Lecture Notes for Econometrics 2002 (first year PhD course in Stockholm)

Paul Söderlind[1]

June 2002 (some typos corrected and some material added later)

[1] University of St. Gallen. *Address:* s/bf-HSG, Rosenbergstrasse 52, CH-9000 St. Gallen, Switzerland. *E-mail:* Paul.Soderlind@unisg.ch. Document name: EcmAll.TeX.

Contents

1 Introduction

1.1 Means and Standard Deviation

The mean and variance of a series are estimated as

$$\bar{x} = \sum_{t=1}^{T} x_t/T \text{ and } \hat{\sigma}^2 = \sum_{t=1}^{T} (x_t - \bar{x})^2/T. \tag{1.1}$$

The standard deviation (here denoted $\text{Std}(x_t)$), the square root of the variance, is the most common measure of volatility.

The mean and standard deviation are often estimated on rolling data windows (for instance, a "Bollinger band" is ±2 standard deviations from a moving data window around a moving average—sometimes used in analysis of financial prices.)

If x_t is iid (independently and identically distributed), then it is straightforward to find the variance of the sample average. Then, note that

$$\begin{aligned} \text{Var}\left(\sum_{t=1}^{T} x_t/T\right) &= \sum_{t=1}^{T} \text{Var}\left(x_t/T\right) \\ &= T \, \text{Var}\left(x_t\right)/T^2 \\ &= \text{Var}\left(x_t\right)/T. \end{aligned} \tag{1.2}$$

The first equality follows from the assumption that x_t and x_s are independently distributed (so the covariance is zero). The second equality follows from the assumption that x_t and x_s are identically distributed (so their variances are the same). The third equality is a trivial simplification.

A sample average is (typically) unbiased, that is, the expected value of the sample average equals the population mean. To illustrate that, consider the expected value of the sample average of the iid x_t

$$\begin{aligned} \mathrm{E} \sum_{t=1}^{T} x_t/T &= \sum_{t=1}^{T} \mathrm{E}\, x_t/T \\ &= \mathrm{E}\, x_t. \end{aligned} \tag{1.3}$$

The first equality is always true (the expectation of a sum is the sum of expectations), and

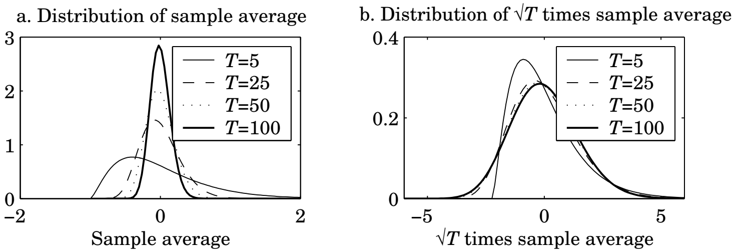


Figure 1.1: Sampling distributions. This figure shows the distribution of the sample mean and of $\sqrt{T}$ times the sample mean of the random variable $z_t - 1$ where $z_t \sim \chi^2(1)$.

the second equality follows from the assumption of identical distributions which implies identical expectations.

1.2 Testing Sample Means

The *law of large numbers* (LLN) says that the sample mean converges to the true population mean as the sample size goes to infinity. This holds for a very large class of random variables, but there are exceptions. A sufficient (but not necessary) condition for this convergence is that the sample average is unbiased (as in (1.3)) and that the variance goes to zero as the sample size goes to infinity (as in (1.2)). (This is also called convergence in mean square.) To see the LLN in action, see *Figure 1.1*.

The *central limit theorem* (CLT) says that $\sqrt{T}\bar{x}$ converges in distribution to a normal distribution as the sample size increases. See *Figure 1.1* for an illustration. This also holds for a large class of random variables—and it is a very useful result since it allows us to test hypothesis. Most estimators (including LS and other methods) are effectively some kind of sample average, so the CLT can be applied.

The basic approach in testing a hypothesis (the "null hypothesis"), is to compare the test statistics (the sample average, say) with how the distribution of that statistics (which is a random number since the sample is finite) would look like if the null hypothesis is true. For instance, suppose the null hypothesis is that the population mean is μ Suppose also that we know that distribution of the sample mean is normal with a known variance h^2 (which will typically be estimated and then treated as if it was known). Under the null hypothesis, the sample average should then be $N(\mu, h^2)$. We would then reject the null hypothesis if the sample average is far out in one the tails of the distribution. A traditional two-tailed test amounts to rejecting the null hypothesis at the 10% significance level if the test statistics is so far out that there is only 5% probability mass further out in that tail (and another 5% in the other tail). The interpretation is that if the null hypothesis is actually true, then there would only be a 10% chance of getting such an extreme (positive or negative) sample average—and these 10% are considered so low that we say that the null is probably wrong.

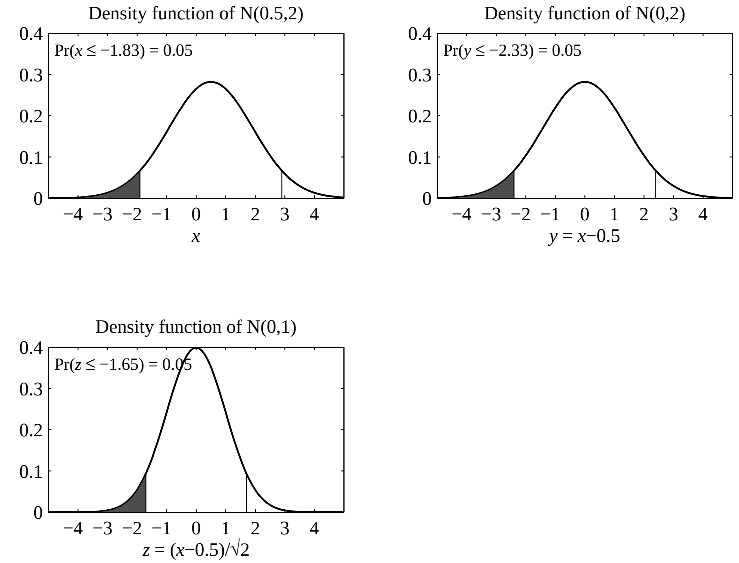


Figure 1.2: Density function of normal distribution with shaded 5% tails.

See *Figure 1.2* for some examples or normal distributions. recall that in a normal distribution, the interval ±1 standard deviation around the mean contains 68% of the probability mass; ±1.65 standard deviations contains 90%; and ±2 standard deviations contains 95%.

In practice, the test of a sample mean is done by "standardizing" the sampe mean so that it can be compared with a standard $N(0,1)$ distribution. The logic of this is as follows

$$\Pr(\bar{x} \geq 2.7) = \Pr(\bar{x} - \mu \geq 2.7 - \mu) \tag{1.4}$$

$$= \Pr\left(\frac{\bar{x} - \mu}{h} \geq \frac{2.7 - \mu}{h}\right). \tag{1.5}$$

If $\bar{x} \sim N(\mu, h^2)$, then $(\bar{x} - \mu)/h \sim N(0,1)$, so the probability of $\bar{x} \geq 2.7$ can be calculated by calculating how much probability mass of the standard normal density function there is above $(2.7 - \mu)/h$.

To construct a two-tailed test, we also need.the probability that $\bar{x}$ is above some number. This number is chosen to make the two-tailed tst symmetric, that is, so that there is as much probability mass below lower number (lower tail) as above the upper number (upper tail). With a normal distribution (or, for that matter, any symmetric distribution) this is done as follows. Note that $(\bar{x}-\mu)/h \sim N(0,1)$ is symmetric around 0. This means that the probability of being above some number, $(C - \mu)/h$, must equal the probability of being below -1 times the same number, or

$$\Pr\left(\frac{\bar{x} - \mu}{h} \geq \frac{C - \mu}{h}\right) = \Pr\left(-\frac{C - \mu}{h} \leq \frac{\bar{x} - \mu}{h}\right). \tag{1.6}$$

A *10% critical value* is the value of $(C - \mu)/h$ that makes both these probabilities equal to 5%—which happens to be 1.645. The easiest way to look up such critical values is by looking at the normal cumulative distribution function—see *Figure 1.2*.

1.3 Covariance and Correlation

The covariance of two variables (here x and y) is typically estimated as

$$\widehat{\text{Cov}}\left(x_t, z_t\right) = \textstyle\sum_{t=1}^{T} \left(x_t - \bar{x}\right)\left(z_t - \bar{z}\right)/T. \tag{1.7}$$

Note that this is a kind of sample average, so a CLT can be used.

The correlation of two variables is then estimated as

$$\widehat{\text{Corr}}\left(x_t, z_t\right) = \frac{\widehat{\text{Cov}}\left(x_t, z_t\right)}{\widehat{\text{Std}}\left(x_t\right)\widehat{\text{Std}}\left(z_t\right)}, \tag{1.8}$$

where $\widehat{\text{Std}}(x_t)$ is an estimated standard deviation. A correlation must be between -1 and 1 (try to show it). Note that covariance and correlation measure the degree of *linear* relation only. This is illustrated in Figure 1.4.

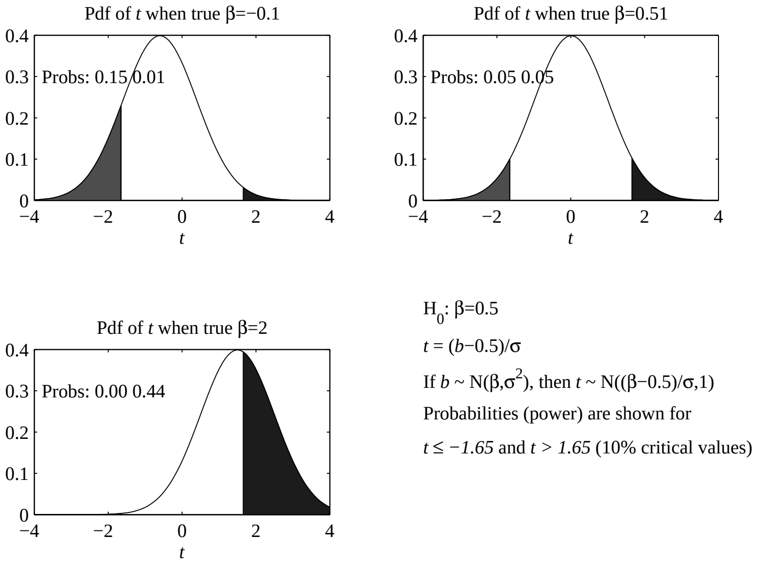


Figure 1.3: Power of two-sided test

The *pth autocovariance* of x is estimated by

$$\widehat{\text{Cov}}\left(x_t, x_{t-p}\right) = \textstyle\sum_{t=1}^{T} \left(x_t - \bar{x}\right)\left(x_{t-p} - \bar{x}\right)/T, \tag{1.9}$$

where we use the same estimated (using all data) mean in both places. Similarly, the *pth autocorrelation* is estimated as

$$\widehat{\text{Corr}}\left(x_t, x_{t-p}\right) = \frac{\widehat{\text{Cov}}\left(x_t, x_{t-p}\right)}{\widehat{\text{Std}}\left(x_t\right)^2}. \tag{1.10}$$

Compared with a traditional estimate of a correlation (1.8) we here impose that the standard deviation of x_t and x_{t-p} are the same (which typically does not make much of a difference).

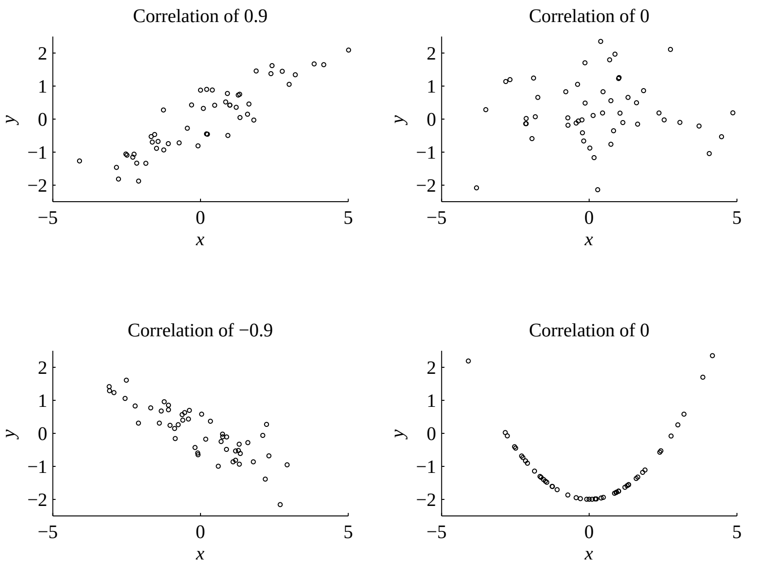


Figure 1.4: Example of correlations on an artificial sample. Both subfigures use the same sample of y.

