\right)^{1/4}.$$

Here, $o(\kappa^{\alpha})$ denotes terms of a higher order of κ^{α}.

Two things are worth pointing out. Investors now indeed trade only infrequently. Yet, in equilibrium, the coincidence of the timing and the size of their trade are guaranteed. Moreover, the transaction costs lower the stock price, giving rise to an illiquidity discount, in addition to the risk discount given by $\bar{a}$.

These results are derived with one pair of investors, who have offsetting trading needs. Extending these results to allow more pairs is straightforward. However, when the heterogeneity in investors' trading needs takes more general forms, the solution to the equilibrium becomes much harder. Yet, as argued in LMW1, the qualitative features of investors' optimal trading policy are quite robust. Thus, we may expect the results on the trading frequency, trade size, and volume to survive. However, the robustness of the results on the equilibrium price with general forms of heterogeneity in trading needs is less clear.

6.2. Volume Under Fixed Transactions Costs

Given the optimal trading policies in equilibrium, we can now analyze how the level of trading volume depends on the transaction costs. Intuitively, an increase in transaction costs must reduce the volume of trade. Our model suggests a specific form for this relation. In particular, the equilibrium trade size is a constant. From our solution to equilibrium, the volume of trade between time intervals t and $t+1$ is given by

$$V_{t+1} = \sum_{\{k:\, t<\tau_k \le t+1\}} \left|\delta_k^i\right|, \tag{6.9}$$

where $i = 1$ or 2. The average trading volume per unit of time is

$$E\left[V_{t+1}\right] = E\left[\sum_k 1_{\{\tau_k \in (t,\, t+1]\}}\right] \delta \equiv \omega\delta,$$

where ω denotes the expected frequency of trade (i.e., the number of trades per unit of time). For convenience, we define another measure of average trading volume as the number of shares traded over the average time between trades, or

$$V = \frac{\delta}{\Delta\tau} = \frac{\sigma_X^2}{\delta}, \tag{6.10}$$

where $\Delta\tau \equiv E\left[\tau_{k+1} - \tau_k\right] \approx \delta^2/\sigma_z^2$ is the average time between trades.[32] From (6.7), we have

$$V = \sigma_Z^2 \phi^{-1} \kappa^{-1/4} \left[1 + O\left(\kappa^{1/4}\right)\right],$$

where $O(\kappa^\alpha)$ denotes terms of the same order of κ^α. Clearly, as κ goes to zero, trading volume goes to infinity. However, we also have

$$\frac{\Delta V}{V} \approx -\frac{1}{4}\frac{\Delta\kappa}{k}.$$

In other words (for positive transaction costs), a 1 percent increase in the transaction cost decreases trading volume by only 0.25 percent. In this sense, within the range of positive transaction costs, an increase in the cost reduces the volume only mildly at the margin.

Figure 6.5 plots the average volume measure V versus different values of transaction cost κ as well as the appropriate power laws. Clearly, as κ approaches zero, volume diverges.

6.3. A Calibration Exercise

Our model shows that even small fixed transaction costs imply a significant reduction in trading volume and an illiquidity discount in asset prices. To further examine the impact of fixed costs in equilibrium, we calibrate our model by using historical data and derive numerical implications for the illiquidity

[32] Of course, V is different from $E[V_{t+1}]$ by Jensen's inequality. The calculation of $\Delta\tau$ is straightforward (see LMW1).

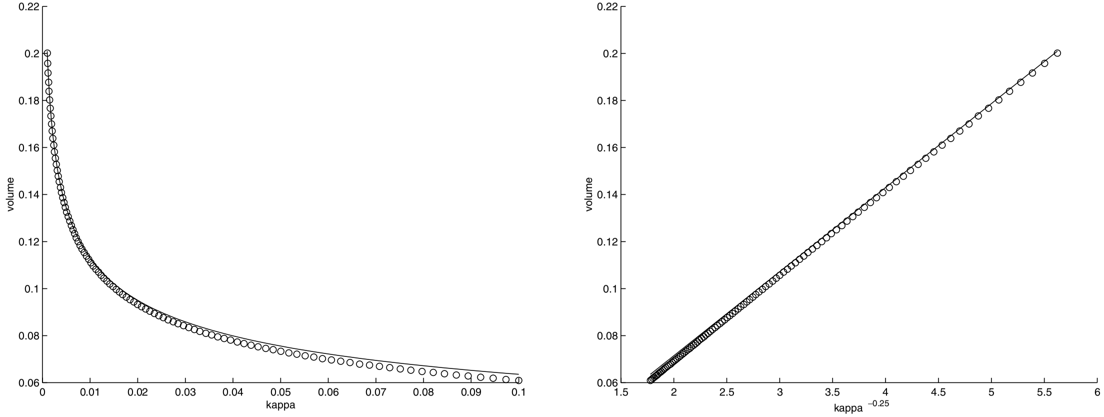


Figure 6.5. Trading volume ν plotted against κ (left) and $\kappa^{-1/4}$ (right). The circles represent the numerical solution. The solid line plots the analytical approximation. The parameter values are $r = 0.037$, $\rho = 0.07$, $\sigma_X = 8.8362$, $\sigma_D = 0.3311$, $\sigma_N = 0.3311$, $\sigma_{DN} = -\sigma_D\sigma_N$, $\gamma = 0.5$, $\bar{\theta} = 12.8225$, $\bar{a}_D = 0.05$, and $P_t = 0.6486$.

discount, trading frequency, and trading volume. From (6.8), for small fixed costs κ we can reexpress the illiquidity discount π as

$$\pi \equiv a - \bar{a} \approx \frac{1}{\sqrt{6}} r^{-1/2} \gamma^{3/2} \sigma_X \bar{a} \kappa^{1/2}. \tag{6.11}$$

Without loss of generality, we set $\sigma_N = 1$; hence the remaining parameters to be calibrated are the interest rate r, the risk discount $\bar{a}$, the volatility of the idiosyncratic nontraded risk σ_X, the agents' coefficient of absolute risk aversion γ, and the fixed transaction cost κ.

The starting point for our calibration exercise is a study by Campbell and Kyle (1993). In particular, they propose and estimate a detrended stock price process of the following form:

$$P_t = A_t - \frac{\lambda}{r} - Z_t, \tag{6.12}$$

where A_t (the present value of future dividends discounted at the risk-free rate) is assumed to follow a Gaussian process, Z_t (fluctuations in stock demand) is assumed to follow an AR(1) Gaussian process, r is the risk-free rate, and λ/r is the risk discount.[33] In Section 2, we saw that, in the absence of transactions costs, our model yields the same price process as (6.12). Moreover, in our model, λ/r is denoted by $\bar{a}$ and Z_t is the aggregate exposure of nontraded risk, which generates changes in stock demand. Therefore, we can obtain values for r, $\bar{a}$, γ, and σ_Z (the instantaneous volatility of Z_t) from their parameter estimates.

Campbell and Kyle based their estimates on annual time series of U.S. real stock prices and dividends from 1871 to 1986. The real stock price of each year is defined by the Standard S&P Composite Stock Price Index in January, normalized by the Producer Price Index in the same month. The real dividend each year is taken to be the annual dividend per share normalized by the Producer Price Index (over this sample period, the average annual dividend growth rate is 0.013). The price and dividend series are then detrended by an exponential detrending factor exp($-0.013\,t$), and the detrended series are used to estimate (6.12) by means of maximum likelihood estimation. In particular, they obtain the following estimates for the price process:

$$\begin{aligned} &r = 0.037, \quad \lambda = 0.0210, \quad \bar{A} = 1.3514, \quad \alpha_Z^{\mathrm{CK}} = 0.0890, \\ &\sigma_Z^{\mathrm{CK}} = 0.1371, \quad \sigma_P^{\mathrm{CK}} = 0.3311, \quad \rho_{PZ}^{\mathrm{CK}} = -0.5176, \end{aligned}$$

where $\bar{A}$ denotes the unconditional mean of A_t^{CK}, α_Z^{CK} and σ_Z^{CK} denote the mean-reversion coefficient and the instantaneous volatility of Z_t^{CK}, σ_P^{CK} denotes the instantaneous volatility of P_t^{CK}, and ρ_{PY}^{CK} denotes the instantaneous correlation between P_t^{CK} and Z_t^{CK}.[34] From these estimates, we are able to compute values

[33] See Campbell and Kyle (1993), Equation (2.3), p. 3.

[34] See Campbell and Kyle's estimates (1993, p. 20) for "Model B."

for the following parameters in our model (in addition to the value of r):

$$\bar{a}_D = 0.0500, \quad \bar{p}_0 = 0.5676, \quad \gamma\sigma_Z = 1.3470, \quad \sigma_D = 0.2853,$$
$$\bar{P} = 0.6486$$

(see LMW1 for the computation of these parameter values of the model from the estimates of Campbell and Kyle). These estimates do not allow us to fully specify the values of γ and σ_Z. However, they do allow us to fix the product of the two. Thus, a choice of γ uniquely specifies the value of σ_Z.

Our model also contains the parameter σ_X, the volatility of idiosyncratic nontraded risk. Because it is the *aggregate* nontraded risk that affects prices, Campbell and Kyle (1993) provide only an estimate for the volatility σ_Z of *aggregate* nontraded risk as a function of the coefficient of absolute risk aversion γ.

Obtaining an estimate for the magnitude of σ_X requires data at a more disaggregated level, which has been performed by Heaton and Lucas (1996) using Panel Study of Income Dynamics (PSID) data. Their analysis shows that the residual variability in the growth rate of individual income – the variability of the component that is uncorrelated with aggregate income – is eight to thirteen times larger than the variability in the growth rate of aggregate income. On the basis of this result, we use values for σ_X to be four times the value of σ_Z to be conservative.[35]

The two remaining parameters to be calibrated are the coefficient of absolute risk aversion γ and the fixed cost κ. Because there is little agreement as to what the natural choices are for these two parameters, we calibrate our model for a range of values for both.

Table 6.9 reports the results of our calibrations. The table contains five subpanels. The first subpanel reports the fixed transaction cost as the percentage of average trade amount, the second subpanel reports the expected time between trades τ (in years), the third subpanel reports the illiquidity discount in the stock price (as a percentage of the price $\bar{P} \equiv \mu_D/r - \bar{a}$ in the absence of transaction costs), the fourth panel reports the return premiums on the stock (caused by its risk as well as illiquidity), and the fifth subpanel reports the average turnover per year, all as functions of the transactions cost κ, which ranges from 1 basis point to 5 percent of $\bar{P}$, and the absolute risk aversion coefficient γ, which ranges from 0.001 to 5.000.[36]

The entries in Table 6.9 show that our model is capable of yielding empirically plausible values for trading frequency, trading volume, and the illiquidity discount. In contrast to much of the existing literature, such as that by Huang (1998), Schroeder (1998), and Vayanos (1998), we find that transactions costs

[35] Other values for the ratio of σ_X and σ_Z are considered in LMW1.

[36] We display transaction costs as a percentage of $\bar{P}$ simply to provide a less scale-dependent measure of their magnitudes. Because κ is a fixed cost, its value is, by definition, scale dependent and must therefore be considered in the complete context of the calibration exercise.

Table 6.9. *Kolmogorov–Smirnov test of the equality of conditional and unconditional one-day return distributions for NYSE/AMEX and NASDAQ stocks*

Statistic	HS	IHS	BTOP	BBOT	TTOP	TBOT	RTOP	RBOT	DTOP	DBOT
					NYSE/AMEX Stocks, 1962–1996					
γ	1.89	1.22	1.15	1.76	0.90	1.09	1.84	2.45	1.51	1.06
p value	0.002	0.104	0.139	0.004	0.393	0.185	0.002	0.000	0.021	0.215
$\gamma\ \tau(\searrow)$	1.49	0.95	0.44	0.62	0.73	1.33	1.37	1.77	0.96	0.78
p-value	0.024	0.327	0.989	0.839	0.657	0.059	0.047	0.004	0.319	0.579
$\gamma\ \tau(\nearrow)$	0.72	1.05	1.33	1.59	0.92	1.29	1.13	1.24	0.74	0.84
p value	0.671	0.220	0.059	0.013	0.368	0.073	0.156	0.090	0.638	0.481
γ Diff.	0.88	0.54	0.59	0.94	0.75	1.37	0.79	1.20	0.82	0.71
p value	0.418	0.935	0.879	0.342	0.628	0.046	0.557	0.111	0.512	0.698
					NASDAQ Stocks, 1962–1996					
γ	2.31	2.68	1.60	1.84	2.81	2.34	2.69	1.90	2.29	2.06
p value	0.000	0.000	0.012	0.002	0.000	0.000	0.000	0.001	0.000	0.000
$\gamma\ \tau(\searrow)$	1.86	1.53	1.35	0.99	1.97	1.95	2.16	1.73	1.38	1.94
p value	0.002	0.019	0.052	0.281	0.001	0.001	0.000	0.005	0.045	0.001
$\gamma\ \tau(\nearrow)$	1.59	2.10	1.82	1.59	1.89	1.18	1.57	1.22	2.15	1.46
p value	0.013	0.000	0.003	0.013	0.002	0.126	0.014	0.102	0.000	0.028
γ Diff.	1.08	0.86	1.10	0.80	1.73	0.74	0.91	0.75	0.76	1.52
p value	0.195	0.450	0.175	0.542	0.005	0.637	0.379	0.621	0.619	0.020

Note: Conditional returns are defined as the daily return three days following the conclusion of an occurrence of one of ten technical indicators: head-and-shoulders (HS), inverted head-and-shoulders (IHS), broadening top (BTOP), broadening bottom (BBOT), triangle top (TTOP), triangle bottom (TBOT), rectangle top (RTOP), rectangle bottom (RBOT), double top (DTOP), and double bottom (DBOT). All returns have been normalized by subtraction of their means and division by their standard deviations; p values are with respect to the asymptotic distribution of the Kolmogorov–Smirnov test statistic. Here $\tau(\searrow)$ and $\tau(\nearrow)$ indicate that the conditional distribution is also conditioned on decreasing and increasing volume trend, respectively.

can have a very large impact on both the trading frequency and the illiquidity discount in the stock price. For example, Schroeder (1998) finds that when faced with a fixed transactions cost of 0.1 percent, individuals in his model trade only once every ten years! In Table 6.9, we see that for a 0.1 percent fixed cost, individuals in our model trade anywhere between $1/0.002 = 600$ and $1/0.148 = 6.8$ times per year as the risk-aversion parameter varies from 0.001 to 5.000, respectively. This contrast between our results and those of the existing literature stems from the fact that our investors have a strong need to trade frequently. The high-frequency changes in their risk exposure to nonfinancial risk imply that not trading can be very costly. Furthermore, not trading means that the risk exposure from holding market-clearing levels of the stock is much greater. Models of transactions costs often fail to account for a high-frequency component in trading needs.[37]

As the risk-aversion parameter increases while holding κ fixed, trading becomes less frequent, the illiquidity discount increases, and the trade size also declines. For example, a risk-aversion parameter of 5.00 and a fixed cost of 1 percent of $\bar{P}$ implies that the investor will trade approximately once every two years, each trade consisting of only 1.297 shares, with an illiquidity discount of 1.547 percent of $\bar{P}$.

What seems very striking is that for reasonable magnitudes of investors' trading needs (measured by σ_X) and transaction costs, our model produces reasonable levels of volume. For example, a risk-aversion parameter of 1 and a fixed cost of 0.50 percent of $\bar{P}$ imply a trading frequency of $1/0.148 = 6.8$ trades per year, an illiquidity discount of 1.97 percent, and a turnover of 352 percent, which is higher than the observed turnover (see Section 4). These results suggest that existing levels of trading frequency and volume in financial markets may not be as unusual or as irrational as many have thought. The calibration results in Table 6.9 show that our dynamic equilibrium model is clearly capable of generating empirically plausible implications.

7. TECHNICAL ANALYSIS

Although the interest in trading volume is a relatively recent development in the academic finance literature, it has been a long-standing tradition among finance professionals engaging in "technical analysis" or "charting," the practice of forecasting future price movements in financial securities based on geometric patterns in the time-series plots of past prices and volume. Historically ridiculed by academics as "voodoo finance," technical analysis has never enjoyed the widespread acceptance among academics and industry practitioners that fundamental analysis and quantitative finance have. However, several

[37] Although many partial equilibrium models do contain a high-frequency component in the uncertainty faced by their investors, such as those by Constantinides (1986) and Amihud and Mendelson (1986a), they still miss these equilibrium effects because they do not take into account the unwillingness of investors to hold large amounts of the risky asset in the presence of transaction costs.

recent academic studies suggest that historical prices may contain incremental information that has not already been incorporated into current market prices, raising the possibility that technical analysis can add value to the investment process.[38] Moreover, a closer reading of the early technical analysis literature, for example, Hamilton (1922), reveals a surprisingly contemporary view of the market forces that influence prices and price dynamics. In particular, the importance of supply and demand, buying and selling pressure, and the risk preferences of market participants was acknowledged by technical analysts long before financial economists developed similar interests (albeit with different tools and terminology). And the emphasis that technical analysts place on trading volume is the motivation for our interest in technical analysis.

In this section, we review the results of Lo, Mamaysky, and Wang (2000b; hereafter, LMW2), in which the information content of technical indicators is measured by first developing an automated procedure for detecting certain types of patterns, such as head-and-shoulders patterns, and then applying this procedure to historical prices of U.S. stocks to measure the impact of these patterns on postpattern return distributions. By comparing the unconditional empirical distribution of daily stock returns to the conditional distribution – conditioned on the occurrence of specific technical indicators such as head-and-shoulders or double-bottom indicators – they find that over the thirty-five-year sample period, several technical indicators do provide incremental information and may have some practical value.

In Subsection 7.1 we describe the pattern-detection algorithm of LMW2, Subsection 7.2 discusses the statistical methods for gauging the information content of the patterns detected, and Subsection 7.3 reports the empirical results of the pattern-detection algorithm applied to a large sample of individual U.S. stocks from 1962 to 1996.

7.1. Automating Technical Analysis

To determine the efficacy of technical analysis, we must be able to apply it in a consistent fashion over an extended period of time and across a broad sample of securities, and then assess its performance statistically. Therefore, we must first develop a method for automating the identification of technical indicators; that is, we require a pattern-recognition algorithm. Once such an algorithm is

[38] For example, in rejecting the random walk hypothesis for weekly U.S. stock indexes, Lo and MacKinlay (1988, 1999) have shown that past prices may be used to forecast future returns to some degree, a fact that all technical analysts take for granted. Studies by Tabell and Tabell (1964), Treynor and Ferguson (1985), Brown and Jennings (1989), Jegadeesh and Titman (1993), Blume, Easley, and O'Hara (1994), Chan, Jegadeesh, and Lakonishok (1996), Lo and MacKinlay (1997), Grundy and Martin (1998), and Rouwenhorst (1998) have also provided indirect support for technical analysis, and more direct support has been given by Pruitt and White (1988), Neftci (1991), Brock, Lakonishok, and LeBaron (1992), Neely, Weller, and Dittmar (1997), Neely and Weller (1998), Chang and Osler (1994), Osler and Chang (1995), and Allen and Karjalainen (1999).

developed, it can be applied to a large number of securities over many time periods to quantify the information content of various technical indicators. Moreover, quantitative comparisons of the performance of several indicators can be conducted, and the statistical significance of such performance can be assessed through Monte Carlo simulation and bootstrap techniques. This is the approach taken by LMW2.[39]

The starting point of LMW2's analysis is the assumption that prices P_t can be represented by the following expression:

$$P_t = m(\cdot) + \epsilon_t, \tag{7.1}$$

where $m(\cdot)$ is a nonlinear function of time (and perhaps other state variables), and ϵ_t is white noise. LMW2 argue that technical analysts estimate $m(\cdot)$ visually by attempting to discern geometric regularities in the raw price series $\{P_t\}$, and that this process is similar in spirit to *smoothing estimators* in which sophisticated forms of local averaging are used to estimate $m(\cdot)$ by averaging out the noise ϵ_t. Specifically, LMW2 propose the following algorithm for detecting the occurrence of various technical patterns:

1. Define each technical pattern in terms of its geometric properties, for example, local extrema (maxima and minima) of $m(\cdot)$.
2. Construct a kernel estimator $\hat{m}(\cdot)$ of a given time series of prices so that its extrema can be determined numerically.
3. Analyze $\hat{m}(\cdot)$ for occurrences of each technical pattern.

LMW2 focus on five pairs of technical patterns that are among the most popular patterns of traditional technical analysis (see, e.g., Edwards and Magee, 1966, Chapters VII–X): head and shoulders (HS) and inverse head and shoulders (IHS), broadening tops (BT) and bottoms (BB), triangle tops (TT) and bottoms (TB), rectangle tops (RT) and bottoms (RB), and double tops (DT) and bottoms (DB). Specifically, denote by $E_1, E_2, \ldots, E_n$ the n extrema of $m(\cdot)$ and by $t_1^*, t_2^*, \ldots, t_n^*$ the dates on which these extrema occur. Then LMW2 propose the following definitions for the HS and IHS patterns:

Definition 7.1. *(HS) HS and IHS patterns are characterized by a sequence of five consecutive local extrema $E_1, \ldots, E_5$ such that*

$$\text{HS} \equiv \begin{cases} E_1 \textit{ a maximum} \\ E_3 > E_1,\ E_3 > E_5 \\ E_1 \textit{ and } E_5 \textit{ within } 1.5 \textit{ percent of their average} \\ E_2 \textit{ and } E_4 \textit{ within } 1.5 \textit{ percent of their average} \end{cases},$$

[39] A similar approach has been proposed by Chang and Osler (1994) and Osler and Chang (1995) for the case of foreign-currency trading rules based on a head-and-shoulders pattern. They develop an algorithm for automatically detecting geometric patterns in price or exchange data by looking at properly defined local extrema.

$$\mathrm{IHS} \equiv \begin{cases} E_1 \textit{ a minimum} \\ E_3 < E_1,\ E_3 < E_5 \\ E_1 \textit{ and } E_5 \textit{ within } 1.5 \textit{ percent of their average} \\ E_2 \textit{ and } E_4 \textit{ within } 1.5 \textit{ percent of their average} \end{cases}.$$

Note that only five consecutive extrema are required to identify an HS pattern, which follows from the formalization of the geometry of an HS pattern: three peaks, with the middle peak higher than the other two. Because consecutive extrema must alternate between maxima and minima for smooth functions,[40] the three-peaks pattern corresponds to a sequence of five local extrema: maximum, minimum, highest maximum, minimum, and maximum. The IHS is simply the mirror image of the HS, with the initial local extrema a minimum.

LMW2 develop similar definitions for broadening, rectangle, and triangle patterns, each with two possible versions depending on whether the initial extremum is a local maximum or minimum, yielding a total of ten patterns in all.

Given a sample of prices $\{P_1, \ldots, P_T\}$, kernel regressions for rolling subsamples or *windows*, and within each window, local extrema of the estimated function $\hat{m}(\tau)$, can be readily identified by finding times τ such that $\mathrm{sgn}(\hat{m}'(\tau)) = -\mathrm{sgn}(\hat{m}'(\tau + 1))$, where $\hat{m}'$ denotes the derivative of $\hat{m}$ with respect to τ and $\mathrm{sgn}(\cdot)$ is the signum function. If the signs of $\hat{m}'(\tau)$ and $\hat{m}'(\tau + 1)$ are $+1$ and -1, respectively, then we have found a local maximum, and if they are -1 and $+1$, respectively, then we have found a local minimum. Once such a time τ has been identified, we proceed to identify a maximum or minimum in the original price series $\{P_t\}$ in the range $[t - 1, t + 1]$, and the extrema in the original price series are used to determine whether or not a pattern has occurred according to the definitions of the ten technical patterns.[41] One useful consequence of this algorithm is that the series of extrema that it identifies contains alternating minima and maxima. That is, if the kth extremum is a maximum, then it is always the case that the $(k + 1)$th extremum is a minimum, and vice versa.

An important advantage of using this kernel regression approach to identify patterns is the fact that it ignores extrema that are "too local." For example, a simpler alternative is to identify local extrema from the raw price data directly, that is, to identify a price P_t as a local maximum if $P_{t-1} < P_t$ and $P_t > P_{t+1}$, and vice versa for a local minimum. The problem with this approach is that it identifies too many extrema, and it also yields patterns that are not visually consistent with the kind of patterns that technical analysts find compelling.

[40] After all, for two consecutive maxima to be local maxima, there must be a local minimum in between, and vice versa for two consecutive minima.

[41] If $\hat{m}'(\tau) = 0$ for a given τ, which occurs if closing prices stay the same for several consecutive days, we need to check whether the price we have found is a local minimum or maximum. We look for the date s such that $s = \inf\{s > \tau\ :\ \hat{m}'(s) \neq 0\}$. We then apply the same method as already discussed, except here we compare $\mathrm{sgn}(\hat{m}'(\tau - 1))$ and $\mathrm{sgn}(\hat{m}'(s))$. See LMW2 for further details.

Once all of the local extrema in a given window have been identified, the presence of the various technical patterns can be determined by using definitions such as Definition 7.1. This procedure is then repeated for the next window and continues until the end of the sample is reached.

7.2. Statistical Inference

Although there have been many tests of technical analysis over the years, most of these tests have focused on the profitability of technical trading rules.[42] Although some of these studies do find that technical indicators can generate statistically significant trading profits, they beg the question of whether or not such profits are merely the equilibrium rents that accrue to investors willing to bear the risks associated with such strategies. Without specifying a fully articulated dynamic general equilibrium asset-pricing model, it is impossible to determine the economic source of trading profits.

Instead, LMW2 propose a more fundamental test in their study, one that attempts to gauge the information content in the technical patterns of Subsection 7.1 by comparing the unconditional empirical distribution of returns with the corresponding conditional empirical distribution, conditioned on the occurrence of a technical pattern. If technical patterns are informative, conditioning on them should alter the empirical distribution of returns; if the information contained in such patterns has already been incorporated into returns, the conditional and unconditional distribution of returns should be close. Although this is a weaker test of the effectiveness of technical analysis – informativeness does not guarantee a profitable trading strategy – it is, nevertheless, a natural first step in a quantitative assessment of technical analysis.

To measure the distance between the two distributions, LMW2 use the Kolmogorov–Smirnov test,[43] which is designed to test the null hypothesis that two samples have the same distribution function, and which is based on the empirical cumulative distribution functions of both samples. Under the null

[42] For example, Chang and Osler (1994) and Osler and Chang (1995) propose an algorithm for automatically detecting HS patterns in foreign exchange data by looking at properly defined local extrema. To assess the efficacy of an HS trading rule, they take a stand on a class of trading strategies and compute the profitability of these across a sample of exchange rates against the U.S. dollar. The null return distribution is computed by a bootstrap that samples returns randomly from the original data so as to induce temporal independence in the bootstrapped time series. By comparing the actual returns from trading strategies to the bootstrapped distribution, the authors find that for two of the six currencies in their sample (the yen and the Deutsche mark), trading strategies based on an HS pattern can lead to statistically significant profits. Also see Neftci and Policano (1984), Pruitt and White (1988), and Brock et al. (1992).

[43] LMW2 also compute chi-squared goodness-of-fit statistics, but we omit them to conserve space. Note that the sampling distribution of the Kolmogorov–Smirnov statistic is derived under the assumption that returns are independently and identically distributed, which is not plausible for financial data. LMW2 attempt to address this problem by normalizing the returns of each security, that is, by subtracting its mean and dividing by its standard deviation (see Subsection 7.3), but this does not eliminate the dependence or heterogeneity and warrants further research.

hypothesis, Smirnov (1939a, 1939b) has derived the limiting distribution of the statistic, and an approximate α-level test of the null hypothesis can be performed by computing the statistic and rejecting the null if it exceeds the upper 100αth percentile for the null distribution (see Hollander and Wolfe, 1973, Table A.23; Csáki, 1984; Press et al., 1986, Chapter 13.5; and LMW2).

7.3. Empirical Results

LMW2 apply the Kolmogorov–Smirnov test to the daily returns of individual NYSE/AMEX and NASDAQ stocks from 1962 to 1996, using data from the CRSP. To ameliorate the effects of nonstationarities induced by changing market structure and institutions, they split the data into NYSE/AMEX stocks and NASDAQ stocks and into seven five-year periods: 1962–1966, 1967–1971, and so on. To obtain a broad cross section of securities, in each five-year subperiod, they randomly select ten stocks from each of five market capitalization quintiles (using mean market capitalization over the subperiod), with the further restriction that at least 75 percent of the price observations must be nonmissing during the subperiod.[44] This procedure yields a sample of fifty stocks for each subperiod across seven subperiods (note that they sample with replacement; hence there may be names in common across subperiods).

For each stock in each subperiod, LMW2 apply the procedure outlined in Subsection 7.1 to identify all occurrences of the ten patterns they define mathematically according to the properties of the kernel estimator. For each pattern detected, they compute the one-day continuously compounded return three days after the pattern has completed. Therefore, for each stock, there are ten sets of such conditional returns, each conditioned on one of the ten patterns of Subsection 7.1.

For each stock, a sample of *unconditional* continuously compounded returns is constructed by using nonoverlapping intervals of length τ, and the empirical distribution function of these returns is compared with those of the conditional returns. To facilitate such comparisons, all returns are standardized – both conditional and unconditional – by subtracting means and dividing by standard deviations; hence,

$$X_{it} = \frac{R_{it} - \text{mean}[R_{it}]}{\text{SD}[R_{it}]}, \tag{7.2}$$

where the means and standard deviations are computed for each individual stock within each subperiod. Therefore, by construction, each normalized return series has zero mean and unit variance.

To increase the power of their goodness-of-fit tests, LMW2 combine the normalized returns of all fifty stocks within each subperiod; hence for each

[44] If the first price observation of a stock is missing, they set it equal to the first nonmissing price in the series. If the tth price observation is missing, they set it equal to the first nonmissing price prior to t.

subperiod they have two samples – unconditional and conditional returns – from which two empirical distribution functions are computed and compared by using the Kolmogorov–Smirnov test.

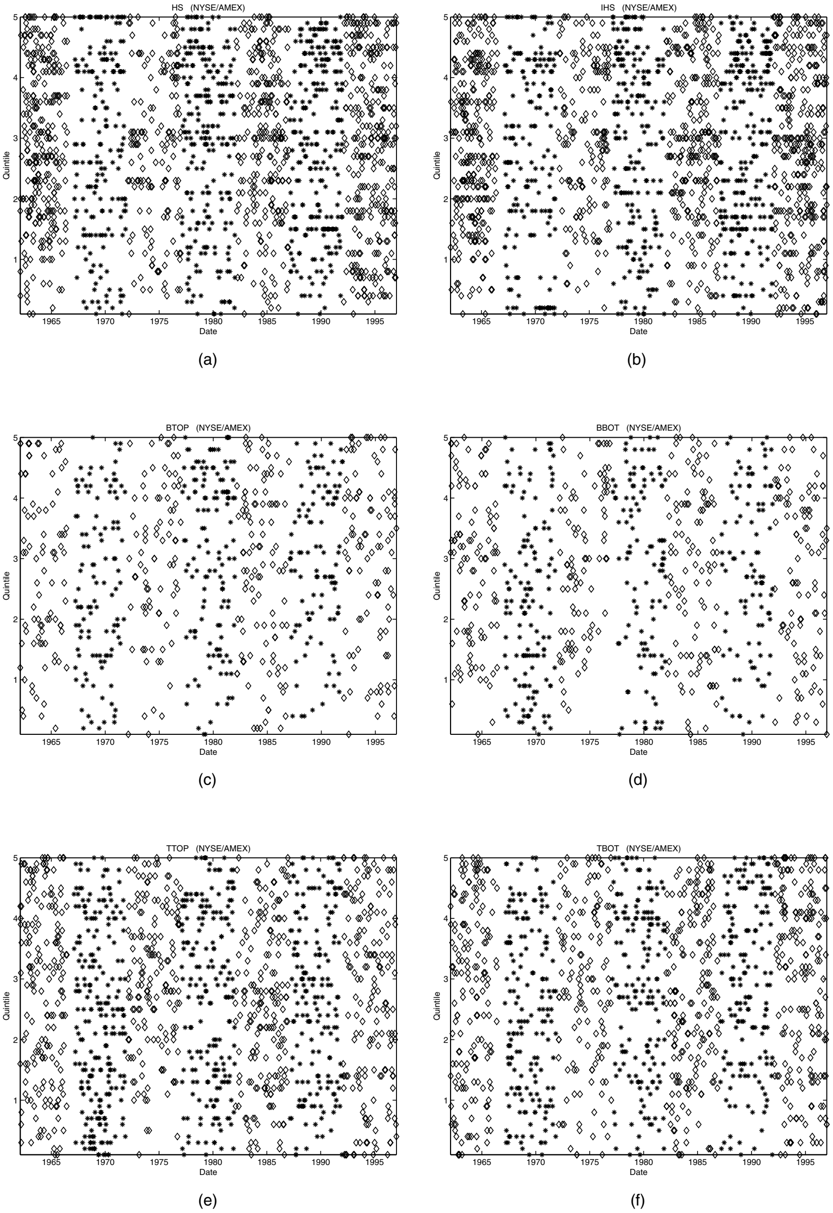
Finally, given the prominent role that volume plays in technical analysis, LMW2 also construct returns conditioned on increasing or decreasing volume. Specifically, for each stock in each subperiod, they compute its average share turnover during the first and second halves of each subperiod, τ_1 and τ_2, respectively.[45] If $\tau_1 > 1.2 \times \tau_2$, they categorize this as a "decreasing volume" event; if $\tau_2 > 1.2 \times \tau_1$, they categorize this as an "increasing volume" event. If neither of these conditions holds, then neither event is considered to have occurred. Using these events, conditional returns can be constructed conditioned on two pieces of information: the occurrence of a technical pattern and the occurrence of increasing or decreasing volume. Therefore, the empirical distribution of unconditional returns can be compared with three conditional-return distributions: the distribution of returns conditioned on technical patterns, the distribution conditioned on technical patterns and increasing volume, and the distribution conditioned on technical patterns and decreasing volume.[46]

To develop some idea of the cross-sectional and time-series distributions of each of the ten patterns, Figures 6.6 and 6.7 plot the occurrences of the patterns for the NYSE/AMEX and NASDAQ samples, respectively, where each symbol represents a pattern detected by the LMW2 algorithm. The vertical axis is divided into five quintiles, the horizontal axis is calendar time, and alternating symbols (diamonds and asterisks) represent distinct subperiods. These graphs show that there are many more patterns detected in the NYSE/AMEX sample than in the NASDAQ sample (Figure 6.6 is more densely populated than Figure 6.7). Also, for the NYSE/AMEX sample, the distribution of patterns is not clustered in time or among a subset of securities, but there seem to be more patterns in the first and last subperiods for the NASDAQ sample.

Table 6.9 contains the results of the Kolmogorov–Smirnov test of the equality of the conditional- and unconditional-return distributions for NYSE/AMEX and NASDAQ stocks from 1962 to 1996. Recall that conditional returns are defined as the one-day return starting three days following the conclusion of an occurrence of a pattern. The p values are with respect to the asymptotic distribution of the Kolmogorov–Smirnov test statistics. The entries in the top panel of Table 6.9 show that for NYSE/AMEX stocks, five of the ten patterns – HS, BBOT, RTOP, RBOT, and DTOP – yield statistically significant test statistics, with p values ranging from 0.000 for RBOT to 0.021 for DTOP patterns. However, for the other five patterns, the p values range from 0.104 for

[45] For the NASDAQ stocks, τ_1 is the average turnover over the first third of the sample, and τ_2 is the average turnover over the final third of the sample.

[46] Of course, other conditioning variables can easily be incorporated into this procedure, though the "curse of dimensionality" imposes certain practical limits on the ability to estimate multivariate conditional distributions nonparametrically.



(a)

(b)

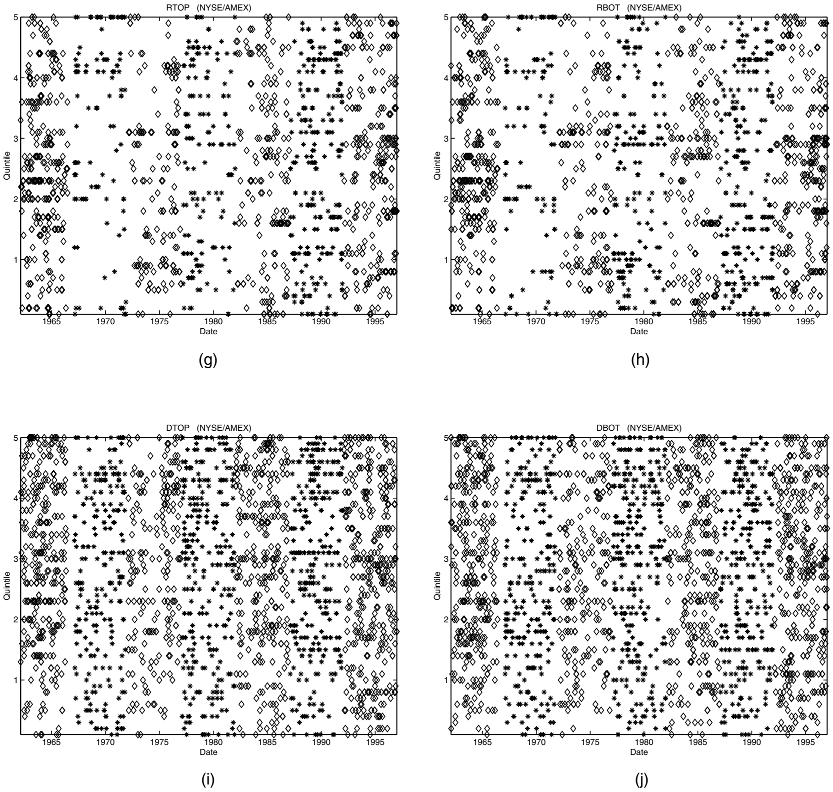
(c)

(d)

(e)

(f)

Figure 6.6. Distribution of patterns in NYSE/AMEX sample.