1.4 Least Squares

Consider the simplest linear model

$$y_t = x_t\beta_0 + u_t, \tag{1.11}$$

where all variables are zero mean scalars and where β_0 is the true value of the parameter we want to estimate. The task is to use a sample $\{y_t, x_t\}_{t=1}^{T}$ to estimate β and to test hypotheses about its value, for instance that $\beta = 0$.

If there were no movements in the unobserved errors, u_t, in (1.11), then any sample would provide us with a perfect estimate of β. With errors, any estimate of β will still leave us with some uncertainty about what the true value is. The two perhaps most important issues in econometrics are how to construct a good estimator of β and how to assess the uncertainty about the true value.

For any possible estimate, $\hat{\beta}$, we get a fitted residual

$$\hat{u}_t = y_t - x_t\hat{\beta}. \tag{1.12}$$

One appealing method of choosing $\hat{\beta}$ is to minimize the part of the movements in y_t that we cannot explain by $x_t\hat{\beta}$, that is, to minimize the movements in $\hat{u}_t$. There are several candidates for how to measure the "movements," but the most common is by the mean of squared errors, that is, $\Sigma_{t=1}^{T}\hat{u}_t^2/T$. We will later look at estimators where we instead use $\Sigma_{t=1}^{T}|\hat{u}_t|/T$.

With the sum or mean of squared errors as the loss function, the optimization problem

$$\min_{\beta} \frac{1}{T}\sum_{t=1}^{T}(y_t - x_t\beta)^2 \tag{1.13}$$

has the first order condition that the derivative should be zero as the optimal estimate $\hat{\beta}$

$$\frac{1}{T}\sum_{t=1}^{T} x_t\left(y_t - x_t\hat{\beta}\right) = 0, \tag{1.14}$$

which we can solve for $\hat{\beta}$ as

$$\hat{\beta} = \left(\frac{1}{T}\sum_{t=1}^{T}x_t^2\right)^{-1}\frac{1}{T}\sum_{t=1}^{T}x_t y_t, \text{ or} \tag{1.15}$$

$$= \widehat{\text{Var}}(x_t)^{-1}\,\widehat{\text{Cov}}(x_t, y_t), \tag{1.16}$$

where a hat indicates a sample estimate. This is the Least Squares (LS) estimator.

1.5 Maximum Likelihood

A different route to arrive at an estimator is to maximize the likelihood function. If u_t in (1.11) is iid $N\left(0, \sigma^2\right)$, then the probability density function of u_t is

$$\text{pdf}(u_t) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-u_t^2/\left(2\sigma^2\right)\right]. \tag{1.17}$$

Since the errors are independent, we get the joint pdf of the $u_1, u_2, \ldots, u_T$ by multiplying the marginal pdfs of each of the errors. Then substitute $y_t - x_t\beta$ for u_t (the derivative of the transformation is unity) and take logs to get the log likelihood function of the sample

$$\ln L = -\frac{T}{2}\ln(2\pi) - \frac{T}{2}\ln\left(\sigma^2\right) - \frac{1}{2}\sum_{t=1}^{T}\left(y_t - x_t\beta\right)^2/\sigma^2. \tag{1.18}$$

This likelihood function is maximized by minimizing the last term, which is proportional to the sum of squared errors - just like in (1.13): LS is ML when the errors are iid normally distributed.

Maximum likelihood estimators have very nice properties, provided the basic distributional assumptions are correct. If they are, then MLE are typically the most efficient/precise estimators, at least asymptotically. ML also provides a coherent framework for testing hypotheses (including the Wald, LM, and LR tests).

1.6 The Distribution of $\hat{\beta}$

Equation (1.15) will give different values of $\hat{\beta}$ when we use different samples, that is different draws of the random variables u_t, x_t, and y_t. Since the true value, β_0, is a fixed constant, this distribution describes the uncertainty we should have about the true value after having obtained a specific estimated value.

To understand the distribution of $\hat{\beta}$, use (1.11) in (1.15) to substitute for y_t

$$\begin{aligned}\hat{\beta} &= \left(\frac{1}{T}\sum_{t=1}^{T}x_t^2\right)^{-1}\frac{1}{T}\sum_{t=1}^{T}x_t\left(x_t\beta_0 + u_t\right)\\ &= \beta_0 + \left(\frac{1}{T}\sum_{t=1}^{T}x_t^2\right)^{-1}\frac{1}{T}\sum_{t=1}^{T}x_t u_t,\end{aligned} \tag{1.19}$$

where β_0 is the true value.

The first conclusion from (1.19) is that, with $u_t = 0$ the estimate would always be perfect — and with large movements in u_t we will see large movements in $\hat{\beta}$. The second conclusion is that not even a strong opinion about the distribution of u_t, for instance that u_t is iid $N\left(0, \sigma^2\right)$, is enough to tell us the whole story about the distribution of $\hat{\beta}$. The reason is that deviations of $\hat{\beta}$ from β_0 are a function of $x_t u_t$, not just of u_t. Of course, when x_t are a set of deterministic variables which will always be the same irrespective of which sample we use, then $\hat{\beta} - \beta_0$ is a time invariant linear function of u_t, so the distribution of u_t carries over to the distribution of $\hat{\beta}$. This is probably an unrealistic case, which forces us to look elsewhere to understand the properties of $\hat{\beta}$.

There are two main routes to learn more about the distribution of $\hat{\beta}$: *(i)* set up a small "experiment" in the computer and simulate the distribution or *(ii)* use the asymptotic distribution as an approximation. The asymptotic distribution can often be derived, in contrast to the exact distribution in a sample of a given size. If the actual sample is large, then the asymptotic distribution may be a good approximation.

A law of large numbers would (in most cases) say that both $\sum_{t=1}^{T}x_t^2/T$ and $\sum_{t=1}^{T}x_t u_t/T$ in (1.19) converge to their expected values as $T \to \infty$. The reason is that both are sample averages of random variables (clearly, both x_t^2 and $x_t u_t$ are random variables). These expected values are $\mathrm{Var}(x_t)$ and $\mathrm{Cov}(x_t, u_t)$, respectively (recall both x_t and u_t have zero means). The key to show that $\hat{\beta}$ is *consistent*, that is, has a probability limit equal to β_0, is that $\mathrm{Cov}(x_t, u_t) = 0$. This highlights the importance of using good theory to derive not only the systematic part of (1.11), but also in understanding the properties of the errors. For instance, when theory tells us that y_t and x_t affect each other (as prices and quantities typically do), then the errors are likely to be correlated with the regressors - and LS is inconsistent. One common way to get around that is to use an instrumental variables technique. More about that later. Consistency is a feature we want from most estimators, since it says that we would at least get it right if we had enough data.

Suppose that $\hat{\beta}$ is consistent. Can we say anything more about the asymptotic distribution? Well, the distribution of $\hat{\beta}$ converges to a spike with all the mass at β_0, but the distribution of $\sqrt{T}\hat{\beta}$, or $\sqrt{T}\left(\hat{\beta} - \beta_0\right)$, will typically converge to a non-trivial normal distribution. To see why, note from (1.19) that we can write

$$\sqrt{T}\left(\hat{\beta} - \beta_0\right) = \left(\frac{1}{T}\sum_{t=1}^{T}x_t^2\right)^{-1}\frac{\sqrt{T}}{T}\sum_{t=1}^{T}x_t u_t. \tag{1.20}$$

The first term on the right hand side will typically converge to the inverse of $\mathrm{Var}(x_t)$, as discussed earlier. The second term is $\sqrt{T}$ times a sample average (of the random variable $x_t u_t$) with a zero expected value, since we assumed that $\hat{\beta}$ is consistent. Under weak conditions, a central limit theorem applies so $\sqrt{T}$ times a sample average converges to a normal distribution. This shows that $\sqrt{T}\hat{\beta}$ has an *asymptotic normal distribution*. It turns out that this is a property of many estimators, basically because most estimators are some kind of sample average. For an example of a central limit theorem in action, see Appendix B

1.7 Diagnostic Tests

Exactly what the variance of $\sqrt{T}(\hat{\beta} - \beta_0)$ is, and how it should be estimated, depends mostly on the properties of the errors. This is one of the main reasons for diagnostic tests. The most common tests are for homoskedastic errors (equal variances of u_t and u_{t-s}) and no autocorrelation (no correlation of u_t and u_{t-s}).

When ML is used, it is common to investigate if the fitted errors satisfy the basic assumptions, for instance, of normality.

1.8 Testing Hypotheses about $\hat{\beta}$

Suppose we now assume that the asymptotic distribution of $\hat{\beta}$ is such that

$$\sqrt{T}\left(\hat{\beta} - \beta_0\right) \overset{d}{\to} N\left(0, v^2\right) \text{ or} \tag{1.21}$$

We could then test hypotheses about $\hat{\beta}$ as for any other random variable. For instance, consider the hypothesis that $\beta_0 = 0$. If this is true, then

$$\Pr\left(\sqrt{T}\hat{\beta}/v < -2\right) = \Pr\left(\sqrt{T}\hat{\beta}/v > 2\right) \approx 0.025, \tag{1.22}$$

which says that there is only a 2.5% chance that a random sample will deliver a value of $\sqrt{T}\hat{\beta}/v$ less than -2 and also a 2.5% chance that a sample delivers a value larger than 2, assuming the true value is zero.

We then say that we reject the hypothesis that $\beta_0 = 0$ at the 5% significance level (95% confidence level) if the test statistics $|\sqrt{T}\hat{\beta}/v|$ is larger than 2. The idea is that, if the hypothesis is true ($\beta_0 = 0$), then this decision rule gives the wrong decision in 5% of the cases. That is, 5% of all possible random samples will make us reject a true hypothesis. Note, however, that this test can only be taken to be an approximation since it relies on the asymptotic distribution, which is an approximation of the true (and typically unknown) distribution.

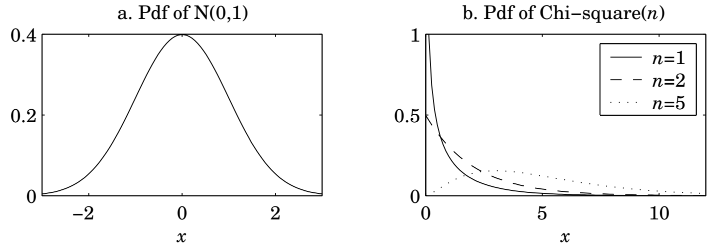


Figure 1.5: Probability density functions

The natural interpretation of a really large test statistics, $|\sqrt{T}\hat{\beta}/v| = 3$ say, is that it is very unlikely that this sample could have been drawn from a distribution where the hypothesis $\beta_0 = 0$ is true. We therefore choose to reject the hypothesis. We also hope that the decision rule we use will indeed make us reject false hypothesis more often than we reject true hypothesis. For instance, we want the decision rule discussed above to reject $\beta_0 = 0$ more often when $\beta_0 = 1$ than when $\beta_0 = 0$.

There is clearly nothing sacred about the 5% significance level. It is just a matter of convention that the 5% and 10% are the most widely used. However, it is not uncommon to use the 1% or the 20%. Clearly, the lower the significance level, the harder it is to reject a null hypothesis. At the 1% level it often turns out that almost no reasonable hypothesis can be rejected.

The t-test described above works only if the null hypothesis contains a single restriction. We have to use another approach whenever we want to test several restrictions jointly. The perhaps most common approach is a *Wald test*. To illustrate the idea, suppose β is an $m \times 1$ vector and that $\sqrt{T}\hat{\beta} \overset{d}{\to} N(\mathbf{0}, V)$ under the null hypothesis , where V is a covariance matrix. We then know that

$$\sqrt{T}\hat{\beta}'V^{-1}\sqrt{T}\hat{\beta} \overset{d}{\to} \chi^2(m). \tag{1.23}$$

The decision rule is then that if the left hand side of (1.23) is larger that the 5%, say, critical value of the $\chi^2(m)$ distribution, then we reject the hypothesis that all elements in β are zero.

A Practical Matters

A.0.1 Software

- Gauss, MatLab, RATS, Eviews, Stata, PC-Give, Micro-Fit, TSP, SAS
- Software reviews in *The Economic Journal* and *Journal of Applied Econometrics*

A.0.2 Useful Econometrics Literature

1. Greene (2000), *Econometric Analysis* (general)
2. Hayashi (2000), *Econometrics* (general)
3. Johnston and DiNardo (1997), *Econometric Methods* (general, fairly easy)
4. Pindyck and Rubinfeld (1998), *Econometric Models and Economic Forecasts* (general, easy)
5. Verbeek (2004), *A Guide to Modern Econometrics* (general, easy, good applications)
6. Davidson and MacKinnon (1993), *Estimation and Inference in Econometrics* (general, a bit advanced)
7. Ruud (2000), *Introduction to Classical Econometric Theory* (general, consistent projection approach, careful)
8. Davidson (2000), *Econometric Theory* (econometrics/time series, LSE approach)
9. Mittelhammer, Judge, and Miller (2000), *Econometric Foundations* (general, advanced)
10. Patterson (2000), *An Introduction to Applied Econometrics* (econometrics/time series, LSE approach with applications)
11. Judge et al (1985), *Theory and Practice of Econometrics* (general, a bit old)
12. Hamilton (1994), *Time Series Analysis*

13. Spanos (1986), *Statistical Foundations of Econometric Modelling,* Cambridge University Press (general econometrics, LSE approach)
14. Harvey (1981), *Time Series Models,* Philip Allan
15. Harvey (1989), *Forecasting, Structural Time Series...* (structural time series, Kalman filter).
16. Lütkepohl (1993), *Introduction to Multiple Time Series Analysis* (time series, VAR models)
17. Priestley (1981), *Spectral Analysis and Time Series* (advanced time series)
18. Amemiya (1985), *Advanced Econometrics,* (asymptotic theory, non-linear econometrics)
19. Silverman (1986), *Density Estimation for Statistics and Data Analysis* (density estimation).
20. Härdle (1990), *Applied Nonparametric Regression*

B A CLT in Action

This is an example of how we can calculate the limiting distribution of a sample average.

Remark B.1 *If* $\sqrt{T}(\bar{x}-\mu)/\sigma \sim N(0,1)$ *then* $\bar{x} \sim N(\mu, \sigma^2/T)$.

Example B.2 *(Distribution of* $\Sigma_{t=1}^{T}(z_t-1)/T$ *and* $\sqrt{T}\Sigma_{t=1}^{T}(z_t-1)/T$ *when* $z_t \sim \chi^2(1)$.*) When* z_t *is iid* $\chi^2(1)$*, then* $\Sigma_{t=1}^{T} z_t$ *is distributed as a* $\chi^2(T)$ *variable with pdf* $f_T()$*. We now construct a new variable by transforming* $\Sigma_{t=1}^{T} z_t$ *as to a sample mean around one (the mean of* z_t*)*

$$\bar{z}_1 = \Sigma_{t=1}^{T} z_t / T - 1 = \Sigma_{t=1}^{T}(z_t-1)/T.$$

Clearly, the inverse function is $\Sigma_{t=1}^{T} z_t = T\bar{z}_1 + T$*, so by the "change of variable" rule we get the pdf of* $\bar{z}_1$ *as*

$$g(\bar{z}_1) = f_T(T\bar{z}_1 + T)\,T.$$

Example B.3 *Continuing the previous example, we now consider the random variable*

$$\bar{z}_2 = \sqrt{T}\bar{z}_1,$$

with inverse function $\bar{z}_1 = \bar{z}_2/\sqrt{T}$. By applying the "change of variable" rule again, we get the pdf of $\bar{z}_2$ as

$$h\left(\bar{z}_2\right) = g(\bar{z}_2/\sqrt{T})/\sqrt{T} = f_T\left(\sqrt{T}\bar{z}_2 + T\right)\sqrt{T}.$$

Example B.4 *When z_t is iid $\chi^2(1)$, then $\Sigma_{t=1}^T z_t$ is $\chi^2(T)$, which we denote $f(\Sigma_{t=1}^T z_t)$. We now construct two new variables by transforming $\Sigma_{t=1}^T z_t$*

$$\begin{aligned}&\bar{z}_1 = \Sigma_{t=1}^T z_t/T - 1 = \Sigma_{t=1}^T\left(z_t - 1\right)/T, \text{ and}\\ &\bar{z}_2 = \sqrt{T}\bar{z}_1.\end{aligned}$$

Example B.5 *We transform this distribution by first subtracting one from z_t (to remove the mean) and then by dividing by T or $\sqrt{T}$. This gives the distributions of the sample mean and scaled sample mean, $\bar{z}_2 = \sqrt{T}\bar{z}_1$ as*

$$\begin{aligned}f\left(\bar{z}_1\right) &= \frac{1}{2^{T/2}\Gamma\left(T/2\right)} y^{T/2-1}\exp\left(-y/2\right) \text{ with } y = T\bar{z}_1 + T, \text{ and}\\ f\left(\bar{z}_2\right) &= \frac{1}{2^{T/2}\Gamma\left(T/2\right)} y^{T/2-1}\exp\left(-y/2\right) \text{ with } y = \sqrt{T}\bar{z}_1 + T.\end{aligned}$$

These distributions are shown in Figure 1.1. It is clear that $f\left(\bar{z}_1\right)$ converges to a spike at zero as the sample size increases, while $f\left(\bar{z}_2\right)$ converges to a (non-trivial) normal distribution.

Example B.6 *(Distribution of $\Sigma_{t=1}^T\left(z_t - 1\right)/T$ and $\sqrt{T}\Sigma_{t=1}^T\left(z_t - 1\right)/T$ when $z_t \sim \chi^2(1)$.) When z_t is iid $\chi^2(1)$, then $\Sigma_{t=1}^T z_t$ is $\chi^2(T)$, that is, has the probability density function*

$$f\left(\Sigma_{t=1}^T z_t\right) = \frac{1}{2^{T/2}\Gamma\left(T/2\right)}\left(\Sigma_{t=1}^T z_t\right)^{T/2-1}\exp\left(-\Sigma_{t=1}^T z_t/2\right).$$

We transform this distribution by first subtracting one from z_t (to remove the mean) and then by dividing by T or $\sqrt{T}$. This gives the distributions of the sample mean, $\bar{z}_1 = \Sigma_{t=1}^T\left(z_t - 1\right)/T$, and scaled sample mean, $\bar{z}_2 = \sqrt{T}\bar{z}_1$ as

$$\begin{aligned}f\left(\bar{z}_1\right) &= \frac{1}{2^{T/2}\Gamma\left(T/2\right)} y^{T/2-1}\exp\left(-y/2\right) \text{ with } y = T\bar{z}_1 + T, \text{ and}\\ f\left(\bar{z}_2\right) &= \frac{1}{2^{T/2}\Gamma\left(T/2\right)} y^{T/2-1}\exp\left(-y/2\right) \text{ with } y = \sqrt{T}\bar{z}_1 + T.\end{aligned}$$

These distributions are shown in Figure 1.1. It is clear that $f\left(\bar{z}_1\right)$ converges to a spike at zero as the sample size increases, while $f\left(\bar{z}_2\right)$ converges to a (non-trivial) normal distribution.

Bibliography

Amemiya, T., 1985, *Advanced Econometrics*, Harvard University Press, Cambridge, Massachusetts.

Davidson, J., 2000, *Econometric Theory*, Blackwell Publishers, Oxford.

Davidson, R., and J. G. MacKinnon, 1993, *Estimation and Inference in Econometrics*, Oxford University Press, Oxford.

Greene, W. H., 2000, *Econometric Analysis*, Prentice-Hall, Upper Saddle River, New Jersey, 4th edn.

Hamilton, J. D., 1994, *Time Series Analysis*, Princeton University Press, Princeton.

Härdle, W., 1990, *Applied Nonparametric Regression*, Cambridge University Press, Cambridge.

Harvey, A. C., 1989, *Forecasting, Structural Time Series Models and the Kalman Filter*, Cambridge University Press.

Hayashi, F., 2000, *Econometrics*, Princeton University Press.

Johnston, J., and J. DiNardo, 1997, *Econometric Methods*, McGraw-Hill, New York, 4th edn.

Lütkepohl, H., 1993, *Introduction to Multiple Time Series*, Springer-Verlag, 2nd edn.

Mittelhammer, R. C., G. J. Judge, and D. J. Miller, 2000, *Econometric Foundations*, Cambridge University Press, Cambridge.

Patterson, K., 2000, *An Introduction to Applied Econometrics: A Time Series Approach*, MacMillan Press, London.

Pindyck, R. S., and D. L. Rubinfeld, 1998, *Econometric Models and Economic Forecasts*, Irwin McGraw-Hill, Boston, Massachusetts, 4ed edn.

Priestley, M. B., 1981, *Spectral Analysis and Time Series*, Academic Press.

Ruud, P. A., 2000, *An Introduction to Classical Econometric Theory*, Oxford University Press.

Silverman, B. W., 1986, *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.

Verbeek, M., 2004, *A Guide to Modern Econometrics*, Wiley, Chichester, 2nd edn.

2 Univariate Time Series Analysis

Reference: Greene (2000) 13.1-3 and 18.1-3
Additional references: Hayashi (2000) 6.2-4; Verbeek (2004) 8-9; Hamilton (1994); Johnston and DiNardo (1997) 7; and Pindyck and Rubinfeld (1998) 16-18

2.1 Theoretical Background to Time Series Processes

Suppose we have a sample of T observations of a random variable

$$\left\{y_t^i\right\}_{t=1}^T = \left\{y_1^i, y_2^i, ..., y_T^i\right\},$$

where subscripts indicate time periods. The superscripts indicate that this sample is from planet (realization) i. We could imagine a continuum of parallel planets where the same time series process has generated different samples with T different numbers (different realizations).

Consider period t. The distribution of y_t across the (infinite number of) planets has some density function, $f_t(y_t)$. The mean of this distribution

$$\mathrm{E}y_t = \int_{-\infty}^{\infty} y_t f_t(y_t)\, dy_t \tag{2.1}$$

is the expected value of the value in period t, also called the *unconditional mean of* y_t. Note that $\mathrm{E}y_t$ could be different from $\mathrm{E}y_{t+s}$. The unconditional variance is defined similarly.

Now consider periods t and $t-s$ jointly. On planet i we have the pair $\left\{y_{t-s}^i, y_t^i\right\}$. The bivariate distribution of these pairs, across the planets, has some density function $g_{t-s,t}(y_{t-s}, y_t)$.[1] Calculate the covariance between y_{t-s} and y_t as usual

$$\begin{aligned} \mathrm{Cov}(y_{t-s}, y_t) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (y_{t-s} - \mathrm{E}y_{t-s})(y_t - \mathrm{E}y_t)\, g_{t-s,t}(y_{t-s}, y_t)\, dy_t dy_{t-s} && (2.2) \\ &= \mathrm{E}(y_{t-s} - \mathrm{E}y_{t-s})(y_t - \mathrm{E}y_t). && (2.3) \end{aligned}$$

[1] The relation between $f_t(y_t)$ and $g_{t-s,t}(y_{t-s}, y_t)$ is, as usual, $f_t(y_t) = \int_{-\infty}^{\infty} g_{t-s,t}(y_{t-s}, y_t)\, dy_{t-s}$.

This is the s^{th} *autocovariance of* y_t. (Of course, $s = 0$ or $s < 0$ are allowed.)

A stochastic process is *covariance stationary* if

$$
\begin{aligned}
\mathrm{E}y_t = \mu \text{ is independent of } t, \quad & (2.4)\\
\mathrm{Cov}\,(y_{t-s}, y_t) = \gamma_s \text{ depends only on } s, \text{ and} \quad & (2.5)\\
\text{both } \mu \text{ and } \gamma_s \text{ are finite.} \quad & (2.6)
\end{aligned}
$$

Most of these notes are about covariance stationary processes, but Section 2.7 is about non-stationary processes.

Humanity has so far only discovered one planet with coin flipping; any attempt to estimate the moments of a time series process must therefore be based on the realization of the stochastic process from planet earth only. This is meaningful only if the process is ergodic for the moment you want to estimate. *A covariance stationary process is said to be ergodic for the mean* if

$$\operatorname{plim} \frac{1}{T} \sum_{t=1}^{T} y_t = \mathrm{E}y_t, \tag{2.7}$$

so the sample mean converges in probability to the unconditional mean. A sufficient condition for ergodicity for the mean is

$$\sum_{s=0}^{\infty} |\mathrm{Cov}\,(y_{t-s}, y_t)| < \infty. \tag{2.8}$$

This means that the link between the values in t and $t - s$ goes to zero sufficiently fast as s increases (you may think of this as getting independent observations before we reach the limit). If y_t is normally distributed, then (2.8) is also sufficient for the process to be ergodic for all moments, not just the mean. *Figure 2.1* illustrates how a longer and longer sample (of one realization of the same time series process) gets closer and closer to the unconditional distribution as the sample gets longer.