(g)

(h)

(i)

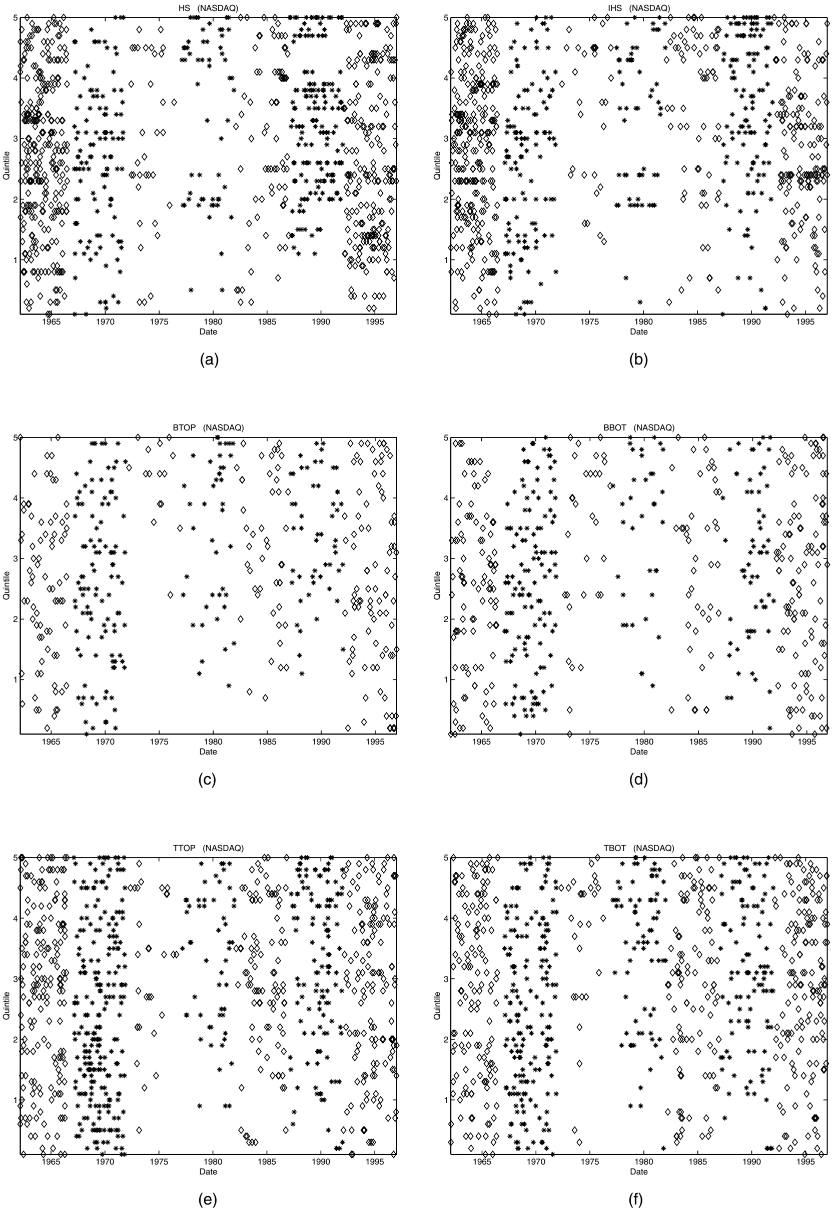
(j)

Figure 6.6. *(continued).*

IHS to 0.393 for DBOT, which implies an inability to distinguish between the conditional and unconditional distributions of normalized returns.

When LMW2 condition on declining volume trend as well as the occurrence of the patterns, the statistical significance declines for most patterns, but increases for TBOT. In contrast, conditioning on increasing volume trend yields an increase in the statistical significance of BTOP patterns. This difference may suggest an important role for volume trend in TBOT and BTOP patterns. The difference between the increasing and decreasing volume-trend conditional distributions is statistically insignificant for almost all the patterns (the sole exception is the TBOT pattern). This drop in statistical significance may be due to a lack of power of the Kolmogorov–Smirnov test given the relatively small sample sizes of these conditional returns.

The bottom panel of Table 6.9 reports corresponding results for the NASDAQ sample, and, in contrast to the NYSE/AMEX results, here all the patterns are statistically significant at the 5 percent level. This is especially significant because the NASDAQ sample exhibits far fewer patterns than the NYSE/AMEX



(a)

(b)

(c)

(d)

(e)

(f)

Figure 6.7. Distribution of patterns in NASDAQ sample.

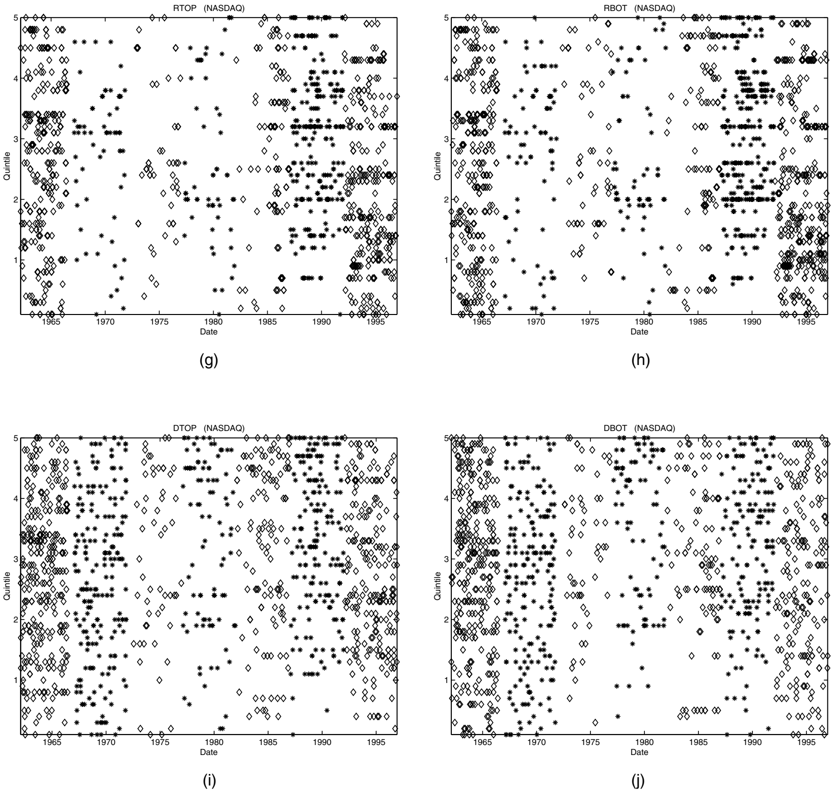


Figure 6.7. *(continued).*

sample (compare Figures 6.6 and 6.7), so the Kolmogorov–Smirnov test is likely to have lower power in this case.

As with the NYSE/AMEX sample, volume trend seems to provide little incremental information for the NASDAQ sample except in one case: increasing volume and BTOP. Except for the TTOP pattern, the Kolmogorov–Smirnov test still cannot distinguish between the decreasing and increasing volume-trend conditional distributions, as the last pair of rows of Table 6.9 indicate.

When applied to many stocks over many time periods, LMW2's approach shows that certain technical patterns do provide incremental information, especially for NASDAQ stocks. Although this does not necessarily imply that technical analysis can be used to generate "excess" trading profits, it does raise the possibility that technical analysis can add value to the investment process. Moreover, the evidence also suggests that volume trend provides incremental information in some cases. Although this hardly seems to be a controversial conclusion (that both prices and quantities contain incremental information for future returns), nevertheless, it comes from a rather surprising source that may contain other insights into the role of trading volume for economic activity.

8. CONCLUSIONS

Trading volume is an important aspect of the economic interactions of investors in financial markets. Both volume and prices are driven by underlying economic forces, and thus convey important information about the workings of the market. Although the literature on financial markets has focused almost exclusively on the behavior of returns based on simplifying assumptions about the market such as perfect competition, lack of frictions, and informational efficiency, we wish to develop a more realistic framework to understand the empirical characteristics of prices and volume.

Here we hope to have made a contribution toward this goal. We first develop a dynamic equilibrium model for asset trading and pricing. The model qualitatively captures the most important motive for investors to participate in the market, namely, to achieve optimal allocations of wealth over time and across different risk profiles. We then explore the implications of the model for the behavior of volume and returns, particularly the cross-sectional behavior of volume and the dynamic volume-return relations. We test these implications empirically and have found them to be generally consistent with the data. Fully realizing that our model merely provides a benchmark at best because it omits many important factors such as asymmetric information, market frictions, and other trading motives, we extend our model to include information asymmetry and transaction costs. We also go beyond the framework of our formal model and analyze the relation between price and volume in heuristic models of the market such as technical analysis. Our empirical analysis of these heuristic models finds some interesting connections between volume and price dynamics.

Our main approach in this paper has been to investigate and study the behavior of price and volume by using a structured equilibrium framework to motivate and direct our empirical analysis. Although this has led to several interesting insights, there are many other directions to be explored. One important direction is to derive and test the implications of the model for identifying the specific risk factors that explain the cross-sectional variation in expected returns. We are currently pursuing this line of research in Lo and Wang (2000b). Another direction is to extend the framework to include other important factors, such as a richer set of trading motives, the actual trading mechanism, price impact, frictions, and other institutional aspects in our analysis. We hope to turn to these issues in the near future.

APPENDIX

In this appendix, we give a proof for Theorem 2.1. We first solve the investors' optimization problem under the stock prices in the form of (2.14) and then show that a and b can be chosen to clear the market.

Define $\theta_t \equiv (1; X_t; Y_t; Z_t)$ to be the state variable for an individual investor, say, i. For simplicity, we omit the superscript i for now. Then

$$d\theta_t = \alpha_\theta \theta_t dt + \sigma_\theta dB_t, \tag{A.1}$$

where $\alpha_\theta \equiv \text{diag}\{0, \alpha_X, \alpha_Y, \alpha_Z\}$ is a diagonal matrix, $\sigma_\theta \equiv (0; \sigma_X; \sigma_Y; \sigma_Z)$. Given the price function in (2.14), the excess dollar return on the stocks can be written as

$$dQ_t = e_Q \theta_t dt + \sigma_Q dB_t, \tag{A.2}$$

where $e_Q \equiv (ra, 0, 0, (r + \alpha_Z)b)$ and $\sigma_Q \equiv \sigma_D - b\sigma_Z$.

Let $J(W, \theta, t)$ denote the value function. We conjecture that it has the following form:

$$J(W, \theta, t) = -e^{-\rho t - r\gamma W - 1/2\theta' v \theta}. \tag{A.3}$$

The Bellman equation then takes the following form:

$$0 = \sup_{c,\, S} - e^{-\rho t - \gamma c_t} + \text{E}[dJ]/dt \tag{A.4a}$$

$$\begin{aligned} &= \sup_{c,\, S} - e^{-\rho t - \gamma c_t} - J\big[\rho + (r\gamma)(rW - c) - 1/2\theta' m \theta \\ &\quad + (r\gamma)S' e_Q \theta - 1/2(r\gamma)^2 S' \sigma_{QQ} S - (r\gamma)^2 S' \sigma_{QN} \iota' \theta \\ &\quad - (r\gamma) S' \sigma_{Q\theta} v \theta\big], \end{aligned} \tag{A.4b}$$

where

$$m \equiv (r\gamma)^2 \sigma_N^2 \iota\iota' + v\sigma_{\theta\theta} v + v\alpha_\theta + \alpha_\theta v + (r\gamma)(\iota\sigma_{N\theta} v + v\sigma_{\theta N}\iota'). \tag{A.5}$$

The first-order condition for optimality gives

$$c = rW - \frac{1}{\gamma}\ln r + \frac{1}{2\gamma}\theta' v \theta, \tag{A.6a}$$

$$S = \frac{1}{r\gamma}[e_Q - (r\gamma)\sigma_{QN}\iota' - \sigma_{Q\theta} v]\theta. \tag{A.6b}$$

Substituting into the Bellman equation, we have

$$\begin{aligned} 0 = v_{00} + \frac{r}{2}\theta' v \theta + \frac{1}{2}\theta' m \theta - \frac{1}{2}\theta(e_Q - r\gamma\sigma_{QN}\iota' - \sigma_{Q\theta} v)' \\ \times (e_Q - r\gamma\sigma_{QN}\iota' - \sigma_{Q\theta} v)\theta, \end{aligned} \tag{A.7}$$

where $v_{00} \equiv r - \rho - r\,\ln r$. This then leads to the following equation for v:

$$\begin{aligned} 0 = \bar{v} + \frac{1}{2}rv + \frac{1}{2}m - \frac{1}{2}(e_Q - r\gamma\sigma_{QN}\iota' - \sigma_{Q\theta} v)' \\ \times (e_Q - r\gamma\sigma_{QN}\iota' - \sigma_{Q\theta} v), \end{aligned} \tag{A.8}$$

where $\bar{v} \equiv v_{00}((1, 0, 0, 0); (0, 0, 0, 0); (0, 0, 0, 0); (0, 0, 0, 0))$.

We now consider market clearing. First,

$$S_t^i = \frac{1}{r\gamma}(\sigma_{QQ})^{-1}\left[ra + (r\ +\ \alpha_Z)bZ_t - (r\gamma)\sigma_{QN}\iota'\theta_t - \sigma_{Q\theta}v\theta_t\right]. \tag{A.9}$$

Second, let $v \equiv (v_0; v_X; v_Y; v_Z)$. Because $\sigma_{Q\theta} = (0, 0, 0, \sigma_{QZ})$, we have $\sigma_{Q\theta}v = \sigma_{QZ}v_Z$. Thus,

$$S_t^i = \frac{1}{r\gamma}(\sigma_{QQ})^{-1}\left[ra + (r\ +\ \alpha_Z)bZ_t - (r\gamma)\sigma_{QN}\left(X_t^i + Y_t^i + Z_t\right)\right.$$
$$\left. - \sigma_{QZ}\left(v_{Z0} + v_{ZX}X_t^i + v_{ZY}Y_t^i + v_{ZZ}Z_t\right)\right]. \tag{A.10a}$$

Third, summing over the investors, we have

$$\iota = \sum_{i=1}^{I} S_t^i = \frac{1}{r\bar{\gamma}}(\sigma_{QQ})^{-1}\left[ra + (r\ +\ \alpha_Z)bZ_t - (r\gamma)\sigma_{QN}Z_t\right.$$
$$\left. - \sigma_{QZ}\left(v_{Z0} + v_{ZZ}Z_t\right)\right], \tag{A.11}$$

where $\bar{\gamma} = \gamma/I$. Thus, we have

$$a = \bar{\gamma}(\sigma_{QQ})^{-1}\iota + \left(\frac{v_{Z0}}{r}\right)\sigma_{QZ}, \tag{A.12a}$$

$$b = \frac{1}{r + \alpha_Z}(v_{ZZ}\sigma_{QZ} + r\gamma\sigma_{QN}). \tag{A.12b}$$

Substituting (A.12) (the equilibrium prices) into the expression for investors' asset demands gives us their equilibrium holdings:

$$S_t^i = \left(\frac{1}{I}\right)\iota - \left(X_t^i + Y_t^i\right)(\sigma_{QQ})^{-1}\sigma_{QN}$$
$$- \frac{1}{r\gamma}\left(v_{ZX}X_t^i + v_{ZY}Y_t^i\right)(\sigma_{QQ})^{-1}\sigma_{QZ}, \tag{A.13}$$

which is the four-fund separation result in Theorem 2.1.

The remaining part of the proof is the existence of a solution to the system of algebraic equations defined by (A.8) and (A.12). The proof of the existence of a solution in the case of a single stock can be found in Huang and Wang (1997), which can be extended to the case of multiple stocks. In particular, Equation (A.8) reduces to a Riccati equation, which has a closed-form solution (under certain parameter restrictions; see Huang and Wang, 1997, for more details). The existence of a solution to (A.12) is then straightforward to establish.

ACKNOWLEDGMENTS

We thank Joon Chae, Jannette Papastaikoudi, Antti Petajisto, Jean-Paul Sursock, and Mila Getmansky for excellent research assistance, and Franklin Allen, Lars

Hansen, and participants of the Eighth World Congress of the Econometric Society for helpful comments. Financial support from the MIT Laboratory for Financial Engineering and the National Science Foundation (Grant SBR–9709976) is gratefully acknowledged.

MIT Sloan School of Management, 50 Memorial Dr., Cambridge, MA 02142–1347, and NBER.

References

Ajinkya, B. B. and P. C. Jain (1989), "The Behavior of Daily Stock Market Trading Volume," *Journal of Accounting and Economics*, 11, 331–359.

Allen, F. and R. Karjalainen (1999), "Using Genetic Algorithms to Find Technical Trading Rules," *Journal of Financial Economics*, 51, 245–271.

Amihud, Y. and H. Mendelson (1986a), "Asset Pricing and the Bid-Ask Spread," *Journal of Financial Economics*, 17, 223–249.

Amihud, Y. and H. Mendelson (1986b), "Liquidity and Stock Returns," *Financial Analysts Journal*, 42, 43–48.

Andersen, T. (1996), "Return Volatility and Trading Volume: An Information Flow Interpretation," *Journal of Finance*, 51, 169–204.

Antoniewicz, R. L. (1993), "Relative Volume and Subsequent Stock Price Movements," Working Paper, Board of Governors of the Federal Reserve System.

Atkins, A. and E. Dyl (1997), "Market Structure and Reported Trading Volume: NASDAQ versus the NYSE," *Journal of Financial Research*, 20, 291–304.

Banz, R. (1981), "The Relation Between Return and Market Value of Common Stocks," *Journal of Financial Economics*, 9, 3–18.

Bamber, L. (1986), "The Information Content of Annual Earnings Releases: A Trading Volume Approach," *Journal of Accounting Research*, 24, 40–56.

Black, F. (1976), "Studies of Stock Price Volatility Changes," in *Proceedings of the 1976 Meetings of the Business and Economic Statistics Section*, Washington, DC: American Statistical Association, 177–181.

Black, F., M. Jensen, and M. Scholes (1972), "The Capital Asset Pricing Model: Some Empirical Tests," in *Studies in the Theory of Capital Markets* (ed. by M. Jensen), New York: Praeger.

Blume, L., D. Easley, and M. O'Hara (1994), "Market Statistics and Technical Analysis: The Role of Volume," *Journal of Finance*, 49, 153–181.

Brock, W., J. Lakonishok, and B. LeBaron (1992), "Simple Technical Trading Rules and the Stochastic Properties of Stock Returns," *Journal of Finance*, 47, 1731–1764.

Brown, D. and R. Jennings (1989), "On Technical Analysis," *Review of Financial Studies*, 2, 527–551.

Brown, K., W. Van Harlow, and S. Tinic (1993), "The Risk and Required Return of Common Stock Following Major Price Innovations," *Journal of Financial and Quantitative Analysis*, 28, 101–116.

Campbell, J., S. Grossman, and J. Wang (1993), "Trading Volume and Serial Correlation in Stock Returns," *Quarterly Journal of Economics*, 108, 905–939.

Campbell, J. and A. Kyle (1993), "Smart Money, Noise Trading, and Stock Price Behavior," *Review of Economic Studies*, 60, 1–34.

Campbell, J., A. Lo, and C. MacKinlay (1996), *The Econometrics of Financial Markets*. Princeton, NJ: Princeton University Press.

Chamberlain, G. (1983), "Funds, Factors, and Diversification in Arbitrage Pricing Models," *Econometrica*, 51, 1305–1323.

Chamberlain, G. and M. Rothschild (1983), "Arbitrage, Factor Structure, and Mean-Variance Analysis on Large Asset Markets," *Econometrica*, 51, 1281–1304.

Chan, L., N. Jegadeesh, and J. Lakonishok (1996), "Momentum Strategies," *Journal of Finance*, 51, 1681–1713.

Chan, L. and J. Lakonishok (1995), "The Behavior of Stock Prices Around Institutional Trades," *Journal of Finance*, 50, 1147–1174.

Chang, K. and C. Osler (1994), "Evaluating Chart-Based Technical Analysis: The Head-and-Shoulders Pattern in Foreign Exchange Markets," Working paper, Federal Reserve Bank of New York.

Constantinides, G. M. (1986), "Capital Market Equilibrium with Transaction Costs," *Journal of Political Economy*, 94(4), 842–862.

Cready, W. M. and R. Ramanan (1991), "The Power of Tests Employing Log-Transformed Volume in Detecting Abnormal Trading," *Journal of Accounting and Economics*, 14, 203–214.

Csáki, E. (1984), "Empirical Distribution Function," in *Handbook of Statistics*, Vol. 4 (ed. by P. Krishnaiah and P. Sen), Amsterdam: Elsevier Science.

Dhillon, U. and H. Johnson (1991), "Changes in the Standard and Poor's 500 List," *Journal of Business*, 64, 75–85.

Fama, E. and K. French (1992), "The Cross-Section of Expected Stock Returns," *Journal of Finance*, 47, 427–465.

Gallant, R., P. Rossi, and G. Tauchen (1992), "Stock Prices and Volume," *Review of Financial Studies*, 5, 199–242.

Goetzmann, W. and M. Garry (1986), "Does Delisting from the S&P 500 Affect Stock Prices?" *Financial Analysis Journal*, 42, 64–69.

Grundy, B. and S. Martin (1998), "Understanding the Nature of the Risks and the Source of the Rewards to Momentum Investing," Unpublished Working Paper, Wharton School, University of Pennsylvania.

Hamilton, J. (1994), *Times Series Analysis*. Princeton, NJ: Princeton University Press.

Hamilton, W. (1922), *The Stock Market Barometer*. New York: Wiley.

Harris, L. and E. Gurel (1986), "Price and Volume Effects Associated with Changes in the S&P 500 List: New Evidence for the Existence of Price Pressures," *Journal of Finance*, 46, 815–829.

Heaton, J. and D. J. Lucas (1996), "Evaluating the Effects of Incomplete Markets on Risk Sharing and Asset Pricing," *Journal of Political Economy*, 104(3), 443–487.

He, H. and J. Wang (1995), "Differential Information and Dynamic Behavior of Stock Trading Volume," *Review of Financial Studies*, 8, 919–972.

Hollander, M. and D. Wolfe (1973), *Nonparametric Statistical Methods*, New York: Wiley.

Hu, S. (1997), "Trading Turnover and Expected Stock Returns: Does It Matter and Why?" Working paper, National Taiwan University.

Huang, C.-F. and H. Pages (1990), "Optimal Consumption and Portfolio Policies with an Infinite Horizon: Existence and Convergence," Working Paper, MIT.

Huang, J. and J. Wang (1997), "Market Structure, Security Prices, and Informational Efficiency," *Macroeconomic Dynamics*, 1, 169–205.

Huang, M. (1998), "Liquidity Shocks and Equilibrium Liquidity Premia," Unpublished Working Paper, Graduate School of Business, Stanford University.

Jacques, W. (1988), "The S&P 500 Membership Anomaly, or Would You Join This Club?" *Financial Analysts Journal*, 44, 73–75.

Jain, P. (1987), "The Effect on Stock Price of Inclusion in or Exclusion from the S&P 500," *Financial Analysts Journal*, 43, 58–65.

Jain, P. and G. Joh (1988), "The Dependence Between Hourly Prices and Trading Volume," *Journal of Financial and Quantitative Analysis*, 23, 269–282.

Jegadeesh, N. and S. Titman (1993), "Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency," *Journal of Finance*, 48, 65–91.

Karpoff, J. (1987), "The Relation Between Price Changes and Trading Volume: A Survey," *Journal of Financial and Quantitative Analysis*, 22, 109–126.

Karpoff, J. and R. Walkling (1988), "Short-Term Trading Around Ex-Dividend Days: Additional Evidence," *Journal of Financial Economics*, 21, 291–298.

Karpoff, J. and R. Walkling (1990), "Dividend Capture in NASDAQ Stocks," *Journal of Financial Economics*, 28, 39–65.

Kwiatkowski, D., P. Phillips, P. Schmidt, and Y. Shin (1992), "Testing the Null Hypothesis of Stationarity Against the Alternative of a Unit Root: How Sure Are We That Economic Time Series Have a Unit Root?" *Journal of Econometrics*, 54, 159–178.

Lakonishok, J. and S. Smidt (1986), "Volume for Winners and Losers: Taxation and Other Motives for Stock Trading," *Journal of Finance*, 41, 951–974.

Lakonishok, J. and T. Vermaelen (1986), "Tax-Induced Trading Around Ex-Dividend Days," *Journal of Financial Economics*, 16, 287–319.

Lamoureux, C. and J. Wansley (1987), "Market Effects of Changes in the Standard & Poor's 500 Index," *Financial Review*, 22, 53–69.

LeBaron, B. (1992), "Persistence of the Dow Jones Index on Rising Volume," Working Paper, University of Wisconsin.

Lim, T., A. Lo, J. Wang, and P. Adamek (1998), "Trading Volume and the MiniCRSP Database: An Introduction and User's Guide," Working Paper LFE–1038–98, MIT Laboratory for Financial Engineering.

Llorente, G., R. Michaely, G. Saar, and J. Wang (2000), "Dynamic Volume-Return Relations for Individual Stocks," Working Paper, MIT.

Lo, A. and C. MacKinlay (1988), "Stock Market Prices Do Not Follow Random Walks: Evidence from a Simple Specification Test," *Review of Financial Studies*, 1, 41–66.

Lo, A. and C. MacKinlay (1997), "Maximizing Predictability in the Stock and Bond Markets," *Macroeconomic Dynamics*, 1, 102–134.

Lo, A. and C. MacKinlay (1999), *A Non-Random Walk Down Wall Street*. Princeton, NJ: Princeton University Press.

Lo, A., H. Mamaysky, and J. Wang (2000a), "Asset Prices and Trading Volume under Fixed Transaction Costs," Working Paper, MIT.

Lo, A., H. Mamaysky, and J. Wang (2000b), "Foundations of Technical Analysis: Computational Algorithms, Statistical Inference and Empirical Implementation," *Journal of Finance*, 55, 1705–1765.

Lo, A. and J. Wang (2000a), Trading Volume: Definitions, Data Analysis, and Implications of Portfolio Theory, *Review of Financial Studies*, 13, 257–300.

Lo, A. and J. Wang (2000b), "Trading Volume: Implications of an Intertemporal Capital Asset-Pricing Model," Work in Progress, MIT.

Lynch-Koski, J. (1996), "A Microstructure Analysis of Ex-Dividend Stock Price Behavior Before and After the 1984 and 1986 Tax Reform Acts," *Journal of Business*, 69, 313–338.

Marsh, T. and R. Merton (1987), "Dividend Behavior for the Aggregate Stock Market," *Journal of Business*, 60, 1–40.

Merton, R. (1971), "Optimal Consumption and Portfolio Rules in a Continuous-Time Model," *Journal of Economic Theory*, 3, 373–413.

Merton, R. (1973), "An Intertemporal Capital Asset Pricing Model," *Econometrica*, 41, 867–887.

Merton, R. (1987), "A Simple Model of Capital Market Equilibrium with Incomplete Information," *Journal of Finance*, 42, 483–510.

Michaely, R. (1991), "Ex-Dividend Day Stock Price Behavior: The Case of the 1986 Tax Reform Act," *Journal of Finance*, 46, 845–860.

Michaely, R. and M. Murgia (1995), "The Effect of Tax Heterogeneity on Prices and Volume Around the Ex-Dividend Day: Evidence from the Milan Stock Exchange," *Review of Financial Studies*, 8, 369–399.

Michaely, R. and J. Vila (1995), "Investors' Heterogeneity, Prices and Volume around the Ex-Dividend Day," *Journal of Financial and Quantitative Analysis*, 30, 171–198.

Michaely, R. and J. Vila (1996), "Trading Volume with Private Valuation: Evidence from the Ex-Dividend Day," *Review of Financial Studies*, 9, 471–509.

Michaely, R., J.-L. Vila, and J. Wang (1996), "A Model of Trading Volume with Tax-Induced Heterogeneous Valuation and Transaction Costs," *Journal of Financial Intermediation*, 5, 340–371.

Morse, D. (1980), "Asymmetric Information in Securities Markets and Trading Volume," *Journal of Financial and Quantitative Analysis*, 15, 1129–1148.

Muirhead, R. (1982), *Aspects of Multivariate Statistical Theory*. New York: Wiley.

Neely, C., P. Weller, and R. Dittmar (1997), "Is Technical Analysis in the Foreign Exchange Market Profitable? A Genetic Programming Approach," *Journal of Financial and Quantitative Analysis*, 32, 405–426.

Neely, C. and P. Weller (1998), "Technical Trading Rules in the European Monetary System," Working Paper, Federal Bank of St. Louis.

Neftci, S. (1991), "Naive Trading Rules in Financial Markets and Wiener-Kolmogorov Prediction Theory: A Study of Technical Analysis," *Journal of Business*, 64, 549–571.

Neftci, S. and A. Policano (1984), "Can Chartists Outperform the Market? Market Efficiency Tests for 'Technical Analyst'," *Journal of Futures Markets*, 4, 465–478.

Ohlson, J. and B. Rosenberg (1976), "The Stationary Distribution of Returns and Portfolio Separation in Capital Markets: A Fundamental Contradiction," *Journal of Financial and Quantitative Analysis*, 11, 393–402.

Osler, C. O. and K. Chang (1995), "Head and Shoulders: Not Just a Flaky Pattern," Staff Report No. 4, Federal Reserve Bank of New York.

Press, W., B. Flannery, S. Teukolsky, and W. Vetterling (1986), *Numerical Recipes: The Art of Scientific Computing*. Cambridge: Cambridge University Press.

Pruitt, S. and J. Wei (1989), "Institutional Ownership and Changes in the S&P 500," *Journal of Finance*, 44, 509–513.

Pruitt, S. and R. White (1988), "The CRISMA Trading System: Who Says Technical Analysis Can't Beat the Market?" *Journal of Portfolio Management*, 14, 55–58.

Reinganum, M. (1992), "A Revival of the Small-Firm Effect," *Journal of Portfolio Management*, 18, 55–62.

Richardson, G., S. Sefcik, and R. Thompson (1986), "A Test of Dividend Irrelevance Using Volume Reaction to a Change in Dividend Policy," *Journal of Financial Economics*, 17, 313–333.

Roll, R. (1984), "A Simple Implicit Measure of the Effective Bid-Ask Spread in an Efficient Market," *Journal of Finance*, 39, 1127–1140.

Ross, S. A. (1989), "Discussion: Intertemporal Asset Pricing," in *Theory of Valuation* (ed. by S. Bhattacharya and G. Constantinides), Totowa, NJ: Rowman & Littlefield, 85–96.

Rouwenhorst, G. (1998), "International Momentum Strategies," *Journal of Finance*, 53, 267–284.

Schroeder, M. (1998), "Optimal Portfolio Selection with Fixed Transaction Costs," Working Paper, Northwestern University.

Shleifer, A. (1986), "Do Demand Curves for Stocks Slope Down?" *Journal of Finance*, 41, 579–590.

Smirnov, N. (1939a), "Sur les Écarts de la Courbe de Distribution Empirique," *Rec. Math.* (*Mat. Sborn.*), 6, 3–26.

Smirnov, N. (1939b), "On the Estimation of the Discrepancy Between Empirical Curves of Distribution for Two Independent Samples," *Bulletin. Math. Univ. Moscow*, 2, 3–14.

Stickel, S. (1991), "The Ex-Dividend Day Behavior of Nonconvertible Preferred Stock Returns and Trading Volume," *Journal of Financial and Quantitative Analysis*, 26, 45–61.

Stickel, S. and R. Verrecchia (1994), "Evidence That Volume Sustains Price Changes," *Financial Analysts Journal*, 50 November–December, 57–67.

Tabell, A. and E. Tabell (1964), "The Case for Technical Analysis," *Financial Analysts Journal*, 20, 67–76.

Tkac, P. (1996), "A Trading Volume Benchmark: Theory and Evidence," Working Paper, Department of Finance and Business Economics, University of Notre Dame.

Treynor, J. and R. Ferguson (1985), "In Defense of Technical Analysis," *Journal of Finance*, 40, 757–773.

Vayanos, D. (1998), "Transaction Costs and Asset Prices: A Dynamic Equilibrium Model," *Review of Financial Studies*, 11(1), 1–58.

Wang, J. (1994), "A Model of Competitive Stock Trading Volume," *Journal of Political Economy*, 102, 127–168.

Woolridge, J. and C. Ghosh (1986), "Institutional Trading and Security Prices: The Case of Changes in the Composition of the S&P 500 Market Index," *Journal of Financial Research*, 9, 13–24.

Wu, G. J. and C. S. Zhou (2001), "A Model of Active Portfolio Management," Working Paper, University of Michigan.

A Discussion of the Papers by John Geanakoplos and by Andrew W. Lo and Jiang Wang

Franklin Allen

1. INTRODUCTION

At first sight, the two papers in this section seem unrelated. The one by John Geanakoplos is about the role of collateral in explaining liquidity crises and crashes. Andrew Lo and Jiang Wiang's paper is concerned with a theoretical and empirical analysis of trading volume. However, on closer inspection, they have important interrelationships and provide an interesting contrast. They both trace their intellectual roots back to the Arrow–Debreu model, yet they represent two very different approaches to financial economics, both of which are widely used.

The papers investigate deviations from the standard perfect-markets assumption. Frictions are incorporated in a different way, though. In the Geanakoplos paper, the problem is preventing default, and collateral is the way this is achieved. In the Lo and Wang paper, there are asymmetric information and fixed transaction costs. What is more important is that the motivations for trade are quite different. In the Geanakoplos paper, it is differences in beliefs; in the Lo and Wang paper, it is different shocks to nonfinancial income. Both of these assumptions are crucial to the results the authors obtain. They represent quite different traditions.

The Geanakoplos paper is a mainstream general equilibrium paper. In the Arrow–Debreu model, the possibility of differences in preferences is an important component of the model. Because beliefs are embedded in preferences, allowing different beliefs is a standard assumption.

In finance, and in particular in asset pricing, allowing different beliefs is currently viewed as a nonstandard assumption. In explaining trading volume, Lo and Wang briefly mention asymmetric information but do not consider differences in beliefs arising from differences in priors, which is distinct.

This difference in standard assumptions is an interesting phenomenon. Brennan (1989) argues that the reason it took so long from the time Markowitz developed mean-variance analysis to the discovery of the Capital Asset Pricing Model was the boldness of the assumption that everybody has the same beliefs. In the 1960s, the notion that people could have different beliefs was quite acceptable in mainstream papers. Lintner's (1969) widely quoted paper on a variant of the CAPM with heterogeneous beliefs and Ross' (1976) arbitrage pricing theory provide good examples. However, since then, views appear to have changed. In recent years there have been few asset-pricing papers that adopt this assumption. Harris and Raviv (1993) provided one of these, but it has not been widely imitated. Morris (1995) provides a very nice summary of the arguments in favor of allowing for differences in prior beliefs. In addition, there is some empirical evidence that differences in beliefs are important in practice. Kandel and Person's (1995) results suggest that trading around earnings announcements is due to differences in priors.

Section 2 of this discussion considers John Geanakoplos's paper, and Section 3 considers the paper by Andrew Lo and Jiang Wang. Section 4 gives conclusions.

2. THE GEANAKOPLOS PAPER

This paper builds a theory to help understand the liquidity crises and crashes that occurred in fixed income markets in 1994 and in 1998. These were characterized by a price crash in defaultable assets that was not followed by an increase in subsequent defaults. There were spillovers to other markets, margin requirements were raised, and borrowing decreased. The paper builds on the research by Dubey, Geanakoplos, and Shubik (2001), Dubey and Geanakoplos (2001a, 2001b), Geanakoplos (1997), and Geanakoplos and Zame (1998).

The starting point of the model is that the possibility of default necessitates the use of collateral. The natural buyers of an asset, who are the people that value it the most, may have to borrow to acquire it and must post collateral as security for their loans. Collateral is liquid wealth that is in the form of physical assets that can be stored. The liquidity cost of buying an asset is the margin requirement for an asset multiplied by its price. When agents choose their bundle of goods, there are then two constraints. The first is the standard budget constraint that requires that the value of a person's expenditures must not exceed her or his wealth. The second is a liquidity constraint. This requires that the liquidity needed to enable a person to purchase her or his bundle must not exceed her or his liquid wealth.

Incorporating such collateral requirements and liquidity constraints into a general equilibrium analysis, Geanakoplos and Zame (1998) have demonstrated existence and constrained efficiency of equilibrium. The current paper focuses on developing an example to show how these features lead to liquidity crises and crashes. With collateral requirements and heterogeneous beliefs, asset prices depend in an important way on the distribution of wealth. If relatively optimistic buyers are wealthy enough, they will be able to borrow and purchase all of the asset and will be the marginal holders. As a result, its price will be high. If bad news about the asset payoffs arrives, its price can fall for two reasons. The first is because of the fall in expected cash flows. The second is that there is a redistribution of wealth and as a result the marginal holders may no longer be from the relatively optimistic group. Instead, the marginal holders may now belong to a relatively pessimistic group and the fall in price may be considerably amplified. The volatility that results may lead to an increase in margin requirements, and this can further amplify the change in price if the marginal holder switches to an even more pessimistic group. As a result, small changes in expectations can have a big effect.

The paper is a nice contribution to the literature on financial crises. In many cases, relatively small changes in fundamentals seem to cause large crises. The paper is part of an important and growing literature that explains how this can happen. One strand of this literature uses models with multiple equilibria, in which a crisis is modeled as a switch in equilibrium. The difficulty with this approach is that there are not many good theories of equilibrium selection. An exception is from Morris and Shin (1998), who are able to show in the context of currency crises how a lack of common knowledge can lead to a unique selection of equilibrium. However, Hellwig (2000) has shown that their result depends in an important way on how a lack of common knowledge is introduced. Choosing a different information structure can lead to a different selection of equilibrium.

The alternative to approaches based on multiple equilibria is the use of models in which there is amplification. Examples of such models are given by Kiyotaki and Moore (1997), Chari and Kehoe (1999), and Allen and Gale (2001). The Geanakoplos paper belongs to this strand of the literature. It is a very plausible explanation of the liquidity crises and crashes in the fixed income market that occurred in 1994 and 1998.

One of the important issues the analysis raises is the role of the central bank in providing liquidity. In practice, its actions in providing liquidity through the banking system appear to be crucial in determining asset prices. This aspect is not included in the model here. An important extension would be to do so.

3. THE LO AND WANG PAPER

There is a great amount of literature on the theory of asset pricing and on tests of these theories. As the authors point out, by contrast, there is a small literature on the volumes that are traded in financial markets. The paper seeks to change this. The authors start by providing a fairly standard model of dynamic trading. They then use this as the basis of their empirical analysis. The effects of incorporating asymmetric information and fixed transaction costs are also investigated. Finally, they consider technical analysis, as this has always placed considerable emphasis on volume. The paper builds on three of the authors' previous contributions (Lo and Wang, 2000 and Lo, Mamaysky, and Wang, 2000a, 2001b).

The dynamic trading model that is developed has risk-averse investors that maximize expected utility. They bear risk from stochastic asset returns and also from nonfinancial income such as labor income that is positively correlated with stock dividends. The model exhibits four-fund separation. The first two funds are the risk-free asset and the market portfolio. The third is a hedging portfolio that allows investors to hedge against nonfinancial risk. The fourth is another hedging portfolio that allows investors to hedge against changes in market risks driven by changes in aggregate exposure to nonfinancial risk. It is also shown that the model leads to a dynamic volume-return relationship in which returns accompanied by high volume are more likely to be reversed in the future. This result relies on the assumption of symmetric information. Asymmetric information can reverse it.

Using a volume measure based on turnover, Lo and Wang's empirical work indicates that cross-sectional variation in turnover does seem related to stock-specific characteristics such as risk, size, price, trading costs, and S&P 500 membership. With regard to $K + 1$ fund separation, they demonstrate that the first K principal components of the covariance matrix of turnover should explain most of the time-series variation in turnover, if turnover is driven by a linear K-factor model. They find that a one-factor model is a reasonable approximation and a two-factor model captures over 90 percent of the time-series variation in turnover. They also find that the dynamic volume-return relationship suggested by their theoretical model is generally supported by their empirical analysis.

In the theoretical model, investors can trade in the market continuously with no transaction costs. When information flow to the market is continuous in such models, trading volume is infinite. In the empirical analysis it is assumed that trading occurs at finite intervals. When fixed transaction costs are incorporated into the analysis, the level of volume depends critically on their size, as does the illiquidity discount in stock prices. This contrasts with the existing literature, which finds that transaction costs do not have a significant impact on trading volume or illiquidity discounts. It is shown that this version of the model is consistent with empirically plausible trading volumes and illiquidity discounts.