2.2 Estimation of Autocovariances

Let y_t be a vector of a covariance stationary and ergodic. The sth covariance matrix is

$$R\,(s) = \mathrm{E}\,(y_t - \mathrm{E}y_t)\,(y_{t-s} - \mathrm{E}y_{t-s})'\,. \tag{2.9}$$

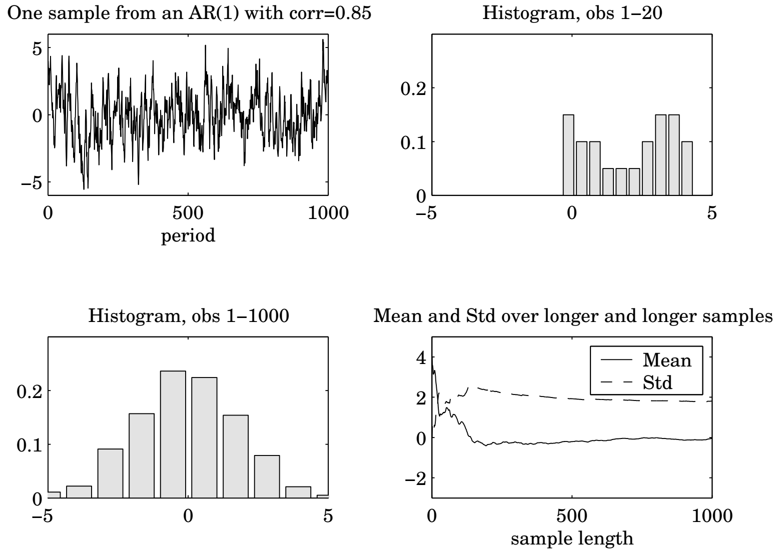


Figure 2.1: Sample of one realization of $y_t = 0.85y_{t-1} + \varepsilon_t$ with $y_0 = 4$ and $\mathrm{Std}(\varepsilon_t) = 1$.

Note that $R\,(s)$ does not have to be symmetric unless $s = 0$. However, note that $R\,(s) = R\,(-s)'$. This follows from noting that

$$
\begin{aligned}
R\,(-s) &= \mathrm{E}\,(y_t - \mathrm{E}y_t)\,(y_{t+s} - \mathrm{E}y_{t+s})'\\
&= \mathrm{E}\,(y_{t-s} - \mathrm{E}y_{t-s})\,(y_t - \mathrm{E}y_t)'\,, \qquad (2.10a)
\end{aligned}
$$

where we have simply changed time subscripts and exploited the fact that y_t is covariance stationary. Transpose to get

$$R\,(-s)' = \mathrm{E}\,(y_t - \mathrm{E}y_t)\,(y_{t-s} - \mathrm{E}y_{t-s})'\,, \tag{2.11}$$

which is the same as in (2.9). If y_t is a scalar, then $R\,(s) = R\,(-s)$, which shows that *auto*covariances are symmetric around $s = 0$.

Example 2.1 *(Bivariate case.) Let $y_t = [x_t, z_t]'$ with $Ex_t = Ez_t = 0$. Then*

$$\hat{R}(s) = E\begin{bmatrix} x_t \\ z_t \end{bmatrix}\begin{bmatrix} x_{t-s} & z_{t-s} \end{bmatrix}$$
$$= \begin{bmatrix} Cov(x_t, x_{t-s}) & Cov(x_t, z_{t-s}) \\ Cov(z_t, x_{t-s}) & Cov(z_t, x_{t-s}) \end{bmatrix}.$$

Note that $R(-s)$ is

$$R(-s) = \begin{bmatrix} Cov(x_t, x_{t+s}) & Cov(x_t, z_{t+s}) \\ Cov(z_t, x_{t+s}) & Cov(z_t, x_{t+s}) \end{bmatrix}$$
$$= \begin{bmatrix} Cov(x_{t-s}, x_t) & Cov(x_{t-s}, z_t) \\ Cov(z_{t-s}, x_t) & Cov(z_{t-s}, x_t) \end{bmatrix},$$

which is indeed the transpose of $R(s)$.

The autocovariances of the (vector) y_t process can be estimated as

$$\hat{R}(s) = \frac{1}{T}\sum_{t=1+s}^{T}(y_t - \bar{y})(y_{t-s} - \bar{y})', \tag{2.12}$$
$$\text{with } \bar{y} = \frac{1}{T}\sum_{t=1}^{T} y_t. \tag{2.13}$$

(We typically divide by T in even if we have only $T - s$ full observations to estimate $R(s)$ from.)

Autocorrelations are then estimated by dividing the diagonal elements in $\hat{R}(s)$ by the diagonal elements in $\hat{R}(0)$

$$\hat{\rho}(s) = \text{diag}\hat{R}(s)/\text{diag}\hat{R}(0) \text{ (element by element).} \tag{2.14}$$

2.3 White Noise

A *white noise time process* has

$$\text{E}\varepsilon_t = 0$$
$$\text{Var}(\varepsilon_t) = \sigma^2, \text{ and}$$
$$\text{Cov}(\varepsilon_{t-s}, \varepsilon_t) = 0 \text{ if } s \neq 0. \tag{2.15}$$

If, in addition, ε_t is normally distributed, then it is said to be Gaussian white noise. The conditions in (2.4)-(2.6) are satisfied so this process is covariance stationary. Moreover, (2.8) is also satisfied, so the process is ergodic for the mean (and all moments if ε_t is normally distributed).

2.4 Moving Average

A q^{th}*-order moving average process* is

$$y_t = \varepsilon_t + \theta_1\varepsilon_{t-1} + ... + \theta_q\varepsilon_{t-q}, \tag{2.16}$$

where the *innovation* ε_t is white noise (usually Gaussian). We could also allow both y_t and ε_t to be vectors; such a process it called a vector MA (VMA).

We have $\text{E}y_t = 0$ and

$$\text{Var}(y_t) = \text{E}\left(\varepsilon_t + \theta_1\varepsilon_{t-1} + ... + \theta_q\varepsilon_{t-q}\right)\left(\varepsilon_t + \theta_1\varepsilon_{t-1} + ... + \theta_q\varepsilon_{t-q}\right)$$
$$= \sigma^2\left(1 + \theta_1^2 + ... + \theta_q^2\right). \tag{2.17}$$

Autocovariances are calculated similarly, and it should be noted that autocovariances of order $q + 1$ and higher are always zero for an MA(q) process.

Example 2.2 *The mean of an MA(1), $y_t = \varepsilon_t + \theta_1\varepsilon_{t-1}$, is zero since the mean of ε_t (and ε_{t-1}) is zero. The first three autocovariance are*

$$Var(y_t) = E(\varepsilon_t + \theta_1\varepsilon_{t-1})(\varepsilon_t + \theta_1\varepsilon_{t-1}) = \sigma^2\left(1 + \theta_1^2\right)$$
$$Cov(y_{t-1}, y_t) = E(\varepsilon_{t-1} + \theta_1\varepsilon_{t-2})(\varepsilon_t + \theta_1\varepsilon_{t-1}) = \sigma^2\theta_1$$
$$Cov(y_{t-2}, y_t) = E(\varepsilon_{t-2} + \theta_1\varepsilon_{t-3})(\varepsilon_t + \theta_1\varepsilon_{t-1}) = 0, \tag{2.18}$$

and Cov$(y_{t-s}, y_t) = 0$ *for* $|s| \geq 2$. *Since both the mean and the covariances are finite and constant across* t, *the MA(1) is covariance stationary. Since the absolute value of the covariances sum to a finite number, the MA(1) is also ergodic for the mean. The first autocorrelation of an MA(1) is*

$$Corr\,(y_{t-1}, y_t) = \frac{\theta_1}{1+\theta_1^2}.$$

Since the white noise process is covariance stationary, and since an MA(q) with $m < \infty$ is a finite order linear function of ε_t, it must be the case that the MA(q) is covariance stationary. It is ergodic for the mean since Cov$(y_{t-s}, y_t) = 0$ for $s > q$, so (2.8) is satisfied. As usual, Gaussian innovations are then sufficient for the MA(q) to be ergodic for all moments.

The effect of ε_t on y_t, $y_{t+1}, ...,$ that is, the *impulse response function,* is the same as the MA coefficients

$$\frac{\partial y_t}{\partial \varepsilon_t} = 1, \frac{\partial y_{t+1}}{\partial \varepsilon_t} = \theta_1, ..., \frac{\partial y_{t+q}}{\partial \varepsilon_t} = \theta_q, \text{ and } \frac{\partial y_{t+q+k}}{\partial \varepsilon_t} = 0 \text{ for } k > 0. \tag{2.19}$$

This is easily seen from applying (2.16)

$$\begin{aligned}
y_t &= \underline{\varepsilon_t} + \theta_1\varepsilon_{t-1} + ... + \theta_q\varepsilon_{t-q} \\
y_{t+1} &= \varepsilon_{t+1} + \theta_1\underline{\varepsilon_t} + ... + \theta_q\varepsilon_{t-q+1} \\
&\vdots \\
y_{t+q} &= \varepsilon_{t+q} + \theta_1\varepsilon_{t-1+q} + ... + \theta_q\underline{\varepsilon_t} \\
y_{t+q+1} &= \varepsilon_{t+q+1} + \theta_1\varepsilon_{t+q} + ... + \theta_q\varepsilon_{t+1}.
\end{aligned}$$

The expected value of y_t, conditional on $\{\varepsilon_w\}_{w=-\infty}^{t-s}$ is

$$\begin{aligned}
\mathrm{E}_{t-s}y_t &= \mathrm{E}_{t-s}\left(\varepsilon_t + \theta_1\varepsilon_{t-1} + ... + \theta_s\varepsilon_{t-s} + ... + \theta_q\varepsilon_{t-q}\right) \\
&= \theta_s\varepsilon_{t-s} + ... + \theta_q\varepsilon_{t-q},
\end{aligned} \tag{2.20}$$

since $\mathrm{E}_{t-s}\varepsilon_{t-(s-1)} = \ldots = \mathrm{E}_{t-s}\varepsilon_t = 0$.

Example 2.3 *(Forecasting an MA(1).) Suppose the process is*

$$y_t = \varepsilon_t + \theta_1\varepsilon_{t-1}, \textit{ with } Var\,(\varepsilon_t) = \sigma^2.$$

The forecasts made in $t = 2$ *then have the follow expressions—with an example using* $\theta_1 = 2, \varepsilon_1 = 3/4$ *and* $\varepsilon_2 = 1/2$ *in the second column*

General	*Example*
y_2	$= 1/2 + 2 \times 3/4 = 2$
$E_2y_3 = E_2\,(\varepsilon_3 + \theta_1\varepsilon_2) = \theta_1\varepsilon_2$	$= 2 \times 1/2 = 1$
$E_2y_4 = E_2\,(\varepsilon_4 + \theta_1\varepsilon_3) = 0$	$= 0$

Example 2.4 *(MA(1) and conditional variances.) From Example 2.3, the forecasting variances are—with the numerical example continued assuming that* $\sigma^2 = 1$

General	*Example*
$Var(y_2 - E_2y_2) = 0$	$= 0$
$Var(y_3 - E_2y_3) = Var(\varepsilon_3 + \theta_1\varepsilon_2 - \theta_1\varepsilon_2) = \sigma^2$	$= 1$
$Var(y_4 - E_2y_4) = Var\,(\varepsilon_4 + \theta_1\varepsilon_3) = \sigma^2 + \theta_1^2\sigma^2$	$= 5$

If the innovations are iid Gaussian, then the distribution of the $s-$period forecast error

$$y_t - \mathrm{E}_{t-s}y_t = \varepsilon_t + \theta_1\varepsilon_{t-1} + ... + \theta_{s-1}\varepsilon_{t-(s-1)}$$

is

$$(y_t - \mathrm{E}_{t-s}y_t) \sim N\left[0, \sigma^2\left(1 + \theta_1^2 + ... + \theta_{s-1}^2\right)\right], \tag{2.21}$$

since $\varepsilon_t, \varepsilon_{t-1}, ..., \varepsilon_{t-(s-1)}$ are independent Gaussian random variables. This implies that the *conditional distribution of* y_t, conditional on $\{\varepsilon_w\}_{w=-\infty}^{s}$, is

$$y_t|\{\varepsilon_{t-s}, \varepsilon_{t-s-1}, \ldots\} \sim N\,[\mathrm{E}_{t-s}y_t, \mathrm{Var}(y_t - \mathrm{E}_{t-s}y_t)] \tag{2.22}$$

$$\sim N\left[\theta_s\varepsilon_{t-s} + ... + \theta_q\varepsilon_{t-q}, \sigma^2\left(1 + \theta_1^2 + ... + \theta_{s-1}^2\right)\right]. \tag{2.23}$$

The conditional mean is the point forecast and the variance is the variance of the forecast error. Note that if $s > q$, then the conditional distribution coincides with the unconditional distribution since ε_{t-s} for $s > q$ is of no help in forecasting y_t.

Example 2.5 *(MA(1) and convergence from conditional to unconditional distribution.) From examples 2.3 and 2.4 we see that the conditional distributions change according to*

(where Ω_2 indicates the information set in $t = 2$)

$$\begin{array}{lr}
\underline{\textit{General}} & \underline{\textit{Example}} \\
y_2|\,\Omega_2 \sim N\,(y_2, 0) & = N\,(2, 0) \\
y_3|\,\Omega_2 \sim N\,(E_2 y_3, Var(y_3 - E_2 y_3)) & = N\,(1, 1) \\
y_4|\,\Omega_2 \sim N\,(E_2 y_4, Var(y_4 - E_2 y_4)) & = N\,(0, 5)
\end{array}$$

Note that the distribution of $y_4|\,\Omega_2$ coincides with the asymptotic distribution.