The final part of the paper considers technical analysis. Lo and Wang first develop an algorithm that can identify technical patterns such as "head and shoulders" and various types of "tops" and "bottoms." The effectiveness of the technical analysis is analyzed indirectly by conditioning on the occurrence of a technical pattern and on declining and increasing volume. The conditional and unconditional distributions are compared, and a difference is interpreted as indicating that technical patterns are informative. When they apply these to many stocks over many time periods, Lo and Wang find that certain technical patterns do provide incremental information, especially for NASDAQ stocks. The volume trends also provide incremental information in some cases.

The work described in the paper is a good contribution to the literature on trading volume. It provides a benchmark analysis. One of the most important questions it raises is how much trading volume is due to the types of factors modeled, such as risk sharing and shocks to nonfinancial income. There are many other plausible alternatives that seem at least as likely to be important as any of these factors.

The first is differences in prior beliefs, which play such an important role in the Geanakoplos paper. Casual observations such as the variation in analysts' recommendations suggest that this factor can explain a significant amount of trading. This motivation for trade contrasts with asymmetric information in that people do not try to deduce informed people's information. They agree to disagree.

Allen (2001) argues that it is not surprising that asset-pricing models based on the assumption that individuals invest their own wealth are unable to explain many asset-pricing anomalies, such as the Internet and technology "bubbles," when in practice institutions own the majority of stocks. A similar argument applies here. The agency issues that arise with delegated portfolio management such as churning are likely to explain a significant proportion of volume. Other important factors include the role of derivative markets and whether they are complements or substitutes and the extent to which markets are effectively arbitraged.

The authors are right to stress the paucity of the asset-quantities literature both in absolute terms and relative to the asset-pricing literature. One would hope that this paper will help to stimulate a larger literature that will identify the relative importance of risk sharing, differences in prior beliefs, institutional investment, and other relevant factors.

4. CONCLUDING REMARKS

As the discussion herein has indicated, both of these papers are well worth reading. The fact that they have common intellectual roots in the Arrow–Debreu model but take such a different approach is also revealing. In many areas of financial economics, particularly asset pricing, there is a great deal of rigidity in what is permissible in terms of standard assumptions. When changes such as the move toward behavorial finance do come, they tend to be large. More gradual changes such as considering the effect of differences in beliefs and the role of agency problems should also be considered.

References

Allen, F. (2001), "Do Financial Institutions Matter?" *Journal of Finance*, 56, 1165–1175.

Allen, F. and D. Gale (2001), "Financial Fragility," Working Paper, Wharton Financial Institutions Center, University of Pennsylvania.

Brennan, M. (1989), "Capital Asset Pricing Model," in *The New Palgrave Dictionary of Economics* (ed. by J. Eatwell, M. Milgate, and P. Newman), New York: Stockton Press.

Chari, V. and P. Kehoe (2000), "Financial Crises as Herds," Working Paper, Federal Reserve Bank of Minneapolis.

Dubey, P. and J. Geanakoplos (2001a), "Signalling and Default: Rothschild–Stiglitz Reconsidered," Discussion Paper 1305, Cowles Foundation.

Dubey, P. and J. Geanakoplos (2001b), "Insurance Contracts Designed by Competitive Pooling," Discussion Paper 1315, Cowles Foundation.

Dubey, P., J. Geanakoplos, and M. Shubik (2001), "Default and Punishment in General Equilibrium," Discussion Paper 1304, Cowles Foundation.

Geanakoplos, J. (1997), "Promises, Promises," in *The Economy as an Evolving Complex System, II* (ed. by W. Arthur, S. Durlauf, and D. Lane), Reading, MA: Addison-Wesley, 285–320.

Geanakoplos, J. and W. Zame (1998), "Default, Collateral, and Crashes," Working Paper, Yale University.

Harris, M. and A. Raviv (1993), "Differences of Opinion Make a Horse Race," *Review of Financial Studies*, 6, 473–506.

Hellwig, C. (2000), "Public Information, Private Information and the Multiplicity of Equilibria in Coordination Games," Working Paper, London School of Economics.

Kandel, E. and N. Pearson (1995), "Differential Interpretation of Public Signals and Trade in Speculative Markets," *Journal of Political Economy*, 103, 831–872.

Kiyotaki, N. and J. Moore (1997), "Credit Chains," *Journal of Political Economy*, 105, 211–248.

Lintner, J. (1969), "The Aggregation of Investor's Diverse Judgments and Preferences in Purely Competitive Security Markets," *Journal of Financial and Quantitative Analysis*, 4, 347–400.

Lo, A., H. Mamaysky, and J. Wiang (2000a), "Asset Prices and Trading Volume Under Fixed Transaction Costs," Working Paper, MIT.

Lo, A., H. Mamaysky, and J. Wiang (2000b), "Foundations of Technical Analysis: Computational Algorithms, Statistical Inference, and Empirical Implementation," *Journal of Finance*, 55, 1705–1765.

Lo, A. and J. Wang (2000), "Trading Volume: Definitions, Data Analysis, and Implications of Portfolio Theory," *Review of Financial Studies*, 13, 257–300.

Morris, S. (1995), "The Common Prior Assumption in Economic Theory," *Economics and Philosophy*, 11, 227–253.

Morris, S. and H. Shin (1998), "Unique Equilibrium in a Model of Self-Fulfilling Currency Attacks," *American Economic Review*, 88, 587–597.

Ross, S. (1976), "The Arbitrage Pricing Theory of Capital Asset Pricing," *Journal of Economic Theory*, 13, 341–360.

CHAPTER 7

Inverse Problems and Structural Econometrics

The Example of Instrumental Variables

Jean-Pierre Florens

1. INTRODUCTION

The development of nonparametric estimation in econometrics has been extremely important over the past fifteen years. Inference was first concentrated on the data's distribution, described, for example, by its density or by its hazard function, or by some characteristics of the conditional distributions, such as the conditional expectations. This approach is typically a reduced-form analysis oriented to sophisticated data description, even if the selection of conditioning variables may depend on a theoretical model. On the other side, the structural econometric analysis is focused on the estimation of the (possibly functional) parameters that describe the economic agent's behavior and that are not, in general, "simple" transformations of the sampling distribution. An excellent state-of-the-art discussion of nonparametric econometrics is given by Pagan and Ullah (1999) (see also Aït Sahalia (1995) for a general "reduced form analysis" in the nonparametric framework).

A first objective of this paper is to introduce a general framework for structural functional inference in connection with the inverse-problems literature. An inverse problem is the resolution of a functional equation with a particular attention to the sensitivity of the solution to possible errors in the specification of the equation due for instance to an estimation procedure (see, e.g., for recent surveys of the literature, Colton et al., 2000).

More specifically, we analyze linear inverse problems in which the parameter of interest is a function φ solution of a linear equation $K_F\varphi = \psi_F$ in which both the linear operator K_F and the right-hand side depend on the (unknown) distribution F of the sample. This linear problem may be ill posed if K_F does not admit a continuous inverse, and this problem must be regularized (see Tikhonov and Arsenin, 1977 or Wahba, 1973).

One of the fundamental questions of structural econometrics is the treatment of endogeneity. This question is addressed in terms that are different from the definition of exogeneity based on the notion of cut (see Engle, Henry, and Richard, 1983 and Florens and Mouchart, 1985). The problem is to define a relation, such as $Y = \varphi(Z) + U$, in the absence of the "exogeneity assumption," $E(U|Z) = 0$. Different possible definitions are given in the paper, and the instrumental variable definition is featured – $E(U|W) = 0$, where W is a set of instruments. This presentation is more in the tradition of Frisch (1934, 1938), Reiersol (1941, 1945), Sargan (1958), Basmann (1959), or Theil (1953).

Practical implementation of the solution of a linear inverse problem is developed, and we finally present a synthesis of some of our previous works on the asymptotic properties of the Tikhonov regularization of the solution of an ill-posed linear inverse problem.

2. FUNCTIONAL STRUCTURAL ECONOMETRICS AND INVERSE PROBLEMS

A structural analysis in nonparametric (i.e., with functional parameters) econometrics can be introduced by considering the following elements.

1. The functional parameter of interest is denoted by φ, and this unknown function is an element of a Banach space Φ.
2. The observation mechanism is characterized by a random element S, which is generally a random vector in $\mathbb{R}^m$ but could also be infinite dimensional. The probability measure of S is defined by a cumulative distribution function F. This cumulative distribution function (CDF) is an element of a topological space $\mathcal{F}$. This econometrician observes a sample $s_1, \dots, s_N$ of S. In the last sections of this paper we essentially consider independently identically distributed (*iid*) samples, but the extension to weakly dependent observations (e.g., strong mixing stationary processes) does not deeply modify our analysis.
3. The economic model defines the parameter of interest φ and connects this parameter to the probability distribution F of the sample by a functional equation:
$$A(\varphi, F) = 0, \tag{2.1}$$
where A is an operator defined on $\Phi \times \mathcal{F}$ and valued in a Banach space $\mathcal{E}$. The main feature of this presentation is that φ is implicitly related to F, which allows us to set fundamental questions of structural econometrics as identification, that is, unicity of the solution of (2.1) for given F, or overidentification, that is, existence of the solution. Statistical nonparametric inference or reduced-form analysis is generally concerned with explicit definitions of the parameter of interest, such as the regression function or the cumulative hazard function.

In this paper, we call the three elements Φ, $\mathcal{F}$, and A the Structural Functional Model. This definition is illustrated by the following examples. In this section, only nonlinear examples are given. Linear examples are considered in Section 3.

Example 2.1 (Conditional-moment condition). *This example covers a large class of particular cases. It gives a natural way to specify a relation between φ and F. Let us assume $S = (Y, Z) \in \mathbb{R}^m$ is a random vector, and h is an operator* *defined on* $\mathbb{R}^m \times \Phi$ *and valued in* $\mathbb{R}^r$. *We assume that h is integrable for any* φ *and we defined A by*

$$A(\varphi, F) = E^F(h(S, \varphi)|Z = z).$$

The usual (conditional) moment condition is obtained where φ *is finite dimensional* ($\Phi \subset \mathbb{R}^k$), *and this example also covers the marginal moment condition* $E^F(h(S, \varphi)) = 0$. *Following the Hansen (1982) paper, a huge literature examines this condition (see, e.g., Hall, 1993).*

Most of this literature considers finite-dimensional parameters and a finite number of moment conditions, but infinite-dimensional extensions are given by Carrasco and Florens (2000a).

Moment or conditional-moment conditions are generally derived from economic models by assuming that the first-order conditions of optimization programs that characterized the behavior of economic agents are satisfied on average (see, e.g., Ogaki, 1993).

Example 2.2 (Surplus analysis and nonlinear differential operators). *Let us assume* $S = (Y, Z, W) \in \mathbb{R}^3$ *and define*

$$m_F(z, w) = E(Y|Z = z, W = w).$$

This conditional-expectation function is assumed to be smooth and the parameter of interest φ *is a differentiable function from* $\mathbb{R}$ *to* $\mathbb{R}$. *This function is assumed to be a solution of a Cauchy–Lipschitz differential equation,*

$$\varphi'(z) = m_F(z, \varphi(z)),$$

under a boundary condition $\varphi(z_0) = a_0$. *In that case,*

$$A(\varphi, F) = \varphi' - m_F(\cdot, \varphi),$$

and $\mathcal{E}$ *is the set of real variable real-valued functions.*

A nonparametric estimation of the surplus function of a consumer gives an example of this functional equation. Following Hausman (1981, 1985) and Hausman and Newey (1995), we see that the surplus function φ *satisfies the equation*

$$\varphi'(z) = m_F(z, w_0 - \varphi(z)),$$

where Y is the consumption of a good, Z is the price, W is the revenue of the consumer, m_F *is the demand function, and* (z_0, w_0) *is an initial value of the price and the revenue. The boundary condition assumes that* $\varphi(z_0) = 0$. *A general treatment of functional parameters solutions of Cauchy–Lipschitz differential equations and others applications is given by Vanhems (2000).*

Example 2.3 (Game theoretic model). *Here we consider incomplete information symmetric games that can be simplified in the following way. A player of a game receives a private signal* $\xi \in \mathbb{R}$ *and plays an action* $S \in \mathbb{R}$. *We consider* *cases in which the ξ are iid generated for all the players and all the games and the distribution of any ξ, characterized by its CDF φ, is common knowledge for the players. Actions are related to signals by a strategy function*

$$S = \sigma_\varphi(\xi),$$

which is obtained, for example, as a Nash equilibrium and depends on the CDF φ. For simplicity, σ_φ is supposed to be one to one and increasing. The econometrician observes a sample of the action S but ignores the signals, and the parameter of interest is φ. The strategy function (as a function of ξ and φ) is known. Let F be the CDF of the actions. This distribution satisfies $F = \varphi \circ \sigma_\varphi^{-1}$ and the operator A can be defined by

$$A(\varphi, F) = \varphi - F \circ \sigma_\varphi.$$

The private value first-price auction model gives a particular case of this class of examples. In this case, the strategy function verifies

$$\sigma_\varphi(\xi) = \xi - \frac{\int_{\xi 0}^{\xi} \varphi(u)^K du}{\varphi(\xi)^K},$$

where the number of bidders is $K + 1$ and $\xi \in [\xi_0, \xi_1] \subset \mathbb{R}$. This example was treated in numerous papers (see Guerre, Perrigne, and Vuong, 2000 for a recent nonparametric analysis). A general treatment of the game theoretic models (including several extensions) is given by Florens, Protopopescu, and Richard (1997).

For a given F, φ is identified if two solutions of (2.1) are necessarily equal and φ is locally identified if, for any solution, there exists a neighborhood in which no other solution exists. Local identification is a useful concept in nonlinear cases. If A is differentiable in the Frechet sense, the implicit function theorem (for a discussion of several differentiability concepts in relation to the implicit function theorem, see Van der Vaart and Wellner, 1996) gives a sufficient condition for local identifiability. If (φ, F) satisfies $A(\varphi, F) = 0$, let us compute the Frechet derivative of A with respect to φ at (φ, F). This derivative is a linear operator from Φ to $\mathcal{E}$, and if this linear operator is one to one, local identification in a neighborhood of φ is warranted. (For application at the game theoretic models, see Florens et al., 1997 and Florens and Sbai, 2000.)

Identifiability or local identifiability is typically a property of F. Its analysis in specific models should exhibit conditions on F that imply identification. It is natural to construct models such that identification is satisfied for the true CDF (i.e., the data-generating process (DGP)). However, in numerous particular cases, identification is not verified for the estimated $\hat{F}_N$ (which is, in general, the empirical CDF or a smooth regularization of the empirical CDF). Linear models will provide examples of this lack of identification, and solutions are given in Section 4.

Existence of a solution to Equation (2.1) is also a property of F. If a solution exists for F in a strict subset of $\mathcal{F}$ only, the model will be said to be overidentified. In that case, it is natural to assume that the true DGP, F_0 satisfies the existence condition, but in general the equation $A(\varphi, \hat{F}_N) = 0$ has no solution where $\hat{F}_N$ is a usual unconstrained estimator.

If there exists a neighborhood of the true F_0 such that a solution of $A(\varphi, F) = 0$ exists for any F_* in this neighborhood and if $\hat{F}_N$ converges to F_0 (relatively to the same topology), then overidentification will necessarily disappear for a finite (possibly large) sample size and is not a major issue (this is, for example, the case in the private value first-price auction model). However, in general, a solution does not exist for any sample size. Two types of treatments to this problem are adopted (see Manski, 1988). The first one consists of a modification of the original definition of the parameter of interest. For example, φ becomes the argmin of $\|A(\varphi, F)\|$ instead of the solution of $A(\varphi, F) = 0$ or is the solution of a new functional equation $A_*(\varphi, F) = 0$, which extends the original one. This solution is essentially adopted in the generalized method of moments (GMM) estimation, and our analysis belongs to this methodology. A second way to beat overidentification is to constrain the estimation of F in order to satisfy existence conditions. This is done in finite-dimensional parameter estimation by using unequal weights to the observations (see Owen, 1990, Qin and Lawless, 1994, and Kitamura and Stutzer, 1997).

3. LINEAR INVERSE PROBLEMS

Here we analyze particular models in which the equation $A(\varphi, F) = 0$ is linear (up to an additive term) in φ.

The presentation is simplified by assuming that Φ is a Hilbert space. Let us consider another Hilbert space Ψ. A linear structural model is defined by the equation

$$A(\varphi, F) = K_F\varphi - \psi_F = 0, \tag{3.1}$$

where $\psi_F \in \Psi$ and K_F is a linear operator from Φ to Ψ. Both the linear operator and the constant term depend in general on $F \in \mathcal{F}$.

Linear operators constitute a very large class of transformations of φ. Important families of operators are integral operators and differential operators, and the properties of Equation (3.1) will depend on topological properties of K_F (e.g., continuity or compactness). This diversity is illustrated by the following examples.

Example 3.1 (Density). *As noticed by Hardle and Linton (1994), density estimation may be seen as a linear inverse problem defined, in the real case* ($S \in \mathbb{R}$), *by*

$$\int_{-\infty}^{s} \varphi(u)du = F(s).$$

In that case φ is the density of F with respect to (wrt) the Lebesgue measure, and K_F is an integral operator. This presentation is interesting because it will be used to point out that density estimation is an ill-posed problem (in a sense that is defined later on).

Example 3.2 (Differential operators). *Let us assume that φ is a continuously differentiable function from $\mathbb{R}$ to $\mathbb{R}$ and that the model is characterized by*

$$\varphi^{(p)} + \alpha_{1F}\varphi^{(p-1)} + \cdots + \alpha_{P_F}\varphi = \psi_F,$$

where $\varphi^{(k)}$ is the kth derivative of φ and α_{JF} are functions dependent on F. The solution is constrained to a set of limit conditions. Extensions to partial differential operators in the case of functions of several variables can also be considered. For example, let us consider the case where $S = (Y, Z, W)$, $\psi_F(z, w) = E(Y|Z = z, W = w)$, and $\varphi(z, w)$ is a smooth function of two variables. If $A(\varphi, F) = (\partial/\partial_z)\,\varphi$ (with a boundary condition $\varphi|z_0, w_0 = y_0$) the solution is the integral of a regression (see Florens and Vanhems, 2000, for an application). Extension to some partial differential equations is given in dynamic models by Banon (1978) and Aït-Sahalia (1996).

Example 3.3 (Backfitting estimation in additive nonparametric regression). *Let $S = (Y, X_1, X_2) \in \mathbb{R} \times \mathbb{R}^p \times \mathbb{R}^q$. The parameter of interest φ is $(\varphi_1, \varphi_2) \in L^2(X_1) \times L^2(X_2)$; $L^2(X_1)$ and $L^2(X_2)$ are the Hilbert spaces of square integrable functions of X_1 and X_2, respectively). The underlying probability measure used for the definition of these spaces is the DGP. The functions φ_1 and φ_2 are defined as the functions that minimize $E(Y - \varphi_1(X_1) - \varphi_2(X_2))^2$ or equivalently that are the solution of the linear inverse problem (see, e.g., Hastie and Tibshirani, 1990):*

$$\begin{aligned}\varphi_1(x_1) + E(\varphi_2(X_2)|X_1 = x_1) &= E(Y|X_1 = x_1),\\ E(\varphi_1(X_1)|X_2 = x_2) + \varphi_2(x_2) &= E(Y|X_2 = x_2).\end{aligned}$$

Example 3.4 (Repeated measurement model). *This example is extremely similar to example 3.3. Suppose we have two ways to measure an unobserved value η. The measurement equations are given by $Y_j = \eta + \varphi(X_j) + u_j (j = 1, 2)$, where u_j is a zero-mean (given the X_js) error term and φ is a bias function depending on observable variables X_j. The order of the measurements is not relevant; (Y_1, X_1, Y_2, X_2) is distributed as (Y_2, X_2, Y_1, X_1). We observe an iid sample of (Y_1, X_1, Y_2, X_1) corresponding to an iid sample of η. The unknown value η is eliminated by difference and if $Y = Y_2 - Y_1$, it follows that $E(Y|X_1 = x_1, X_2 = x_2) = \varphi(x_2) - \varphi(x_1)$, where φ is a square integrable function. This relation implies that φ is a solution of*

$$\varphi(x) - E(\varphi(X_1)|X_2 = x) = E(Y|X_2 = x),$$

which is a particular case of (3.1). The function φ is then used to forecast η by $Y_j - \varphi(X_j)$. For details and applications, see Gaspar and Florens (1998).

Note that if the joint distribution of (X_1, X_2) has a density $f(x_1, x_2)$ wrt the Lebesgue measures, the previous equation may be rewritten as

$$\varphi(x) - \int \varphi(u)\frac{f(u,x)}{f(x)}du = r_F(x),$$

where $r_F(x) = E(Y|X_1 = x)$. This equation is a Fredholm type II equation (see, e.g., Tricomi, 1985 and Debrath and Mikusinski, 1999). The system of equations that characterize φ_1 and φ_2 in Example 3.3 is also a Fredholm type II integral equation.

Example 3.5. *The following example is motivated by the extension of GMM to a continuous number of moment conditions. It also applies to regressions with a continuous number of regressors.*

Let us consider $u(S,t)$ a function of $t \in [0,1]$ dependent on the random element S and $h(S,\tau,t)$ a function of $(\tau,t) \in [0,1] \times [0,1]$, also S dependent. The parameter of interest is a function $\varphi(t)$ defined on $[0,1]$, real valued and a solution of

$$\int E^F(h(S,\tau,t))\varphi(\tau)d\tau = E^F(u(S,t)).$$

This equation is a Fredholm type I integral equation. It is the natural extension of a linear equation system from finite dimension to infinite dimension. Despite this simplicity, this type of equation raises a complex question, as we see in Section 4. This equation is motivated by models with continuous numbers of regressors. Consider, for example, the model

$$Y = \int_0^1 X(t)\beta(t)dt + U,$$

where the regressors are indexed by $t \in [0,1]$. In this model the random elements are equal to $(Y, X(\cdot))$, where Y is real and X is a random element of the set $L^2_{[0,1]}$ of the square integrable function defined on $[0,1]$ provided with the uniform measure. The model assumes $E(Y|X) = \langle X, \beta\rangle$, where β is an element of $L^2_{[0,1]}$. This moment condition implies

$$\int E(X(\tau)X(t))\beta(t)dt = E(YX(\tau)),$$

and may be treated in a particular case of the previous relation.

The GMM with a continuous number of moment conditions also gives a motivation for this kind of equation. Let us consider a moment condition

$$E(h(S,\theta,t)) = 0,$$

where S is a random element, θ is a vector of a parameter, and $t \in [0,1]$ indexes the moment conditions.

The overidentification issue is solved by replacing this equation h_j by the minimization of

$$\int E(h(S,\theta,t))E(h(S,\theta,\tau))k(t,\tau)dsd\tau,$$

where k is a weighting linear operator. Optimal GMM are considered by Carrasco and Florens (2000a) and are shown to be the solution of the minimization of

$$\int E(h(S,\theta,t))\varphi(\theta,t)dt,$$

where $\varphi(\theta,t)$ is solution of

$$\int E(h(S,\theta,t)u(S,\theta,\tau))\varphi(\theta,t)dt = E(h(S,\theta,\tau)).$$

The formalization of the inversion of the variance of the moment conditions then leads to a linear integral equation that is a part of the implementation of optimal GMM.

The class of linear inverse problems is very large; it covers cases with very different statistical properties.

Identification and overidentification can be reformulated in the linear case. The function φ is identified if K_F is one to one and this property is equivalent to $\mathcal{N}(K_F) = \{0\}$, where $\mathcal{N}(K_F)$ is the null set of K_F. A solution of Equation (3.1) exists if ψ_F belongs to the range of K_F, denoted $\mathcal{R}(K_F)$.

The main question raised by a linear equation is the existence of the inverse of K_F and if it has continuity. Intuitively we want to estimate φ by $\hat{\varphi}_N = K_{\hat{F}_N}^{-1}\psi_{\hat{F}_N}$. This computation requires an inversion of $K_{\hat{F}_N}$ and the continuity of $K_{\hat{F}_N}^{-1}$ because even if $\psi_{\hat{F}_N}$ and $K_{\hat{F}_N}$ converge to ψ_{F_0} and K_{F_0}, this continuity is necessary for the consistency of $\hat{\varphi}_{\hat{F}_N}$. This continuity is not always satisfied because a linear operator is not necessarily continuous in the infinite dimension case.

A linear inverse problem is said to be well posed if K_F^{-1} exists and is continuous. (This notion is from Hadamard; see, e.g., Nashed and Wahba, 1974 and Tikhonov and Arsenin, 1977.) This problem is ill posed otherwise. As we will see later on, some important econometric issues, such as instrumental variables estimation, define ill-posed inverse problems.

4. ILL-POSED LINEAR INVERSE PROBLEMS

Let two Hilbert spaces Φ and Ψ and $\mathcal{F}$ be a family of CDF of a random element S. We simplify our presentation by considering Hilbert spaces and not Banach spaces. Hilbert spaces are self-adjoint, and we can use an orthonormal basis and spectral decomposition of operators. From a statistical viewpoint, convergences will be defined *wrt.* a norm associated with a scalar product, and normal distributions in Hilbert spaces are more easy to deal with than in

Banach spaces. In many examples, Φ and Ψ are L^2 type function spaces, and their topological structure is dependent on a probability measure. Suppose that the definition of the sets Φ and Ψ is constructed in such a way that these sets do not depend on the probability F in $\mathcal{F}$ (e.g., all the F have a support included in a compact set of $\mathbb{R}^m$, and Φ and Ψ are spaces of continuous functions), but the scalar product is relative to the true DGP F_0 in $\mathcal{F}$.

We consider a linear inverse problem $K_F\varphi = \psi_F$, where K_F is a linear operator from Φ to Ψ and ψ_F is an element of Ψ.

We restrict our analysis to an important but specific case of operators.

Hypothesis 4.1. *$\forall F \in \mathcal{F}$, K_F is a compact operator.*

Recall that K_F is compact if the closure of the image of the closed unit sphere is compact. We give in the application to instrumental variables an interpretable sufficient condition that implies compactness of an operator.

A compact operator is bounded ($\sup_{\|\varphi\|\leq 1} \|K_F\varphi\|$ finite) or equivalently continuous. Its dual operator K_F^* (from Ψ to Φ), characterized by $\langle K_F\varphi, \psi\rangle = \langle \varphi, K_F^*\psi\rangle$, is also compact, and the two self-adjoint operators $K_F^*K_F$ (from Φ to Φ) and $K_F K_F^*$ (from Ψ to Ψ) are also compact.

Compact operators have only a discrete spectrum. More precisely, there exist two orthonormal families $(\varphi_{jF})_{j=0,1,2,\ldots}$ and $(\psi_{jF})_{j\,=\,0,1,2,\ldots}$ of Φ and Ψ and a sequence of decreasing positive numbers $\lambda_{0F} \geq \lambda_{1F} \geq \cdots > 0$ such that

$$\begin{aligned}
K_F^*K_F\varphi_{jF} &= \lambda_{jF}^2\varphi_{jF}, \quad K_F K_F^*\psi_{jF} = \lambda_{JF}^2\psi_{jF},\\
K_F\varphi_{jF} &= \lambda_{jF}\psi_{jF}, \quad K_F^*\psi_{jF} = \lambda_{jF}\varphi_{jF},\\
\forall\varphi \in \Phi,\ \varphi &= \sum_{j=0}^{\infty}\langle\varphi, \varphi_{jF}\rangle\varphi_{jF} + \bar{\varphi}_F, \quad \text{where } K_F\bar{\varphi}_F = 0,\\
\forall\psi \in \Psi,\ \Psi &= \sum_{j=0}^{\infty}\langle\psi, \psi_{jF}\rangle\psi_{jF} + \bar{\psi}_F, \quad \text{where } K_F^*\bar{\psi}_F = 0.
\end{aligned} \tag{4.1}$$

The spectrums of $K_F^*K_F$ and of $K_F K_F^*$ are discrete and included in $\{\lambda_{0F}^2, \lambda_{1F}^2, \ldots\} \cup \{0\}$. If K_F is one to one, the spectrum of $K_F^*K_F$ reduces to the family of λ_{jF}^2, but zero may be an eigenvalue of $K_F K_F^*$.

Let us come back to the equation $K_F\varphi = \psi_F$. A unique solution exists if K_F is one to one. Compact operators have a range that is, in general, strictly smaller than the space Ψ (in particular if $\Phi = \Psi$, a compact operator can be onto Φ only if Φ has a finite dimension; see Wahba, 1973) and then a solution to $K_F\varphi = \psi_F$ does not exist in general. As before, we denote $\mathcal{F}_0$ as the set of F such that a unique solution exists and the true CDF F_0 is assumed to be an element of $\mathcal{F}_0$. If $F \in \mathcal{F}_0$, then zero is not an eigenvalue of $K_F^*K_F$. In that case, we can compute the solution by using the decompositions given in (4.1).

First, let us write

$$K_F\varphi_F = \sum_{j=0}^{\infty}\lambda_{jF}\langle\varphi, \varphi_{jF}\rangle\psi_{jF},$$

because as ψ_F is an element of the range of K_F, $\bar{\psi}_F$ must be zero. Then, using the unicity of decomposition on φ_{jF}, we have

$$\lambda_{jF}\langle\varphi, \varphi_{jF}\rangle = \langle\psi, \psi_{jF}\rangle$$

and

$$\varphi_F = \sum_{j=0}^{\infty} \frac{1}{\lambda_j}\langle\psi_F, \psi_{jF}\rangle\varphi_{iF}. \tag{4.2}$$

A solution exists if and only if this series converges.

If K_F is not one to one or if ψ_F does not belong to the range of K_F, inversion of K_F may be replaced by generalized inversion. Equivalently, it can be proved (see, e.g., Luenberger, 1969) that if ψ_F belongs to $\mathcal{R}(K_F) + \mathcal{N}(K_F^*)$, there exists a unique function φ of minimal norm that minimizes $\|K_F\varphi - \psi_F\|$. This solution may be decomposed into

$$\varphi_F = \sum_{j/\lambda_j \neq 0} \frac{1}{\lambda_j}\langle\psi_F, \psi_{jF}\rangle\varphi_{jF}. \tag{4.3}$$

This series converges under the assumption $\psi_F \in \mathcal{R}(K_F) + \mathcal{N}(K_F^*)$. Let $\mathcal{F}_*$ be the set of F such that $\psi_F \in \mathcal{R}(K_F) + \mathcal{N}(K_F^*)$. $\mathcal{F}_*$ contains $\mathcal{F}_0$ because if $F \in \mathcal{F}_0$, $\psi_F \in \mathcal{R}(K_F)$. However, the condition $\psi_F \in \mathcal{R}(K_F) + \mathcal{N}(K_F^*)$ is not always satisfied. It is always true that $\Psi = \overline{\mathcal{R}(K_F)} + \mathcal{N}(K_F^*)$, but $\mathcal{R}(K_F)$ is not closed in general.

As we will see in the examples, usual estimators of F define operators $K_{\hat{F}_N}$ with a finite-dimensional range. This range is then closed and $\hat{F}_N$ is an element of $\mathcal{F}_*$.

The inverse of a compact operator and the generalized inverse are not continuous operators. A small perturbation of ψ_F in the direction of a ψ_{jF} corresponding to a small λ_{jF} will generate a large perturbation of φ. Then, even if K_F is known and if only ψ_F is estimated, the estimation of φ obtained by replacing ψ_F by $\psi_{\hat{F}_n}$ is generally not consistent. Examples given later on will illustrate this problem.

A regularization is then necessary to obtain consistent estimation. In this paper we privilege the so-called Tikhonov regularization methods. Other approaches play similar roles, such as the spectral cutoff regularization or the Landweber–Fridman iterations, which are defined but not studied from a statistical viewpoint.

Tikhonov regularization (see Tikhonov and Arsenin, 1977, Groetsch, 1984, and Kress, 1999) generalizes to infinite dimension the well-known ridge regression method used to deal with colinearity problems.[1] The initial linear equation

[1] Using standard notation, the ridge regression estimator of a linear model $y = X\beta + u$ is defined by $\hat{\beta}_\alpha = (\alpha NI + X'X)^{-1}X'_y$, where α is a positive number and I is the identity matrix. This estimator is used when $X'X$ is singular or quasi-singular. A Bayesian analysis of linear models provides a natural interpretation of this estimator as a posterior mean of β.

$K_F\varphi = \psi_F$ is replaced by a modified equation

$$(\alpha I + K_F^* K_F)\varphi = K_F^* \psi_F,$$

where α is a strictly positive number, and I is the identity operator on Φ. If α is not an eigenvalue of $K_F^* K_F$, then the linear operator $\alpha I + K_F^* K_F$ has a continuous inverse on the range of K_F^* and the solution of this equation has the following Fourier decomposition:

$$\varphi_F^\alpha = \sum_{j=0}^{\infty} \frac{\lambda_{jF}}{\alpha + \lambda_{jF}^2} \langle \psi_F, \psi_{jF} \rangle \varphi_{jF}.$$

If F is estimated by $\hat{F}_N$, previous formulas defined $\varphi_{\hat{F}_N}^\alpha$, and we will see that the norm of $\varphi_{F_0} - \varphi_{\hat{F}_N}^\alpha$ decreases to zero if α goes to zero at a suitable speed.

An equivalent interpretation of Tikhonov regularization is the following: The minimization of $\| K_F\varphi - \psi_F \|^2$ that defines the generalized inverse is replaced by the minimization of $\| K_F\varphi - \psi_F \|^2 + \alpha \|\varphi\|^2$, and α can be interpreted as a penalization parameter. This approach is extensively used in spline estimation, for example (see Wahba, 1990). A more efficient estimation may be found from the L^2-norm analysis. The Tikhonov method uses all the eigenvalues of $K_F^* K_F$ but prevents their convergence to zero by adding the positive value α. A spectral cutoff method controls the decrease of the λ_{jF}s by retaining only the eigenvalues greater than a given ρ:

$$\varphi_F^\rho = \sum_{\lambda_{jF} > \rho} \frac{1}{\lambda_{jF}} \langle \psi_F, \psi_{jF} \rangle \varphi_{jF}. \tag{4.4}$$

The Tikhonov regularization requires the inversion of $\alpha I + K^* K$, and the spectral cutoff regularization requires the computation of the spectrum. These two computations may be difficult. Another regularization scheme involves only successive applications of an operator and may be implemented recursively.

Let us consider a, a positive number that $a < 1/\|K\|^2$. We call the Landweber–Fridman regularization the value

$$\varphi_F^m = \sum_{j=0}^{m} (I - aK_F^* K_F)^j K^* \psi_F.$$

This function may be computed through the following recursive relation:

$$\varphi_F^\ell = (I - aK_F^* K_F)\varphi_F^{\ell-1} + aK^*\psi_F,$$

starting from $\varphi_F^0 = 0$ and using until $\ell = m$.

Most compact operators are integral operators operating on functions of real variables. In those cases, K_F is characterized by its kernel $k_F(s,t)$ - (s,t) are vectors of real numbers and

$$K_F\varphi = \int k_F(\tau, t)\varphi(\tau) d\tau. \tag{4.5}$$

The compactness of K_F is equivalent in that case to a more interpretable condition on k_F (k_F^2 must be integrable *wrt* z and t). Operators such as $I - K_F$, that is,

$$(I - K_F)\varphi = \varphi(t) - \int k_F(\tau, t)\varphi(\tau)d\tau,$$

are not compact operators and their inverses are continuous. Then, the inverse problems presented in Examples 3.3 (backfitting) and 3.4 (measurement) are not ill posed and may be solved without regularization. We illustrate by developing previous Example 3.5, a case of an ill-posed problem.

Example 4.1. *Let us assume that* $(s_1, \dots, s_N)$ *is an iid sample of* $S \in \mathbb{R}^m$*, and the parameter of interest is a real-valued continuous function* $\varphi(t)$ $(t \in [0, 1])$ *solution of*

$$\int_0^1 E^F(v(S, \tau)v(S, t))\varphi(\tau)d\tau = E^F(u(S, t)).$$

The function h *of Example 3.5. now has the product form* $h(S, \tau, t) = v(S, \tau)v(S, t)$*. If* v *is a zero-mean process, the* K_F *operator is the covariance operator of* v*. As we have seen, this example covers the case of a continuous number of regressors.*

If $k_F(\tau, t) = E^F(v(S, \tau)v(S, t))$ *is a continuous function of* $(\tau, t) \in [0, 1] \times [0, 1]$*, it is square integrable and the operator* K_F *is a Hilbert–Schmidt operator and then is compact (see Dunford and Schwartz, 1963). The kernel* k_F *is symmetric. Then* K_F *is self-adjoint* $(K_F = K_F^*)$*.*

Take, for example, $v(S, t) = S - t$*, where* S *is a zero-mean square integrable random variable. Then* $k_F(\tau, t) = E^F((S - \tau)(S - t)) = \tau t + V (V = var(S))$*. This operator is not one to one (two functions* φ_1 *and* φ_2 *such that* $\int \tau\varphi_1(\tau)d\tau = \int \tau\varphi_2(\tau)d\tau$ *and* $\int \varphi_1(\tau)d\tau = \int \varphi_2(\tau)d\tau$ *have the same image). The range of* K_F *is the set of affine functions. A one-to-one example is given by the covariance operator of a Brownian motion: let* $S = (W_t)_{t\in[0,1]}$ *be a Brownian motion. Assume that* $v(S, t) = W_t$*. Then* $k_p(s, t) = s \wedge t$*, whose null set is* $\{0\}$*, and*

$$\mathcal{R}_F(K_F) = \{\psi/\psi \in \mathcal{C}^1[0, 1]\psi(0) = 0 \textit{ and } \psi'(1) = 0\}.$$

A natural estimator of k_F *is obtained by estimating* F *by the empirical probability measure, that is,*

$$k_{\hat{F}_N}(\tau, s) = \frac{1}{N}\sum_{n=1}^{N} v(s_n, \tau)v(s_n, t).$$

This kernel defines a so-called Pincherle–Goursat integral operator (or degenerated kernel; see Tricomi, 1985). This operator maps a function φ *onto a* *linear combination of the $v(S_n, t)$:*

$$K_{\hat{F}_N}\varphi = \frac{1}{N}\sum_{n=1}^{N} v(s_n, t)\int_0^1 v(s_n, \tau)\varphi(\tau)d\tau,$$

and this range is the N-dimensional space spanned by the $v(s_n, t)$ (assumed to be linearly independent). Then, even if K_F is one to one for the true value F_0, the estimated operator $K_{\hat{F}_N}$ is not one to one and only N eigenvalues of $K_{\hat{F}_N}K_{\hat{F}_N}$ are not equal to zero. Moreover, the estimator of the right-hand side of the equation is equal to

$$\psi_{\hat{F}_N} = \frac{1}{N}\sum_{n=1}^{N} u(s_n, t),$$

and is not, in general, in the range of $K_{\hat{F}_N}$. The generalized inverse solution reduces in that case to solve the linear system $A\underline{\varphi} = b$, where A is the $N \times N$ matrix of general element $(1/N)\int_0^1 v(s_j, \xi)v(s_n, \xi)d\xi$, b is the vector of general element $(1/N)\sum_n \int v(s_j, \xi)u(s_n, \xi)d\xi$, and $\underline{\varphi}$ is the vector of $\int \varphi(\tau)v(s_n, \tau)d\tau$.

This procedure is analogous to the estimation of a model with incidental parameters (i.e., a model where a new parameter appears with each new observation) and the solution of the equation $A\underline{\varphi} = b$ cannot be provided a consistent estimator.

A Tikhonov regularization of this inverse problem leads us to solve the following functional equation:

$$\alpha\varphi(t) + \frac{1}{N^2}\sum_{j=1}^{n} v(s_j, t)\sum_{n=1}^{N}\int v(s_j, \xi)v(s_n, \xi)d\xi \times \int \varphi(\tau)v(s_n, \tau)d\tau$$
$$= \frac{1}{N^2}\sum_{j=1}^{n} v(s_j, t)\sum_{n=1}^{N}\int v(s_j, \xi)u(s_n, \xi)d\xi.$$

This functional equation can be solved in two steps. First, multiplying by $v(s_\ell, t)$ and integrating wrt t gives a linear $N \times N$ system where unknown variables are the $\int \varphi(\tau)v(s_n, \tau)d\tau$. This system can be solved and $\varphi(t)$ is then obtained from the given expression. This example shows that even if expressions in terms of Fourier decomposition are useful for analyzing the properties of the estimator, practical computations may be realized by inversion of finite-dimensional linears systems.

5. RELATION BETWEEN ENDOGENOUS VARIABLES

Let us assume that the observed random vector S can be decomposed into $(Y, Z, X, W) \in \mathbb{R} \times \mathbb{R}^p \times \mathbb{R}^k \times \mathbb{R}^q$. The assumptions derived from the economic models are the following. First, X and W are exogenous. This means that no information on the parameter of interest is carried by the marginal distribution generating X and W or, equivalently, that the parameter of interest may be deduced without loss of information from the conditional distribution of Y and Z given X and W. The second economic assumption says that the parameter of interest is a function of φ (or a transformation of this function) and X that satisfies a relation of Z:

$$Y = \varphi(Z, X) + U. \tag{5.1}$$

Equation (3.1) involves a normalization (a general function would be $\nu(Y, Z, X, U) = 0$), an additive structure for the residuals and an exclusion of W variables.

These assumptions are not sufficient to characterize φ in an unambiguous way, and they have to be completed by an assumption on the residual. This assumption must preserve the endogeneity of both Y and Z. Three different hypotheses have been used in the literature. All of these define φ as the solution of a linear inverse problem, and we call the three possible characterizations of φ the instrumental variables (IV) definition, the local instrumental variable (LIV) definition, or the control function (CF) definition.

The IV assumption is written as[2]

$$E^F(Y - \varphi(Z, X)|X, W) = E^F(Y - \varphi(Z, X)|X). \tag{5.2}$$

The usual assumption on IV regression assumed that the conditional expectation of U given all the exogenous variables (X, W) is zero. Following Heckman and Vytlacil (1999), we relax this condition, and $E(U|X, W)$ may be a function of X.

The main interest of this assumption is to consider a case in which (W, X) is not exogenous if φ is the parameter of interest (because $E(U|X, W) \neq 0$) but (X, W) becomes exogenous if the derivatives of φ with respect to Z are the parameters of interest (for an application, see Dearden, Ferri, and Meghir, 2002).

The function φ is a solution of a linear inverse problem

$$K_F^{\mathrm{IV}}\varphi = \psi_F^{\mathrm{IV}},$$

where

$$K_F^{\mathrm{IV}}\varphi = E(\varphi(Z, X)|X, W) - E(\varphi(Z, X)|X),$$

and

$$\psi_F^{\mathrm{IV}} = E(Y|X, W) - E(Y|X).$$

Using conventional notation for the densities of probability measures, $K_F^{\mathrm{IV}}\varphi$ may be written as

$$\left(K_F^{\mathrm{IV}}\varphi\right)(x, w) = \int \varphi(z, x)\{f(z|x, w) - f(z|x)\}dz,$$

and is an integral operator whose kernel is equal to $f(z|x, w) - f(z|x)$.

[2] To simplify our presentation, we can assume that all CDF we consider have the same compact support in $\mathbb{R}^{1+p+k+q}$ and are all equivalent (i.e., have the same null sets) to the Lebesgue measure on this compact. We assume that we restrict our presentation to continuous functions of a random vector. Then almost surely (AS) equalities become equality everywhere.

This linear operator is not one to one because functions of X only are elements of the null space of K_F^{IV}. If the econometrician is interested in the relation between Z and Y, it is sufficient to know φ up to functions of X (see the literature on average treatment effect, or ATE, by Imbens and Angrist, 1994 and Heckman and Vytlacil, 1999). Under regularity assumptions, this means that the partial derivatives of φ *wrt* z are identified.

The identification issue is then to describe models in which $\mathcal{N}(K_F^{\mathrm{IV}})$ reduces to $L^2(X)$. This condition is equivalent to the property that "any function of (Z, X) whose expectation given (X, W) is zero is in $L^2(X)$." This kind of condition was introduced in the analysis of relations between sufficient and ancillary statistics (see Lehman and Scheffe, 1950). Connection with identification of IV models and interpretation of this condition as a rank condition were pointed out by Newey and Powell (1989) and by Darolles, Florens, and Renault (2000). Extensive analysis of this concept, under the heading "strong identification," can be found in Mouchart and Rolin (1984) and in Florens, Mouchart, and Rolin (1990).

For the LIV assumption, if we assume differentiability of φ and of conditional expectations we consider, φ satisfies the LIV hypothesis if

$$E\left(\frac{\partial\varphi}{\partial z_j}(z, x)|X = x, W = w\right) = \frac{\dfrac{\partial}{\partial w_l}E(Y|X = x, W = w)}{\dfrac{\partial}{\partial W_l}E(Z_j|X = x, W = w)},$$

$$\forall j = 1, \ldots, p, \quad l = 1, \ldots, q. \qquad (5.3)$$

This definition naturally extends the linear case and can be interpreted easily. Discrete z was originally considered by Heckman and Vytlacil (1999) and discrete z and variations of w (instead of derivatives) were introduced by Imbens and Angrist (1994) and were called the local ATE (LATE).

This equation introduces an overidentification constraint because the right-hand side must be identical for any $l = 1, \ldots, q$. This condition is satisfied if $E(Y|X, W) = E(Y|X, m(X, W))$.

The function φ is the solution of a linear inverse problem, where $K_F = T_F D$, with $D\varphi$ as the vector of partial derivatives of φ *wrt* the coordinates of Z and T_F as the conditional expectation operator $(\lambda(Z, X) \to T_F\lambda = E(\lambda(Z, X)|X, W))$.

This operator K_F cannot be one to one, and, under a regularity condition,[3] it contains all the functions of X. Conversely, if Z is strongly identified by W given X, then T_F is one to one and the null set of K_F reduces to $L^2(X)$.

For the CF assumption, we assume there exists a function $V(Z, X, W)$ such that the information contained by Z, X, W and by V, X, W are identical, say,

[3] The distribution of (Z, X) must be such that the derivative *wrt* z_j of a function AS equal to a function of X must be zero, or equivalently if a function of Z is AS equal to a function of X if and only it is AS constant: This property is called measurable separability by Florens et al. (1990).

$V = Z - m(X, Z)$, and

$$E(U|Z, X, W) = E(U|V, X).$$

Consequently, if $h(V, X) = E(U|V, X)$, we have

$$E(Y|Z, X, W) = \varphi(Z, X) + h(V, X). \tag{5.4}$$

This assumption was used in several parametric contexts (see Heckman, 1979) and was systematically analyzed by Newey, Powell, and Vella (1999).

This model is an additive regression model, which implies that φ is a solution of the following set of equations:

$$\begin{aligned} \varphi(Z, X) + E(h(V, X)|ZX) &= E(Y|Z, X), \\ E(\varphi(Z, X)|V, X) + h(V, X) &= E(Y|V, X). \end{aligned}$$

Then φ is a solution of

$$\begin{aligned} \varphi(Z, X) - E(E(\varphi(V, X)|Z, X) = E(Y|Z, X) \\ -E(E(Y|V, X)|Z, X). \end{aligned} \tag{5.5}$$

Equation (5.5) can be rewritten as $K_F\varphi = \psi_F$, where $K_F = I - A_F^* A_F (A_F : L^2(Z, X) \ni \lambda \to E(\lambda|V, X) \in L^2(V, X)$, and $A_F^* : L^2(V, X) \ni \mu \to E(\mu|Z, X) \in L^2(Z, X))$.

The operator K_F cannot be one to one because here also its null space contains the functions of X.

As pointed out by Newey et al. (1999), $\mathcal{N}(K_F)$ contains only functions of X if V and Z are measurably separated given X (see Florens et al., 1990), that is, if any function of V and X AS equal to a function of Z and X is AS equal to a function of X. This condition is not always satisfied and can also be interpreted as a rank condition.

Remark: If F is dominated by the Lebesgue measure, we have seen that the IV assumption implies that φ satisfies a Fredholm type I equation. In the LIV case, $D\varphi$ is also solution of this type of equation:

$$\int \frac{\partial \varphi}{\partial z_j}(z, x) f(z, x|x, w) dz = \psi_F(x, w),$$

where ψ_F is the right-hand side of Equation (5.3).

In the CF approach, φ is a solution of a Fredholm type II equation:

$$\varphi(z, x) - \int \varphi(z, x) k(\xi, z, x) = \psi_F,$$

where now ψ_F is the right-hand side of Equation (5.5), and

$$k(\xi, z, x) = \int f(z, x|v, x) f(v, x|z, x) d\xi.$$

As we see in the next section, the properties of the solution are very different in this last case than in the first two cases.

It is easy to verify that if (Y, Z, W) are jointly normal, then these three problems give identical (linear) solutions. In nonlinear models, this equivalence is no longer true, and one can easily construct a model where the solutions are different (see, e.g., Florens et al., 2000, for equalities conditions).

6. IV ESTIMATION

To simplify the argument, we concentrate our analysis on the specific case in which no exogenous variables appear in the function φ. Then, the IV assumption becomes $E(U|W) =$ constant and φ can be identified only up to a constant term. It is natural in this context to assume that $E(U) = 0$ in order to eliminate this identification problem, and the case we consider now assumes

$$E(Y - \varphi(Z)|W) = 0. \tag{6.1}$$

We complete this assumption by using the following hypothesis on the joint probability measure on (Z, W). This hypothesis is fundamental for our spectral decomposition approach (for a different point of view of spectral decomposition of the conditional expectation operator, see Chen, Hansen, and Scheinkman, 2000).

Assumption 6.1. *The joint distribution of (Z, W) is dominated by the product of its marginal probabilities and its density is square integrable wrt the product measure.*

In the case of a probability measure dominated by the Lebesque measure, this condition is equivalent to

$$\int \frac{f^2(z, w)}{f(z)f(w)} dzdw < \infty.$$

Let us now denote the two conditional expectation operators by T_F and T_F^*, the dual operator:

$$\begin{aligned} T_F &: L^2(Z) \to L^2(W) \quad T_F\varphi = E(\varphi|W) \quad \varphi \in L^2(Z), \\ T_F^* &: L^2(W) \to L^2(Z) \quad T^*\psi = E(\psi|Z) \quad \psi \in L^2(W). \end{aligned}$$

The original problem may be denoted as $T_F\varphi = r_F$, where $r_F = E(Y|W) \in L^2(W)$.

Under Assumption 6.1, T_F is a compact operator (see Breiman and Friedman, 1985) and the analysis developed in Section 4 applies. Here IV estimation is an ill-posed inverse problem and requires a regularization procedure.

The same argument applies to LIV estimation. Take as the parameter of interest the vector of partial derivatives, $D\varphi$. This vector of functions is also a solution to an ill-posed inverse problem, $T_F D\varphi = \psi_F$, where ψ_F is defined in Equation (5.3), and where the linear operator is compact.

Under an assumption $m(Z, V)$ analogous to the assumption on (Z, W), CF estimation leads to a well-posed inverse problem and does not require regularization. Indeed, φ is solution of $(I - A_F^* A_F)\varphi = \psi_F$; see Equation (5.5). The function ψ_F is in the domain of $(I - A_F^* A_F)$, and the inverse operator is bounded and then continuous. This can be seen by using a spectral decomposition of $A_F^* A_F$ whose eigenvalues are denoted by μ_j^2, and whose eigenvectors are denoted by ε_j. Then

$$\varphi = \sum_{j=1}^{\infty} \frac{1}{1-\mu_j^2} \langle \psi_F, \varepsilon_j \rangle \varepsilon_j.$$

The sum starts at $j = 1$ because ε_0 is the constant function equal to one and $\langle \psi_F, \varepsilon_F \rangle = 0$ because ψ_F is a zero-mean vector.

This series converges in norm L^2 because

$$\begin{aligned} \sum_{j=1}^{\infty} \left(\frac{1}{1-\mu_j^2} \right)^2 \langle \psi_F, \varepsilon_j \rangle^2 &\leq \left(\frac{1}{1-\mu_1^2} \right)^2 \sum_{j=1}^{\infty} \langle \psi_F, \varepsilon_j \rangle^2 \\ &\leq \left(\frac{1}{1-\mu_1^2} \right)^2 \|\psi_F\|^2. \end{aligned}$$

Finally, $\sup \|(I - A_F^* A_F^*)^{-1} \psi_F\|$, where $\|\psi\| \leq 1$ and $\psi \in$ domain $(I - A_F^* A_F^*)^{-1}$ = set of the zero-mean vector, is smaller than $|1/(1-\mu_1)|$, which means that the inverse operator is continuous.

We conclude this section with a short description of the practical implementation of the estimation of φ in the case of IV assumption. The sample is $(y_n, z_n, w_n)_{n=1,\ldots,N}$, and the equation $(\alpha_N I + T_{\hat{F}_N}^* T_{\hat{F}_N})\varphi = T_{\hat{F}_N}^* r_{\hat{F}_N}$ may be simplified into

$$\begin{aligned} &\alpha_N \varphi(z) + \frac{1}{\sum_{\ell=1}^{N} H_N(z - z_\ell)} \sum_{\ell=1}^{N} \frac{\sum_{n=1}^{N} \varphi(z_n) H_N(w_\ell - w_n)}{\sum_{n=1}^{N} H_N(w_\ell - w_n)} H_N(z - z_\ell) \\ &\quad = \frac{1}{\sum_{\ell=1}^{N} H_N(z - z_\ell)} \sum_{\ell=1}^{N} \frac{\sum_{n=1}^{N} y_n H_N(w_\ell - w_n)}{\sum_{n=1}^{N} H_N(w_\ell - w_n)} H_N(z - z_\ell), \end{aligned} \tag{6.2}$$

where H_N is the usual smoothing kernel (conventionally the same letter is used for different kernels applied to the ws or the zs). This functional equation gives $\varphi(z)$ for any z, knowing $\varphi(z_n) n = 1, \ldots, N$. Then in a first step, rewrite

Equation (6.2) for $z = z_1, \ldots z_N$. This provides an $N \times N$ linear system that can be solved to obtain $\varphi(z_n)$. The choice of the α_N parameter is very important, and we see in the next section what the contraints are on its speed of convergence. We also discuss the choice of this parameter.

This approach avoids any computation of eigenvalues or eigenvectors, but they are implicitly present in the resolution of the linear system. Using the same methodology as Darolles, Florens, and Gouriéroux (1998), one can check that the estimator we have defined may be rewritten as

$$\varphi_{\hat{F}_N}^{\alpha N} = \sum_{j=0}^{N-1} \frac{\hat{\lambda}_{j\hat{F}_N}}{\alpha_N + \hat{\lambda}_{\hat{F}_N}^2} \left(\frac{1}{N} \sum_{n=1}^{N} y_n \varphi_{j\hat{F}_N}(z_n) \right) \varphi_{j\hat{F}_N}(z), \tag{6.3}$$

where $\hat{\lambda}_{j\hat{F}_N}^2$ are the N nonnull eigenvalues of $T_{\hat{F}_N}^* T_{\hat{F}_N}$, and $\varphi_{j\hat{F}_N}$ is their corresponding eigenvector.

7. ASYMPTOTIC THEORY FOR TIKHONOV REGULARIZATION OF ILL-POSED LINEAR INVERSE PROBLEMS

Here we concentrate our presentation on new questions raised by the linear inverse problem $K_F \varphi = \psi_F$, where K_F is a compact operator. We will then assume asymptotic behavior of the elements of the equation (which can be difficult to verify in particular models), and we will show how they are transformed by the resolution. As previously announced, we will develop a Hilbert space approach, both for consistency and for asymptotic normality.

Let φ_{F_0} be the unique solution of $K_{F_0} \varphi = \psi_{F_0}$, where F_0 is the true DGP, which is an element of $\mathcal{F}_0$.

We denote by $\varphi_{F_0}^{\alpha}$ the solution of

$$\left(\alpha I + K_{F_0}^* K_{F_0}\right)\varphi = K_{F_0}^* \psi_{F_0} = K_{F_0}^* K_{F_0} \varphi_0,$$

for any $\alpha > 0$. Given a sample $(s_1, \ldots, s_N)$, $\hat{F}_N$ is an estimator of F and $K_{\hat{F}_N}$ and $\psi_{\hat{F}_N}$ is the corresponding estimation of K_F and ψ_F.

The properties of this estimation mechanism are given by the following assumptions:

Assumption 7.1. *$\exists a_N$ sequence in $\mathbb{R}$ $a_n \to \infty$ such that*[4]

$$\left\| K_{\hat{F}_N}^* K_{\hat{F}_N} - K_F^* K_F \right\| \sim O\left(\frac{1}{a_N}\right).$$

In this assumption the norm of an operator A from Φ to Φ is defined by $\sup_{\|\varphi\| \leq 1} \|A\varphi\|$, and the norm on Φ is the Hilbert norm possibly dependent on F_0.

[4] All the equivalences are in probability *wrt* the DGP. Almost sure equivalences will give AS convergence in Theorem 7.1.

Assumption 7.2. *$\exists b_N$ sequence in $\mathbb{R}$ $b_N \to \infty$ such that*

$$\left\| K^*_{\hat F_N} \psi_{\hat F_N} - K^*_{\hat F_N} K_{\hat F_N} \varphi_0 \right\| \sim O\left(\frac{1}{b_N}\right).$$

This assumption replaces the assumption on $\psi_{\hat F_N}$. Intuitively, $\psi_{\hat F_N}$ converges to ψ_{F_0} equal to $K_{F_0}\varphi_0$, but as K^*_F is a compact operator, taking the image of $\psi_{\hat F_N} - K_{\hat F_N}\varphi_0$ by $K^*_{\hat F_N}$ regularizes the estimation and may improve the speed of convergence.

Assumption 7.3. *$\alpha_N \to 0$, $1/\alpha_N a_N \sim O(1)$, and $\alpha_N b_N \to \infty$.*

Theorem 7.1. *Under Assumptions* 7.1, 7.2, *and* 7.3, $\|\varphi^{\alpha_N}_{\hat F_N} - \varphi\| \to 0$ *in probability.*

Proof. This proof is standard if the operator K_F is known and where the only error is on ψ_F (see Groetsch, 1984 or Kress, 1999). Extension to the estimation error on K_F generalizes the arguments developed in Carrasco and Florens (2000a) and in Darolles et al. (2000). The main steps of the proofs are the following:

(i)

$$\left\|\varphi^{\alpha_N}_{\hat F_N} - \varphi_{F_0}\right\| \leq \left\|\varphi^{\alpha_N}_{\hat F_N} - \varphi^{\alpha_N}_{F_0}\right\| + \left\|\varphi^{\alpha_N}_{F_0} - \varphi_{F_0}\right\|,$$

and $\|\varphi^{\alpha_N}_{F_0} - \varphi_{F_0}\| \to 0$ if $\alpha_N \to 0$ (see any of the just-described references).

(ii)

$$\begin{aligned}
\varphi^{\alpha_N}_{\hat F_N} - \varphi^{\alpha_N}_{F_0} &= \left(\alpha_N I + K^*_{\hat F_F} K_{\hat F_N}\right)^{-1} K_{\hat F_N}\psi_{\hat F_N} \\
&\quad - \left(\alpha_N I + K^*_{F_0} K_{F_0}\right)^{-1} K^*_{F_0} K_{F_0}\varphi_0 \\
&= \left(\alpha_N I + K^*_{\hat F_N} K_{\hat F_N}\right)^{-1}\left(K^*_{\hat F_N}\psi_{\hat F_N} - K^*_{\hat F_N} K_{\hat F_N}\varphi_{F_0}\right) \\
&\quad + \alpha_N\left[\left(\alpha_N I + K^*_{\hat F_N} K_{\hat F_N}\right)^{-1} - \left(\alpha_N I + K^*_{F_0} K_{F_0}\right)^{-1}\right]\varphi_{F_0}.
\end{aligned}$$

The last equality follows from the identity

$$(\alpha I + A)^{-1} A = I - \alpha(\alpha I + A)^{-1}.$$

Then $\|\varphi^{\alpha_N}_{\hat F_N} - \varphi^{\alpha_N}_0\| \leq I + II$, where *I* and *II* are defined and analyzed separately.

(iii)

$$\begin{aligned}
I &= \left\|\left(\alpha_N I + K^*_{\hat F_N} K_{\hat F_N}\right)^{-1}\left(K^*_{\hat F_N}\psi_{\hat F_N} - K^*_{\hat F_N} K_{\hat F_N}\varphi_{F_0}\right)\right\| \\
&\leq \left\|\left(\alpha_N I + K^*_{\hat F_N} K_{\hat F_N}\right)^{-1}\right\| \left\|K^*_{\hat F_N}\psi_{\hat F_N} - K^*_{\hat F_N} K_{\hat F_N}\varphi_{F_0}\right\|.
\end{aligned}$$

The first term is majored by $1/\alpha_N$ (see Groetsch, 1984), and the second is $0(1/b_N)$ by Assumption 7.2. By Assumption 7.3, $\alpha_N b_N \to \infty$ and $I \to 0$.

(iv)

$$
\begin{aligned}
II &= \alpha_N \left\| \left[\left(\alpha_N I + K^*_{\hat{F}_N} K_{\hat{F}_N}\right)^{-1} - \left(\alpha_N I + K^*_{F_N} K_{F_0}\right)^{-1} \right] \varphi_{F_0} \right\| \\
&= \left\| \alpha_N \left(\alpha_N I + K^*_{F_0} K^*_{F_0}\right)^{-1} \varphi_{F_0} \right\| \times \left\| K^*_{\hat{F}_N} K_{\hat{F}_N} - K^*_{F_0} K_{F_0} \right\| \\
&\quad \times \left\| \left(\alpha I + K^*_{F_0} K_{F_0}\right)^{-1} \right\|.
\end{aligned}
$$

The first term is equal to $\|\varphi - \varphi^{\alpha_N}\|$ and has a zero limit. The second term is by Assumption 7.1 and is equivalent to $1/a_N$, and the last term is smaller than $1/\alpha_N$. As $1/\alpha_N a_N \sim O(1)$, $II \to 0$. ■

Example 7.1. *Consider Example 4.1. Following, for example, Carrasco and Florens (2000a), we have $\|K_{\hat{F}_N} - K_{F_0}\| \sim O(1/\sqrt{N})$. Using the property $K^*_F = K^*_F$ and a first-order approximation, we find that it follows that $\|K^2_{\hat{F}_N} - K^2_{F_0}\|$ is also equivalent to $1/\sqrt{N}$. Moreover,*

$$
\begin{aligned}
\left\| K_{\hat{F}_N} \psi_{\hat{F}_N} - K^2_{\hat{F}_N} \varphi_0 \right\| \leq \left\| K_{\hat{F}_N} \right\| \Big\{ \left\| \psi_{\hat{F}_N} - K_{F_0} \varphi_0 \right\| \\
+ \left\| \hat{K}_{\hat{F}_n} - K_{F_0} \right\| \left\| \varphi_{F_0} \right\| \Big\},
\end{aligned}
$$

which implies $b_n = \sqrt{N}$ because $\|\psi_{\hat{F}_N} - K_{F_0}\varphi_0\| \sim O(1/\sqrt{N})$.

Then the two conditions are satisfied if $\alpha_n \sqrt{N} \to \infty$.

Example 7.2. *Consider the case of IV estimation. It is proved in Darolles et al. (2000) that $1/a_N = 1/\sqrt{Nh^p_N} + h^\rho_N$, where h_N is the bandwidth of the kernel smoothing, p is the dimension of z, and ρ is the minimum between the order of the kernel and the degree of smoothness of the density of the DGP. Moreover, $1/b_N = 1/\sqrt{N} + h^\rho_N$. Then the estimator is consistent if $h^{2\rho}_N/\alpha^2_N \to 0$ and $1/\alpha^2_N N h^\rho_N \sim O(1)$.*

The decomposition of $\varphi^{\hat{\alpha}_N}_{\hat{P}_N} - \varphi_{F_0}$ considered in the proof of Theorem 7.1 can be used to determine an optimal speed of convergence to zero of α_N and to give a bound on the speed of convergence of $\|\varphi^{\alpha_N}_{\hat{F}_N} - \varphi_{F_0}\|$. This analysis requires an assumption of the behavior of the regularization bias $\|\varphi^{\alpha_N}_{F_0} - \varphi_{F_0}\|$, which satisfies

$$
\left\| \varphi^{\alpha_N}_{F_0} - \varphi_{F_0} \right\| = \alpha_N \left(\alpha_N I + K^*_{F_0} K_{F_0}\right)^{-1} \varphi_{F_0} \tag{7.1}
$$

$$
= \alpha^2_N \sum_{j=0}^{\infty} \frac{1}{(\alpha_N + \lambda_{jF_0})^2} \langle \varphi_{F_0}, \varphi_{jF} \rangle \varphi_{jF_0}. \tag{7.2}
$$

We will assume that φ_{F_0} is such that $\|\varphi^{\alpha_n}_{F_0} - \varphi_{F_0}\|^2 \sim O(\alpha^\beta)$.

This condition associates φ_{F_0} and K_{F_0} and is basically a condition on the relative rate of decline of the Fourier coefficients of φ_{F_0} on the basis $\varphi_{jF}(\langle \varphi_{F_0}, \varphi_{jF_0} \rangle)$ and of the eigenvalues $\lambda^2_{jF_0}$ of the operator.

Darolles et al. (2000) show that $\beta \in [0, 2]$ and give characteristics of particular cases of β. In case of IV, the β coefficient may be interpreted as a measure of the strength or weakness of the instruments. Then

$$\left\|\varphi_{\hat{F}_N}^{\alpha_N} - \varphi_{F_0}\right\|^2 = 0\left(\frac{1}{\alpha_N^2 aN} + \frac{1}{\alpha_N^2 bN}\alpha_N^\beta + \alpha_N^\beta\right),$$

and an optimal choice of α_n is obtained if the behavior of the first and last terms are equal. Then,

$$\alpha_N = a_N^{[-1/(\beta+2)]}.$$

We need to verify that, under this choice, the second term converges to zero, if it is the case that $a_N^{[\beta/(\beta+2)]}$ gives a lower bound of the speed of convergence. In the applications given herein, this bound is $n^{[\beta/(\beta+2)]}$ (under a suitable choice of the bandwidth if a kernel estimation is necessary).

The last element to be considered is the asymptotic normality of our estimator. This normality follows from the next hypothesis:

Assumption 7.4.

$$b_N\left(K_{\hat{F}_N}^* \psi_{\hat{F}_N} - K_{\hat{F}_N}^* K_{\hat{F}_N}\varphi_{F_0}\right) \Rightarrow N(0, \Omega).$$

This convergence is assumed to be a functional convergence in the Hilbert space Φ and Ω is a covariance operator in this space.

Let us assume first that K_{F_0} is known and that the parameter α is kept constant. Under these two conditions we have

$$b_n\left(\varphi_{\hat{F}_N} - \varphi_{F_0}^\alpha\right) = \left(\alpha I + K_{F_0}^*\right)^{-1}\left(b_n\left(K_{F_0}^*\psi_{\hat{F}_N} - K_F^* K_{F_0}\varphi\right)\right),$$

which converges in Φ to a zero-mean normal probability whose covariance operator is equal to

$$\left(\alpha I + K_{F_0}^* K_{F_0}\right)^{-1}\Omega\left(\alpha I + K_{F_0}^* K_{F_0}\right)^{-1}.$$

Indeed, standard matrix computation can be extended to continuous operators.

The extension of this result to the case of an unknown operator K_F, with α constant, modifies this result in the following way: Let

$$B_N^\alpha = \alpha\left[\left(\alpha I + K_{\hat{F}_N}^\alpha K_{\hat{F}_N}\right)^{-1} - \left(\alpha I + K_{F_0}^\alpha K_{F_0}\right)\right]\varphi.$$

We have obviously, from part (ii) of the proof of Theorem 7.1, the following:

$$\begin{aligned} b_N\left(\varphi_{\hat{F}_N}^\alpha - \varphi_{F_0}^\alpha - B_N^\alpha\right) = {} & \left(\alpha I + K_{\hat{F}_N}^* K_{\hat{F}_N}\right)^{-1} b_N \\ & \times \left(K_{\hat{F}_N}^* \psi_{\hat{F}_N}\psi_{\hat{F}_N} - K_{\hat{F}_N}^* K_{\hat{F}_N}\varphi_0\right), \end{aligned}$$

and this term converges to the same normal probability measure in Φ as if K_F if known. However, a bias term has been introduced in the left-hand side term. In the proof of Theorem 7.1, we have checked that in the case of α fixed $\|B_N^\alpha\|$

converges to zero at speed $1/a_n$. The bias term can be neglected if b_n/a_n has a zero limit, that is, if the operator converges at a higher speed than the right-hand side of the equation.

If $\alpha_N \to 0$, we cannot expect asymptotic normality in a functional sense. In particular, the limit when α_N decreases to zero of the covariance operator Ω is not bounded and is not a covariance operator of a Hilbert-valued normal element. Then we will look for pointwise normality instead of functional normality in the following sense. Let ζ be an element of ϕ. We will analyze asymptotic normality of

$$\nu_N(\zeta)\big\langle \varphi^{\alpha_N}_{\hat{F}_N} - \tilde{\varphi}, \zeta \big\rangle,$$

where $\tilde{\varphi}$ is a suitable function and $\nu_N(\zeta) \to \infty$.

This class of results is obtained by using the following methodology.

First, let us denote by ξ_N the random element $b_N(K^*_{\hat{F}_N}\psi_{\hat{F}_N} - K^*_{\hat{F}_N}\varphi_{F_0})$ and by ξ its limit ($\xi \sim N(0, \Omega)$). For a given N, $M_N = (\alpha_N I + K^*_{F_0}K^*_{F_0})^{-1}$ and

$$\varepsilon = \frac{\langle M_N\xi, \zeta\rangle}{\langle \zeta, M_N\Omega M_N\zeta\rangle^{1/2}} \sim N(0, 1) \quad \forall N,$$

because M_N is bounded and $M_N\xi \sim N(0, M_N\Omega M_N)$.

Second, let us first assume that K_{F_0} is known. Then

$$\frac{b_N\big\langle \varphi^{\alpha_N}_{\hat{F}_N} - \varphi^{\alpha_N}_{F_0}, \zeta\big\rangle}{\big\langle \zeta, M_N\Omega M_N\zeta\big\rangle^{1/2}} = \varepsilon + \frac{\langle \xi_N - \xi, M_N\zeta\rangle}{\langle \zeta, M_N\Omega M_N\zeta\rangle}.$$

Moreover,

$$\frac{\langle \xi_N - \xi, M_N\rangle^2}{\langle \zeta, M_N\Omega M_N\zeta\rangle} \leq \|\xi_N - \xi\|^2 \frac{\|M_N\zeta\|^2}{\langle \zeta, M_N\Omega M_N\zeta\rangle^{1/2}}.$$

This term converges to zero if $\|M_N\zeta\|^2/\langle \zeta, M_N\Omega M_N\zeta\rangle$ is bounded because $\|\xi_N - \xi\| \to 0$ in probability. We introduce this condition as a hypothesis.

Assumption 7.5. $\zeta \in \Phi$ *is such that* $\|M_N\zeta\|^2/\langle \zeta, M_N\Omega M_n\zeta\rangle \sim 0(1)$.

Remark that if ζ belongs to the finite-dimensional subspace generated by $\varphi_0, \ldots, \varphi_{N_0}$ (where $\lambda_j \neq 0\ \forall j = 0, \ldots, N_0$), Assumption 7.5 is satisfied.

We note by

$$\nu_N(\zeta) = \frac{b_N^2}{\langle \zeta, M_N\Omega M_N\zeta\rangle}$$

the speed of convergence. We may conclude that

$$\sqrt{\nu_N(\zeta)}\big\langle \hat{\varphi}^{\alpha_N}_{\hat{F}_N} - \varphi^{\alpha_N}_{F_0}, \zeta\big\rangle \Rightarrow N(0, 1).$$

Third, if K_{F_0} is not known, let us consider

$$\sqrt{\nu_N(\zeta)}\big\langle \varphi^{\alpha_N}_{\hat{F}_N} - \varphi^{\alpha_N}_{F_0} - B^{\alpha_N}_N, \zeta\big\rangle = \varepsilon + A_1 + A_2, +A_3,$$

where

$$A_1 = \frac{\langle \xi_N - \xi, M_N \zeta \rangle}{\langle \zeta, M_N \Omega M_N \zeta \rangle^{1/2}}, \quad A_2 = \frac{\langle \xi, (\hat{M}_N - M_N)\zeta \rangle}{\langle \zeta, M_N \Omega M_N \zeta \rangle^{1/2}},$$

where $\hat{M}_N = (\alpha_N I + K^*_{\hat{F}_N} K_{\hat{F}_N})^{-1}$, and

$$A_3 = \frac{\langle \xi_N - \xi, (\hat{M}_N - M_N)\zeta \rangle}{\langle \zeta, M_N \Omega M_N \zeta \rangle^{1/2}}.$$

We have shown in the previous case that under Assumption 7.5, A_1 converges to zero. The term A_2 has the same behavior as

$$\frac{\|\xi\| \|M_N\| \|K^*_{\hat{F}_N} K_{\hat{F}_N} - K^*_{F_0} K_{F_0}\| \|M_N \zeta\|}{\langle \zeta, M_N \Omega M_N \rangle^{1/2}} \leq \frac{\|M_N \zeta\|}{\langle \zeta, M_N \Omega M_N \rangle^{1/2}} \frac{1}{\alpha_N a_N} \|\zeta\|,$$

because $\|M_N\| \leq 1/\alpha_N$ and Hypothesis 7.1.

We then reenforce Hypothesis 7.3.

Assumption 7.6. $\alpha_N a_N \to \infty$.

This assumption implies that $A_2 \to 0$ and an analogous proof shows that $A_3 \to 0$.

Then, under the previous assumptions,

$$\sqrt{\nu_N(\zeta)} \big\langle \hat{\varphi}_{\hat{F}_N} - \varphi^{\alpha_N}_{F_0} - B^{\alpha_N}_N, \zeta \big\rangle \Rightarrow N(0, 1).$$

Fourth, the next step consists of finding assumptions that transform the centering function. First we look for an elimination of the bias term $B^{\alpha_N}_N$.

$$\begin{aligned} \left| \sqrt{\nu_N(\zeta)} B^{\alpha_N}_N \right| &= \frac{b_N \alpha_N}{\langle \zeta, M_N \Omega M_N \rangle^{1/2}} \langle (\hat{M}_N - M_N) \varphi_{F_0}, \zeta \rangle \\ &\leq b_N \|\alpha_N M_N \varphi\| \left\| K^*_{\hat{F}_N} \hat{K}_{\hat{F}_N} - K^*_{F_0} K_{F_0} \right\| \\ &\quad \times \frac{\|M_N \zeta\|}{\langle \zeta, M_N \Omega M_N \zeta \rangle^{1/2}} \|\alpha_N M_N \varphi\| \\ &= \left\| \varphi^{\alpha_N}_{F_0} - \varphi_{F_0} \right\| \to 0. \end{aligned}$$

We just have to impose that the product of the others terms is bounded. Using Assumption 7.2, we find that a general assumption is the following.

Assumption 7.7. $b_N / a_N \|M_N \zeta\| / \langle \zeta, M_n \Omega M_N \zeta \rangle^{1/2} \sim 0(1)$.

This assumption is satisfied under Assumption 7.5 if $b_N / a_N \sim 0(1)$, but this hypothesis could be too strong. If $b_N / a_N \to \infty$, more assumptions are needed in order to eliminate the bias term.

Then under Assumptions 7.1–7.7 we get

$$\nu_N(\zeta)\big\langle \hat{\varphi}_{\hat{F}_N} - \varphi_{F_0}^{\alpha_N}, \zeta\big\rangle \Rightarrow N(0, 1).$$

Fifth, and finally, we want to replace $\varphi_{F_0}^{\alpha_N}$ by φ_{F_0} in the previous convergence. Recalling that $\|\varphi_{F_0}^{\varphi_N} - \varphi_{F_0}\| \sim 0(\alpha_N)$, we require the following assumption.

Assumption 7.8. $\alpha_N^2 \nu_N(\zeta) \to 0.$

Under Assumptions 7.1–7.6 and 7.8 we obtain

$$\sqrt{\nu_N(\zeta)}\big\langle \varphi_{\hat{F}_N}^{\alpha_N} - \varphi_{F_0}, \zeta\big\rangle \Rightarrow N(0, 1),$$

if K_{F_0} is known, and

$$\sqrt{\nu_N(\zeta)}\big\langle \varphi_{\hat{F}_N}^{\alpha_N} - \varphi_{F_0}, -B_{F_0}^{\alpha_N}\zeta\big\rangle \Rightarrow N(0, 1),$$

in the general case.

If, moreover, Assumption 7.7 is satisfied, pointwise asymptotic normality without bias is satisfied:

$$\sqrt{\nu_N(\zeta)}\big\langle \varphi_{\hat{F}_N}^{\alpha_N} - \varphi_{F_0}, \zeta\big\rangle \Rightarrow N(0, 1).$$

In the case developed in Examples 4.1 and 7.1, all the assumptions can be satisfied and this last pointwise normality is verified. In the case of IV estimation (Example 7.2), Assumption 7.7 is not true and a bias correction term must be introduced in order to get asymptotic normality.

8. CONCLUSIONS

This paper proposed a general framework for structural functional estimation, and some results related to the linear compact case are given. Application to IV estimations motivates this analysis. Numerous questions are not considered. In particular, the choice of the regularization α_N in relation to optimal speed of convergence and to minimax estimation is not treated in this paper (some steps in that direction are made in Carrasco and Florens, 2000b). The general methodology presented in the paper may be extended to nonlinear inverse problems, to some well-posed inverse problems, and to dynamic cases. A deep discussion about the different definitions of relations between endogenous variables is necessary for obtaining unambiguous nonparametric estimations (see Blundell and Powell, 1999 and Florens et al., 2000).