Estimation of MA processes is typically done by setting up the likelihood function and then using some numerical method to maximize it.

2.5 Autoregression

A p^{th}-order autoregressive process is

$$y_t = a_1 y_{t-1} + a_2 y_{t-2} + ... + a_p y_{t-p} + \varepsilon_t. \tag{2.24}$$

A VAR(p) is just like the AR(p) in (2.24), but where y_t is interpreted as a vector and a_i as a matrix.

Example 2.6 *(VAR(1) model.) A VAR(1) model is of the following form*

$$\begin{bmatrix} y_{1t} \\ y_{2t} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} y_{1t-1} \\ y_{2t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix}.$$

All stationary AR(p) processes can be written on MA(∞) form by repeated substitution. To do so we rewrite the AR(p) as a first order vector autoregression, VAR(1). For instance, an AR(2) $x_t = a_1 x_{t-1} + a_2 x_{t-2} + \varepsilon_t$ can be written as

$$\begin{bmatrix} x_t \\ x_{t-1} \end{bmatrix} = \begin{bmatrix} a_1 & a_2 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} x_{t-1} \\ x_{t-2} \end{bmatrix} + \begin{bmatrix} \varepsilon_t \\ 0 \end{bmatrix}, \text{ or} \tag{2.25}$$

$$y_t = A y_{t-1} + \varepsilon_t, \tag{2.26}$$

where y_t is an 2×1 vector and A a 4×4 matrix. This works also if x_t and ε_t are vectors and. In this case, we interpret a_i as matrices and 1 as an identity matrix.

Iterate backwards on (2.26)

$$\begin{aligned}
y_t &= A\,(A y_{t-2} + \varepsilon_{t-1}) + \varepsilon_t \\
&= A^2 y_{t-2} + A\varepsilon_{t-1} + \varepsilon_t \\
&\vdots \\
&= A^{K+1} y_{t-K-1} + \sum_{s=0}^{K} A^s \varepsilon_{t-s}.
\end{aligned} \tag{2.27}$$

Remark 2.7 *(Spectral decomposition.) The n eigenvalues (λ_i) and associated eigenvectors (z_i) of the $n \times n$ matrix A satisfy*

$$(A - \lambda_i I_n)\, z_i = \mathbf{0}_{n \times 1}.$$

If the eigenvectors are linearly independent, then

$$A = Z \Lambda Z^{-1}, \textit{ where } \Lambda = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix} \textit{ and } Z = \begin{bmatrix} z_1 & z_2 & \cdots & z_n \end{bmatrix}.$$

Note that we therefore get

$$A^2 = AA = Z\Lambda Z^{-1} Z \Lambda Z^{-1} = Z \Lambda \Lambda Z^{-1} = Z \Lambda^2 Z^{-1} \Rightarrow A^q = Z \Lambda^q Z^{-1}.$$

Remark 2.8 *(Modulus of complex number.) If $\lambda = a + bi$, where $i = \sqrt{-1}$, then $|\lambda| = |a + bi| = \sqrt{a^2 + b^2}$.*

Take the limit of (2.27) as $K \to \infty$. If $\lim_{K\to\infty} A^{K+1} y_{t-K-1} = 0$, then we have a moving average representation of y_t where the influence of the starting values vanishes asymptotically

$$y_t = \sum_{s=0}^{\infty} A^s \varepsilon_{t-s}. \tag{2.28}$$

We note from the spectral decompositions that $A^{K+1} = Z \Lambda^{K+1} Z^{-1}$, where Z is the matrix of eigenvectors and Λ a diagonal matrix with eigenvalues. Clearly, $\lim_{K\to\infty} A^{K+1} y_{t-K-1} = 0$ is satisfied if the eigenvalues of A are all less than one in modulus and y_{t-K-1} does not grow without a bound.

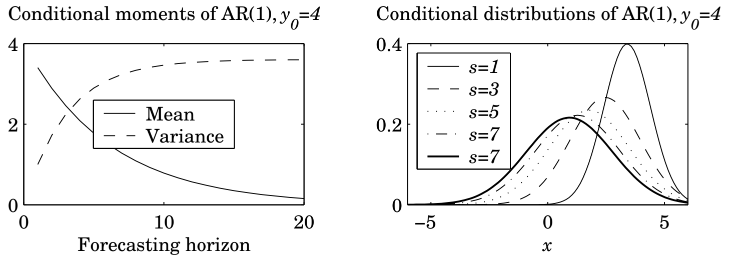


Figure 2.2: Conditional moments and distributions for different forecast horizons for the AR(1) process $y_t = 0.85y_{t-1} + \varepsilon_t$ with $y_0 = 4$ and $\text{Std}(\varepsilon_t) = 1$.

Example 2.9 *(AR(1).) For the univariate AR(1) $y_t = ay_{t-1} + \varepsilon_t$, the characteristic equation is $(a - \lambda)\, z = 0$, which is only satisfied if the eigenvalue is $\lambda = a$. The AR(1) is therefore stable (and stationarity) if $-1 < a < 1$. This can also be seen directly by noting that $a^{K+1}y_{t-K-1}$ declines to zero if $0 < a < 1$ as K increases.*

Similarly, most finite order MA processes can be written ("inverted") as AR(∞). It is therefore common to approximate MA processes with AR processes, especially since the latter are much easier to estimate.

Example 2.10 *(Variance of AR(1).) From the MA-representation $y_t = \sum_{s=0}^{\infty} a^s \varepsilon_{t-s}$ and the fact that ε_t is white noise we get $Var(y_t) = \sigma^2 \sum_{s=0}^{\infty} a^{2s} = \sigma^2 / \left(1 - a^2\right)$. Note that this is minimized at $a = 0$. The autocorrelations are obviously $a^{|s|}$. The covariance matrix of $\{y_t\}_{t=1}^{T}$ is therefore (standard deviation×standard deviation×autocorrelation)*

$$\frac{\sigma^2}{1-a^2}\begin{bmatrix} 1 & a & a^2 & \cdots & a^{T-1} \\ a & 1 & a & \cdots & a^{T-2} \\ a^2 & a & 1 & \cdots & a^{T-3} \\ \vdots & & & \ddots & \\ a^{T-1} & a^{T-2} & a^{T-3} & \cdots & 1 \end{bmatrix}.$$

Example 2.11 *(Covariance stationarity of an AR(1) with $|a| < 1$.) From the MA-representation $y_t = \sum_{s=0}^{\infty} a^s \varepsilon_{t-s}$, the expected value of y_t is zero, since $E\varepsilon_{t-s} = 0$. We know that Cov(y_t, y_{t-s})= $a^{|s|}\sigma^2 / \left(1 - a^2\right)$ which is constant and finite.*

Example 2.12 *(Ergodicity of a stationary AR(1).) We know that Cov(y_t, y_{t-s})= $a^{|s|}\sigma^2 / \left(1 - a^2\right)$, so the absolute value is*

$$|Cov(y_t, y_{t-s})| = |a|^{|s|}\, \sigma^2 / \left(1 - a^2\right)$$

Using this in (2.8) gives

$$\begin{aligned} \sum_{s=0}^{\infty} |Cov\,(y_{t-s}, y_t)| &= \frac{\sigma^2}{1-a^2} \sum_{s=0}^{\infty} |a|^s \\ &= \frac{\sigma^2}{1-a^2}\frac{1}{1-|a|} \text{ (since } |a| < 1) \end{aligned}$$

which is finite. The AR(1) is ergodic if $|a| < 1$.

Example 2.13 *(Conditional distribution of AR(1).) For the AR(1) $y_t = ay_{t-1} + \varepsilon_t$ with $\varepsilon_t \sim N\left(0, \sigma^2\right)$, we get*

$$\begin{aligned} E_t y_{t+s} &= a^s y_t, \\ Var\,(y_{t+s} - E_t y_{t+s}) &= \left(1 + a^2 + a^4 + ... + a^{2(s-1)}\right)\sigma^2 \\ &= \frac{a^{2s}-1}{a^2-1}\sigma^2. \end{aligned}$$

The distribution of y_{t+s} conditional on y_t is normal with these parameters. See Figure 2.2 for an example.

2.5.1 Estimation of an AR(1) Process

Suppose we have sample $\{y_t\}_{t=0}^{T}$ of a process which we know is an AR(p), $y_t = ay_{t-1} + \varepsilon_t$, with normally distributed innovations with unknown variance σ^2.

The pdf of y_1 conditional on y_0 is

$$\text{pdf}\,(y_1|y_0) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_1 - ay_0)^2}{2\sigma^2}\right), \tag{2.29}$$

and the pdf of y_2 conditional on y_1 and y_0 is

$$\text{pdf}\,(y_2|\,\{y_1, y_0\}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_2 - ay_1)^2}{2\sigma^2}\right). \tag{2.30}$$

Recall that the joint and conditional pdfs of some variables z and x are related as

$$\text{pdf}\,(x,z) = \text{pdf}\,(x|z) * \text{pdf}\,(z)\,. \tag{2.31}$$

Applying this principle on (2.29) and (2.31) gives

$$\begin{aligned}\text{pdf}\,(y_2, y_1|y_0) &= \text{pdf}\,(y_2|\,\{y_1, y_0\})\,\text{pdf}\,(y_1|y_0)\\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^2 \exp\left(-\frac{(y_2 - ay_1)^2 + (y_1 - ay_0)^2}{2\sigma^2}\right).\end{aligned} \tag{2.32}$$

Repeating this for the entire sample gives the likelihood function for the sample

$$\text{pdf}\left(\{y_t\}_{t=0}^{T}\middle|\, y_0\right) = \left(2\pi\sigma^2\right)^{-T/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{t=1}^{T}(y_t - a_1 y_{t-1})^2\right). \tag{2.33}$$

Taking logs, and evaluating the first order conditions for σ^2 and a gives the usual OLS estimator. Note that this is MLE *conditional on* y_0. There is a corresponding exact MLE, but the difference is usually small (the asymptotic distributions of the two estimators are the same under stationarity; under non-stationarity OLS still gives consistent estimates). The MLE of Var(ε_t) is given by $\sum_{t=1}^{T} \hat{v}_t^2/T$, where $\hat{v}_t$ is the OLS residual.

These results carry over to any finite-order VAR. The MLE, conditional on the initial observations, of the VAR is the same as OLS estimates of each equation. The MLE of the ij^{th} element in Cov(ε_t) is given by $\sum_{t=1}^{T} \hat{v}_{it}\hat{v}_{jt}/T$, where $\hat{v}_{it}$ and $\hat{v}_{jt}$ are the OLS residuals.

To get the exact MLE, we need to multiply (2.33) with the unconditional pdf of y_0 (since we have no information to condition on)

$$\text{pdf}\,(y_0) = \frac{1}{\sqrt{2\pi\sigma^2/(1-a^2)}} \exp\left(-\frac{y_0^2}{2\sigma^2/(1-a^2)}\right), \tag{2.34}$$

since $y_0 \sim N(0, \sigma^2/(1-a^2))$. The optimization problem is then non-linear and must be solved by a numerical optimization routine.

2.5.2 Lag Operators*

A common and convenient way of dealing with leads and lags is the *lag operator*, L. It is such that

$$\mathrm{L}^s y_t = y_{t-s} \text{ for all (integer) } s.$$

For instance, the ARMA(2,1) model

$$y_t - a_1 y_{t-1} - a_2 y_{t-2} = \varepsilon_t + \theta_1 \varepsilon_{t-1} \tag{2.35}$$

can be written as

$$\left(1 - a_1\mathrm{L} - a_2\mathrm{L}^2\right) y_t = (1 + \theta_1\mathrm{L})\,\varepsilon_t, \tag{2.36}$$

which is usually denoted

$$a\,(\mathrm{L})\,y_t = \theta\,(\mathrm{L})\,\varepsilon_t. \tag{2.37}$$

2.5.3 Properties of LS Estimates of an AR(p) Process*

Reference: Hamilton (1994) 8.2

The LS estimates are typically biased, but consistent and asymptotically normally distributed, provided the AR is stationary.