ACKNOWLEDGMENTS

This was an invited lecture to the 2000 World Congress of the Econometric Society. This paper is strongly related to research conducted with M. Carrasco, S. Darolles, and E. Renault. I also acknowledge C. Gouriéroux, L. P. Hansen, J. Heckman, C. Meghir, C. Protopopescu, J. F. Richard, E. Sbaï, A. Vanhems, and E. Vytlacil for useful discussions on this topic. The revision of this paper has benefited from numerous remarks and comments by L. P. Hansen. Author contact: University of Toulouse (IDEI-GREMAQ), Manufacture des Tabacs, 21, Allée de Brienne, F-31000 Toulouse, France. E-mail: florens@cict.fr.

References

Aït-Sahalia, Y. (1995), "The Delta and Bootstrap Methods for Nonparametric Kernel Functionals," Discussion Paper, MIT.

Aït-Sahalia, Y. (1996), "Nonparametric Pricing of Interest Rate Derivative Securities," *Econometrica*, 64, 527–560.

Banon, G. (1978), "Nonparametric Identification for Diffusion Processes," *SIAM Journal of Control and Optimization*, 16, 380–395.

Basmann, R. L. (1959), "A Generalized Classical Method of Linear Estimation of Coefficients in Structural Equations," *Econometrica*, 25, 77–83.

Blundell, R. and J. Powell (1999), "Endogeneity in Single Index Models," Discussion Paper, UCL.

Carrasco, M. and J. P. Florens (2000a), "Generalization of GMM to a Continuum of Moments Conditions," *Econometric Theory*, 16, 797–834.

Carrasco, M. and J. P. Florens (2000b), "Efficient GMM Estimation Using the Empirical Characteristic Function," GREMAQ-University of Toulouse.

Chen, X., L. P. Hansen, and J. Scheinkman (2000), "Principal Components and the Long Run," Discussion Paper, University of Chicago.

Colton, D., H. W. Engle, J. R. McLaughin, and W. Rundell (Eds.) (2000), *Surveys on Solution Methods for Inverse Problems*. New York: Springer.

Darolles, S., J. P. Florens, and C. Gouriéroux (1998a), "Kernel Based Nonlinear Canonical Analysis," Discussion Paper 9858, CREST.

Darolles, S., J. P. Florens, and E. Renault (1998b), "Nonlinear Principal Components and Inference on a Conditional Expectation Operator," Discussion Paper, CREST.

Darolles, S., J. P. Florens, and E. Renault (2000), "Nonparametric Instrumental Regression" Discussion paper. GREMAQ, University of Toulouse.

Dearden, L., J. Ferri, and C. Meghir (2002), "The Effect of School Quality on Educational Attainment and Wages," Discussion Paper, IFS. *Review of Economics and Statistics*.

Debrath, L. and P. Mikusinski (1999), *Hilbert Spaces with Applications*. London: Academic Press.

Dunford, N. and J. Schwartz (1963), *Linear Operators 2*. New York: Wiley.

Engle, R. H., D. F. Hendry, and J. F. Richard (1983), "Exogeneity," *Econometrica*, 51(2), 277–304.

Florens, J. P., J. Heckman, C. Meghir, and E. Vytlacil (2000), "Instrumental Variables, Local Instrumental Variables and Control Functions," Discussion Paper, GREMAQ-University of Toulouse.

Florens, J. P. and M. Mouchart (1985), "Conditioning in Dynamic Models," *Journal of Time Series Analysis*, 53(1), 15–35.

Florens, J. P., M. Mouchart, and J. M. Rolin (1990), *Elements of Bayesian Statistics*. New York: Dekker.

Florens, J. P., and E. Sbaï (2000), "Identification in Empirical Games," Discussion Paper, GREMAQ-University of Toulouse.

Florens, J. P., C. Protopopescu, and J. F. Richard (1997), "Identification and Estimation of a Class of Game Theoretic Models," Discussion Paper, GREMAQ-University of Toulouse.

Florens, J. P. and A. Vanhems (2000), "Estimation non Paramétrique de l'Épaisseur de la Couche Imosphérique: Application aux Mesures du Satellite Topex-Poséïdon," Discussion Paper, GREMAQ-University of Toulouse.

Frisch, R. (1934), "Statistical Confluence Analysis by Means of Complete Regression Systems," Discussion Paper, Universitets Økonomiske Institutt, Oslo.

Frisch, R. (1938), "Statistical Versus Theoretical Relations in Economic Macrodynamics," Discussion Paper, Cambridge University.

Gaspar, P. and J. P. Florens (1998), "Estimation of the Sea State Bias in Radar Altimeter Measurements of Sea Level: Results from a Nonparametric Method," *Journal of Geophysical Research*, 103(15), 803–814.

Groetsch, C. (1984), *The Theory of Tikhonov Regularization for Fredholm Equations of the First Kind*. London: Pitman.

Guerre, E., I. Perrigne, and Q. Vuong (2000), "Optimal Nonparametric Estimation of First-Price Auctions," *Econometrica*, 68(3), 525–574.

Hall, A. (1993), "Some Aspects of the Generalized Method of Moments Estimation," in *Handbook of Statistics*, Vol. 11, (ed. by G. S. Maddala, C. R. Rao and H. D. Vinod), Amsterdam: North-Holland, 393–417.

Hansen, L. (1982), "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50, 1029–1054.

Hardle, W. and O. Linton (1994), "Applied Nonparametric Methods," *Handbook of Econometrics*, 4, 2295–2339.

Hastie, T. J. and R. J. Tibshirani (1990), *Generalized Additive Models*. London: Chapman & Hall.

Hausman, J. (1981), "Exact Consumer's Surplus and Deadweight Loss," *American Economic Review*, 71, 662–676.

Hausman, J. (1985), "The Econometrics of Nonlinear Budget Sets," *Econometrica*, 53, 1255–1282.

Hausman, J. and W. K. Newey (1995), "Nonparametric Estimation of Exact Consumers Surplus and Deadweight Loss," *Econometrica*, 63, 1445–1476.

Heckman, J. (1979), "Sample Selection Bias as a Specification Error," *Econometrica*, 47, 153–161.

Heckman, J. and V. Vytlacil (1999), "Local Instrumental Variables," Working Paper, University of Chicago.

Imbens, G. and J. Angrist (1994), "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467–476.

Kitamura, Y. and M. Stutzer (1997), "An Information Theoretic Alternative to Generalized Method of Moments Estimation," *Econometrica*, 65(4), 861–874.

Kress, R. (1999), *Linear Integral Equations*, second edition, New York: Springer-Verlag.

Lehman, E. L. and H. Scheffe (1950), "Completeness, Similar Regions, and Unbiased Tests. Part I," *Sankhya*, 10, 219–236.

Luenberger, D. G. (1969), *Optimization by Vector Space Methods*. New York: Wiley.

Manski, C. (1988), "Analog Estimation Methods in Econometrics," London: Chapman & Hall.

Mouchart, M. and J. M. Rolin (1984), "A Note on Conditional Independence," *Statistica*, 45(3), 427–430.

Nashed, M. Z. and G. Wahba (1974), "Generalized Inverse in Reproducing Kernel Spaces: An Approach to Regularization of Linear Operator Equations," *SIAM Journal of Mathematical Analysis*, 5(6), 974–987.

Newey, W. and J. Powell (1989), "Instrumental Variables for Nonparametric Models," Discussion Paper, MIT.

Newey, W., J. Powell, and F. Vella (1999), "Nonparametric Estimation of Triangular Simultaneous Equations Models," *Econometrica*, 67, 565–604.

Ogaki, M. (1993), "Generalized Method of Moments: Econometric Applications," in *Handbook of Statistics*, Vol. 11, (ed. by G. S. Maddala, C. R. Rao, and H. D. Vinod), Amsterdam: North-Holland, 455–488.

Owen, A. (1990), "Empirical Likelihood Ratio Confidence Regions," *The Annals of Statistics*, 18 (1), 90–120.

Pagan, A. and A. Ullah (1999), *Nonparametric Econometrics*. Cambridge: Cambridge University Press.

Qin, J. and J. Lawless (1994), "Empirical Likelihood and General Estimating Equations," *The Annals of Statistics*, 22 (1), 300–325.

Reiersol, O. (1941), "Confluence Analysis of Lag Moments and Other Methods of Confluence Analysis," *Econometrica*, 9, 1–24.

Reiersol, O. (1945), "Confluence Analysis by Means of Instrumental Sets of Variables," *Arkiv for Mathematik, Astronomie och Fysik*, 32.

Sargan, J. D. (1958), "The Estimation of Economic Relationship Using Instrumental Variables," *Econometrica*, 26, 393–415.

Theil, H. (1953), "Repeated Least Squares Applied to Complete Equations System," mimeo, Central Planning Bureau, The Hague.

Tikhonov, A. and V. Arsenin (1977), *Solutions of Ill-Posed Problems*. Washington, DC: Winston & Sons.

Tricomi, F. G. (1985), *Integral Equations*. New York: Dover.

Van der Vaart, A. W. and J. A. Wellner (1996), *Weak Convergence and Empirical Processes*. New York: Springer.

Vanhems, A. (2000), "Nonparametric Solutions to Random Ordinary Differential Equations of First Orders," Discussion Paper, GREMAQ-University of Toulouse.

Wahba, G. (1973), "Convergence Rates of Certain Approximate Solutions of Fredholm Integrals of the First Kind," *Journal of Approximation Theory*, 7, 167–185.

Wahba, G. (1990), *Spline Models for Observational Data*. Philadelphia: SIAM.

CHAPTER 8

Endogeneity in Nonparametric and Semiparametric Regression Models

Richard Blundell and James L. Powell

1. INTRODUCTION

The analysis of data with endogenous regressors – that is, observable explanatory variables that are correlated with unobservable error terms – is arguably the main contribution of econometrics to statistical science. Although "endogeneity" can arise from a number of different sources, including mismeasured regressors, sample selection, heterogeneous treatment effects, and correlated random effects in panel data, the term originally arose in the context of "simultaneity," in which the explanatory variables were, with the dependent variable, determined through a system of equations, so that their correlation with error terms arose from feedback from the dependent to the explanatory variables. Analysis of linear supply-and-demand systems (with normal errors) yielded the familiar rank and order conditions for identification, two- and three-stage estimation methods, and analysis of structural interventions. Although these multistep estimation procedures have been extended to nonlinear parametric models with additive nonnormal errors (e.g., Amemiya, 1974 and Hansen 1982), extensions to nonparametric and semiparametric models have only recently been considered.

The aim of this chapter is to examine the existing literature on estimation of some "nonparametric" models with endogenous explanatory variables, and to compare the different identifying assumptions and estimation approaches for particular models and determine their applicability to others. To maintain a manageable scope for the chapter, we restrict our attention to nonparametric and semiparametric extensions of the usual simultaneous equations models (with endogenous regressors that are continuously distributed). We consider the identification and estimation of the "average structural function" and argue that this parameter is one parameter of central interest in the analysis of semiparametric and nonparametric models with endogenous regressors. The two leading cases we consider are additive nonparametric specifications in which the regression function is unknown, and nonadditive models in which there is some known transformation function that is monotone but not invertible. An important example of the latter, and one that we use as an empirical illustration, is the binary response model with endogenous regressors. We do not explicitly consider the closely related problems of selectivity, heterogeneous treatment effects, correlated random effects, or measurement error (see Heckman et al., 1998, Angrist, Imbens, and Rubin, 1996, and Arellano and Honoré, 1999 for lucid treatments of these topics). Moreover, we consider only recent work on nonparametric and semiparametric variants of the two-stage least-squares (2SLS) estimation procedure; Matzkin (1994) gives a broader survey of identification and estimation of nonlinear models with endogenous variables. Also, for convenience, we restrict attention to randomly sampled data, though most of our discussion applies in non-*iid* contexts, provided the structural equations and stochastic restrictions involve only a finite number of observable random variables.

In the next subsections, a number of different generalizations of the linear structural equation are presented, and the objects of estimation (the parameters of interest) are defined and motivated. The sections that follow consider how two common interpretations of the 2SLS estimator for linear equations – the "instrumental variables" and "control function" approaches – may or may not be applicable to nonparametric generalizations of the linear model and to their semiparametric variants. The discussion then turns to a particular semiparametric model, the binary response model with linear index function and nonparametric error distribution, and describes in detail how estimation of the parameters of interest can be constructed by using the control function approach. This estimator is applied to the empirical problem of the relation of labor force participation to nonlabor income studied in Blundell and Powell (1999). The results point to the importance of the semiparametric approach developed here and the strong drawbacks of the linear probability model and other parametric specifications.

1.1. Structural Equations

A natural starting point for investigation of endogeneity is the classical linear structural equation

$$y = \mathbf{x}'\boldsymbol{\beta} - u, \tag{1.1}$$

where $(y, \mathbf{x}')$ represents a data point of dimension $[1 \times (k+1)]$, $\boldsymbol{\beta}$ is a conformable parameter vector, and u is an unobservable disturbance term. The explanatory variables $\mathbf{x}$ are assumed to include a subset of continuous *endogenous* variables, meaning that

$$E(\mathbf{x}u) \neq 0. \tag{1.2}$$

This is the standard single linear equation treated in the literature on simultaneous equations, and it is considered here as a base case for comparison with other nonlinear setups. To identify and estimate the coefficient vector $\boldsymbol{\beta}$ in this setup, the model must be completed by imposing restrictions on the unobservable error terms u that are consistent with (1.2); the traditional approach assumes that some observable vector $\mathbf{z}$ of instrumental variables is available, satisfying the moment condition

$$E(\mathbf{z}u) = 0, \tag{1.3}$$

which leads to the well-known 2SLS estimator of $\boldsymbol{\beta}$ (Basmann, 1959 and Theil, 1953). The algebraic form of the 2SLS estimator can be derived from a number of different estimation principles based on (1.3), or on stronger conditions that imply it. As we will see, some of these estimation approaches, under suitably strengthened stochastic restrictions, can be extended to the nonparametric and semiparametric generalizations of the linear model that are considered here, but in certain important cases this turns out not to be so.

At an opposite extreme from the standard linear model, the relation between y and its observable and unobservable determinants $\mathbf{x}$ and u could be assumed to be of a general form

$$y = H(\mathbf{x},u), \tag{1.4}$$

which may represent a single equation of a nonlinear simultaneous equation system, possibly with a limited or qualitative dependent variable. Of course, without further restrictions on H, this function would be unidentified even under strong conditions on the unobservables (like independence of $\mathbf{x}$ and u), but it is useful to view the various nonparametric and semiparametric models that follow as special cases of this setup.

One important class of structural functions, treated in more detail in the paragraphs that follow, assumes H to be additively separable,

$$y = g(\mathbf{x}) + u, \tag{1.5}$$

which would be the nonparametric regression model if the expectation of u given $\mathbf{x}$ could be assumed to be zero (i.e., if $\mathbf{x}$ were exogenous). Identification and estimation of g when a subset of $\mathbf{x}$ is endogenous and instrumental variables $\mathbf{z}$ are available is the subject of a number of recent studies, including those by Newey and Powell (1989), Newey, Powell, and Vella (1999), Darolles, Florens, and Renault (2000), and Ng and Pinkse (1995).

A nonseparable variant of (1.4) would be Matzkin's (1991) nonparametric version of Han's (1987) generalized regression model,

$$y = t(g(\mathbf{x}), u), \tag{1.6}$$

in which h is a known function that is monotone, but not invertible, in its first argument (a "single index" if g is one dimensional), and g is an unknown function satisfying appropriate normalization and identification restrictions. A leading special case of this specification is the nonparametric binary choice model (Matzkin, 1992), in which H is an indicator function for positivity of the sum of $g(\mathbf{x})$ and u:

$$t(g(\mathbf{x}), u) = 1(g(\mathbf{x}) + u > 0).$$

We group all of the models (1.4)–(1.6) in the *nonparametric* category, where the term refers to the lack of parametric structure to the structural function H or h or the "regression" function g. *Semiparametric* models restrict H (and possibly the distribution of u) to have finite-dimensional parametric components; that is,

$$y = h(\mathbf{x}, \boldsymbol{\beta}, u). \tag{1.7}$$

For example, another special case of (1.4) is Han's (1987) original model, where the single-index function g is assumed to be linear in $\mathbf{x}$,

$$y = t(\mathbf{x}'\boldsymbol{\beta}, u); \tag{1.8}$$

estimation of this model when $\mathbf{x}$ is endogenous is considered by, for example, Lewbel (1998) and Blundell and Powell (1999), which focuses on the binary response version of this linear index model. Yet another semiparametric special case,

$$y = s(\mathbf{x}, \boldsymbol{\beta}, g(\cdot)) + u, \tag{1.9}$$

where $\boldsymbol{\beta}$ is a finite parameter vector and h is a known function, has been considered in the recent work by Ai and Chen (2000). Although estimation of the coefficient vector $\boldsymbol{\beta}$ is typically the main objective of a semiparametric analysis of such models, estimation of the distribution of u, or at least certain functionals of it, will also be needed to evaluate the response of the dependent variable y to possible exogenous movements in the explanatory variables.

1.2. Parameters of Interest

For a nonparametric model, the parameters of interest are actually unknown functions that summarize important characteristics of the structural function H and the distribution of the errors u; these parameters will be identified if they can be extracted from the distributions of the observable random variables. From a random sample of observations on the dependent variable y, regressors $\mathbf{x}$, and instrumental variables $\mathbf{z}$, the joint distribution of y, $\mathbf{x}$, $\mathbf{z}$, is by definition identified, and conditional distributions and moments can be consistently estimated by using standard nonparametric methods. In particular, the conditional expectation of functions of y, given either $\mathbf{x}$ or $\mathbf{z}$ or both, can be estimated without imposing additional restrictions (besides, say, smoothness and finite moment restrictions) on the joint distribution of the observable data, and these conditional expectations clearly summarize key features of the structural function and error distribution. However, as is clear from the well-worn supply and demand examples, knowledge only of the conditional distributions of observables is insufficient for analysis of the results of certain types of structural interventions that affect the distribution of the regressors $\mathbf{x}$ but not the structural error terms u. Thus, the expectation of y given the instruments $\mathbf{z}$, called the *reduced form for y*, may be of interest if the values of the instrumental variables are control variables for the policymaker, but for interventions that alter the explanatory variables **x** directly, independently of the error terms u, knowledge only of the joint distribution of the observables will be insufficient. Similarly, potential results of interventions that directly affect some components of a vector-valued structural function H – for example, a change in the supply function in a supply-and-demand system caused by rationing – clearly could not be analyzed solely from knowledge of the reduced form for y prior to the intervention, nor could the results of policies intended to change the distribution of the unobservable component u.

For an analysis of such policies, it would be most useful to know the form of the structural function $H(\mathbf{x}, u)$ from (1.4), along with the joint distribution of the errors u and **x**, **z**, but these may not be identifiable, at least for models without additive errors. An alternative summary version of the structural function H that can be more straightforward to estimate is the *average structural function (ASF),* where the average is taken over the *marginal* distribution of the error terms u,

$$G(\mathbf{x}) \equiv \int H(\mathbf{x}, u) dF_u, \tag{1.10}$$

for F_u, the marginal cumulative distribution function of u. In models with additively separable errors, that is,

$$H(\mathbf{x},u) = g(\mathbf{x}) + u, \tag{1.11}$$

as in (1.5), the ASF $G(\mathbf{x})$ reduces to the usual regression function $g(\mathbf{x})$, which would correspond to $E[y|\mathbf{x}]$ if the error terms u had a conditional mean zero given **x**. More generally, the ASF G would be the counterfactual conditional expectation of y given **x** if the endogeneity of **x** were absent, that is, if the regressors **x** could be manipulated independently of the errors, which would be considered invariant to the structural change. For the heterogeneous treatment effect model (see Heckman and Robb, 1985 and Imbens and Angrist, 1994), the ASF is directly related to the average treatment effect recovered from an experimental design – specifically, the average treatment effect for a binary regressor would be $G(1) - G(0)$.

In some structural interventions, where the regressors **x** can be manipulated directly, knowledge of the function G, or its derivatives, would be sufficient to assess the impact of the policy. For example, in the classical supply-and-demand example, with y corresponding to quantity demanded and **x** representing price, the ASF would suffice to determine expected demand if the market supply function were replaced with a fixed price (by fiat or by the world market), and if the distribution of u were assumed to be invariant to this structural change. And, for the additively separable model (1.11), the ASF embodies the direct effect of the regressors for a particular observation, holding the error terms fixed; in this case the individual-specific, not just average, effects of changes in **x** can be analyzed if G is known.

However, for interventions that do not directly determine the endogenous regressors, knowledge of the ASF is not enough; additional structural information about the structural function H and the distribution of **x** (and possibly **z**) would be required to evaluate the policy effect. For example, for the effects of imposition of a sales tax to be analyzed, which would rescale $\mathbf{x}$ in the structural function by a fixed amount exceeding unity, the "inverse supply function" relating price $\mathbf{x}$ to quantity supplied y and other observable and unobservable covariates would have to be specified to account for the joint determination of price and quantity following the structural intervention. More generally, policies that alter some components of the function H rather than manipulating the argument $\mathbf{x}$ require specification, identification, and consistent estimation of all components of H (including equations for all endogenous regressors) and the distribution of u. As shown by Roehrig (1988) and Imbens and Newey (2000), nonparametric identification of a fully specified system of simultaneous equations is possible under strong restrictions on the forms of the structural function – for example, invertibility of $H(\mathbf{x}, u)$ in the unobservable component u – but such restrictions may be untenable for limited dependent variable models such as the binary response model analyzed in the paragraphs that follow. Thus, the average structural function G may be the only feasible measure of the direct effect of $\mathbf{x}$ on y for limited dependent variable models and for "limited information" settings in which the structural relations for the endogenous regressors in a single structural equation are incompletely specified.

Of course, the expected value of the dependent variable y need not be the only summary measure of interest; a complete evaluation of policy could require the entire distribution of y given an "exogenous" $\mathbf{x}$. Because that distribution can be equivalently characterized by the expectation of all measurable functions of y, a more ambitious objective is calculation of the ASF for any transformation $\tau(y)$ of y with finite first moment,

$$G_\tau(\mathbf{x}) \equiv \int \tau(H(\mathbf{x}, u))dF_u.$$

For those sets of stochastic restrictions on u that require additively separable form (1.11) for identification and estimation of the ASF, this collection of expectations can be evaluated directly from the marginal distribution of $u = y - g(\mathbf{x})$; for those stochastic restrictions that do require additivity of errors for identification of G, the collection of functions G_τ (and thus the structural distribution of y) can be derived by redefinition of the dependent variable to $\tau(y)$.

For semiparametric problems of the form (1.7), the finite-dimensional parameter vector $\boldsymbol{\beta}$ is typically of interest in its own right, because economic hypotheses can impose testable restrictions on the signs or magnitudes of its components. A primary goal of the statistical analysis of semiparametric models is to construct consistent estimators of $\boldsymbol{\beta}$ that converge at the parametric rate (the inverse of the square root of the sample size), with asymptotically normal distributions and consistently estimable asymptotic covariance matrices. For some sets of restrictions on error distributions, this objective may be feasible even when estimation of the ASF $G(\mathbf{x})$ is not. For example, for the semiparametric model (1.7), the form of the reduced form for y,

$$E[y|\mathbf{z}] = E[h(\mathbf{x}'\boldsymbol{\beta}, u)|\mathbf{z}],$$

can, under appropriate restrictions, be exploited to obtain estimates of the single-index coefficients $\boldsymbol{\beta}$ even if G is not identified, as noted as follows.

Even for a fully nonparametric model, some finite-dimensional summary measures of G may be of interest. In particular, the "average derivative" of $G(\mathbf{x})$ with respect to **x** (Stoker, 1986) can be an important measure of marginal effects of an exogenous shift in the regressors. Altonji and Ichimura (2000) consider the estimation of the derivative y with respect to x for the case when y is censored. They are able to derive a consistent estimator for the nonadditive case. Unlike the estimation of single-index coefficients, which are generally identified only up to a scale factor, estimation of the average derivatives of the ASF G is problematic unless G itself is identifiable.

2. NONPARAMETRIC ESTIMATION UNDER ALTERNATIVE STOCHASTIC RESTRICTIONS

As in the traditional treatment of linear simultaneous equations, we will assume that there exists a $1 \times m$ vector **z** of instrumental variables, typically with $m \geq k$. The particular stochastic restrictions on **z**, **x**, and u will determine what parameters are identified and what estimators are available in each of the model specifications (1.4)–(1.9). Each of the stochastic restrictions is a stronger form of the moment condition (1.3), and each can be used to motivate the familiar 2SLS estimator in the linear model with additive errors (1.1) under the usual rank condition, but their applicability to the nonparametric and semiparametric models varies according to the form of the structural function H.

2.1. Instrumental Variables Methods

2.1.1. The Linear Model

The instrumental variables (IV) version of the standard 2SLS estimator is the sample analog to the solution of a weaker implication of (1.3), namely,

$$0 = E(P[\mathbf{x}|\mathbf{z}]u) \equiv E(\Pi'\mathbf{z}u) = E(\Pi'\mathbf{z}(y - \mathbf{x}'\boldsymbol{\beta})), \tag{2.1}$$

where $P[x|z]$ is the population least-squares projection of **x** on **z**, with

$$\Pi \equiv \{E(\mathbf{z}\mathbf{z}')\}^{-1}E(\mathbf{z}'\mathbf{x}). \tag{2.2}$$

Replacing population expectations with sample averages in (2.1) yields the 2SLS estimator

$$\widehat{\boldsymbol{\beta}}_{2\mathrm{SLS}} = (\widehat{\mathbf{X}}'\mathbf{X})^{-1}\widehat{\mathbf{X}}'\mathbf{y}, \tag{2.3}$$

with

$$\widehat{\mathbf{X}} = \mathbf{Z}\widehat{\Pi} \quad \text{and} \quad \widehat{\Pi} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}, \tag{2.4}$$

and where **X**, **Z**, and **y** are the $N \times k$, $N \times m$, and $N \times 1$ data arrays corresponding respectively to **x**, **z**, and y, for a sample of size N. When the linear form of the residual $u = y - \mathbf{x}'\beta$ is replaced with a nonlinear, parametric version $u = m(y, \mathbf{x}, \boldsymbol{\beta})$, extension of this estimation approach yields the generalized IV estimator (GIVE) of Sargan (1958), the nonlinear two-stage least-squares (NLLS) estimator of Amemiya (1974), and the generalized method of moments (GMM) estimator (Hansen, 1982).[1]

Another closely related formulation of 2SLS exploits a different implication of (1.3), namely,

$$0 = P[u|\mathbf{z}] = P[y|\mathbf{z}] - P[\mathbf{x}|\mathbf{z}]'\beta, \tag{2.5}$$

where the population linear projection coefficients of u and y on $\mathbf{z}$ are defined analogously to (2.2). Replacing $P[y|\mathbf{z}]$ and $P[\mathbf{x}|\mathbf{z}]$ with their sample counterparts and applying least squares yields Basmann's (1959) version of 2SLS,

$$\widehat{\beta}_{2\text{SLS}} = (\widehat{\mathbf{X}}'\widehat{\mathbf{X}})^{-1}\widehat{\mathbf{X}}'\widehat{\mathbf{y}}, \tag{2.6}$$

where now

$$\widehat{\mathbf{y}} = \mathbf{Z}\widehat{\pi} \equiv \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}.$$

Although logically distinct from the IV interpretation of 2SLS in (2.3), extension of the estimation approaches in either (2.3) or (2.6) yields the same NLLS and GMM estimators in the nonlinear parametric case, and we refer to generalization of either approach to the nonparametric or semiparametric structural equations as an *IV method.*

2.1.2. *Extensions to Additive Nonparametric Models*

To extend the IV methods to nonparametric settings, we find it natural to strengthen the unconditional moment restriction $E(\mathbf{z}u) = 0$ to a conditional mean restriction

$$E(u|\mathbf{z}) = 0, \tag{2.7}$$

just as the assumption of $E[\mathbf{x}u] = 0$ is strengthened to $E[u|\mathbf{x}] = 0$ for a nonparametric regression model. For the additive structural function (1.11), identification and estimation of $g(\mathbf{x})$ was considered by Newey and Powell (1989) and Darolles et al. (2000). Substitution of the error term $u = y - g(\mathbf{x})$ into condition (2.7) yields a relationship between the reduced form $E[y|\mathbf{z}]$ and the structural function g:

$$\begin{aligned} E[y|\mathbf{z}] &= E[g(\mathbf{x})|\mathbf{z}] \\ &= \int g(\mathbf{x})dF_{\mathbf{x}|\mathbf{z}}, \end{aligned} \tag{2.8}$$

[1] When β is overidentified, that is, the dimension m of the instruments $\mathbf{x}$ exceeds the dimension k of β, asymptotically efficient estimation would be based on a different implication of (1.3), in which Π is replaced by a different $(m \times k)$ matrix, as noted by Hansen (1982).

where $F_{\mathbf{x}|\mathbf{z}}$ is the conditional cumulative distribution function (CDF) of $\mathbf{x}$ given $\mathbf{z}$. The reduced form for y, $E[y|\mathbf{z}]$, and the conditional distribution of $\mathbf{x}$ given $\mathbf{z}$ are functionals of the joint distribution of the observable variables y, $\mathbf{x}$, and $\mathbf{z}$ are identified; identifiability of the structural function g therefore reduces to the uniqueness of the solution of the integral equation (2.8). And, as noted in the Newey–Powell and Darolles–Florens–Renault manuscripts, this in turn reduces to the question of statistical completeness of the family of conditional distributions $F_{\mathbf{x}|\mathbf{z}}$ in the "parameter" $\mathbf{z}$. (See, e.g., Ferguson, 1967, Section 3.6, for a definition of completeness and its connection to minimum variance unbiased estimation in parametric problems.) Although conditions for completeness of $F_{\mathbf{x}|\mathbf{z}}$ are known for certain parametric classes of distributions (e.g., exponential families), and generally the "order condition" $\dim(\mathbf{z}) \geq \dim(\mathbf{x})$ must be satisfied, in a nonparametric estimation setting, uniqueness of the solution of (2.8) must be imposed as a primitive assumption. Darolles et al. (2000) give a more thorough discussion of the conditions for existence and uniqueness of the solution of (2.8) and its variants.

In the special case in which $\mathbf{x}$ and $\mathbf{z}$ have a joint distribution that is discrete with finite support, conditions for identification and consistent estimation of the ASF $g(\mathbf{x})$ are straightforward to derive. Suppose $\{\boldsymbol{\xi}_j, j = 1, \ldots, J\}$ are the set of possible values for $\mathbf{x}$ and $\{\zeta_l, l = 1, \ldots, L\}$ are the support points for $\mathbf{z}$, and let

$$\begin{aligned} \boldsymbol{\pi} &\equiv \operatorname{vec}(E[y|\mathbf{z} = \zeta_l]), \qquad (2.9) \\ \mathbf{P} &\equiv [P_{jl}] \equiv [\Pr\{\mathbf{x} = \boldsymbol{\xi}_j | \mathbf{z}_l = \zeta_j\}] \end{aligned}$$

denote the vector of reduced-form values $E[y|\mathbf{z}]$ and the matrix of conditional probabilities that $\mathbf{x} = \boldsymbol{\xi}_j$ given that $\mathbf{z} = \zeta_l$, respectively; these would clearly be identified, and consistently estimable, from a random sample of observations on y, $\mathbf{x}$, and $\mathbf{z}$. If $\mathbf{g} \equiv \operatorname{vec}(g(\boldsymbol{\xi}_j))$ denotes the vector of possible values of $g(\mathbf{x})$, the question of identifiability of $\mathbf{g}$ using (2.8) is the question of uniqueness of the solution to the set of linear equations

$$\boldsymbol{\pi} = \mathbf{P}\mathbf{g}, \qquad (2.10)$$

so that $\mathbf{g}$ is identified if and only if $\operatorname{rank}\{\mathbf{P}\} = J = \dim\{\mathbf{g}\}$, which requires the order condition $L = \dim\{\boldsymbol{\pi}\} \geq \dim\{\mathbf{g}\} = J$. When $\mathbf{g}$ is identified, it may be consistently estimated (at a parametric rate) by replacing $\boldsymbol{\pi}$ and $\mathbf{P}$ by estimators using the empirical CDF of the observed vectors in (2.10), and solving for $\hat{\mathbf{g}}$ in the just-identified case $J = L$, or using a minimum chi-square procedure when $\mathbf{g}$ is overidentified ($J < L$).[2] More details for this finite-support case are given by Das (1999).

[2] Note that when $J = K = 2$, this is the "treatment effect" in the homogeneous treatment effect model case with additive errors. The heterogeneous treatment effect case is a specific form of the general nonadditive model.

2.1.3. The Ill-Posed Inverse Problem

Unfortunately, the simple structure of this finite-support example does not easily translate to the general case, in which $\mathbf{x}$ and $\mathbf{z}$ may have continuously distributed components. Unlike in typical nonparametric estimation problems, where identification results can be easily translated into consistent estimators of the identified functions under smoothness or monotonicity restrictions, identification of g and consistent estimators of the components $E[y|\mathbf{z}]$ and $F_{\mathbf{x}|\mathbf{z}}$ are not, by themselves, sufficient for a solution of a sample analogue of (2.8) to be a consistent estimator of g. First, it is clear that, unlike the standard nonparametric regression problem, the function $g(\mathbf{x})$ (and the reduced form and conditional distribution function) must be estimated for all values of $\mathbf{x}$ in the support of the conditional distribution, and not just at a particular value $\mathbf{x}_0$ of interest; thus, consistency of g must be defined in terms of convergence of a suitable measure of distance between the functions $\hat{g}(\cdot)$ and $g(\cdot)$ (e.g., the maximum absolute difference over possible $\mathbf{x}$ or the integrated squared difference) to zero in probability. Moreover, the integral equation (2.8), a generalization of the Fredholm integral equation of the first kind, is a notorious example of an ill-posed inverse problem: The integral $T_{\mathbf{z}}(g) \equiv \int g(\mathbf{x}) dF_{\mathbf{x}|\mathbf{z}}$, although continuous in g for the standard functional distance measures, has an inverse that is not continuous in general, even if the inverse is well defined. That is, even if a unique solution $\hat{g}$ of the sample version

$$\widehat{E}[y|\mathbf{z}] = \int g(\mathbf{x}) d\widehat{F}_{\mathbf{x}|\mathbf{z}} \tag{2.11}$$

of (2.8) exists, that solution $\tilde{g} \equiv \hat{T}_{\mathbf{z}}^{-1}(\hat{E}[y|\mathbf{z}])$ is not continuous in the argument $\hat{E}[y|\mathbf{z}]$, so consistency of the reduced-form estimator (and the estimator of the conditional distribution of $\mathbf{x}$ given $\mathbf{z}$) does not imply consistency of $\hat{g}$. Heuristically, the reduced form $E[y|\mathbf{z}]$ can be substantially smoother than the structural function $g(\mathbf{x})$, so that very different structural functions can yield very similar reduced forms; the ill-posed inverse problem is a functional analogue to the problem of multicollinearity in a classical linear regression model, where large differences in regression coefficients can correspond to small differences in fitted values of the regression function. Such ill-posed inverse problems are well known in applied mathematics and statistics, arising, for example, in the problem of estimation of the density of an unobservable variable x that measured with error; that is, observations are available only on $y = x + u$, where u is an unobservable error term with known density function (the deconvolution problem). O'Sullivan (1986) surveys the statistical literature on ill-posed inverse problems and describes the general "regularization" approaches to construction of consistent estimators for such problems. If the joint distribution of $\mathbf{x}$ and $\mathbf{z}$ is approximated by a distribution with a finite support (as common for "binning" approaches to nonparametric estimation of conditional distributions), then the ill-posed inverse problem would manifest itself as an extreme sensitivity of the "transition" matrix $\mathbf{P}$ to choice of J and L, and its near-singularity as J and L increase to infinity.

2.1.4. Consistent Estimation Methods

Newey and Powell (1989) impose further restrictions on the set of possible structural functions g to obtain a consistent estimator, exploiting the fact that the inverse of a bounded linear functional such as $T_{\mathbf{z}}(g)$ will be continuous if the domain of the functional is compact. For this problem, compactness of the set of possible g functions with respect to, say, the "sup norm" measure of distance between functions, can be ensured by restricting the "Sobolev norm" of all possible g functions to be bounded above by a known constant. This Sobolev norm (denoted here as $\|g\|_S$) is a different but related distance measure on functions that involves a sum of the integrated squared values of $g(\mathbf{x})$ and a certain number of its derivatives. In effect, the requirement that $\|g\|_S$ is bounded, which ensures that g is sufficiently smooth, counteracts the ill-posed inverse problem by substantially restricting the possible candidates for the inverse function $\hat{T}_{\mathbf{z}}^{-1}(\hat{E}[y|\mathbf{z}])$.

To obtain a computationally feasible estimation procedure, Newey and Powell assume that the structural function g can be well approximated by a function that is linear in parameters,

$$g(\mathbf{x}) \cong g_J(\mathbf{x}) \equiv \sum_{j=1}^{J} \alpha_j \rho_j(\mathbf{x}), \tag{2.12}$$

where the $\{\rho_j\}$ are known, suitably chosen "basis functions" (like polynomials or trigonometric functions) that yield an arbitrarily close approximation to g as the number of terms J in the sum is increased. For this approximation, the corresponding approximation to the reduced form $E[y|\mathbf{z}] = E[g(\mathbf{x})|\mathbf{z}]$ is

$$E[g_J(\mathbf{x})|\mathbf{z}] = \sum_{j=1}^{J} \alpha_j E[\rho_j(\mathbf{x})|\mathbf{z}], \tag{2.13}$$

which is itself linear in the same parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_J)'$, so that constrained least-squares regression of y on nonparametric estimates $\hat{E}[\rho_j(\mathbf{x})|\mathbf{z}]$ of the conditional means of the basis functions $\rho_j(\mathbf{x})$ can be used to estimate the coefficients of the approximate structural function g_J under the compactness restriction. Furthermore, the square of the Sobolev norm $\|g_J\|_S$ of the linear approximating function g_J can be written as a quadratic form,

$$\|g_J\|_S^2 = \tfrac{1}{2}\boldsymbol{\alpha}'\mathbf{S}_J\boldsymbol{\alpha}, \tag{2.14}$$

where the matrix $\mathbf{S}_J$ is a known matrix constructed by using integrals involving the basis functions $\rho_j(\mathbf{x})$ and their derivatives. Minimization of the sum of squared differences between observed values of y and the estimators

$\{\hat{E}[\rho_j(\mathbf{x})|\mathbf{z}]\}$, subject to the restriction that the quadratic form in (2.14) is bounded above by a known constant B, yields an estimator of the coefficient vector $\boldsymbol{\alpha}$ that is of "penalized least squares" form:

$$\hat{\alpha} = (\widehat{\mathbf{R}}'\widehat{\mathbf{R}} + \hat{\lambda}\mathbf{S}_J)^{-1}\widehat{\mathbf{R}}'\mathbf{y}, \tag{2.15}$$

where $\hat{\mathbf{R}}$ is the matrix of the first-stage estimators $\{\hat{E}[\rho_j(\mathbf{x})|\mathbf{z}]\}$ for the sample, and $\hat{\lambda}$ is a Lagrange multiplier for the constraint $\boldsymbol{\alpha}'\mathbf{S}_J\boldsymbol{\alpha} \leq 2B$. Imposition of the compactness condition thus introduces an adjustment for multicollinearity (through the term $\hat{\lambda}\mathbf{S}_J$) to the otherwise-familiar 2SLS formula, to account for the near-singularity of the fitted values in the first stage, which is at the heart of the ill-posed inverse problem.