As usual the LS estimate is

$$\begin{aligned}\hat{\beta}_{LS} - \beta &= \left[\frac{1}{T}\sum_{t=1}^{T} x_t x_t'\right]^{-1} \frac{1}{T}\sum_{t=1}^{T} x_t\varepsilon_t, \text{ where} \\ x_t &= \left[\begin{array}{cccc} y_{t-1} & y_{t-2} & \cdots & y_{t-p}\end{array}\right].\end{aligned} \tag{2.38}$$

The first term in (2.38) is the inverse of the sample estimate of covariance matrix of x_t (since $\mathrm{E}y_t = 0$), which converges in probability to Σ_{xx}^{-1} (y_t is stationary and ergodic for all moments if ε_t is Gaussian). The last term, $\frac{1}{T}\sum_{t=1}^{T} x_t\varepsilon_t$, is serially uncorrelated, so we can apply a CLT. Note that $\mathrm{E}x_t\varepsilon_t\varepsilon_t'x_t'$ $=\mathrm{E}\varepsilon_t\varepsilon_t'\mathrm{E}x_t x_t' = \sigma^2\Sigma_{xx}$ since u_t and x_t are independent. We therefore have

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T} x_t\varepsilon_t \to^d N\left(0, \sigma^2\Sigma_{xx}\right). \tag{2.39}$$

Combining these facts, we get the asymptotic distribution

$$\sqrt{T}\left(\hat{\beta}_{LS}-\beta\right) \to^{d} N\left(0, \Sigma_{xx}^{-1}\sigma^{2}\right). \tag{2.40}$$

Consistency follows from taking plim of (2.38)

$$\begin{aligned}\operatorname{plim}\left(\hat{\beta}_{LS}-\beta\right) &= \Sigma_{xx}^{-1}\operatorname{plim}\frac{1}{T}\sum_{t=1}^{T}x_t\varepsilon_t\\ &= 0,\end{aligned}$$

since x_t and ε_t are uncorrelated.

2.5.4 Autoregressions versus Autocorrelations*

It is straightforward to see the relation between autocorrelations and the AR model when the AR model is the true process. This relation is given by the *Yule-Walker equations*.

For an AR(1), the autoregression coefficient is simply the first autocorrelation coefficient. For an AR(2), $y_t = a_1 y_{t-1} + a_2 y_{t-2} + \varepsilon_t$, we have

$$\begin{aligned}\begin{bmatrix}\operatorname{Cov}(y_t, y_t)\\ \operatorname{Cov}(y_{t-1}, y_t)\\ \operatorname{Cov}(y_{t-2}, y_t)\end{bmatrix} &= \begin{bmatrix}\operatorname{Cov}(y_t, a_1 y_{t-1} + a_2 y_{t-2} + \varepsilon_t)\\ \operatorname{Cov}(y_{t-1}, a_1 y_{t-1} + a_2 y_{t-2} + \varepsilon_t)\\ \operatorname{Cov}(y_{t-2}, a_1 y_{t-1} + a_2 y_{t-2} + \varepsilon_t)\end{bmatrix}\\ &= \begin{bmatrix}a_1\operatorname{Cov}(y_t, y_{t-1}) + a_2\operatorname{Cov}(y_t, y_{t-2}) + \operatorname{Cov}(y_t, \varepsilon_t)\\ a_1\operatorname{Cov}(y_{t-1}, y_{t-1}) + a_2\operatorname{Cov}(y_{t-1}, y_{t-2})\\ a_1\operatorname{Cov}(y_{t-2}, y_{t-1}) + a_2\operatorname{Cov}(y_{t-2}, y_{t-2})\end{bmatrix}, \text{ or}\end{aligned}$$

$$\begin{bmatrix}\gamma_0\\ \gamma_1\\ \gamma_2\end{bmatrix} = \begin{bmatrix}a_1\gamma_1 + a_2\gamma_2 + \operatorname{Var}(\varepsilon_t)\\ a_1\gamma_0 + a_2\gamma_1\\ a_1\gamma_1 + a_2\gamma_0\end{bmatrix}. \tag{2.41}$$

To transform to autocorrelation, divide through by γ_0. The last two equations are then

$$\begin{bmatrix}\rho_1\\ \rho_2\end{bmatrix} = \begin{bmatrix}a_1 + a_2\rho_1\\ a_1\rho_1 + a_2\end{bmatrix} \text{ or } \begin{bmatrix}\rho_1\\ \rho_2\end{bmatrix} = \begin{bmatrix}a_1/\left(1-a_2\right)\\ a_1^2/\left(1-a_2\right) + a_2\end{bmatrix}. \tag{2.42}$$

If we know the parameters of the AR(2) model (a_1, a_2, and $\operatorname{Var}(\varepsilon_t)$), then we can solve for the autocorrelations. Alternatively, if we know the autocorrelations, then we can solve for the autoregression coefficients. This demonstrates that testing that all the autocorrelations are zero is essentially the same as testing if all the autoregressive coefficients are zero. Note, however, that the transformation is non-linear, which may make a difference in small samples.

2.6 ARMA Models

An ARMA model has both AR and MA components. For instance, an ARMA(p,q) is

$$y_t = a_1 y_{t-1} + a_2 y_{t-2} + ... + a_p y_{t-p} + \varepsilon_t + \theta_1\varepsilon_{t-1} + ... + \theta_q\varepsilon_{t-q}. \tag{2.43}$$

Estimation of ARMA processes is typically done by setting up the likelihood function and then using some numerical method to maximize it.

Even low-order ARMA models can be fairly flexible. For instance, the ARMA(1,1) model is

$$y_t = a y_{t-1} + \varepsilon_t + \theta\varepsilon_{t-1}, \text{ where } \varepsilon_t \text{ is white noise.} \tag{2.44}$$

The model can be written on MA(∞) form as

$$y_t = \varepsilon_t + \sum_{s=1}^{\infty} a^{s-1}(a+\theta)\varepsilon_{t-s}. \tag{2.45}$$

The autocorrelations can be shown to be

$$\rho_1 = \frac{(1+a\theta)(a+\theta)}{1+\theta^2+2a\theta}, \text{ and } \rho_s = a\rho_{s-1} \text{ for } s = 2, 3, \ldots \tag{2.46}$$

and the conditional expectations are

$$\mathrm{E}_t\, y_{t+s} = a^{s-1}(a y_t + \theta\varepsilon_t)\; s = 1, 2, \ldots \tag{2.47}$$

See Figure 2.3 for an example.

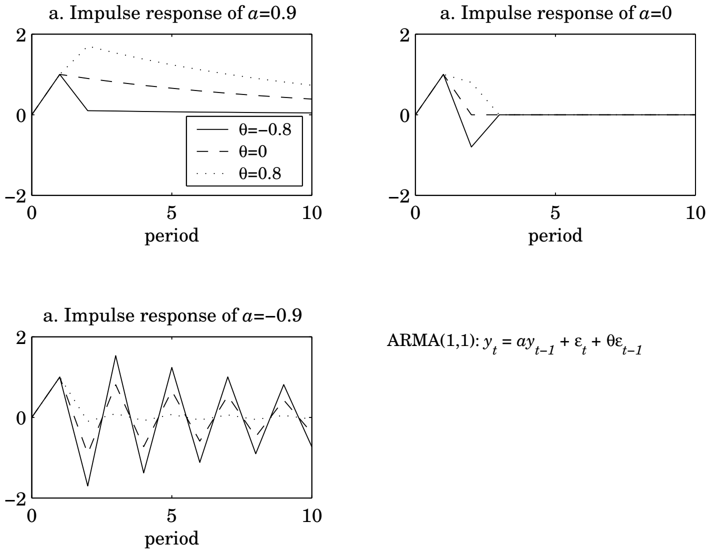


Figure 2.3: Impulse response function of ARMA(1,1)

2.7 Non-stationary Processes

2.7.1 Introduction

A *trend-stationary process* can be made stationary by subtracting a linear trend. The simplest example is

$$y_t = \mu + \beta t + \varepsilon_t \tag{2.48}$$

where ε_t is white noise.

A *unit root* process can be made stationary only by taking a difference. The simplest example is the *random walk* with drift

$$y_t = \mu + y_{t-1} + \varepsilon_t, \tag{2.49}$$

where ε_t is white noise. The name "unit root process" comes from the fact that the largest eigenvalues of the canonical form (the VAR(1) form of the AR(p)) is one. Such a process is said to be integrated of order one (often denoted I(1)) and can be made stationary by taking first differences.

Example 2.14 *(Non-stationary AR(2).) The process $y_t = 1.5y_{t-1} - 0.5y_{t-2} + \varepsilon_t$ can be written*

$$\begin{bmatrix} y_t \\ y_{t-1} \end{bmatrix} = \begin{bmatrix} 1.5 & -0.5 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} y_{t-1} \\ y_{t-2} \end{bmatrix} + \begin{bmatrix} \varepsilon_t \\ 0 \end{bmatrix},$$

where the matrix has the eigenvalues 1 and 0.5 and is therefore non-stationary. Note that subtracting y_{t-1} from both sides gives $y_t - y_{t-1} = 0.5\,(y_{t-1} - y_{t-2}) + \varepsilon_t$, so the variable $x_t = y_t - y_{t-1}$ is stationary.

The *distinguishing feature of unit root processes* is that *the effect of a shock never vanishes*. This is most easily seen for the random walk. Substitute repeatedly in (2.49) to get

$$\begin{aligned} y_t &= \mu + (\mu + y_{t-2} + \varepsilon_{t-1}) + \varepsilon_t \\ &\vdots \\ &= t\mu + y_0 + \sum_{s=1}^{t} \varepsilon_s. \end{aligned} \tag{2.50}$$

The effect of ε_t never dies out: a non-zero value of ε_t gives a permanent shift of the level of y_t. This process is clearly non-stationary. A consequence of the permanent effect of a shock is that the variance of the conditional distribution grows without bound as the forecasting horizon is extended. For instance, for the random walk with drift, (2.50), the distribution conditional on the information in $t = 0$ is $N\left(y_0 + t\mu, s\sigma^2\right)$ if the innovations are Gaussian. This means that the expected change is $t\mu$ and that the conditional variance grows linearly with the forecasting horizon. The unconditional variance is therefore infinite and the standard results on inference are not applicable.

In contrast, the conditional distributions from the trend stationary model, (2.48), is $N\left(st, \sigma^2\right)$.

A process could have two unit roots (integrated of order 2: I(2)). In this case, we need to difference twice to make it stationary. Alternatively, a process can also be explosive, that is, have eigenvalues outside the unit circle. In this case, the impulse response function diverges.

Example 2.15 *(Two unit roots.) Suppose y_t in Example (2.14) is actually the first difference of some other series, $y_t = z_t - z_{t-1}$. We then have*

$$\begin{aligned} z_t - z_{t-1} &= 1.5\,(z_{t-1} - z_{t-2}) - 0.5\,(z_{t-2} - z_{t-3}) + \varepsilon_t \\ z_t &= 2.5 z_{t-1} - 2 z_{t-2} + 0.5 z_{t-3} + \varepsilon_t, \end{aligned}$$

which is an AR(3) with the following canonical form

$$\begin{bmatrix} z_t \\ z_{t-1} \\ z_{t-2} \end{bmatrix} = \begin{bmatrix} 2.5 & -2 & 0.5 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} z_{t-1} \\ z_{t-2} \\ z_{t-3} \end{bmatrix} + \begin{bmatrix} \varepsilon_t \\ 0 \\ 0 \end{bmatrix}.$$

The eigenvalues are 1, 1, and 0.5, so z_t has two unit roots (integrated of order 2: I(2) and needs to be differenced twice to become stationary).

Example 2.16 *(Explosive AR(1).) Consider the process $y_t = 1.5 y_{t-1} + \varepsilon_t$. The eigenvalue is then outside the unit circle, so the process is explosive. This means that the impulse response to a shock to ε_t diverges (it is 1.5^s for s periods ahead).*

2.7.2 Spurious Regressions

Strong trends often causes problems in econometric models where y_t is regressed on x_t. In essence, if no trend is included in the regression, then x_t will appear to be significant, just because it is a proxy for a trend. The same holds for unit root processes, even if they have no deterministic trends. However, the innovations accumulate and the series therefore tend to be trending in small samples. A warning sign of a spurious regression is when $R^2 > DW$ statistics.

For trend-stationary data, this problem is easily solved by detrending with a linear trend (before estimating or just adding a trend to the regression).

However, this is usually a poor method for a unit root processes. What is needed is a first difference. For instance, a first difference of the random walk is

$$\begin{aligned} \Delta y_t &= y_t - y_{t-1} \\ &= \varepsilon_t, \end{aligned} \tag{2.51}$$

which is white noise (any finite difference, like $y_t - y_{t-s}$, will give a stationary series), so we could proceed by applying standard econometric tools to Δy_t.

One may then be tempted to try first-differencing all non-stationary series, since it may be hard to tell if they are unit root process or just trend-stationary. For instance, a first difference of the trend stationary process, (2.48), gives

$$y_t - y_{t-1} = \beta + \varepsilon_t - \varepsilon_{t-1}. \tag{2.52}$$

Its unclear if this is an improvement: the trend is gone, but the errors are now of MA(1) type (in fact, non-invertible, and therefore tricky, in particular for estimation).

2.7.3 Testing for a Unit Root I*

Suppose we run an OLS regression of

$$y_t = a y_{t-1} + \varepsilon_t, \tag{2.53}$$

where the true value of $|a| < 1$. The asymptotic distribution is of the LS estimator is

$$\sqrt{T}\,(\hat{a} - a) \sim N\left(0, 1 - a^2\right). \tag{2.54}$$

(The variance follows from the standard OLS formula where the variance of the estimator is $\sigma^2 \left(X'X/T\right)^{-1}$. Here plim $X'X/T =$Var(y_t) which we know is $\sigma^2/\left(1 - a^2\right)$).

It is well known (but not easy to show) that when $a = 1$, then $\hat{a}$ is biased towards zero in small samples. In addition, the asymptotic distribution is no longer (2.54). In fact, there is a discontinuity in the limiting distribution as we move from a stationary/to a non-stationary variable. This, together with the small sample bias means that we have to use simulated critical values for testing the null hypothesis of $a = 1$ based on the OLS estimate from (2.53).

The approach is to calculate the test statistic

$$t = \frac{\hat{a} - 1}{\text{Std}(\hat{a})},$$

and reject *the null of non-stationarity* if t is less than the critical values published by Dickey and Fuller (typically more negative than the standard values to compensate for the small sample bias) or from your own simulations.

In principle, distinguishing between a stationary and a non-stationary series is very difficult (and impossible unless we restrict the class of processes, for instance, to an AR(2)), since any sample of a non-stationary process can be arbitrary well approximated by *some* stationary process et vice versa. The lesson to be learned, from a practical point of view, is that *strong persistence in the data* generating process (stationary or not) *invalidates the usual results on inference*. We are usually on safer ground to apply the unit root results in this case, even if the process is actually stationary.

2.7.4 Testing for a Unit Root II*

Reference: Fuller (1976), *Introduction to Statistical Time Series*; Dickey and Fuller (1979), "Distribution of the Estimators for Autoregressive Time Series with a Unit Root," *Journal of the American Statistical Association*, **74**, 427-431.

Consider the AR(1) with intercept

$$y_t = \gamma + \alpha y_{t-1} + u_t, \text{ or } \Delta y_t = \gamma + \beta y_{t-1} + u_t, \text{ where } \beta = (\alpha - 1). \tag{2.55}$$

The *DF test* is to test the null hypothesis that $\beta = 0$, against $\beta < 0$ using the usual t statistic. However, under the null hypothesis, the distribution of the t statistics is far from a student-t or normal distribution. Critical values, found in Fuller and Dickey and Fuller, are lower than the usual ones. Remember to add any nonstochastic regressors that in required, for instance, seasonal dummies, trends, etc. If you forget a trend, then the power of the test goes to zero as $T \to \infty$. The critical values are lower the more deterministic components that are added.

The *asymptotic* critical values are valid even under heteroskedasticity, and non-normal distributions of u_t. However, no autocorrelation in u_t is allowed for. In contrast, the simulated *small sample* critical values are usually only valid for iid normally distributed disturbances.

The *ADF test* is a way to account for serial correlation in u_t. The same critical values apply. Consider an AR(1) $u_t = \rho u_{t-1} + e_t$. A Cochrane-Orcutt transformation of (2.55) gives

$$\Delta y_t = \gamma (1 - \rho) + \tilde{\beta} y_{t-1} + \rho (\beta + 1) \Delta y_{t-1} + e_t, \text{ where } \tilde{\beta} = \beta (1 - \rho). \tag{2.56}$$

The test is here the t test for $\tilde{\beta}$. The fact that $\tilde{\beta} = \beta (1 - \rho)$ is of no importance, since $\tilde{\beta}$ is zero only if β is (as long as $\rho < 1$, as it must be). (2.56) generalizes so one should include

p lags of Δy_t if u_t is an AR(p). The test remains valid even under an MA structure if the number of lags included increases at the rate $T^{1/3}$ as the sample lenngth increases. In practice: add lags until the remaining residual is white noise. The size of the test (probability of rejecting H_0 when it is actually correct) can be awful in small samples for a series that is a I(1) process that initially "overshoots" over time, as $\Delta y_t = e_t - 0.8e_{t-1}$, since this makes the series look mean reverting (stationary). Similarly, the power (prob of rejecting H_0 when it is false) can be awful when there is a lot of persistence, for instance, if $\alpha = 0.95$.

The power of the test depends on the span of the data, rather than the number of observations. Seasonally adjusted data tend to look more integrated than they are. Should apply different critical values, see Ghysel and Perron (1993), *Journal of Econometrics*, **55**, 57-98. A break in mean or trend also makes the data look non-stationary. Should perhaps apply tests that account for this, see Banerjee, Lumsdaine, Stock (1992), *Journal of Business and Economics Statistics*, **10**, 271-287.

Park (1990, "Testing for Unit Roots and Cointegration by Variable Addition," *Advances in Econometrics*, 8, 107-133) sets up a framework where we can use both non-stationarity as the null hypothesis and where we can have stationarity as the null. Consider the regression

$$y_t = \sum_{s=0}^{p} \beta_s t^s + \sum_{s=p+1}^{q} \beta_s t^s + u_t, \tag{2.57}$$

where the we want to test if H_0: $\beta_s = 0, s = p+1, ..., q$. If $F(p, q)$ is the Wald-statistics for this, then $J(p, q) = F(p, q)/T$ has some (complicated) asymptotic distribution under the null. You reject non-stationarity if $J(p, q) <$ critical value, since $J(p, q) \to^p 0$ under (trend) stationarity.

Now, define

$$G(p, q) = F(p, q) \frac{\text{Var}(u_t)}{\text{Var}\left(\sqrt{T}\bar{u}_t\right)} \sim \chi^2_{p-q} \text{ under } \text{H}_0 \text{ of stationarity,} \tag{2.58}$$

and $G(p, q) \to^p \infty$ under non-stationarity, so we reject stationarity if $G(p, q) >$ critical value. Note that $\text{Var}(u_t)$ is a traditional variance, while $\text{Var}\left(\sqrt{T}\bar{u}_t\right)$ can be estimated with a Newey-West estimator.

2.7.5 Cointegration*

Suppose y_{1t} and y_{2t} are both (scalar) unit root processes, but that

$$\begin{aligned} z_t &= y_{1t} - \beta y_{2t} \qquad (2.59) \\ &= \begin{bmatrix} 1 & -\beta \end{bmatrix} \begin{bmatrix} y_{1t} \\ y_{2t} \end{bmatrix} \end{aligned}$$

is stationary. The processes y_t and x_t must then share the same *common stochastic trend*, and are therefore *cointegrated* with the cointegrating vector $\begin{bmatrix} 1 & -\beta \end{bmatrix}$. Running the regression (2.59) gives an estimator $\hat{\beta}_{LS}$ which converges much faster than usual (it is "superconsistent") and is not affected by any simultaneous equations bias. The intuition for the second result is that the simultaneous equations bias depends on the simultaneous reactions to the shocks, which are stationary and therefore without any long-run importance.

This can be generalized by letting y_t be a vector of n unit root processes which follows a VAR. For simplicity assume it is a VAR(2)

$$y_t = A_1 y_{t-1} + A_2 y_{t-2} + \varepsilon_t. \qquad (2.60)$$

Subtract y_t from both sides, add and subtract $A_2 y_{t-1}$ from the right hand side

$$\begin{aligned} y_t - y_{t-1} &= A_1 y_{t-1} + A_2 y_{t-2} + \varepsilon_t - y_{t-1} + A_2 y_{t-1} - A_2 y_{t-1} \\ &= (A_1 + A_2 - I)\, y_{t-1} - A_2 (y_{t-1} - y_{t-2}) + \varepsilon_t \qquad (2.61) \end{aligned}$$

The left hand side is now stationary, and so is $y_{t-1} - y_{t-2}$ and ε_t on the right hand side. It must therefore be the case that $(A_1 + A_2 - I)\, y_{t-1}$ is also stationary; it must be n linear combinations of the cointegrating vectors. Since the number of cointegrating vectors must be less than n, the rank of $A_1 + A_2 - I$ must be less than n. To impose this calls for special estimation methods.

The simplest of these is *Engle and Granger's two-step procedure*. In the first step, we estimate the cointegrating vectors (as in 2.59) and calculate the different z_t series (fewer than n). In the second step, these are used in the *error correction form* of the VAR

$$y_t - y_{t-1} = \gamma z_{t-1} - A_2 (y_{t-1} - y_{t-2}) + \varepsilon_t \qquad (2.62)$$

to estimate γ and A_2. The relation to (2.61) is most easily seen in the bivariate case. Then, by using (2.59) in (2.62) we get

$$y_t - y_{t-1} = \begin{bmatrix} \gamma & -\gamma\beta \end{bmatrix} y_{t-1} - A_2 (y_{t-1} - y_{t-2}) + \varepsilon_t, \qquad (2.63)$$

so knowledge (estimates) of β (scalar), γ (2 × 1), A_2 (2 × 2) allows us to "back out" A_1.

Bibliography

Greene, W. H., 2000, *Econometric Analysis*, Prentice-Hall, Upper Saddle River, New Jersey, 4th edn.

Hamilton, J. D., 1994, *Time Series Analysis*, Princeton University Press, Princeton.

Hayashi, F., 2000, *Econometrics*, Princeton University Press.

Johnston, J., and J. DiNardo, 1997, *Econometric Methods*, McGraw-Hill, New York, 4th edn.

Pindyck, R. S., and D. L. Rubinfeld, 1998, *Econometric Models and Economic Forecasts*, Irwin McGraw-Hill, Boston, Massachusetts, 4ed edn.

Verbeek, M., 2004, *A Guide to Modern Econometrics*, Wiley, Chichester, 2nd edn.

3 The Distribution of a Sample Average

Reference: Hayashi (2000) 6.5

Additional references: Hamilton (1994) 14; Verbeek (2004) 4.10; Harris and Matyas (1999); and Pindyck and Rubinfeld (1998) Appendix 10.1; Cochrane (2001) 11.7

3.1 Variance of a Sample Average

In order to understand the distribution of many estimators we need to get an important building block: the variance of a sample average.

Consider a *covariance stationary vector process* m_t with zero mean and $\text{Cov}(m_t, m_{t-s}) = R(s)$ (which only depends on s). That is, we allow for serial correlation in m_t, but no heteroskedasticity. This is more restrictive than we want, but we will return to that further on.

Let $\bar{m} = \sum_{t=1}^{T} m_t/T$. The sampling variance of a mean estimator of the zero mean random variable m_t is defined as

$$\text{Cov}(\bar{m}) = \text{E}\left[\left(\frac{1}{T}\sum_{t=1}^{T} m_t\right)\left(\frac{1}{T}\sum_{\tau=1}^{T} m_\tau\right)'\right]. \tag{3.1}$$

Let the covariance (matrix) at lag s be

$$\begin{aligned} R(s) &= \text{Cov}(m_t, m_{t-s}) \\ &= \text{E}\, m_t m_{t-s}', \end{aligned} \tag{3.2}$$

since $\text{E}\, m_t = 0$ for all t.

Example 3.1 *(m_t is a scalar iid process.) When m_t is a scalar iid process, then*

$$\begin{aligned} Var\left(\frac{1}{T}\sum_{t=1}^{T} m_t\right) &= \frac{1}{T^2}\sum_{t=1}^{T} Var(m_t) \text{ /*independently distributed*/} \\ &= \frac{1}{T^2} T\, Var(m_t) \text{ /*identically distributed*/} \\ &= \frac{1}{T} Var(m_t). \end{aligned}$$

This is the classical iid case. Clearly, $\lim_{T\Rightarrow\infty} Var(\bar{m}) = 0$. By multiplying both sides by T we instead get $Var\left(\sqrt{T}\bar{m}\right) = Var(m_t)$, which is often more convenient for asymptotics.

Example 3.2 *Let x_t and z_t be two scalars, with samples averages $\bar{x}$ and $\bar{z}$. Let $m_t = \begin{bmatrix} x_t & z_t \end{bmatrix}'$. Then $Cov(\bar{m})$ is*

$$\begin{aligned} Cov(\bar{m}) &= Cov\left(\begin{bmatrix} \bar{x} \\ \bar{z} \end{bmatrix}\right) \\ &= \begin{bmatrix} Var(\bar{x}) & Cov(\bar{x}, \bar{z}) \\ Cov(\bar{z}, \bar{x}) & Var(\bar{z}) \end{bmatrix}. \end{aligned}$$

Example 3.3 *($Cov(\bar{m})$ with $T = 3$.) With $T = 3$, we have*

$$Cov(T\bar{m}) =$$
$$E(m_1 + m_2 + m_3)(m_1' + m_2' + m_3') =$$
$$\underbrace{E(m_1m_1' + m_2m_2' + m_3m_3')}_{3R(0)} + \underbrace{E(m_2m_1' + m_3m_2')}_{2R(1)} + \underbrace{E(m_1m_2' + m_2m_3')}_{2R(-1)} + \underbrace{Em_3m_1'}_{R(2)} + \underbrace{Em_1m_3'}_{R(-2)}.$$

The general pattern in the previous example is

$$\text{Cov}(T\bar{m}) = \sum_{s=-(T-1)}^{T-1} (T - |s|)\, R(s). \tag{3.3}$$

Divide both sides by T

$$\text{Cov}\left(\sqrt{T}\bar{m}\right) = \sum_{s=-(T-1)}^{T-1} \left(1 - \frac{|s|}{T}\right) R(s). \tag{3.4}$$

This is the exact expression for a given sample size.