Newey and Powell (1989) give conditions under which the resulting estimator of the structural function g,

$$\hat{g}(\mathbf{x}) \equiv \sum_{j=1}^{J} \hat{\alpha}_j \rho_j(\mathbf{x}), \tag{2.16}$$

is consistent; in addition to the compactness restrictions on the set of possible structural functions, these conditions restrict the form of the basis functions ρ_j and require that the number of terms J in the approximating function increase to infinity with the sample size. However, unlike some other nonparametric regression methods based on series approximations, J can be arbitrarily large for finite samples, and its value need not be related to the sample size to ensure convergence of bias and variance to zero, but instead is governed by the trade-off between numerical precision of the series approximation to g and computational convenience. The Newey and Powell manuscript does not discuss the rate of convergence or asymptotic distribution of $\hat{g}$, nor appropriate choice of the constraint constant B or, equivalently, the Lagrange multiplier $\hat{\lambda} = \hat{\lambda}(B)$, which acts as a smoothing parameter in the second-stage estimator.

A conceptually simple variant of this estimation strategy can be based on a finite-support approximation to the joint distribution of $\mathbf{x}$ and $\mathbf{z}$. Once the data are binned into partitions with representative values $\{\boldsymbol{\xi}_j\}$ and $\{\boldsymbol{\zeta}_l\}$, the linear relation (2.10) between the vector $\mathbf{g}$ of structural function values and the reduced-form vector $\boldsymbol{\pi}$ and transition matrix $\mathbf{P}$ will hold approximately (with the approximation improving as the number of bins increases), and the components $\boldsymbol{\pi}$ and $\mathbf{P}$ can be estimated using bin averages and frequencies. Though the estimated transition matrix $\hat{\mathbf{P}}$ may be nearly singular even if the approximating bins are chosen with $L \gg J$ for J large, the structural function vector $\mathbf{g}$ could be estimated by ridge regression, that is,

$$\hat{\mathbf{g}} = (\hat{\mathbf{P}}'\hat{\mathbf{P}} + \lambda\mathbf{S})^{-1}\hat{\mathbf{P}}'\hat{\pi}, \tag{2.17}$$

for some nonsingular matrix $\mathbf{S}$ and smoothing parameter λ that shrinks to zero as the sample size increases. This can be viewed as a histogram version of the series estimator proposed by Newey and Powell, which uses bin indicators as the basis functions and kernel regression (with uniform kernels) in the first stage.

Darolles et al. (2000) take a different approach to the estimation of the structural function g in (2.8). They embed the problem of the solution of (2.8) in the mean-squared error minimization problem, defining the structural function g as

$$g(\cdot) \equiv \underset{\phi(\cdot)}{\operatorname{argmin}}\, E\left[\left\| E[y|\mathbf{z}] - \int \phi(\mathbf{x}) dF_{\mathbf{x}|\mathbf{z}} \right\|^2\right], \tag{2.18}$$

and note that the "normal equations" for this functional minimization problem are of the form

$$\begin{aligned} E[E[y|\mathbf{z}]|\mathbf{x}] &\equiv \tau(\mathbf{x}) \\ &= E[E[g(\mathbf{x})|\mathbf{z}]|\mathbf{x}] \\ &\equiv T^*(g)(\mathbf{x}). \end{aligned} \tag{2.19}$$

That is, they transform (2.8) into another integral equation by taking conditional expectations of the reduced form $E[y|\mathbf{z}]$ given the original explanatory variables $\mathbf{x}$; an advantage of this formulation is that the transformation $T^*(g) = \tau$ has the same argument ($\mathbf{x}$) as the structural function g, as is standard for the literature on the solution of Fredholm integral equations. Although the ill-posed inverse problem persists, a standard solution method for this formulation is Tikhonov regularization, which replaces the integral equation (2.19) with the approximate problem

$$\tau(\mathbf{x}) = T^*(g^\lambda)(\mathbf{x}) + \lambda g^\lambda(\mathbf{x}), \tag{2.20}$$

for λ being a small, nonnegative smoothing parameter. Although (2.20) reduces to (2.19) as $\lambda \to 0$, it is a Fredholm integral equation of the *second* kind, which is free of the ill-posed inverse problem, when λ is nonzero. Again approximating the solution function g^λ as a linear combination of basis functions,

$$g^\lambda(\mathbf{x}) \cong g^\lambda_J(\mathbf{x}) \equiv \sum_{j=1}^{J} \alpha^\lambda_j \rho_j(\mathbf{x}), \tag{2.21}$$

as in (2.12), a further approximation to the equation (2.19) is

$$E[E[y|\mathbf{z}]|\mathbf{x}] \cong \sum_{j=1}^{J} \alpha^\lambda_j \{E[E[\rho_j(\mathbf{x})|\mathbf{z}]|\mathbf{x}] + \lambda\rho_j(\mathbf{x})\}. \tag{2.22}$$

This suggests a two-stage strategy for estimation of the $\boldsymbol{\alpha}^\lambda$ coefficients: In the first stage, obtain nonparametric estimators of the components $\tau(\mathbf{x}) = E[E[y|\mathbf{z}]|\mathbf{x}]$ and the doubly averaged basis functions $\{T^*(\rho_j)(\mathbf{x}) = E[E[\rho_j(\mathbf{x})|\mathbf{z}]|\mathbf{x}]\}$ using standard nonparametric estimation methods; then, in the second stage, regress the fitted $\hat{\tau}(\mathbf{x})$ on the constructed regressors $\{\hat{T}^*(\rho_j)(\mathbf{x}) - \lambda\rho_j(\mathbf{x}), j = 1, \ldots, J\}$. The terms $\lambda\rho_j(\mathbf{x})$ serve to attenuate the severe multicollinearity of the doubly averaged basis functions in this second-stage regression.

Darolles, Florens, and Renault take the basis functions ρ_j to be the eigenfunctions of the estimated double-averaging operator $\hat{T}^*$, that is, the solutions to the functional equations $\hat{T}^*(\rho_j) = \nu_j \rho_j$ for scalar eigenvalues ν_j, which simplifies both computation of the estimator and derivation of the asymptotic theory. For this choice of basis function, they derive the rate of convergence and asymptotic normal distribution of the estimator $\tilde{g}^\lambda(\mathbf{x}) = \sum_{j=1}^{J} \hat{\alpha}_j^\lambda \rho_j(\mathbf{x})$ under certain regularity conditions. The rate of convergence is comparable to, but slower than, the rate of convergence of standard nonparametric estimators of the reduced form $E[y|\mathbf{z}]$, as a result of the bias introduced by approximating (2.19) by (2.20) for nonzero λ. Their manuscript also proposes an alternative estimator of g based on regression of $\hat{\tau}(\mathbf{x})$ on a subset of the doubly averaged basis functions with eigenvalues bounded away from zero, that is, $\{\hat{T}^*(\rho_j)(\mathbf{x}) : |\nu_j| > b_n\}$ for some $b_n \to 0$, and extends the identification analysis to permit the structural function g to be additively separable in its endogenous and exogenous components.

2.1.5. Nonadditive Models

Both the Newey–Powell and Darolles–Florens–Renault approaches exploit the additive separability of the error terms u in the structural function for y; for models with nonadditive errors, that is, $H(\mathbf{x}, u) \neq g(\mathbf{x}) + u$, the IV assumption imposed in these papers apparently does not suffice to identify the ASF G of (1.10). Of course, it is clear that, for a nonadditive model, the conditional mean assumption (2.7) would not suffice to yield identification even for parametric structural functions, but imposition of the still-stronger assumption of independence of u and $\mathbf{z}$, denoted here as

$$u \perp\!\!\!\perp \mathbf{z} \tag{2.23}$$

[which implies (2.7), and thus (1.3), provided u has finite expectation which can be normalized to zero], will still not suffice in general for identification of the ASF G. This is evident from inspection of the reduced form $E[y|\mathbf{z}]$ in the nonadditive case:

$$\begin{aligned} E[y|\mathbf{z}] &= E[H(\mathbf{x}, u)|\mathbf{z}] \\ &= \int H(\mathbf{x}, u) dF_{u,\mathbf{x}|\mathbf{z}} \\ &= \int \left[\int H(\mathbf{x}, u) dF_{u|\mathbf{x},\mathbf{z}} \right] dF_{\mathbf{x}|\mathbf{z}} \\ &\neq \int \left[\int H(\mathbf{x}, u) dF_u \right] dF_{\mathbf{x}|\mathbf{z}} \\ &= E[G(\mathbf{x})|\mathbf{z}]. \end{aligned} \tag{2.24}$$

That is, independence of u and $\mathbf{z}$ does not imply independence of u and $\mathbf{x}$, $\mathbf{z}$, or even conditional independence of u and $\mathbf{z}$ given $\mathbf{x}$. In the additive case

$H(\mathbf{x}, u) = g(\mathbf{x}) + u$, conditional expectations of each component require only the conditional distributions of u given $\mathbf{z}$ and of $\mathbf{x}$ given $\mathbf{z}$, and not the joint distribution of u, $\mathbf{x}$ given $\mathbf{z}$, which is not identified under (2.23) without further conditions on the relation of $\mathbf{x}$ to $\mathbf{z}$. Furthermore, because (2.24) also holds in general for any function of y, restriction (2.23) does not yield restrictions on the conditional distribution of the observable y given $\mathbf{z}$ that might be used to identify the ASF G for general nonseparable structural functions H.

Of course, failure of the reduced-form relation (2.24) to identify the ASF G does not directly imply that it could not be identified by using some other functionals of the joint distribution of the observables y, $\mathbf{x}$, and $\mathbf{z}$, and it is difficult to provide a constructive proof of nonidentification of the ASF under the independence restriction (2.23) at this level of generality (i.e., with structural function H and the joint distribution of u, $\mathbf{x}$, and $\mathbf{z}$ otherwise unspecified). Still, the general nonidentification result can be illustrated by considering a simple (slightly pathological) binary response example in which the ASF is unidentified under (2.23). Suppose H is binary, with y generated as

$$y = 1(x + u \geq 0), \tag{2.25}$$

for a scalar regressor x generated by a multiplicative model

$$x = z \cdot e, \tag{2.26}$$

for some scalar instrumental variable z with $\Pr\{z \geq 0\} = 1$. For this example, the ASF is

$$G(x) \equiv 1 - F_u(x), \tag{2.27}$$

with F_u the marginal CDF of u. Now suppose the errors u and e are generated as

$$\begin{aligned} u &\equiv \varepsilon \cdot \operatorname{sgn}(\eta), \\ e &\equiv \eta \cdot \operatorname{sgn}(\varepsilon), \end{aligned} \tag{2.28}$$

with ε, η, and z being mutually independently distributed, and

$$\operatorname{sgn}(\eta) \equiv 1 - 2 \cdot 1(\eta < 0).$$

This model is pathological because $\operatorname{sgn}(u) = \operatorname{sgn}(x)$ by construction, and thus $y = 1(x \geq 0)$, whenever $z \neq 0$. Still, the independence condition (2.23) is satisfied, the dependent variable y is nonconstant if the support of e includes positive and negative values, and the endogenous regressor x has a conditional expectation that is a nontrivial function of the instrument z when $E[e] = E[\eta] \cdot (1 - 2 \cdot \Pr\{\varepsilon < 0\}) \neq 0$. Nevertheless, the ASF $G(x)$ is identified only at $x = 0$, when zero is in the support of z (with $G(0) = E[y|z = 0]$) and is not identified elsewhere.

This example demonstrates that, without further restrictions on the form of the structural function H and/or the conditional distribution of u given $\mathbf{x}$

and $\mathbf{z}$, the assumption of independence of the structural error u and the instruments $\mathbf{z}$ is insufficient to identify the ASF in nonadditive models even when the endogenous regressors $\mathbf{x}$ are not independent of the instruments $\mathbf{z}$. The nonidentification of the ASF here is a consequence of the dependence of the support of the endogenous variable x (either zero or the positive or negative half-line) on the realized value of the error term e. In general, if the nature of the endogeneity of $\mathbf{x}$ is restricted by assuming

$$\mathbf{x} = h(\mathbf{z}, \mathbf{e}) \tag{2.29}$$

for some function h that is invertible in the error terms $\mathbf{e}$, that is,

$$\mathbf{e} = k(\mathbf{z}, \mathbf{x}), \tag{2.30}$$

and if the support of $\mathbf{x}$ given $\mathbf{e}$ is independent of $\mathbf{e}$. Imbens (2000) has shown how the ASF G is identified under the independence restriction (2.23) and these additional restrictions on the nature of the endogeneity of $\mathbf{x}$. Imbens' identification argument is based on the control function approach described in more detail later.

As an alternative to imposing such restrictions on the nature of the endogeneity of $\mathbf{x}$, additional structure on the form of the structural function H – such as invertibility of the structural function $H(\mathbf{x}, u)$ in the error term u – may yield more scope for identification of the ASF when the stochastic restrictions involve only the conditional distribution of u given $\mathbf{z}$. For example, suppose there is some invertible transformation $t(y)$ of y for which the additive form (1.11) holds:

$$t(y) = g(\mathbf{x}) + u, \tag{2.31}$$

where u satisfies the conditional mean restriction (2.7). If the transformation t were known, then estimation of g could proceed by using the Newey–Powell or Darolles–Florens–Renault approaches, and the ASF could be estimated by averaging the estimator of $H(\mathbf{x}, u) = t^{-1}(g(\mathbf{x}) + u)$ over the marginal empirical distribution of $u = t(y) - g(\mathbf{x})$. When t is unknown, the conditional mean restriction (2.7) yields an integral equation

$$\begin{aligned} 0 &= E[u|\mathbf{z}] \\ &= \int t(y) dF_{y|\mathbf{z}} - \int g(\mathbf{x}) dF_{\mathbf{x}|\mathbf{z}}, \end{aligned} \tag{2.32}$$

which has multiple solutions, such as $t(y) \equiv g(\mathbf{x}) \equiv k$ for any constant k. Still, with appropriate normalizations on the unknown functions t and g, like $E[t(y)] = 0$ and $E[(t(y))^2] = 1$, it may be possible to extend the estimation approaches for the ill-posed inverse problem to joint estimation of t and g, though this may require overidentification, that is, $m = \dim(\mathbf{z}) > \dim(\mathbf{x}) = k$.

For the semiparametric problems (1.7), the parametric components $\boldsymbol{\beta}$ of the structural function may well be identified and consistently estimable, at the parametric (root-N) rate, even if the ASF G is not identified. Ai and Chen (2000) propose sieve estimation of semiparametric models of the form (1.9) under the assumption that the instrumental variables $\mathbf{z}$ are independent of the error terms u; although the estimator of the infinite-dimensional nuisance function $h(\cdot)$ is not generally consistent with respect to the usual distance measures (such as integrated square differences), the corresponding estimator $\hat{\boldsymbol{\beta}}$ of the parametric component $\boldsymbol{\beta}$ is root-N consistent and asymptotically normal under the regularity conditions they impose.

Lewbel (1998, 2000) considers the single-index generalized regression model (1.8), constructing consistent estimators of the index coefficients $\boldsymbol{\beta}$ under the assumption that one of the components of the explanatory variables $\mathbf{x}$, say x_1, is continuously distributed and independent of the structural error u – and is thus a component of the set of instruments $\mathbf{z}$ satisfying (2.23) given earlier. Provided there exists an exogenous variable x_1 that satisfies these conditions, Lewbel's approach permits a weaker stochastic restriction than independence of $\mathbf{z}$ (including the special regressor $\mathbf{x}_1$) and u – namely, that u need only be independent of x_1 conditionally on the other components of $\mathbf{x}$ and of the instrument vector $\mathbf{z}$. The conditional mean restriction $E[u|\mathbf{z}] = 0$ can also be weakened to the moment restriction $E[\mathbf{z}u] = 0$ in this setup. The conditional independence restriction is similar to the restrictions imposed for the control function methods described later. Nevertheless, even if the coefficient vector $\boldsymbol{\beta}$ were known a priori, the endogeneity of the remaining components of $\mathbf{x}$, and thus of the index $\mathbf{x}'\boldsymbol{\beta}$, would yield the same difficulties in identification of the ASF G as in (2.24).

2.1.6. Fitted-Value Methods

When the conditional mean (2.7) or independence (2.23) restrictions of the IV approach does not suffice to identify the ASF G in a nonparametric model, the researcher can either abandon the ASF concept and focus on alternative summary measures that are identified, or impose stronger restrictions on the structural function or error distributions to achieve identification of the ASF. Though imposing additional restrictions on the structural function H (such as additivity of the error terms u) can clearly help achieve identifiability of G, such restrictions may be implausible when the range of y is restricted (e.g., when y is binary), and it is more customary to strengthen the restrictions on the conditional error distribution u given the instruments $\mathbf{z}$ to identify the parameters of interest.

One alternative set of restrictions and estimation procedures are suggested by Theil's (1953) version of the 2SLS estimator for simultaneous equations. Defining the first-stage residuals $\mathbf{v}$ as the difference between the regressors $\mathbf{x}$ and their linear projections onto $\mathbf{z}$,

$$\mathbf{v} \equiv \mathbf{x} - P[\mathbf{x}|\mathbf{z}] \equiv \mathbf{x} - \Pi'\mathbf{z}, \tag{2.33}$$

where Π is defined in (2.2), the condition (1.3), when combined with the definition of $\mathbf{v}$ and the linear structural function (1.1), yields the restriction

$$\begin{aligned} 0 &= E[P[\mathbf{x}|\mathbf{z}](u + \mathbf{v}'\boldsymbol{\beta})] \\ &= E[(\Pi'\mathbf{z})(y - (\mathbf{z}'\Pi)\boldsymbol{\beta}), \end{aligned} \tag{2.34}$$

so that the structural coefficients $\boldsymbol{\beta}$ are the least-squares regression coefficients of the regression of the dependent variable y on the fitted values $\Pi'\mathbf{z}$ of the regressors $\mathbf{x}$. The sample analogue of the population regression coefficients of y on $\Pi'\mathbf{z}$ is Theil's version of 2SLS,

$$\widehat{\boldsymbol{\beta}}_{2SLS} = (\widehat{\mathbf{X}}'\widehat{\mathbf{X}})^{-1}\widehat{\mathbf{X}}'\mathbf{y}, \tag{2.35}$$

where $\widehat{\mathbf{X}} = \mathbf{Z}\widehat{\Pi}$ is defined as in (2.4). The motivation for this form of 2SLS is the replacement of the endogenous regressors $\mathbf{x}$ with that part of $\mathbf{x}$ (its linear projection on $\mathbf{z}$) that is uncorrelated with the error u in the linear structural equation.

In a nonparametric setting, it is natural to define the first-stage residuals $\mathbf{v}$ as deviations from conditional expectations, rather than linear projections:

$$\begin{aligned} \mathbf{v} &\equiv \mathbf{x} - E[\mathbf{x}|\mathbf{z}] \\ &\equiv \mathbf{x} - \Pi(\mathbf{z}). \end{aligned} \tag{2.36}$$

By construction, $E[\mathbf{v}|\mathbf{z}] = \mathbf{0}$, and, as for the IV approaches, the moment condition (1.3) would be replaced by the stronger conditional mean restriction (2.7), or the still-stronger assumption of independence of the errors and the instruments,

$$(u, \mathbf{v}) \perp\!\!\!\perp \mathbf{z}. \tag{2.37}$$

A nonparametric generalization of Theil's version of 2SLS would estimate $E[\mathbf{x}|\mathbf{z}] = \Pi(\mathbf{z})$ by a suitable nonparametric method in the first stage, and then substitute the fitted values $\tilde{\Pi}(\mathbf{z})$ into the structural function in a second-stage estimation procedure. As noted by Amemiya (1974), though, substitution of fitted values into nonlinear structural functions generally yields inconsistent estimates of the structural parameters, even in parametric problems; estimation methods that use substitution of fitted values into the structural function rely heavily on linearity of the regression function, so that the model can be written in terms of a composite error $u + \mathbf{v}'\boldsymbol{\beta}$ with similar stochastic properties to the structural error u. For the general (nonadditive) structural function H of (1.4), substitution of the reduced form into the structural function yields $y = H(\Pi(\mathbf{z}) + \mathbf{v}, u)$, and, analogously to (2.24), the reduced form for y bears no obvious relation to the ASF G under condition (2.37). Even when the structural function H is additive, $H(\mathbf{x}, u) = g(\mathbf{x}) + u$, the reduced form for y can be written as

$$E[y|\mathbf{z}] = E[\mathbf{y}|\Pi(\mathbf{z})] = \int g(\Pi(\mathbf{z}) + \mathbf{v}) dF_{\mathbf{v}},$$

so insertion of the first-stage equation for $\mathbf{x}$ into the structural function yields an ill-posed inverse relation between the reduced form for y and the ASF g. Thus the fitted-value approach inherits similar limitations to the IV approach, though it may simplify the resulting integral equation, which depends on $\Pi(\mathbf{z})$ rather than $\mathbf{z}$ and involves the marginal distribution of $\mathbf{v}$ rather than the conditional distribution of $\mathbf{x}$ given $\mathbf{z}$.

Of course, for structural equations in which the conditional expectations $E[\mathbf{x}|\mathbf{z}] = \Pi(\mathbf{z})$ are the "right" explanatory variables, the fitted-value estimation method, using nonparametric estimates of $\Pi(\mathbf{z})$, is an obvious way to proceed.[3] And, as was true for the IV approach, for semiparametric problems, consistent estimation of the parametric component $\boldsymbol{\beta}$ may be feasible even when the ASF G is not identified. For example, for the generalized regression model $y = h(\mathbf{x}'\boldsymbol{\beta}, u)$ of (1.8), the reduced form for y given $\mathbf{z}$ is of single-index form when the errors $u, \mathbf{v}$ are independent of the instrument vector $\mathbf{z}$: $E[y|\mathbf{z}] \equiv G^*(\Pi(\mathbf{z})'\boldsymbol{\beta})$ for some function G^*, so that a nonparametric estimator of the first-stage regression function $\Pi(\mathbf{z})$ can be combined with a standard estimation method for single-index regression models.[4]

2.2. Control Function Methods

2.2.1. The Linear Model

Although insertion of the fitted values from the first-stage nonparametric regression of $\mathbf{x}$ on $\mathbf{z}$ is not generally helpful in identification and estimation of the ASF G, alternative assumptions and procedures involving the use of the residuals $\mathbf{v}$ from this first-stage regression to control for endogeneity of the regressors $\mathbf{x}$ do yield identification of the ASF even for the general, nonadditive structural function (1.4). This control function approach has its antecedent in another algebraically equivalent interpretation of the 2SLS estimator $\hat{\boldsymbol{\beta}}_{2SLS}$ as the coefficients on $\mathbf{x}$ in a least-squares regression of y on $\mathbf{x}$ and the residuals $\hat{\mathbf{v}}$ from a linear regression of $\mathbf{x}$ on $\mathbf{z}$:

$$\begin{pmatrix} \widehat{\boldsymbol{\beta}}_{2SLS} \\ \widehat{\boldsymbol{\rho}}_{2SLS} \end{pmatrix} = (\widehat{\mathbf{W}}'\widehat{\mathbf{W}})^{-1}\widehat{\mathbf{W}}'\mathbf{y}, \tag{2.38}$$

where

$$\widehat{\mathbf{W}} = [\mathbf{X}\ \widehat{\mathbf{V}}] \text{ and } \widehat{\mathbf{V}} = \mathbf{X} - \widehat{\mathbf{X}} = \mathbf{X} - \mathbf{Z}\widehat{\Pi},$$

[3] Some asymptotic results for such estimators were given by Ahn and Manski (1993) and by Ahn (1995), which showed, for example, that the rate of convergence of the restricted reduced-form estimator $\hat{E}[y|\hat{\Pi}(\mathbf{z})]$ is the smaller of the rate of convergence of $\hat{\Pi}$ to Π and of $\hat{E}[y|\Pi(\mathbf{z})]$ to $E[y|\Pi(\mathbf{z})]$ where Π is known.

[4] See, for example, Ichimura (1993) and others described in Horowitz (1993) and Powell (1994).

and where $\hat{\boldsymbol{\rho}}_{2SLS}$ are the coefficients on the first-stage residuals $\hat{\mathbf{V}}$.[5] This construction exploits another consequence of the moment condition $E(\mathbf{z}u) = \mathbf{0}$, that

$$\begin{aligned} P[u|\mathbf{x}, \mathbf{z}] &= P[u|\Pi'\mathbf{z} + \mathbf{v}, \mathbf{z}] \\ &= P[u|\mathbf{v}, \mathbf{z}] \\ &= P[u|\mathbf{v}] \\ &\equiv \mathbf{v}'\boldsymbol{\rho} \end{aligned} \tag{2.39}$$

for some coefficient vector $\boldsymbol{\rho}$; the third equality follows from the orthogonality of both error terms u and $\mathbf{v}$ with $\mathbf{z}$. Thus, this particular linear combination of the first-stage errors $\mathbf{v}$ is a function that controls for the endogeneity of the regressors $\mathbf{x}$. It also follows from this formulation that

$$P[u|\mathbf{x}, \mathbf{z}] = P[u|\mathbf{x}, \mathbf{v}], \tag{2.40}$$

which can be used as the basis for a test of overidentification if $\dim(\mathbf{z}) > \dim(\mathbf{v}) = \dim(\mathbf{x})$.

This approach treats endogeneity as an omitted variable problem, where the inclusion of estimates of the first-stage errors $\mathbf{v}$ (the part of the regressors $\mathbf{x}$ that is correlated with $\mathbf{z}$) as a covariate corrects the inconsistency of least-squares regression of y on $\mathbf{x}$, in the same way that the Heckman (1979) two-step estimator corrects for selectivity bias through introduction of an appropriately estimated regressor derived from a parametric form for the error distribution. The control function approach to correct for endogeneity has been extended to nonlinear parametric models by Blundell and Smith (1986, 1989), who show how introduction of first-stage residuals into single-equation Probit or Tobit procedures yields consistent estimators of the underlying regression coefficients when some of the regressors are endogenous.

2.2.2. *Extensions to Additive Nonparametric Models*

Application of the control function approach to nonparametric and semiparametric settings requires strengthening of the linear projection restrictions (2.39) and (2.40) to conditional mean restrictions,

$$\begin{aligned} E(u|\mathbf{x}, \mathbf{z}) &= E(u|\mathbf{x}, \mathbf{v}) \\ &= E(u|\mathbf{v}), \end{aligned} \tag{2.41}$$

[5] It has been difficult to locate a definitive early reference to the control function version of 2SLS. Dhrymes (1970, equation 4.3.57) shows that the 2SLS coefficients can be obtained by a least-squares regression of $\mathbf{y}$ on $\hat{\mathbf{X}}$ and $\hat{\mathbf{V}}$, while Telser (1964) shows how the seemingly unrelated regressions model can be estimated by using residuals from other equations as regressors in a particular equation of interest. Heckman (1978) references this paper in his comprehensive discussion of estimating simultaneous models with discrete endogenous variables.

or, for nonadditive models, the stronger conditional independence assumptions

$$\begin{aligned} u|\mathbf{x}, \mathbf{z} &\sim u|\mathbf{x}, \mathbf{v} \\ &\sim u|\mathbf{v}. \end{aligned} \tag{2.42}$$

While the control variates[6] $\mathbf{v}$ are typically taken to be deviations of $\mathbf{x}$ from its conditional mean $E[\mathbf{x}|\mathbf{z}]$, as in (2.36), this is not required; more generally, $\mathbf{v}$ can be any function of the observable random vectors

$$\mathbf{v} = \boldsymbol{\nu}(y, \mathbf{x}, \mathbf{z}) \tag{2.43}$$

that is identified and consistently estimable, provided (2.41) or (2.42) holds for this formulation (and $\mathbf{v}$ is not a nontrivial function of $\mathbf{x}$). This permits the control function approach to be applied to some systems of nonlinear simultaneous equations for which the appropriate reduced form is difficult or impossible to derive, such as the coherent simultaneous binary response models considered in the paragraphs that follow. In comparison to the identifying assumptions (2.7) or (2.23) for the IV approaches, the corresponding assumptions (2.41) or (2.42) for the control function approach are no more nor less general. Both sets of assumptions are implied by the independence restriction (2.37), which may be plausible for certain applications, and which permits a choice between the two estimation approaches.

Estimation of the ASF g in the additive model (1.11) under the conditional mean exclusion restriction (2.41) was considered by Newey, Powell, and Vella (1999) and by Ng and Pinkse (1995) and Pinkse (2000). Applications can be found in Blundell, Browning, and Crawford (2000) and Blundell and Duncan (1998), for example. When the errors are additive, substitution of $u = y - g(\mathbf{x})$ into (2.41) yields a generalized additive regression form for y:

$$\begin{aligned} E[y|\mathbf{x}, \mathbf{v}] &= E[(g(\mathbf{x}) + u)|\mathbf{x}, \mathbf{v}] \\ &= g(\mathbf{x}) + \eta(\mathbf{v}), \end{aligned} \tag{2.44}$$

for some control function η. With a suitable normalization, say, $E[\eta(\mathbf{v})] = 0$, the ASF g can be estimated using standard additive nonparametric regression methods applied to the regression of y on $\mathbf{x}$ and the first-stage residuals $\hat{\mathbf{v}}$. Both Newey et al. and Ng and Pinkse propose an estimation of (2.44) using a series approximation for the functions g and η :

$$g(\mathbf{x}) + \eta(\mathbf{v}) \cong \sum_{j=1}^{J} \alpha_j \rho_j(\mathbf{x}) + \sum_{l=1}^{L} \gamma_l \psi_l(\mathbf{v}), \tag{2.45}$$

where $\{\rho_j\}$ and $\{\psi_l\}$ are appropriate basis functions, and where the number of terms J and L for each approximating series increases to infinity as the sample size increases. The second stage using this series approximation is a least-squares regression of y on the basis functions $\{\rho_j(\mathbf{x})\}$ and $\{\psi_l(\hat{\mathbf{v}})\}$, and

[6] This use of the term "control variate" is logically distinct from, but similar in spirit to, its use in the literature on Monte Carlo methods; see, for example, Hammersley and Handscomb (1964), Section 5.5.

the estimator of g is given by (2.16), assuming $\rho_1(\mathbf{x}) \equiv 1-$ which enforces the normalization $E[\eta(\mathbf{v})] = 0$. The cited manuscripts give regularity conditions for consistency of the estimator $\hat{g}$, and derive its rate of convergence, which is the same as the rate for a direct nonparametric regression of y on the regressors $\mathbf{x}$; Newey et al. also give conditions under which the estimator $\hat{g}(\mathbf{x})$ is asymptotically normal.

2.2.3. *Nonadditive Models*

Unlike the IV approach, a stronger independence condition (2.42) of the conditional mean exclusion restrictions (2.41) for the control function approach does lead to a consistent estimator of the ASF G when the structural function H is nonadditive, as in (1.4). Blundell and Powell (1999) point out how averaging the conditional mean of y given $\mathbf{x}$ and $\mathbf{v}$ over the marginal distribution of the first-stage errors $\mathbf{v}$ gives the ASF G for the nonadditive model. Because

$$\begin{aligned} E[y|\mathbf{x}, \mathbf{v}] &= E[H(\mathbf{x}, u)|\mathbf{x}, \mathbf{v}] \\ &= \int H(\mathbf{x}, u) dF_{u|\mathbf{x},\mathbf{v}} \\ &= \int H(\mathbf{x}, u) dF_{u|\mathbf{v}} \\ &\equiv H^*(\mathbf{x}, \mathbf{v}), \end{aligned} \tag{2.46}$$

under the strong exclusion restriction (2.42), it follows that the generalized control function H^* can be integrated over the marginal distribution of the (observable) reduced-form errors to obtain the ASF:

$$\begin{aligned} \int H^*(\mathbf{x}, \mathbf{v}) dF_{\mathbf{v}} &= \int \left[\int H(\mathbf{x}, u) dF_{u|\mathbf{v}} \right] dF_{\mathbf{v}} \\ &= \int H(\mathbf{x}, u) dF_u \\ &\equiv G(\mathbf{x}). \end{aligned} \tag{2.47}$$

In a sense, the control function exclusion restriction (2.42) permits replacement of the unidentified structural errors u with the identified control variable $\mathbf{v}$ through iterated expectations, so that averaging the structural function H over the marginal distribution of the structural errors u is equivalent to averaging the (identified) intermediate regression function H^* over the marginal distribution of $\mathbf{v}$. The intermediate structural function H^* is a nonadditive generalization of the previous control function $\eta(\mathbf{v}) = E[u|\mathbf{v}]$ for additive models.

For the binary response example described in (2.25) through (2.28) herein, the first-stage residuals are of the form

$$\begin{aligned} v &= z \cdot (e - E[e]) \\ &= z \cdot [\eta \cdot \mathrm{sgn}(\varepsilon) - E[\eta] \cdot (\Pr\{\varepsilon \geq 0\} - \Pr\{\varepsilon < 0\})], \end{aligned} \tag{2.48}$$

and the conditional exclusion restriction (2.42) holds only if the structural error u is degenerate, that is, $u = 0$ with probability one. In this case, x is exogenous, and the intermediate structural function reduces to

$$\begin{aligned} H^*(x, v) &= E[1(x \geq 0)|x, v] \\ &= 1(x \geq 0), \end{aligned} \tag{2.49}$$

which trivially integrates to the true ASF $G(x) = 1(x \geq 0)$. Thus, imposition of the additional restriction (2.42) serves to identify the ASF here. An alternative control variate to the first-stage errors v would be

$$v^* \equiv \text{sgn}(x) = \text{sgn}(u), \tag{2.50}$$

which satisfies (2.42) when the structural error u is nondegenerate. Because v^* is functionally related to x, the intermediate structural function $H^*(x, v^*) = E[y|x, v^*]$ is not identified with this control variate.

Translating the theoretical formulation (2.47) to a sampling context leads to a "partial mean" (Newey, 1994b) or "marginal integration" (Linton and Nielson, 1995 and Tjostheim and Auestad, 1996) estimator for the ASF G under the conditional independence restrictions (2.42) of the control function approach. That is, after obtaining a first-stage estimator $\hat{\mathbf{v}}$ of the control variate $\mathbf{v}$, which would be the residual from a nonparametric regression of $\mathbf{x}$ on $\mathbf{z}$ when $\mathbf{v}$ is defined by (2.36), one can obtain an estimator $\hat{H}^*$ of the function H^* in (2.46) by a nonparametric regression of y on $\mathbf{x}$ and $\hat{\mathbf{v}}$. A final estimation step would average H^* over the observed values of $\hat{\mathbf{v}}$,

$$\widehat{G}(\mathbf{x}) = \int \widehat{E}(y|\mathbf{x}, \mathbf{v}) d\widehat{F}_{\hat{\mathbf{v}}} \equiv \int \widehat{H}^*(\mathbf{x}, \mathbf{v}) d\widehat{F}_{\hat{\mathbf{v}}}\ , \tag{2.51}$$

where $\widehat{F}_{\hat{\mathbf{v}}}$ is the empirical CDF of the residuals $\hat{\mathbf{v}}$. Alternatively, if $\mathbf{v}$ were assumed to be continuously distributed with density $f_{\mathbf{v}}$, the ASF G could be estimated by integrating $\hat{H}^*$ over a nonparametric estimator $\hat{f}_{\hat{\mathbf{v}}}$ of $f_{\mathbf{v}}$,

$$\widetilde{G}(\mathbf{x}) = \int \widehat{E}(y|\mathbf{x}, \mathbf{v}) \widehat{f}_{\hat{\mathbf{v}}} d\mathbf{v}. \tag{2.52}$$

Either the partial mean (2.51) or marginal integration (2.52) is an alternative to the series estimators based on (2.45) for the additive structural function (1.11), but this latter approach is not applicable to general structural functions, because the intermediate regression function H^* need not be additive in its $\mathbf{x}$ and $\mathbf{v}$ components.

2.2.4. Support Restrictions

The identification requirements for the ASF G are simpler to interpret for the control function approach than the corresponding conditions for identification using the IV approaches, because they are conditions for identification of the nonparametric regression function H^*, which is based on observable random vectors. For example, in order for the ASF $G(\mathbf{x})$ to be identified from the partial-mean formulation (2.47) for a particular value $\mathbf{x}_0$ of $\mathbf{x}$, the support of the conditional distribution of $\mathbf{v}$ given $\mathbf{x} = \mathbf{x}_0$ must be the same as the support of the marginal distribution of $\mathbf{v}$; otherwise, the regression function $H^*(\mathbf{x}_0, \mathbf{v})$ will not be well defined for all $\mathbf{v}$, nor will the integral of H^* over the marginal distribution of $\mathbf{v}$. For those components of $\mathbf{x}$ that are exogenous, that is, those components of $\mathbf{x}$ that are also components of the instrument vector $\mathbf{z}$, the corresponding components of $\mathbf{v}$ are identically zero, both conditionally on $\mathbf{x}$ and marginally, so this support requirement is automatically satisfied. However, for the endogenous components of $\mathbf{x}$, the fact that $\mathbf{x}$ and $\mathbf{v}$ are functionally related through the first-stage relation $\mathbf{v} = \mathbf{x} - \Pi(\mathbf{z})$, or the more general form (2.43), means that the support condition generally requires that $\mathbf{v}$, and thus $\mathbf{x}$, must be continuously distributed, with unbounded support (conditionally on the instruments $\mathbf{z}$) if its marginal distribution is nondegenerate. Similar reasoning, imposing the requirement that $H^*(\mathbf{x}, \mathbf{v})$ be well defined on the support of $\mathbf{x}$ for all possible $\mathbf{v}$, and noting that

$$E[y|\mathbf{x}, \mathbf{v}] = E[y|\Pi(\mathbf{z}), \mathbf{v}] \tag{2.53}$$

implies that the first-stage regression function $\Pi(\mathbf{z}) = E[\mathbf{x}|\mathbf{z}]$ must also be continuously distributed, with full-dimensional support, for the nondegenerate components of $\mathbf{v} = \mathbf{x} - E[\mathbf{x}|\mathbf{z}]$.

The requirement that the endogenous components of $\mathbf{x}$ be continuously distributed is the most important limitation of the applicability of the control function approach to estimation of nonparametric and semiparametric models with endogenous regressors. When the structural equations for the endogenous regressors have limited or qualitative dependent variables, the lack of an invertible representation (2.43) of the underlying error terms for such models generally makes it impossible to construct a control variate $\mathbf{v}$ for which the conditional independence restrictions (2.41) or (2.42) are plausible. The requirement of an additive (or invertible) first-stage relation for the regressors $\mathbf{x}$ in the control function approach is comparable with the requirement of an additive (or invertible) structural function H for the identification of the ASF G using the IV approach.

In contrast, when the errors are invertible and the support of $\mathbf{x}$ does not depend on them, Imbens (2000) has shown how a particular control variate – the conditional cumulative distribution of $\mathbf{x}$ given $\mathbf{z}$, evaluated at the observed values of those random variables – can be used to identify the ASF G under the independence restriction (2.23), because it satisfies the conditional independence restriction (2.42). Also, when it is applicable, the control function approach to estimation with endogenous regressors is compatible with other estimation strategies that use control function methods to adjust for different sources of specification bias, such as selection bias (e.g., Ahn and Powell, 1993, Das, Newey, and Vella, 1998, Heckman, 1978, Heckman and Robb, 1985, Honoré and Powell, 1997, and Vytlacil, 1999) or correlated random effects in panel data models (Altonji and Matzkin, 1997).

3. BINARY RESPONSE LINEAR INDEX MODELS

3.1. Model Specification and Estimation Approach

Although the control function approach adopted by Blundell and Powell (1999) applies to fully nonparametric problems (as described herein), their discussion focuses attention on estimation of the parametric and nonparametric components of a particular semiparametric single-index model (1.8), the binary response model with linear index,

$$y = 1\{\mathbf{x}'\boldsymbol{\beta} + u > 0\}, \tag{3.1}$$

where the conditional independence assumption (2.42) is assumed to hold for $\mathbf{v} \equiv \mathbf{x} - E[\mathbf{x}|\mathbf{z}]$. For this linear binary response model, the ASF G is the marginal CDF of $-u$ evaluated at the linear index $\mathbf{x}'\boldsymbol{\beta}$,

$$G(\mathbf{x}) = F_{-u}(\mathbf{x}'\boldsymbol{\beta}), \tag{3.2}$$

which is interpreted as the counterfactual conditional probability that $y = 1$ given $\mathbf{x}$, if $\mathbf{x}$ were exogenous, that is, if the conditional distribution of u given $\mathbf{x}$ were assumed to be identical to its true marginal distribution. Under the conditional independence restriction (2.42), the intermediate regression function H^* is the conditional CDF of $-u$ given $\mathbf{v}$, again evaluated at $\mathbf{x}'\boldsymbol{\beta}$:

$$H^*(\mathbf{x}, \mathbf{v}) = F_{-u|\mathbf{v}}(\mathbf{x}'\boldsymbol{\beta}|\mathbf{v}). \tag{3.3}$$

Given a random sample of observations on y, $\mathbf{x}$, and $\mathbf{z}$ from the model (3.1) under the exclusion restriction (2.42), the estimation approach proposed by Blundell and Powell for the parameters of interest in this model follows three main steps. The first step uses nonparametric regression methods – specifically, the Nadaraya–Watson kernel regression estimator – to estimate the error term $\mathbf{v}$ in the reduced form, as well as the unrestricted conditional mean of y given $\mathbf{x}$ and $\mathbf{v}$,

$$E[y|\mathbf{x}, \mathbf{v}] \equiv H^*(\mathbf{w}), \tag{3.4}$$

where $\mathbf{w}$ is the $1 \times (k + q)$ vector

$$\mathbf{w} = (\mathbf{x}', \mathbf{v}')'. \tag{3.5}$$

This step can be viewed as an intermediate structural estimation step, which imposes the first exclusion restriction of (2.42) but not the second. The remaining estimation steps use semiparametric "pairwise differencing" or "matching" methods to obtain an estimator of the index coefficients $\boldsymbol{\beta}$, followed by partial-mean estimation of the ASF G.

3.1.1. The Semiparametric Estimator of the Index Coefficients

After using kernel regression methods to obtain an estimator $\hat{H}^*(\mathbf{x}, \hat{\mathbf{v}}) = \hat{E}[y|\mathbf{x}, \hat{\mathbf{v}}]$ of the intermediate regression function H^*, Blundell and Powell use a semiparametric estimation method to extract an estimator of $\boldsymbol{\beta}$ from the relation

$$\begin{aligned} H^*(\mathbf{w}) &= E[y|\mathbf{x}, \mathbf{v}] \\ &= E[y|\mathbf{x}'\boldsymbol{\beta}, \mathbf{v}] \\ &\equiv \Gamma(\mathbf{x}'\boldsymbol{\beta}, \mathbf{v}), \end{aligned} \tag{3.6}$$

which is a consequence of the single-index form of the binary response model (3.1). Although a number of standard methods for estimation of the coefficients of the single-index regression model $E[y|\mathbf{x}] = \Gamma(\mathbf{x}'\boldsymbol{\beta})$ could be extended to the multi-index model[7] (3.6), the particular estimator of $\boldsymbol{\beta}$ adopted by Blundell and Powell (1999) is an adaptation of a method proposed by Ahn, Ichimura, and Powell (1996), which imposes additional regularity conditions of both continuity and monotonicity of $H^*(\lambda, \mathbf{v}) = E[y|\mathbf{x}'\boldsymbol{\beta} = \lambda, \mathbf{v}]$ in its first argument. These conditions follow from the assumption that u is continuously distributed, with support on the entire real line, conditional on $\mathbf{v}$, because of (3.3).

Because the structural index model is related to the conditional mean of y given $\mathbf{w}' = (\mathbf{x}', \mathbf{v}')$ by the relation

$$H^*(\mathbf{w}) = \Gamma(\mathbf{x}'\beta_0, \mathbf{v}), \tag{3.7}$$

invertibility of $\Gamma(\cdot)$ in its first argument implies

$$\mathbf{x}'\beta_0 - \psi(g(\mathbf{w}), \mathbf{v}) = 0, (w.p.1), \tag{3.8}$$

where

$$\psi(\cdot, \mathbf{v}) \equiv \Gamma^{-1}(\cdot, \mathbf{v}), \tag{3.9}$$

that is, $\Gamma(\psi(g, \mathbf{v}), \mathbf{v}) \equiv g$. So, if two observations (with subscripts i and j) have identical conditional means, that is, $H^*(\mathbf{w}_i) = H^*(\mathbf{w}_j)$, and identical reduced-form error terms ($\mathbf{v}_i = \mathbf{v}_j$), it follows from the assumed invertibility of Γ that their indices $\mathbf{x}_i\beta_0$ and $\mathbf{x}_j\beta_0$ are also identical:

$$\begin{aligned} &(\mathbf{x}_i - \mathbf{x}_j)\beta_0 = \psi(g(\mathbf{w}_i), \mathbf{v}_i) - \psi(g(\mathbf{w}_j), \mathbf{v}_j) = \mathbf{0} \\ &\quad \text{if} \quad g(\mathbf{w}_i) = g(\mathbf{w}_j), \quad \mathbf{v}_i = \mathbf{v}_j. \end{aligned} \tag{3.10}$$

For any nonnegative function $\omega(\mathbf{w}_i, \mathbf{w}_j)$ of the conditioning variables $\mathbf{w}_i$ and $\mathbf{w}_j$, it follows that

$$\begin{aligned} 0 &= E[\omega(\mathbf{w}_i, \mathbf{w}_j) \cdot ((\mathbf{x}_i - \mathbf{x}_j)\beta_0)^2 \mid g(\mathbf{w}_i) = g(\mathbf{w}_j), \mathbf{v}_i = \mathbf{v}_j] \\ &\equiv \beta_0' \Sigma_w \beta_0, \end{aligned} \tag{3.11}$$

[7] See Horowitz (1993) and Powell (1994).

where

$$\Sigma_w \equiv E[\omega(\mathbf{w}_i, \mathbf{w}_j) \cdot (\mathbf{x}_i - \mathbf{x}_j)'(\mathbf{x}_i - \mathbf{x}_j) \mid g(\mathbf{w}_i) = g(\mathbf{w}_j), \mathbf{v}_i = \mathbf{v}_j]. \tag{3.12}$$

That is, the nonnegative-definite matrix Σ_w is singular, and, under the identifying assumption that Σ_w has rank $k - 1 = \dim(\mathbf{w})$ – which requires that any nontrivial linear combination $(\mathbf{x}_i - \mathbf{x}_j)a$ of the difference in regressors has nonzero variance when $a \neq 0$, $a \neq \beta_0$, $\mathbf{v}_i = \mathbf{v}_j$, and $(\mathbf{x}_i - \mathbf{x}_j)\beta_0 = 0$ – the unknown parameter vector β_0 is the eigenvector (with an appropriate normalization) corresponding to the unique zero eigenvalue of Σ_w.

Given the preliminary nonparametric estimators $\widehat{\mathbf{v}}_i$ and $\widehat{g}(\widetilde{\mathbf{w}}_i)$ of $\mathbf{v}_i$ and $g(\mathbf{w}_i)$ defined herein, and assuming smoothness (continuity and differentiability) of the inverse function $\psi(\cdot)$ in (3.9), one can obtain a consistent estimator of Σ_w for a particular weighting function $\omega(\mathbf{w}_i, \mathbf{w}_j)$ by a pairwise differencing or matching approach, which takes a weighted average of outer products of the differences $(\mathbf{x}_i - \mathbf{x}_j)$ in the $\binom{n}{2}$ distinct pairs of regressors, with weights that tend to zero as the magnitudes of the differences $|\widehat{g}(\widetilde{\mathbf{w}}_i) - \widehat{g}(\widetilde{\mathbf{w}}_j)|$ and $|\mathbf{v}_i - \mathbf{v}_j|$ increase. The details of this semiparametric estimation procedure for $\boldsymbol{\beta}$ are developed in Blundell and Powell (1999), who demonstrate consistency of the resulting estimator $\hat{\boldsymbol{\beta}}$ and characterize the form of its asymptotic (normal) distribution.

3.1.2. The Partial-Mean Estimator of the ASF

Once the consistent estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is obtained, the remaining parameter of interest for this model is $G(\mathbf{x}'\boldsymbol{\beta})$, the marginal probability that $y_{1i} = 1$ given an exogenous $\mathbf{x}$. The conditional cumulative distribution function $F_{-u|\mathbf{v}}(\mathbf{x}'\boldsymbol{\beta}|\ \mathbf{v}) \equiv \Gamma(\mathbf{x}'\boldsymbol{\beta}, \mathbf{v})$ is first estimated by using a kernel regression estimator $\hat{E}[y|\mathbf{x}'\hat{\boldsymbol{\beta}}, \hat{\mathbf{v}}]$; the Blundell–Powell approach then estimates $G(\overline{\lambda})$ from the sample average of $\hat{\Gamma}(\overline{\lambda}, \hat{\mathbf{v}}_i)$,

$$\widehat{G}(\overline{\lambda}) = \sum_{i=1}^{n} \widehat{\Gamma}(\overline{\lambda}, \widehat{\mathbf{v}}_i)\tau_i, \tag{3.13}$$

where τ_i is some “trimming” term that downweights observations for which Γ is imprecisely estimated. Consistency of this approach requires adapting the arguments in Newey (1994b) and Linton and Nielson (1995) for the case of the averaging over the estimated residual $\widehat{\mathbf{v}}_i$.

4. COHERENCY AND ALTERNATIVE SIMULTANEOUS REPRESENTATIONS

One interpretation of the linear index binary response model described in Section 3 is as the “triangular form” of some underlying joint decision problem. For simplicity, suppose the explanatory variables $\mathbf{x}$ can be partitioned as

$$\mathbf{x}' = (\mathbf{z}_1', y_2), \tag{4.1}$$

where y_2 is a single continuously distributed endogenous regressor; also, suppose the instrument vector $\mathbf{z}$ is also partitioned into subvectors corresponding to the "included" and "excluded" components of $\mathbf{x}$,

$$\mathbf{z}' = (\mathbf{z}_1', \mathbf{z}_2'). \tag{4.2}$$

Then, for a random sample $\{y_i, \mathbf{x}_i, \mathbf{z}_i\}_{i=1}^n$ of observations on $y \equiv y_1$, $\mathbf{x}$, and $\mathbf{z}$, we can express the model (3.1) as

$$y_{1i} = 1\{y_{1i}^* > 0\}, \tag{4.3}$$

for a latent dependent variable y_{1i}^* of the form

$$y_{1i}^* = \mathbf{z}_{1i}'\boldsymbol{\beta}_1 + y_{2i}\beta_2 + u_i. \tag{4.4}$$

If the simultaneity between y_{2i} and y_{1i} can be written in terms of a structural equation for y_{2i} in terms of the *latent* variable y_{1i}^*, that is,

$$y_{2i} = \mathbf{z}_{2i}'\gamma_1 + y_{1i}^*\gamma_2 + \varepsilon_i, \tag{4.5}$$

for some error term ε_i, then substitution of (4.4) in (4.5) delivers the first-stage regression model

$$y_{2i} = \mathbf{z}_i'\boldsymbol{\Pi} + \mathbf{v}_i \tag{4.6}$$

for some coefficient matrix $\boldsymbol{\Pi}$. This triangular structure has y_2 first being determined by $\mathbf{z}$ and the error terms $\mathbf{v}$, while y_1 is then determined by y_2, $\mathbf{z}$, and the structural error u.

In some economic applications, however, joint decision making may be in terms of the observed outcomes rather than latent outcomes implicit in (4.4) and (4.5). For example, consider the joint determination of savings (or consumption) and labor market participation. Let y_1 denote the discrete work decision and let y_2 denote other income including savings. Suppose that work involves a fixed cost α_1. In this case the structural relationship for other income (y_{2i}) will depend on the discrete employment decision (y_{1i}), *not* the latent variable (y_{1i}^*). It will not be possible, therefore, to solve explicitly the reduced form for y_2. Note that, for theoretical consistency, the fixed cost α_2 will also have to be subtracted from the income (or consumption) variable in the participation equation for those in work. As a result, for those who are employed, other income is defined net of fixed costs,

$$\widetilde{y}_{2i} \equiv y_{2i} - \alpha_2 y_{1i}. \tag{4.7}$$

We may therefore wish to replace (4.5) with a model incorporating feedback between the *observed* dependent variables y_1 and y_2,

$$y_{2i} = \mathbf{z}_{2i}'\gamma_1 + y_{1i}\alpha_2 + \varepsilon_i; \tag{4.8}$$

that is, the realization $y_1 = 1$ results in a discrete shift $y_{1i}\alpha_2$ in other income. Because of the nonlinearity in the binary response rule (4.3), there is no explicit reduced form for this system. Indeed, Heckman (1978), in his extensive analysis of simultaneous models with dummy endogenous variables, shows that (4.3), (4.4), and (4.8) are only a statistically "coherent" system, that is, one that processes a unique (if not explicit) reduced form, when $\gamma_2 = 0$.[8]

To provide a fully simultaneous system in terms of observed outcomes, and one that is also coherent, Heckman (1978) further shows that there must be a structural jump in the equation for y_{1i}^*,

$$y_{1i}^* = y_{1i}\alpha_1 + \mathbf{z}_{1i}'\boldsymbol{\beta}_1 + y_{2i}\beta_2 + u_i, \tag{4.9}$$

with the added restriction

$$\alpha_1 + \alpha_2\beta_2 = 0. \tag{4.10}$$

Heckman (1978) labels this the "principle assumption." To derive this condition, notice that from (4.3), (4.8), and (4.9), we can write

$$y_{1i}^* = 1\{y_{1i}^* > 0\}(\alpha_1 + \alpha_2\beta_2) + \mathbf{z}_{1i}'\boldsymbol{\beta}_1 + \mathbf{z}_{2i}'\gamma_1\beta_2 + u_i + \varepsilon_i\beta_2, \tag{4.11}$$

or

$$y_{1i}^* \lesseqgtr 0 \Leftrightarrow 1\{y_{1i}^* > 0\}(\alpha_1 + \alpha_2\beta_2) + \mathbf{z}_{1i}'\boldsymbol{\beta}_1 + \mathbf{z}_{2i}'\gamma_1\beta_2 + u_i + \varepsilon_i\beta_2 \lesseqgtr 0. \tag{4.12}$$

Thus, for a consistent probability model with general distributions for the unobservables and exogenous covariates, we require the coherency condition (4.10).

Substituting for α_1 from (4.10) into (4.9), we have

$$y_{1i}^* = (y_{2i} - y_{1i}\alpha_2)\beta_2 + \mathbf{z}_{1i}'\boldsymbol{\beta}_1 + u_i. \tag{4.13}$$

Note that this adjustment to y_{2i}, which guarantees statistical coherency, is *identical* to the condition for theoretical consistency in the fixed-cost model in which fixed cost α_2 is removed from other income for those who participate, as in (4.7).

Blundell and Smith (1994) derive a control function like the estimator for this setup under joint normality assumptions.[9] However, the semiparametric approach developed in the previous section naturally extends to this case. Noting that $\widetilde{y}_{2i} \equiv y_{2i} - \alpha_2 y_{1i}$, we find that the coherency condition implies that the model can be rewritten as

$$y_{1i} = 1\{\mathbf{z}_{1i}'\boldsymbol{\beta}_1 + \widetilde{y}_{2i}\beta_2 + u_i > 0\}, \tag{4.14}$$

[8] See Gourieroux, Laffont, and Monfort (1980) for further discussion of coherency conditions, and see Lewbel (1999) for a recent statement of this result.

[9] Blundell and Smith (1986) also develop an exogeneity test based on this estimator and consider results for the equivalent Tobit model. Rivers and Vuong (1988) provide a comprehensive treatment of limited information estimators for this class of parametric limited dependent variable models. They label the Blundell–Smith estimator the "two-stage conditional maximum likelihood" (2SCML) estimator and consider alternative limited information maximum likelihood (LIML) estimators. The efficiency and small sample properties of the 2SCML estimator are also considered. These are further refined in Blundell and Smith (1989). See also the important earlier related work of Amemiya (1978) and Lee (1981, 1993), which builds on the Heckman (1978) estimator.

and

$$\widetilde{y}_{2i} = \mathbf{z}_{2i}'\gamma_2 + \varepsilon_i. \tag{4.15}$$

This specification could easily be generalized to allow for a more complex relationship in more complicated models of nonseparable decision making.[10]

If α_2 were known, then Equations (4.15) and (4.14) are analogous to (4.3), (4.4), and (4.6). Consequently, a semiparametric estimator using the control function approach would simply apply the estimation approach described in this chapter to the conditional model. Following the previous discussion, assumption (2.42) would be replaced by the modified conditional independence restrictions

$$\begin{aligned} u|\mathbf{z}_1, y_2, \mathbf{z}_2 &\sim u|\mathbf{z}_1, \widetilde{y}_2, \varepsilon \\ &\sim u|\varepsilon. \end{aligned} \tag{4.16}$$

The conditional expectation of the binary variable y_1 given the regressors $\mathbf{z}_1$, $\widetilde{y}_2$ and errors ε would then take the form

$$\begin{aligned} E[y_1|\mathbf{z}_1, \widetilde{y}_{2i}, \varepsilon] &= \Pr[-u \le \mathbf{z}_{1i}'\boldsymbol{\beta}_1 + \widetilde{y}_{2i}\beta_2 | \mathbf{z}_1, \widetilde{y}_{2i}, \varepsilon] \\ &= F_{-u|\varepsilon}(\mathbf{z}_{1i}'\boldsymbol{\beta}_1 + \widetilde{y}_{2i}\beta_2|\varepsilon) \\ &\equiv \Gamma(\mathbf{z}_{1i}'\boldsymbol{\beta}_1 + \widetilde{y}_{2i}\beta_2, \varepsilon). \end{aligned} \tag{4.17}$$

Finally, note that although α_2 is unknown, given sufficient exclusion restrictions on $\mathbf{z}_{2i}$, a root-n consistent estimator for α_2 can be recovered from (linear) 2SLS estimation of (4.8). More generally, if the linear form $\mathbf{z}_{2i}'\gamma_1$ of the regression function for y_2 is replaced by a nonparametric form $\gamma(\mathbf{z}_{2i})$ for some unknown (smooth) function γ, then a $\sqrt{n}$-consistent estimator of α_2 in the resulting partially linear specification for y_{2i} could be based on the estimation approach proposed by Robinson (1988), using nonparametric estimators of instruments $(\mathbf{z}_{1i} - E[\mathbf{z}_{1i}|\mathbf{z}_{2i}])$ in an IV regression of y_{2i} on y_{1i}.

5. AN APPLICATION

The empirical application presented here is taken from the Blundell and Powell (1999) study. In that paper we considered the participation in work by men without college education in a sample of British families with children. Employment in this group in Britain is surprisingly low. More than 12 percent of these men do not work, and this number approaches 20 percent for those men with lower levels of education. Largely as a consequence of the low participation rate, this

[10] Note that to test this alternative specification against the triangular specification (4.3), (4.4), and (4.5), one may estimate

$$y_{2i} = \mathbf{z}_{2i}\gamma_1 + y_{1i}\alpha_2 + \widehat{y}_{2i}\delta_2 + w_i$$

by instrumental variables using $\mathbf{z}_i$ as instruments, and then test the null hypothesis $\delta_2 = 0$, where $\widehat{y}_{2i}$ is the prediction of y_{2i} under reduced-form specification (4.6).

group is subject to much policy discussion. We model the participation decision (y_1) in terms of a simple structural binary response framework that controls for market wage opportunities and the level of other income sources in the family. Educational level (z_1) is used as a proxy for market opportunities and is treated as exogenous for participation. However, other income (y_2), which includes the earned income of the spouse, is allowed to be endogenous for the participation decision.

As an instrument (z_{21}) for other family income, we use a welfare benefit entitlement variable. This instrument measures the transfer income the family would receive if neither spouse was working and is computed by using a benefit simulation routine designed for the evaluation of welfare benefits for households in the British data used here. The value of this variable depends on the local benefit rules, the demographic structure of the family, the geographic location, and housing costs. As there are no earnings-related benefits in operation in Britain over the period under study, we may be willing to assume it is exogenous for the participation decision. Moreover, although it will be a determinant of the reduced form for participation and other income, for the structural model herein, it should not enter the participation decision conditional on the inclusion of other income variables.

5.1. The Data

The sample consists of married couples drawn from the British Family Expenditure Survey (FES). The FES is a repeated continuous cross-sectional survey of households that provides consistently defined micro data on family incomes, employment status and education, consumption, and demographic structure. We consider the period 1985–1990. The sample is further selected according to the gender, educational attainment, and date of birth cohort of the head of household. We choose male heads of households, born between 1945 and 1954, who did not receive college education. We also choose a sample from the Northwest region of Britain. These selections are primarily to focus on the income and education variables.

For the purposes of modeling, the participating group consists of employees; the nonparticipating group includes individuals categorized as searching for work as well as the unoccupied. The measure of education used in our study is the age at which the individual left full-time education. Individuals in our sample are classified in two groups: those who left full-time education at age 16 or lower (the lower education base group), and those who left at age 17 or 18. Those who left at age 19 or older are excluded from this sample.

Our measure of exogenous benefit income is constructed for each family as follows: a tax and benefit simulation model[11] is used to construct a simulated

[11] The Institute for Fiscal Studies (IFS) tax and benefit simulation model is TAXBEN (www.ifs.org.uk), designed for the British FES data used in this paper. For an extensive discussion of the use of this data source in the study of male participation, see Blundell, Reed, and Stoker (1999).

Table 8.1. *Descriptive statistics*

Variable	Mean	Std. Dev.
Work (y_1)	0.871	0.387
Education > 16 (z_1)	0.196	0.396
ln (other inc.) (y_2)	5.016	0.434
ln (benefit inc.) (z_{21})	3.314	0.289
Education (sp.) (z_{22})	0.204	0.403
Age	39.191	10.256

Note: The number of observations is 1,606.

budget constraint for each individual family given information about age, location, benefit eligibility, and so on. The measure of out-of-work income is largely composed of income from state benefits; only small amounts of investment income are recorded. State benefits include eligible unemployment benefits,[12] housing benefits, child benefits, and certain other allowances. Because our measure of out-of-work income will serve to identify the structural participation equation, it is important that variation in the components of out-of-work income over the sample is exogenous for the decision to work. In the UK, the level of benefits that individuals receive out of work varies with age, time, and household size, and (in the case of housing benefit) by region. The housing benefit varies systematically with time, location, and cohort.

After making the sample selections described herein, our sample contains 1,606 observations. A brief summary of the data is provided in Table 8.1.[13] The 87.1 percent employment figure for men in this sample is reduced to less than 82 percent for the lower education group that makes up more than 75 percent of our sample. As mentioned earlier, this lower education group refers to those who left formal schooling at 16 years of age or younger and will be the group on which we focus in much of this empirical application. The kernel density estimate of log other income for the low education subsample is given in Figure 8.1.

5.2. A Model of Participation in Work and Other Family Income

To motivate the specification, suppose that observed participation is described by a simple threshold model of labor supply. In this model, the desired supply of hours of work for individual i can be written as

$$h_i^* = \delta_0 + \delta_1' \mathbf{z}_{1i} + \delta_2 \ln w_i + \delta_3 \ln \mu_i + \zeta_i, \tag{5.1}$$

where $\mathbf{z}_{1i}$ includes various observable social demographic variables, $\ln w_i$ is the log hourly wage, $\ln \mu_i$ is the log of "virtual" other income, and ζ_i is some

[12] The unemployment benefit included an earnings-related supplement in 1979, but this was abolished in 1980.

[13] See Blundell and Powell (1999) for further details.

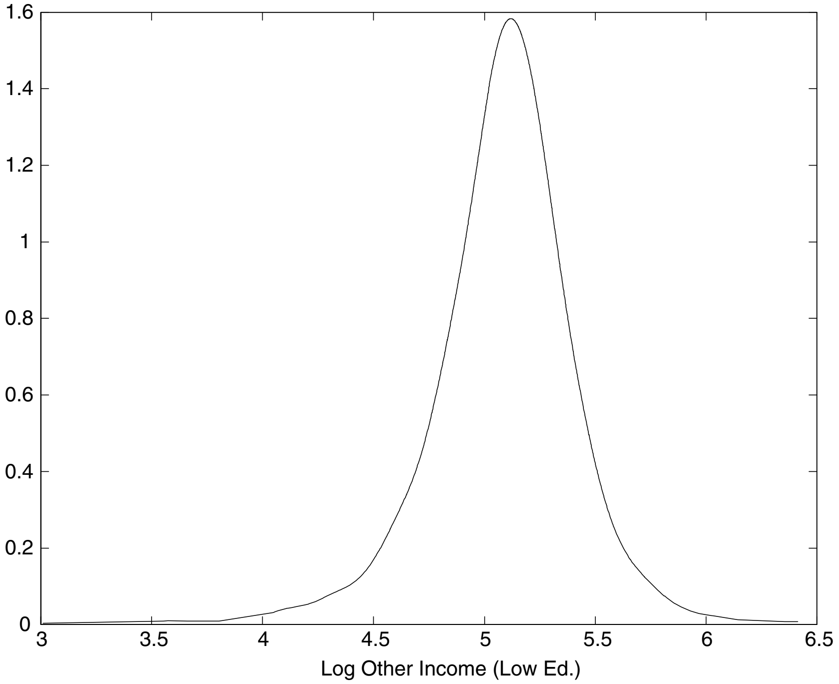


Figure 8.1. Density of Log Other Income for the Low Education Subsample.

unobservable heterogeneity. As $\ln w_i$ is unobserved for nonparticipants, we replace it in (5.1) by the wage equation

$$\ln w_i = \theta_0 + \theta_1' \mathbf{z}_{1i} + \omega_i, \tag{5.2}$$

where $\mathbf{z}_{1i}$ now contains the education level for individual i as well as other determinants of the market wage. Labor supply (5.1) becomes

$$h_i^* = \phi_0 + \phi_1' \mathbf{z}_{1i} + \phi_2 \ln \mu_i + \nu_i. \tag{5.3}$$

Participation in work occurs according to the binary indicator

$$y_{1i} = 1\{h_i^* > h_i^0\}, \tag{5.4}$$

where

$$h_i^0 = \gamma_0 + \gamma_1' \mathbf{z}_{1i} + \xi_i \tag{5.5}$$

is some measure of reservation hours.

Combining these equations, we find that participation is now described by

$$\begin{aligned} y_{1i} &= 1\{\phi_0 + \phi_1' \mathbf{z}_{1i} + \phi_2 \ln \mu_i + \nu_i > \gamma_0 + \gamma_1' \mathbf{z}_{1i} + \xi_i\} && (5.6) \\ &= 1\{\beta_0 + \boldsymbol{\beta}_1' \mathbf{z}_{1i} + \boldsymbol{\beta}_2' y_{2i} + u_i > 0\}, && (5.7) \end{aligned}$$

where y_{2i} is the log other income variable ($\ln \mu_i$). This other income variable is assumed to be determined by the reduced form

$$\begin{aligned} y_{2i} &= E[y_{2i}|\mathbf{z}_i] + v_i \\ &= \Pi(\mathbf{z}_i) + v_i, \end{aligned} \tag{5.8}$$

and $\mathbf{z}_i' = [\mathbf{z}_{1i}', \mathbf{z}_{2i}']$.

In the empirical application, we have already selected households by cohort, region, and demographic structure. Consequently, we are able to work with a fairly parsimonious specification in which $\mathbf{z}_{1i}$ simply contains the education level indicator. The excluded variables $\mathbf{z}_{2i}$ contain the log benefit income variable (denoted z_{21i}) and the education level of the spouse (z_{22i}).

5.3. Empirical Results

In Table 8.2 we present the empirical results for the joint normal–simultaneous Probit model. This consists of a linear reduced form for the log other income variable and a conditional Probit specification for the participation decision. Given the selection by region, cohort, demographic structure, and time period, the reduced form simply contains the education variables and the log exogenous benefit income variable. The results show a strong role for the benefit income variable in the determination of other income.

The first column of Probit results refers to the model without adjustment for the endogeneity of other income. These results show a positive and significant coefficient estimate for the education dummy variable and a small but significantly negative estimated coefficient on other income. The other income coefficient in Table 8.2 is the coefficient normalized by the education coefficient for comparability with the results from the semiparametric specification to be presented later. The impact of adjusting for endogeneity is quite dramatic. The income coefficient is now considerably larger and quite significant. The estimated education coefficient remains positive and significant.

Table 8.2. *Results for the simultaneous Probit specification*

	Reduced Form		Standard Probit		Simult. Probit	
Variable	y_2 Coeff.	Std. Error	Pr [Work] Coeff.	Std. Error	Pr [Work$\mid v$] Coeff.	Std. Error
Education (z_1)	0.0603	0.0224	1.007	0.1474	1.4166	0.1677
ln (other inc.) (y_2)	—	—	−0.3364	0.1293	−2.8376	0.5124
ln (benefit inc.) (z_{21})	0.0867	0.0093	—	—	—	—
Education (sp.) (z_{22})	0.0799	0.0219	—	—	—	—
Exog. test					5.896 (t)	
R^2	0.0708		0.0550		0.0885	
F	30.69(3)		67.84 ($\chi^2_{(2)}$)		109.29 ($\chi^2_{(3)}$)	

Table 8.3. *Semiparametric results, parametric results, and bootstrap distributions*

Specification	$\hat{\beta}_2$	$\sigma_{\hat{\beta}_2}$	10%	25%	50%	75%	90%
Semi-P (with $\widehat{v}$)	−2.2590	0.5621	−4.3299	−3.6879	−2.3275	−1.4643	−1.0101
Semi-P (without $\widehat{v}$)	−0.1871	0.0812	−0.2768	−0.2291	−0.1728	−0.1027	−0.0675
Probit (with $\widehat{v}$)	−2.8376	0.5124	−3.8124	−3.3304	−2.9167	−2.4451	−1.8487
Probit (without $\widehat{v}$)	−0.3364	0.1293	−0.4989	−0.4045	−0.3354	−0.2672	−0.1991
Lin. prob. (with $\widehat{v}$)	−3.1241	0.4679	−3.8451	−3.3811	−3.1422	−2.8998	−2.5425
Lin. prob. (without $\widehat{v}$)	−0.4199	0.1486	−0.6898	−0.5643	−0.4012	−0.3132	−0.2412

Table 8.3 presents the semiparametric estimation results for the linear index coefficients. Bandwidths were chosen according to the $1.06\sigma_z n^{-1/5}$ rule (see Silverman, 1986).[14] The education coefficient in the binary response specification is normalized to unity and so the β_1 estimates in Table 8.3 correspond to the ratio of the other income to the education coefficients. We present the standard Probit results in Table 8.2 for comparison (the mean of the bootstrap estimates for the education coefficient was 0.989, and for the income coefficient it was −0.327). The bootstrap figures relate to 500 bootstrap samples of size $n = 1{,}606$; the standard errors for the semiparametric methods are computed from a standardized interquartile range for the bootstrap distribution, and they are calculated using the usual asymptotic formulas for the Probit and linear probability estimators.

Figure 8.2 graphs the estimate of the ASF, $G(\mathbf{x}'\boldsymbol{\beta})$, derived from the semiparametric estimation with and without controls for the endogeneity of log other income. These plots cover the 5 percent to 95 percent range of the log other income distribution for the lower education group.

In Figure 8.3, we compare these semiparametric results with the results of estimating $G(\mathbf{x}'\boldsymbol{\beta})$ using the Probit and linear probability models. This data set is likely to be a particularly good source on which to carry out this evaluation. First, we know that the correction for endogeneity induces a large change in the β coefficients. Second, the proportion participating in the sample is around 85 percent, which suggests that the choice of probability model should matter as the tail probabilities in the Probit and linear probability models will behave quite differently. They show considerable sensitivity of the estimated $G(\mathbf{x}'\boldsymbol{\beta})$, after allowing for endogeneity, across these alternative parametric models. Both the linear probability model and the Probit model estimates result in estimated probabilities that are very different from those implied by the semiparametric approach. Figure 8.3 shows this most dramatically. For example, the linear probability model estimates a probability that is more than ten percentage points higher at the 20th percentile point of the log other income distribution.

[14] Blundell and Powell (1999) provide results for a similar model specification and also present sensitivity results for this bandwidth choice. In particular, sensitivity to the choice of a smaller bandwidth is investigated and found not to change the overall results.

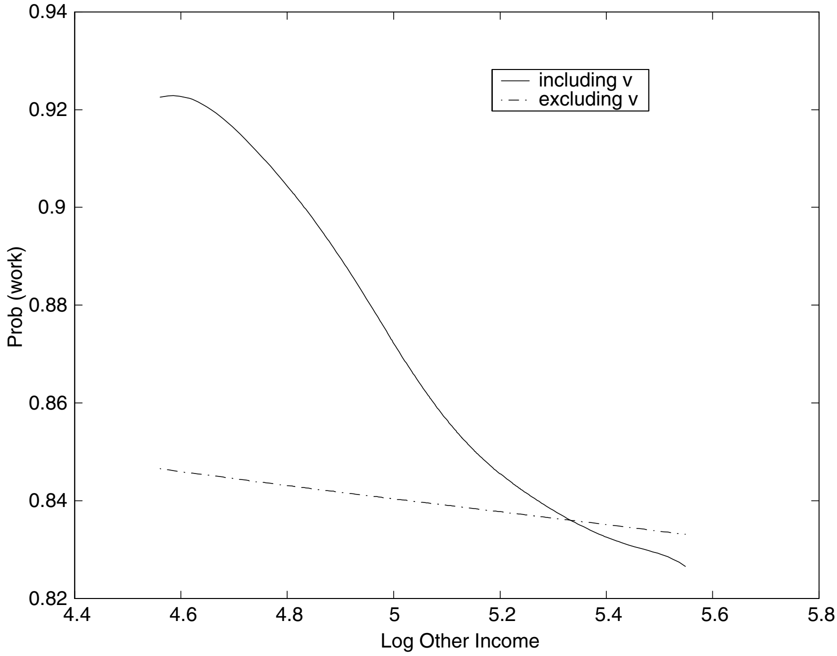


Figure 8.2. Semiparametric Model with and without Controls for Endogeneity.

This example points to the attractiveness of the semiparametric approach developed in this chapter. For this data set, we have found relatively small reductions in precision from adopting the semiparametric control function approach while finding quite different estimated responses in comparison with those from the parametric counterparts.[15] The results show a strong effect of correcting for endogeneity and indicate that adjusting for endogeneity using the standard parametric models, the Probit and linear probability models, can give a highly misleading picture of the impact on participation of an exogenous change in other income. This is highlighted in Figure 8.3, where it is shown that the bias correction for endogeneity in the linear probability model was sufficient to produce predicted probabilities larger than unity over a large range of the income distribution. The Probit model did not fare much better. The semiparametric approach showed a strong downward bias in the estimated income responses when endogeneity of other income was ignored. The corrected semiparametric estimates appear plausible, and, although there were no shape restrictions imposed, the estimated ASF was monotonically declining in other income over the large range of the income distribution.

[15] In Blundell and Powell (1999), we present the analogous analysis using the low-education subsample only. For this sample, the education dummy is equal to zero for all observations and is therefore excluded. Because x is now simply the log other income variable, this analysis is purely *nonparametric*. The results show a slightly shallower slope.

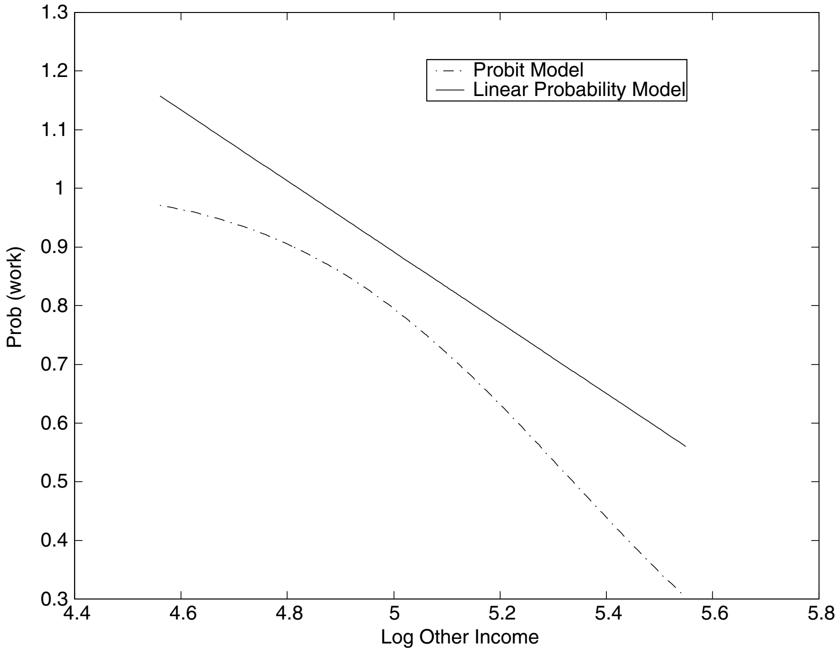


Figure 8.3. Linear Probability and Probit Models with Controls for Endogeneity.

5.4. The Coherency Model

In Blundell and Powell (1999), we also use this application to assess the alternative "coherency" model of Section 4, in which participation itself directly enters the equation determining other income. We interpret this as a fixed-cost model in which other income is dependent on whether or not fixed costs are paid, which in turn depends on participation. In this case, no explicit reduced form for y_2, other income, exists. The model for other income may be written as

$$y_{2i} = \gamma_0 + y_{1i}\alpha_2 + \mathbf{z}_{2i}'\gamma_1 + \varepsilon_i, \tag{5.9}$$

where we are assuming that y_2 relates to the *level* of other income. Participation is now described by

$$y_{1i} = 1\{\beta_0 + y_{1i}\alpha_1 + \mathbf{z}_{1i}'\boldsymbol{\beta}_1 + y_{2i}\beta_2 + u_i > 0\},$$

with the added coherency restriction

$$\alpha_1 + \beta_2\alpha_2 = 0. \tag{5.10}$$

Together these imply

$$y_{1i} = 1\{\beta_0 + \widetilde{y}_{2i}\beta_2 + \mathbf{z}_{1i}'\boldsymbol{\beta}_1 + u_i > 0\}, \tag{5.11}$$

Table 8.4. *Results for the coherency specification*

			Probit		Probit	
Variable	y_2 Coeff.	Std. Error	Pr[Work] Coeff.	Std. Error	Pr[Work$\mid\varepsilon$] Coeff.	Std. Error
Work (y_1)	58.034	8.732	—		—	
Education (z_1)	—		1.6357	0.2989	1.6553	0.3012
Adjusted income ($\widetilde{y}_2$)	—	—	−0.7371	0.0643	−0.5568	0.1433
Benefit inc. (z_{21})	0.4692	0.1453	—	—	—	—
Education (sp.) (z_{22})	0.1604	0.0421	—	—	—	—
$\sigma_{u\varepsilon} = 0$ (t test)	—		—		2.556	

with

$$\widetilde{y}_{2i} = (y_{2i} - y_{1i}\alpha_2),$$

where we note that this fixed-cost adjustment to other income y_{2i} guarantees statistical coherency.

From the discussion in Section 4, we note that the conditional expectation of the binary response variable y_{1i}, given the regressors $\mathbf{z}_1$, $\widetilde{y}_{2i}$ and errors ε, may be expressed as

$$\begin{aligned} E[y_1|\mathbf{z}_1, \widetilde{y}_{2i}, \varepsilon] &= \Pr[-u \leq \mathbf{z}'_{1i}\boldsymbol{\beta}_1 + \widetilde{y}_{2i}\beta_2 | \mathbf{z}_1, \widetilde{y}_{2i}, \varepsilon] \\ &= F(\mathbf{z}'_{1i}\boldsymbol{\beta}_1 + \widetilde{y}_{2i}\beta_2, \varepsilon). \qquad (5.12) \end{aligned}$$

Provided $\widetilde{y}_{2i}$ and ε_i can be measured, estimation follows the same procedure as in the triangular case.

The first column of Table 8.4 presents the estimates of the parameters of the structural equation for y_2 (5.9) in this coherency specification. These are recovered from IV estimation using the education of the husband as an excluded variable. The estimated "fixed cost of work" parameter seems reasonable; recall that the income variable has a mean of approximately £165 per week. The two sets of Probit results differ according to whether or not they control for ε. Notice that, having removed the direct simultaneity of y_1 on y_2 through the adjustment $\widetilde{y}_2$, we find much less evidence of endogeneity bias. Indeed, the coefficients on the adjusted other income variable in the two columns are quite similar (these are normalized relative to the education coefficient). If anything, after adjusting for fixed costs, we find that controlling for ε leads to a downward correction to the income coefficient.

The comparable results for the semiparametric specification are presented in Table 8.5. In these, the linear structural model estimates for the y_2 equation are used exactly as in Table 8.4. They show a very similar pattern with only a small difference in the other income coefficient between the specification that controls for ε and the one that does not. Again, the $\widetilde{y}_2$ adjustment

Table 8.5. *Semiparametric results for the coherency specification*

Variable	Semi-P Pr[Work] Coeff.	 Std. Error	Semi-P Pr[Work$\|\varepsilon$] Coeff.	 Std. Error
Adjusted income ($\widetilde{y}_2$)	-1.009	0.0689	-0.82256	0.2592

seems to capture much of the endogeneity between work and income in this coherency specification.

6. SUMMARY AND CONCLUSIONS

This chapter has considered nonparametric and semiparametric methods for estimating regression models of the form $y = H(\mathbf{x}, u)$, where the $\mathbf{x}$ contains continuous endogenous regressors and where u represents unobserved heterogeneity. It has been assumed that there exists a set of instrumental variables $\mathbf{z}$ with $\dim(\mathbf{z}) \geq \dim(\mathbf{x})$. This general specification was shown to cover a number of nonlinear models of interest in econometrics. The leading cases we considered were additive nonparametric specifications $y = g(\mathbf{x}) + u$, in which $g(\mathbf{x})$ is unknown, and nonadditive models $y = H(g(\mathbf{x}), u)$, in which $g(\mathbf{x})$ is unknown but H is a known function that is monotone but not invertible. An important example of the latter, and one that we used as an empirical illustration, is the binary response model with endogenous regressors. We have focused on identification and estimation in these leading nonparametric regression models and have defined the parameter of interest to be the ASF, $G(\mathbf{x}) \equiv \int H(\mathbf{x}, u) dF_u$, where the average is taken over the *marginal* distribution of the error terms u and where F_u denotes the marginal CDF of u.

In each of these leading cases, and their semiparametric variants, we have considered how three common estimation approaches for linear equations – the instrumental variables, fitted value, and control function approaches – may or may not be applicable. In the case where H and g are linear, *iid* distributed errors the covariance restriction $E(\mathbf{z}u) = 0$ and the rank condition are sufficient to guarantee identification and generate consistent and analytically identical estimators from each of these approaches. In the nonlinear models considered here, this is no longer the case.

In additive nonparametric specifications, we have considered restrictions on the model specification that are sufficient to identify $g(\mathbf{x})$, the ASF in this case. The relationship between the reduced form $E[y|\mathbf{z}]$ and the structural function g is given by $E[y|\mathbf{z}] = \int g(\mathbf{x}) dF_{\mathbf{x}|\mathbf{z}}$, where $F_{\mathbf{x}|\mathbf{z}}$ is the conditional CDF of $\mathbf{x}$ given $\mathbf{z}$. Unlike in typical nonparametric estimation problems, identification of g faces an ill-posed inverse problem and consistent estimators of the components

$E[y|\mathbf{z}]$ and $F_{\mathbf{x}|\mathbf{z}}$ are not, by themselves, sufficient for consistent IV estimation of g. We have reviewed and assessed a number of approaches to IV estimation that have been proposed in the literature to overcome this problem. For the nonadditive case, IV estimation faces more severe difficulties. Without some further specific structure on H, such as invertibility, estimation by IV does not look promising. For our leading case in this nonadditive setting, the binary response model, H is not invertible.

Apart from some very specific cases, we have argued that the fitted-value approach is not well suited to estimation of parameters of interest in these nonlinear models. However, the control function approach has been shown to provide an attractive solution. This approach treats endogeneity as an omitted variable problem, in which the inclusion of estimates of the first-stage errors $\mathbf{v}$ as a covariate corrects the inconsistency in $E(y|\mathbf{x})$. It has been shown to extend naturally under the conditional independence assumption that the distribution of u given $\mathbf{x}$ and $\mathbf{z}$ is the same as the conditional distribution of u given $\mathbf{v}$. This exclusion restriction permits replacement of the unidentified structural errors u with the identified control function $\mathbf{v}$ through iterated expectations, so that averaging the structural function H over the marginal distribution of the structural errors u is equivalent to averaging the (identified) intermediate regression function of y on $\mathbf{x}$ and $\mathbf{v}$ over the marginal distribution of $\mathbf{v}$. We have derived a general approach to identification and estimation of the ASF $G(\mathbf{x})$ by this control function approach and have highlighted the importance of support restrictions on the distribution of the endogenous components of $\mathbf{x}$ and $\mathbf{z}$.

We then considered the particular case of the linear index binary response model. In this semiparametric model, we have described in detail how estimation of the parameters of interest can be constructed using the control function approach. We considered a specific semiparametric matching estimator of the index coefficients that exploits both continuity and monotonicity implicit in the binary response model formulation. We have also shown how the partial-mean estimator from the nonparametric regression literature can be used to estimate the ASF directly. The control function estimator, for this semiparametric model, can easily be adapted to the case in which the model specification is not triangular and certain coherency conditions are required to be satisfied.

Finally, we have studied the response of labor force participation to nonlabor income, viewed as an endogenous regressor, using these techniques. The procedures we have developed appear to work well and suggest that the usual distributional assumptions underlying Probit and linear probability specifications could be highly misleading in binary response models with endogenous regressors. The application found relatively small reductions in precision from adopting the semiparametric approach. The semiparametric approach showed a strong downward bias in the estimated income responses when endogeneity of other income was ignored. The corrected semiparametric estimates appeared plausible, and, although there were no shape restrictions imposed, the estimated ASF was monotonically declining in other income over the large range of the income distribution.

ACKNOWLEDGMENTS

Prepared for the 2000 World Congress of the Econometric Society. We are grateful to Manuel Arellano, David Card, Kenneth Chay, Lars Hansen, James Heckman, Hidehiko Ichimura, Guido Imbens, Michael Jansson, Arthur Lewbel, Oliver Linton, Costas Meghir, Aviv Nevo, Whitney Newey, Thomas Rothenberg, Adrian Pagan, Joris Pinkse, Paul Ruud, James Stock, Thomas Stoker, Frank Vella, Edward Vytlacil, and seminar participants at Berkeley, Centre de Recherche en Economié et Statistique (CREST), MIT, Northwestern, Princeton, Rutgers, University of British Columbia (UBC), and University College London (UCL) for helpful comments on this and related work. We thank Chuan Goh for able research assistance. R. Blundell acknowledges financial support for this research from the Leverhulme Foundation and the Economic and Social Research Council (ESRC). The usual disclaimer applies.

References

Ahn, H. (1995), "Non-Parametric Two Stage Estimation of Conditional Choice Probabilities in a Binary Choice Model Under Uncertainty," *Journal of Econometrics*, 67, 337–378.

Ahn, H. and C. F. Manski (1993), "Distribution Theory for the Analysis of Binary Choice under Uncertainty with Nonparametric Estimation of Expectations," *Journal of Econometrics*, 56, 291–321.

Ahn, H. and J. L. Powell (1993), "Semiparametric Estimation of Censored Selection Models with a Nonparametric Selection Mechanism," *Journal of Econometrics*, 58, 3–29.

Ahn, H., H. Ichimura, and J. L. Powell (1996), "Simple Estimators for Monotone Index Models," manuscript, Department of Economics, U.C. Berkeley.

Ai, C. and X. Chen (2000), "Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions," Working Paper, LSE.

Altonji, J. and H. Ichimura (2000), "Estimating Derivatives in Nonseparable Models with Limited Dependent Variables," Working Paper, University College London.

Altonji, J. and R. L. Matzkin (1997), "Panel Data Estimators for Nonseparable Models with Endogenous Regressors," Working Paper, Northwestern University.

Amemiya, T. (1974), "The Nonlinear Two-Stage Least-Squares Estimator," *Journal of Econometrics*, 2, 105–110.

Amemiya, T. (1978), "The Estimation of a Simultaneous Equation Generalised Probit Model," *Econometrica*, 46, 1193–1205.

Andrews, D. (1994), "Asymptotics for Semiparametric Econometric Models via Stochastic Equicontinuity," *Econometrica*, 62, 43–72.

Angrist, J. (1999), "Estimation of Limited-Dependent Variable Models with Dummy Endogenous Regressors: Simple Strategies for Empirical Practice," mimeo, MIT.

Angrist, J., G. Imbens, and D. R. Rubin (1996), "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 91, 444–455.

Arellano, M. and B. Honoré (1999), "Panel Data Models: Some Recent Developments," in *Handbook of Econometrics*, Vol. 5, (ed. by J. Heckman and E. Leamer), Amsterdam: Elsevier–North-Holland.

Basmann, R. L. (1959), "A Generalised Classical Method of Linear Estimation of Coefficients in a Structural Equation," *Econometrica*, 25, 77–83.

Blundell, R., M. Browning, and I. Crawford (2000), "Nonparametric Engel Curves and Revealed Preference," Working Paper, UCL. Forthcoming *Econometrica*.

Blundell, R. W. and A. Duncan (1998), "Kernel Regression in Empirical Microeconomics," *Journal of Human Resources*, Special Issue on Empirical Microeconometrics.

Blundell, R. W. and J. L. Powell (1999), "Endogeneity in Semiparametric Binary Response Models," mimeo, U. C. Berkeley, reproduced as CeMMAP Working Paper CWP05/01, available at http://cemmap.ifs.org.uk/docs/cwp0501.pdf.

Blundell, R. W. and R. J. Smith (1986), "An Exogeneity Test for a Simultaneous Tobit Model," *Econometrica*, 54, 679–685.

Blundell, R. W. and R. J. Smith (1989), "Estimation in a Class of Simultaneous Equation Limited Dependent Variable Models," *Review of Economic Studies*, 56, 37–58.

Blundell, R. W. and R. J. Smith (1993), "Simultaneous Microeconometric Models with Censored or Qualitative Dependent Variables," in *Handbook of Statistics*, Vol. II, (ed. by G. S. Maddala, C. R. Rao, and H. D. Vinod), Amsterdam: North-Holland.

Blundell, R. W. and R. J. Smith (1994), "Coherency and Estimation in Simultaneous Models with Censored or Qualitative Dependent Variables," *Journal of Econometrics*, 64, 355–373.

Blundell, R. W., H. Reed, and T. Stoker (1999) "Interpreting Movements in Aggregate Wages: The Role of Participation," May, Working Paper 99/13, IFS.

Chamberlain, G. (1987), "Asymptotic Efficiency in Estimation with Conditional Moment Restrictions," *Journal of Econometrics*, 34, 305–334.

Chamberlain, G. (1992), "Efficiency Bounds for Semiparametric Regression," *Econometrica*, 60, 567–596.

Chen, X. and X. Shen (1998), "Sieve Extremum Estimates for Weakly Dependent Data," *Econometrica*, 66, 289–314.

Chen, X. and H. White (1998), "Improved Rates and Asymptotic Normality for Nonparametric Neural Network Estimators," *IEEE Transactions on Information Theory*, 45, 682–691.

Chen, X., L. P. Hansen, and J. Scheinkman (1997), "Shape-Preserving Estimation of Diffusions," Working Paper, University of Chicago, Department of Economics.

Dagenais, M. G. (1999), "Inconsistency of a Proposed Nonlinear Instrumental Variables Estimator for Probit and Logit Models with Endogenous Regressors," *Economics Letters*, 63, 19–21.

Das, M. (1999), "Instrumental Variables Estimation of Models with Discrete Endogenous Regressors," Working Paper, Columbia University, Department of Economics.

Das, M., W. K. Newey, and F. Vella (1998), "Nonparametric Estimation of the Sample Selection Model," Working Paper, MIT, Department of Economics.

Darolles, S., J.-P. Florens, and E. Renault (2000, April), "Nonparametric Instrumental Regression," mimeo, GREMAQ, University of Toulouse.

Dhrymes, P. J. (1970), *Econometrics: Statistical Foundations and Applications*. New York: Springer-Verlag.

Durbin, J. (1954), "Errors in Variables," *Review of the International Statistical Institute*, 22, 23–32.

Fenton, V. and A. R. Gallant (1996), "Convergence Rate of SNP Density Estimators," *Econometrica*, 64, 719–727.

Ferguson, T. S. (1967), *Mathematical Statistics: A Decision Theoretic Approach*. New York: Krieger.

Gourieroux, C., J.-J. Laffont, and A. Montfort (1980), "Coherency Conditions in Simultaneous Linear Equation Models with Endogenous Switching Regimes," *Econometrica*, 48, 675–695.

Hammersley, J. M. and D. C. Handscomb (1964), *Monte Carlo Methods*. London: Chapman & Hall.

Han, A. K.(1987), "Non-Parametric Analysis of a Generalized Regression Model: The Maximum Rank Correlation Estimator," *Journal of Econometrics*, 35, 303–316.

Hansen, L. P. (1982), "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50, 1029–1054.

Härdle, W. (1990), *Applied Nonparametric Regression*. Cambridge: Cambridge University Press.

Härdle W. and O. Linton (1995), "Nonparametric Regression Analysis," in *Handbook of Econometrics*, Vol. IV, (ed. by R. F. Engle and D. L. McFadden), Amsterdam: North-Holland.

Heckman, J. J. (1976), "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models," *Annals of Economic and Social Measurement*, 5, 475–492.

Heckman, J. J. (1978), "Dummy Endogenous Variable in a Simultaneous Equations System," *Econometrica*, 46, 931–959.

Heckman, J. J. (1979), "Sample Selection Bias as a Specification Error," *Econometrica*, 47, 153–162.

Heckman, J. J. (1997), "Instrumental Variables," *Journal of Human Resources*, 32, 441–462.

Heckman, J. J. and B. E. Honoré (1990), "The Empirical Content of the Roy Model," *Econometrica*, 58, 1121–1149.

Heckman, J. J., H. Ichimura, J. Smith, and P. Todd (1998), "Characterizing Selection Bias Using Experimental Data," *Econometrica*, 66, 1017–1018.

Heckman, J. J. and R. Robb (1985), "Alternative Methods for Evaluating the Impact of Interventions," in *Longitudinal Analysis of Labor Market Data*, (ed. by J. Heckman and B. Singer), Econometric Society Monograph 10, Cambridge: Cambridge University Press.

Heckman, J. J. and E. Vytlacil (1999), "Local Instrumental Variables," in *Nonlinear Statistical Inference: Essays in Honor of Takeshi Amemiya*, (ed. by C. Hsiao, K. Morimune, and J. Powell), Cambridge: Cambridge University Press.

Heckman, J. J. and E. Vytlacil (2000), "The Relationship between Treatment Parameters with a Latent Variable Framework," *Economics Letters*, 66, 33–39.

Honoré, B. E. and J. L. Powell (1994), "Pairwise Difference Estimators of Linear, Censored and Truncated Regression Models," *Journal of Econometrics*, 64, 241–278.

Honoré, B. E. and J. L. Powell (1997), "Pairwise Difference Estimators for Nonlinear Models," Working Paper, Princeton University.

Horowitz, J. (1993), "Semiparametric and Nonparametric Estimation of Quantal Response Models," in *Handbook of Statistics*, 11, (ed. by G. S. Maddala, C. R. Rao, and H. D. Vinod), Amsterdam: North-Holland.

Horowitz, J. (1996), "Semiparametric Estimation of a Regression Model with an Unknown Transformation of the Dependent Variable," *Econometrica*, 64(1), 103–137.

Horowitz, J. (1998), *Semiparametric Methods in Econometrics*. New York: Springer-Verlag.

Ichimura, H. (1993), "Local Quantile Regression Estimation of Binary Response Models with Conditional Heteroscedasticity," Working paper, University of Minnesota.

Ichimura, H. (1993), "Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single Index Models," *Journal of Econometrics*, 58, 71–120.

Ichimura, H. and L. F. Lee (1991), "Semiparametric Least Squares Estimation of Multiple Models: Single Equation Estimation," in *Nonparametric and Semiparametric Models in Econometrics and Statistics*, (ed. by W. Barnett, J. Powell, and G. Tauchen), Cambridge: Cambridge University Press.

Imbens, G. W. and J. Angrist (1994), "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467–476.

Imbens, G. W. and W. K. Newey (2000), "Nonparametric Identification of Triangular Simultaneous Equation Models without Additivity," Working Paper, University of California at Los Angeles.

Klein, R. W. and R. H. Spady (1993), "An Efficient Semiparametric Estimator for Discrete Choice Models," *Econometrica*, 61, 387–421.

Lee, L.-F. (1981), "Simultaneous Equation Models with Discrete and Censored Dependent Variables," in *Structural Analysis of Discrete Data With Economic Applications*, (ed. by C. Manski and D. McFadden), Cambridge, MA: MIT Press.

Lee, L.-F. (1993), "Simultaneous Equation Models with Discrete and Censored Dependent Variables," in *Handbook of Statistics*, Vol. 11, (ed. by G. S. Maddala, C. R. Rao, and H. D. Vinod), Amsterdam: North-Holland.

Lewbel, A. (1998), "Semiparametric Latent Variable Estimation with Endogenous or Mismeasured Regressors," *Econometrica*, 66, 105–121.

Lewbel, A. (1999),"Coherent Specification of Simultaneous Systems Containing a Binary Choice Equation," mimeo, Boston College.

Lewbel, A. (2000), "Semiparametric Qualitative Response Model Estimation with Unknown Heteroscedasticity or Instrumental Variables," *Journal of Econometrics*, 97, 145–177.

Linton, O. and J. P. Nielson (1995), "A Kernel Method of Estimating Nonparametric Structured Regression Based on a Marginal Distribution," *Biometrika*, 82, 93–100.

Maddala, G. S. (1983), *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge: Cambridge University Press.

Matzkin, R. L. (1991), "A Nonparametric Maximum Rank Correlation Estimator," in *Nonparametric and Semiparametric Methods in Economics and Statistics*, (ed. by W. A. Barnett, J. L. Powell, and G. E. Tauchen), Cambridge: Cambridge University Press.

Matzkin, R. L. (1992), "Nonparametric and Distribution-Free Estimation of the Binary Choice and Threshold-Crossing Models," *Econometrica*, 60, 239–270.

Matzkin, R. L. (1994), "Restrictions of Economic Theory in Nonparametric Methods," in *Handbook of Econometrics*, Vol. IV, (ed. by R. F. Engle and D. L. McFadden), Amsterdam: Elsevier–North-Holland.

Nelson, F. and L. Olsen (1978), "Specification and Estimation in a Simultaneous Equation Model with Limited Dependent Variables," *International Economic Review*, 19, 695–705.

Newey, W. K. (1985), "Semiparametric Estimation of Limited Dependent Variable Models with Endogenous Explanatory Variables," *Annals de l'INSEE*, 59/60, 219–237.

Newey, W. K. (1987), "Efficient Estimation of Limited Dependent Variable Models with Endogenous Explanatory Variables," *Journal of Econometrics*, 32, 127–237.

Newey, W. K. (1990), "Efficient Instrumental Variables Estimation of Nonlinear Models," *Econometrica*, 58, 809–837.

Newey, W. K. (1993), "Efficient Estimation of Models with Conditional Moment Restrictions, in *Handbook of Statistics*, Vol. 11, (ed. by G. S. Maddala, C. R. Rao, and H. D. Vinod), Amsterdam: North-Holland.

Newey, W. K. (1994a), "The Asymptotic Variance of Semiparametric Estimators," *Econometrica*, 62, 1349–1382.

Newey, W. K. (1994b), "Kernel Estimation of Partial Means and a General Variance Estimator," *Econometric Theory*, 10, 233–253.

Newey, W. K. (1997), "Convergence Rates and Asymptotic Normality for Series Estimators," *Journal of Econometrics*, 79, 147–168.

Newey, W. K. and J. L. Powell (1989), "Nonparametric Instrumental Variables Estimation," Working Paper, MIT.

Newey, W. K., J. L. Powell, and F. Vella (1999), "Nonparametric Estimation of Triangular Simultaneous Equations Models," *Econometrica*, 67, 565–603.

Ng, S. and J. Pinkse (1995), "Nonparametric Two Step Estimation of Unknown Regression Functions when the Regressors and the Regressor Error Are Not Independent," manuscript, University of Montreal.

O'Sullivan, F. (1986), "Ill-Posed Inverse Problems (with Discussion)," *Statistical Science*, 4, 503–527.

Pakes, A. and S. Olley (1995), "A Limit Theorem for a Smooth Class of Semiparametric Estimators," *Journal of Econometrics*, 65, 295–332.

Pagan, A. R. (1986), "Two Stage and Related Estimators and Their Applications," *Review of Economic Studies*, 53, 513–538.

Pinkse, J. (2000), "Nonparametric Two-Step Regression Estimation when Regressors and Error Are Dependent,"*Canadian Journal of Statistics*, 28(2), 289–300.

Powell, J. L. (1998), "Semiparametric Estimation of Censored Selection Models," in *Nonlinear Statistical Inference: Essays in Honor of Takeshi Amemiya*, (ed. by C. Hsiao, K. Morimune, and J. L. Powell), Cambridge: Cambridge University Press.

Powell, J. L. (1994), "Estimation of Semiparametric Models," in *Handbook of Econometrics*, Vol. IV, (ed. by R. F. Engle and D. L. McFadden), Amsterdam: Elsevier–North-Holland.

Powell, J., J. Stock, and T. Stoker (1989), "Semiparametric Estimation of Index Coefficients," *Econometrica*, 57, 1403–1430.

Rivers, D. and Q. H. Vuong, (1988), "Limited Information Estimators and Exogeneity Tests for Simultaneous Probit Models," *Journal of Econometrics*, 39, 347–366.

Robinson, P. (1987), "Asymptotically Efficient Estimation in the Presence of Heteroskedasticity of Unknown Form," *Econometrica*, 55, 875–892.

Robinson, P. (1988), "Root-*N*-Consistent Semiparametric Regression," *Econometrica* 56, 931–954.

Robinson, P. (1991), "Best Nonlinear Three-Stage Least Squares Estimation of Certain Econometric Models," *Econometrica* 59, 755–786.

Roehrig, C. S. (1988), "Conditions for Identification in Nonparametric and Parametric Models," *Econometrica*, 55, 875–891.

Sargan J. D. (1958), “The Estimation of Economic Relationships Using Instrumental Variables,” *Econometrica*, 26, 393–415.

Shen, X. (1997), “On Methods of Sieves and Penalization,” *The Annals of Statistics* 25(6), 2555–2591.

Silverman, B. (1986), *Density Estimation for Statistics and Data Analysis*. London: Chapman & Hall.

Stoker, T. M. (1986), “Consistent Estimation of Scaled Coefficients,” *Econometrica*, 54, 1461–1481.

Telser, L. G. (1964), “Iterative Estimation of a Set of Linear Regression Equations,” *Journal of the American Statistical Association*, 59, 845–862.

Theil, H. (1953), “Repeated Least Squares Applied to Complete Equation Systems,” The Hague: Central Planning Bureau.

Tjostheim, D. and R. H. Auestad (1996), “Nonparametric Identification of Nonlinear Time Series: Projections,” *Journal of the American Statistical Association*, 89, 1398–1409.

Vytlacil, E. (1999), “Semiparametric Identification and the Average Treatment Effect in Non-Separable Models,” mimeo, University of Chicago.

Endogeneity and Instruments in Nonparametric Models

A Discussion of the Papers by Jean-Pierre Florens and by Richard Blundell and James L. Powell

Manuel Arellano

The chapters by Jean-Pierre Florens and by Richard Blundell and James Powell consider the extension of nonlinear models with endogenous regressors to semiparametric and nonparameteric contexts. In my comments, I focus on two aspects. First, I refer to nonlinear implicit structural equations. Second, I provide some comments on the connections between the control variable and the instrumental variable approaches.

1. NONLINEAR IMPLICIT STRUCTURAL EQUATIONS

The literature on parametric nonlinear structural models, beginning with the work of Amemiya (1977), considered the implicit equation formulation

$$f(y, x) \equiv f(w) = u, \tag{1.1}$$

$$E(u|z) = 0, \tag{1.2}$$

where $w = (y, x')$ is a data point and u is an unobservable structural error that is mean independent of a vector z of instrumental variables.

In a semiparametric structural context, this formulation may also provide a useful setup. The model $y = g(x) + u$ in which $g(x)$ is treated as a nonparametric function is a special case of (1.1) with $f(y, x) = y - g(x)$. Such a model is reminiscent of nonparametric regression, but in structural applications other specializations of (1.1) and (1.2) may be relevant.

An example is a time-series consumption Euler equation of the form

$$U'(c_{t+1})r_{t+1} - U'(c_t) = u_{t+1}, \tag{1.3}$$

where $U'(\cdot)$ denotes the marginal utility of consumption and r_{t+1} is the stochastic return on a financial asset (Hansen and Singleton, 1982). Parameters of interest here could be the coefficients of relative risk aversion for different values of c estimated in a nonparametric way (semiparametric models of this type were considered by Gallant and Tauchen, 1989).

In this example there is no left-hand-side variable, but the model imposes a particular structure in the implicit function, namely additivity and monotonicity in $U'(\cdot)$. Structural models often impose not only instrumental variable (IV) conditions but also restrictions on the shape of functions, which may give rise to alternative economically based types of regularization for specific models.

Another situation of interest is when the starting point is an invertible response function,

$$y = H(x, u), \tag{1.4}$$

which can be represented as $f(y, x) = u$, together with the assumption

$$E[c(u)|z] = 0$$

for some function $c(\cdot)$. Identification of $f^*(y, x) = c\,[f(y, x)]$ up to scale affords calculation of the following derivative effects:

$$m\,(y, x) = -\left[\frac{\partial f(y, x)}{\partial y}\right]^{-1} \frac{\partial f(y, x)}{\partial x}. \tag{1.5}$$

The average structural function is given by

$$G\,(x) = E_u\{m[H(x, u), x]\}$$

and may or may not be identified, depending on the nature of the restrictions imposed on $f(y, x)$.

1.1. Discrete Case

If (w, z) is a discrete random vector with finite support, the analysis of the implicit equation IV model is straightforward. The model specifies

$$\sum_{j=1}^{J} f(\xi_j)\,\Pr(w = \xi_j|z = \zeta_\ell) = 0 \quad (\ell = 1, \ldots, L), \tag{1.6}$$

where the supports of w and z are $\{\xi_1, \ldots, \xi_J\}$ and $\{\zeta_1, \ldots, \zeta_L\}$, respectively.

In matrix form we have

$$P\theta = 0, \tag{1.7}$$

where P is an $L \times J$ matrix of conditional probabilities, and θ is the $J \times 1$ vector with elements $\theta_j = f(\xi_j)$. The order condition for identification of θ up to scale is $L \geq J - 1$, and the rank condition is $\mathrm{rank}(P) = J - 1$.

This is a standard generalized method of moments (GMM) problem: Let $r_j = \mathbf{1}(w = \xi_j)$ and $m_\ell = \mathbf{1}(z = \zeta_\ell)$. Then we can write

$$E[m_\ell\,(\theta_1 r_1 + \cdots + \theta_J r_J)] = 0 \quad (\ell = 1, \ldots, L), \tag{1.8}$$

which is in the form of a system of L simultaneous equations with instruments m_ℓ in equation ℓ.

The discreteness of endogenous and conditioning variables plays fundamentally different roles in this context. As another example, in a model that includes a subset of z in f so that $f(w, z_1) = u$, if the endogenous variables w are discrete with finite support but $z = (z_1, z_2)$ are continuous, this is equivalent to considering the following semiparametric conditional moment restriction:

$$E\left[\sum_{j=1}^{J} \theta_j(z_1) r_j | z\right] = 0, \tag{1.9}$$

where $w \in \{\xi_1, \ldots, \xi_J\}$, $\theta_j(z_1) = f(\xi_j, z_1)$, and $r_j = \mathbf{1}(w = \xi_j)$.

1.2. Testing for Overidentification and Underidentification

Models (1.1) and (1.2) can be regarded as a restriction on the distribution of w given z,

$$\int f(w) dF(w|z) = 0.$$

Sometimes the focus is in testing the restrictions on $F(w|z)$ rather than in the estimation of $f(w)$ or other average effects. From this point of view, $f(w)$ becomes a nuisance parameter function.

Clearly, in the discrete case, an invariant chi-square test statistic of the overidentifying restrictions (with $L - J + 1$ degrees of freedom) is readily available – but of no use in the continuous case. This is given by

$$\min_{\theta} n \hat{p}'(I \otimes \theta)[(I \otimes \theta')\hat{V}(I \otimes \theta)]^{-1}(I \otimes \theta')\hat{p}, \tag{1.10}$$

where $\hat{p} = \text{vec}(\hat{P})$ denotes a vector of sample frequencies, and $\hat{V}$ is the estimated sampling variance of $\hat{p}$.

Testing for underidentification in the discrete case is also straightforward. One would test the null hypothesis of underidentification, $\text{rank}(P) < J - 1$, against the alternative of identification, $\text{rank}(P) = J - 1$. A test statistic of this kind provides a natural diagnostic of the extent to which structural parameter estimates are well identified.

2. CONTROL FUNCTIONS AND INSTRUMENTAL VARIABLES

2.1. Additive Errors

Newey, Powell, and Vella (1999) considered a nonparametric structural equation together with an explicit reduced form for the endogenous explanatory variables:

$$y = g(x) + u, \tag{2.1}$$

$$x = \pi(z) + v, \tag{2.2}$$

and the assumptions

$$E(u|z, v) = E(u|v), \tag{2.3}$$

$$E(v|z) = 0. \tag{2.4}$$

These assumptions were chosen for convenience. In effect, they imply

$$\begin{aligned} E(y|x, v) &= g(x) + E(u|x, v) \\ &= g(x) + E(u|z, v) = g(x) + E(u|v) = g(x) + h(v). \end{aligned} \tag{2.5}$$

In this way, the problem of nonparametric estimation of $g(x)$ is assimilated to the problem of estimating the regression function $E(y|x, v)$ subject to an additive structure.

Assumptions (2.3) and (2.4) do not imply that $E(u|z) = 0$. The situation is that

$$E(u|z) = E[E(u|z, v)\,|z] = E[E(u|v)\,|z] = E[h(v)|z].$$

A sufficient condition for $E[h(v)|z] = 0$ is that v is independent of z. The mean independence condition does not guarantee that $E[h(v)|z] = 0$ unless $h(v)$ is linear in v.

Alternatively, suppose we begin with the assumptions $E(u|z) = 0$ and $E(v|z) = 0$. Then, in general, (2.3) or (2.5) is not satisfied.

Expression (2.5) makes it clear that the control function assumption can be very useful in applied work, but one should insist that the IV condition $E(u|z) = 0$ also holds. Having a structural equation in which the instruments are correlated with the errors because of a simplifying assumption will, in general, jeopardize the interpretability of the structural parameters.

From the point of view of econometric practice, it is better to regard the control function assumption as a specialization of the IV assumption than to pretend that one is no more or no less general than the other. I regard the control function approach as one in which estimation of the structural function $g(x)$ is *helped* by an explicit semiparametric modeling of the reduced form for x. In practice this will typically require aiming for a reduced form with errors that are statistically independent of instruments.

For example, suppose that for a scalar x, v is heteroskedastic (and hence not independent of z) with $\sigma^2(z) = E(v^2|z)$, but $v^\dagger = \sigma^{-1}(z)v$ is independent of z. In such case, the assumption

$$E(u|z, v^\dagger) = E(u|v^\dagger) \tag{2.6}$$

will be compatible with $E(u|z) = 0$, but (2.3) will imply, in general, correlation between u and z.

The control v can be generalized further, for example to include some kind of Box–Cox transformation of x. The general idea is that the approach works well when there is a reduced-form equation for x or some transformation of x with errors that are independent of z.

The IV assumption in the model with additive errors constrains only the marginal distributions of y and x given z, whereas the control variable assumption places restrictions on their joint distribution. Suppose we have two independent samples on (y, z) and (x, z), respectively, but the joint distribution of y and x is not observed (as in Angrist and Krueger, 1992, or Arellano and Meghir, 1992). It is interesting to compare the IV and control function assumptions in the two-sample estimation context to highlight the different data requirements in the two approaches. In the IV approach, only the marginal distributions of y and x given z are needed for identification, at least conceptually. Thus $y|z$ can be obtained from one sample and $x|z$ from the other. In the control function approach, however, (y, x, z) have to be observed in the same sample to ensure the identification of the nonparametric regression of y on x and v.

2.2. Discrete Choice

Blundell and Powell (1999) (BP) show how the control function approach can be particularly helpful in models with nonadditive errors. They consider a discrete choice model of the form

$$y = \mathbf{1}(x\beta + u > 0), \tag{2.7}$$

$$x = \pi(z) + v, \tag{2.8}$$

$$E(v|z) = 0, \tag{2.9}$$

together with the assumption

$$u|x, v \sim u|v. \tag{2.10}$$

In this way,

$$\Pr(y = 1|x, v) = \Pr(-u \leq x\beta|x, v) = \Pr(-u \leq x\beta|v).$$

Thus

$$E(y|x, v) = F\left(x\beta, v\right),$$

where $F(\cdot, v)$ is the conditional cumulative distribution function (CDF) of $-u$ given v.

As in the additive-error case of Newey, Powell, and Vella (NPV), the problem of estimating a structural equation is assimilated to the problem of estimating the regression function $E(y|x, v)$ subject to restrictions. In the case of NPV, it was sufficient to assume that u was mean independent of x given v, and $E(y|x, v)$ had an additive structure. In the discrete choice case, full independence of x given v is required, and $E(y|x, v)$ has a multiple index structure. The difference between the two models exists because (2.7) is not additive or invertible in u.

An interesting feature of the BP method is that the marginal CDF of u evaluated at $x\beta$ can be obtained by averaging $F(x\beta, v)$ over v whose CDF is identified:

$$\Pr(-u \leq x\beta) \equiv G(x\beta) = \int F(x\beta, v)\, dF_v. \tag{2.11}$$

This is useful because the function $G(x\beta)$ is arguably a parameter of interest for policy evaluation in this context.

Turning to a discussion of the assumptions, if (u, v) are independent of z, then

$$u|z, v \sim u|v. \tag{2.12}$$

Moreover, in view of (2.8), $u|x, v \sim u|z, v \sim u|v$. However, (2.9), (2.10), and (2.12) by themselves do not imply independence or even lack of correlation between u and the instruments z. Note that if the BP assumption (2.12) holds, in general u will not be independent of z unless v is independent of z:

$$\begin{aligned} F(u|z) &= \int F(u|z, v) dF_v(v|z) = \int F(u|v) dF_v(v|z) \\ &\neq \int F(u|v) dF_v(v) = F(u). \end{aligned}$$

So, the same remarks as for NPV apply in this context. If $x\beta + u$ represents a latent structural equation, one would expect to select instruments on a priori grounds that suggest some form of independence with the unobservable structural disturbance u.

In particular, in the BP empirical data, the conditional variance of the log other income variable may vary with the education of the spouse. If so, it would be nice to obtain a (possibly nonparametric) estimate of $\text{var}(y_2|z) \equiv \sigma^2(z)$. Then consider $v^{\dagger} = v/\sigma(z)$ as an alternative control function. In this way we could assess the impact of the correction for heteroskedasticity in the β coefficients and the estimated average structural function.

The conclusion is that the control function approach is best regarded not as a competing identification strategy to IV assumptions but as a complementary modeling strategy for the reduced form of the model. This strategy is specially useful in discrete choice and related models, for which IV assumptions by themselves do not appear to be sufficient to identify parameters of interest.

References

Amemiya, T. (1977), "The Maximum Likelihood and the Nonlinear Three-Stage Least Squares Estimator in the General Nonlinear Simultaneous Equation Model," *Econometrica*, 45, 955–968.

Angrist, J. and A. Krueger (1992), "The Effect of Age at School Entry on Educational Attainment: An Application of Instrumental Variables with Moments from Two Samples," *Journal of the American Statistical Association*, 87, 328–336.

Arellano, M. and C. Meghir (1992), "Female Labour Supply and On-the-Job Search: An Empirical Model Estimated Using Complementary Data Sets," *Review of Economic Studies*, 59, 537–559.

Blundell, R. W. and J. L. Powell (1999), "Endogeneity in Single Index Models," July, unpublished manuscript, Berkeley.

Gallant, A. R. and G. Tauchen (1989), "Seminonparametric Estimation of Conditionally Constrained Heterogeneous Processes: Asset Pricing Applications," *Econometrica*, 57, 1091–1120.

Hansen, L. P. and K. J. Singleton (1982), "Generalized Instrumental Variables Estimators of Nonlinear Rational Expectations Models," *Econometrica*, 50, 1269–1286.

Newey, W. K., J. L. Powell, and F. Vella (1999), "Nonparametric Estimation of Triangular Simultaneous Equations Models," *Econometrica*, 67, 565–603.

Name Index