In many cases, we use the asymptotic expression (limiting value as $T \to \infty$) instead. If $R(s) = 0$ for $s > q$ so m_t is an MA(q), then the limit as the sample size goes to infinity is

$$\text{ACov}\left(\sqrt{T}\bar{m}\right) = \lim_{T\to\infty} \text{Cov}\left(\sqrt{T}\bar{m}\right) = \sum_{s=-q}^{q} R(s), \tag{3.5}$$

where ACov stands for the asymptotic variance-covariance matrix. This continues to hold even if $q = \infty$, provided $R(s)$ goes to zero sufficiently quickly, as it does in stationary VAR systems. In this case we have

$$\text{ACov}\left(\sqrt{T}\bar{m}\right) = \sum_{s=-\infty}^{\infty} R(s). \tag{3.6}$$

Estimation in finite samples will of course require some cut-off point, which is discussed below.

The traditional estimator of $\text{ACov}\left(\sqrt{T}\bar{m}\right)$ is just $R(0)$, which is correct when m_t has no autocorrelation, that is

$$\text{ACov}\left(\sqrt{T}\bar{m}\right) = R(0) = \text{Cov}(m_t, m_t) \text{ if } \text{Cov}(m_t, m_{t-s}) = 0 \text{ for } s \neq 0. \tag{3.7}$$

By comparing with (3.5) we see that this underestimates the true variance if autocovariances are mostly positive, and overestimates if they are mostly negative. The errors can be substantial.

Example 3.4 *(Variance of sample mean of AR(1).) Let $m_t = \rho m_{t-1} + u_t$, where $Var(u_t) = \sigma^2$. Note that $R(s) = \rho^{|s|}\sigma^2/\left(1-\rho^2\right)$, so*

$$\begin{aligned} AVar\left(\sqrt{T}\bar{m}\right) &= \sum_{s=-\infty}^{\infty} R(s) \\ &= \frac{\sigma^2}{1-\rho^2}\sum_{s=-\infty}^{\infty}\rho^{|s|} = \frac{\sigma^2}{1-\rho^2}\left(1+2\sum_{s=1}^{\infty}\rho^s\right) \\ &= \frac{\sigma^2}{1-\rho^2}\frac{1+\rho}{1-\rho}, \end{aligned}$$

which is increasing in ρ (provided $|\rho| < 1$, as required for stationarity). The variance of $\bar{m}$ is much larger for ρ close to one than for ρ close to zero: the high autocorrelation create long swings, so the mean cannot be estimated with any good precision in a small sample. If we disregard all autocovariances, then we would conclude that the variance of $\sqrt{T}\bar{m}$ is $\sigma^2/\left(1-\rho^2\right)$, which is smaller (larger) than the true value when $\rho > 0$ ($\rho < 0$). For instance, with $\rho = 0.85$, it is approximately 12 times too small. See Figure 3.1 for an illustration.

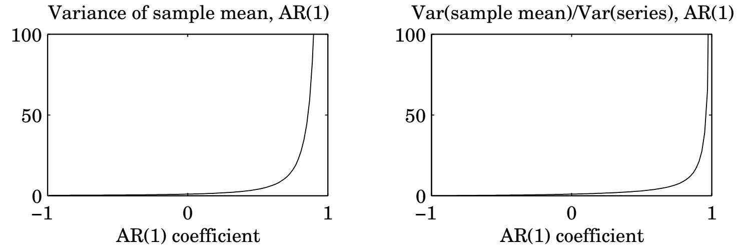


Figure 3.1: Variance of $\sqrt{T}$ times sample mean of AR(1) process $m_t = \rho m_{t-1} + u_t$.

Example 3.5 *(Variance of sample mean of AR(1), continued.) Part of the reason why $Var(\bar{m})$ increased with ρ in the previous examples is that $Var(m_t)$ increases with ρ. We can eliminate this effect by considering how much larger $AVar(\sqrt{T}\bar{m})$ is than in the iid case, that is, $AVar(\sqrt{T}\bar{m})/Var(m_t) = (1+\rho)/(1-\rho)$. This ratio is one for $\rho = 0$ (iid data), less than one for $\rho < 0$, and greater than one for $\rho > 0$. This says that if relatively more of the variance in m_t comes from long swings (high ρ), then the sample mean is more uncertain. See Figure 3.1 for an illustration.*

Example 3.6 *(Variance of sample mean of AR(1), illustration of why $\lim_{T\to\infty}$ of (3.4).) For an AR(1) (3.4) is*

$$\begin{aligned} Var\left(\sqrt{T}\bar{m}\right) &= \frac{\sigma^2}{1-\rho^2}\sum_{s=-(T-1)}^{T-1}\left(1-\frac{|s|}{T}\right)\rho^{|s|} \\ &= \frac{\sigma^2}{1-\rho^2}\left[1+2\sum_{s=1}^{T-1}\left(1-\frac{s}{T}\right)\rho^s\right] \\ &= \frac{\sigma^2}{1-\rho^2}\left[1+2\frac{\rho}{1-\rho}+2\frac{\rho^{T+1}-\rho}{T(1-\rho)^2}\right]. \end{aligned}$$

The last term in brackets goes to zero as T goes to infinity. We then get the result in Example 3.4.

3.2 The Newey-West Estimator

3.2.1 Definition of the Estimator

Newey and West (1987) suggested the following estimator of the covariance matrix in (3.5) as (for some $n < T$)

$$\begin{aligned}\widehat{\text{ACov}}\left(\sqrt{T}\bar{m}\right) &= \sum_{s=-n}^{n}\left(1-\frac{|s|}{n+1}\right)\hat{R}(s)\\ &= \hat{R}(0)+\sum_{s=1}^{n}\left(1-\frac{s}{n+1}\right)\left(\hat{R}(s)+\hat{R}(-s)\right)\text{, or since } \hat{R}(-s)=\hat{R}'(s),\\ &= \hat{R}(0)+\sum_{s=1}^{n}\left(1-\frac{s}{n+1}\right)\left(\hat{R}(s)+\hat{R}'(s)\right)\text{, where} \qquad (3.8)\end{aligned}$$

$$\hat{R}(s)=\frac{1}{T}\sum_{t=s+1}^{T}m_t m_{t-s}' \text{ (if } \mathrm{E}\, m_t = 0). \qquad (3.9)$$

The tent shaped (Bartlett) weights in (3.8) guarantee a positive definite covariance estimate. In contrast, equal weights (as in (3.5)), may give an estimated covariance matrix which is not positive definite, which is fairly awkward. Newey and West (1987) showed that this estimator is consistent if we let n go to infinity as T does, but in such a way that $n/T^{1/4}$ goes to zero.

There are several other possible estimators of the covariance matrix in (3.5), but simulation evidence suggest that they typically do not improve a lot on the Newey-West estimator.

Example 3.7 *(m_t is MA(1).) Suppose we know that $m_t = \varepsilon_t + \theta\varepsilon_{t-1}$. Then $R(s) = 0$ for $s \geq 2$, so it might be tempting to use $n = 1$ in (3.8). This gives $\widehat{ACov}\left(\sqrt{T}\bar{m}\right) = \hat{R}(0) + \frac{1}{2}[\hat{R}(1) + \hat{R}'(1)]$, while the theoretical expression (3.5) is ACov$= R(0) + R(1) + R'(1)$. The Newey-West estimator puts too low weights on the first lead and lag, which suggests that we should use $n > 1$ (or more generally, $n > q$ for an MA(q) process).*

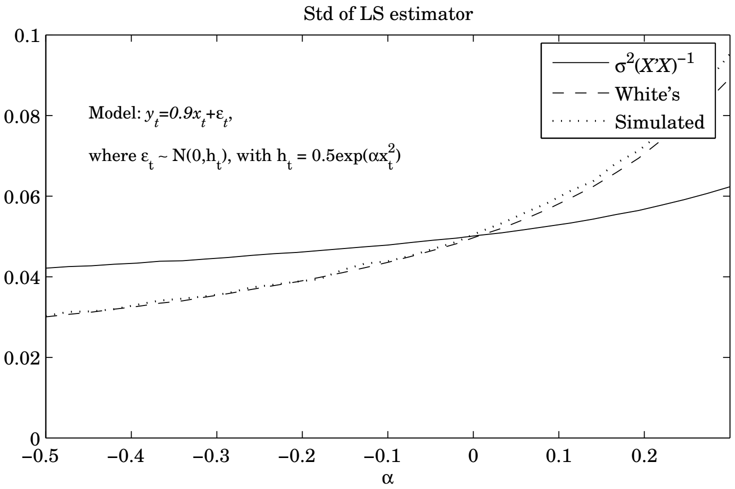


Figure 3.2: Variance of OLS estimator, heteroskedastic errors

It can also be shown that, under quite general circumstances, *$\hat{S}$ in (3.8)-(3.9) is a consistent estimator of ACov$\left(\sqrt{T}\bar{m}\right)$, even if m_t is heteroskedastic* (on top of being autocorrelated). (See Hamilton (1994) 10.5 for a discussion.)

3.2.2 How to Implement the Newey-West Estimator

Economic theory and/or stylized facts can sometimes help us choose the lag length n. For instance, we may have a model of stock returns which typically show little autocorrelation, so it may make sense to set $n = 0$ or $n = 1$ in that case. A popular choice of n is to round $(T/100)^{1/4}$ down to the closest integer, although this does not satisfy the consistency requirement.

It is important to note that definition of the covariance matrices in (3.2) and (3.9) assume that m_t has zero mean. If that is not the case, then the mean should be removed in the calculation of the covariance matrix. In practice, you remove the same number, estimated on the whole sample, from both m_t and m_{t-s}. It is often recommended to

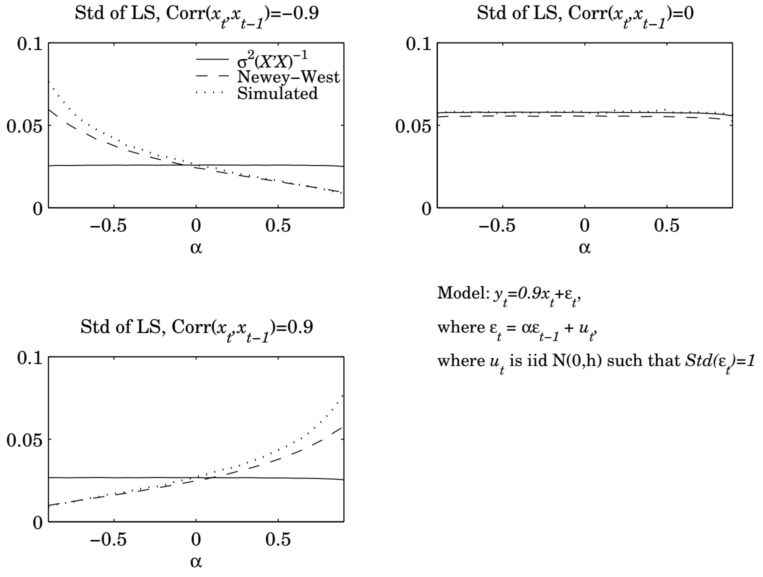


Figure 3.3: Variance of OLS estimator, autocorrelated errors

remove the sample means even if theory tells you that the true mean is zero.

3.3 Summary

$$\text{Let } \bar{m} = \frac{1}{T}\sum_{t=1}^{T} m_t \text{ and } R(s) = \text{Cov}(m_t, m_{t-s}). \text{ Then}$$

$$\text{ACov}\left(\sqrt{T}\bar{m}\right) = \sum_{s=-\infty}^{\infty} R(s)$$

$$\text{ACov}\left(\sqrt{T}\bar{m}\right) = R(0) = \text{Cov}(m_t, m_t) \text{ if } R(s) = \mathbf{0} \text{ for } s \neq 0$$

$$\text{Newey-West} : \widehat{\text{ACov}}\left(\sqrt{T}\bar{m}\right) = \hat{R}(0) + \sum_{s=1}^{n}\left(1 - \frac{s}{n+1}\right)\left(\hat{R}(s) + \hat{R}'(s)\right).$$

Bibliography

Cochrane, J. H., 2001, *Asset Pricing*, Princeton University Press, Princeton, New Jersey.

Hamilton, J. D., 1994, *Time Series Analysis*, Princeton University Press, Princeton.

Harris, D., and L. Matyas, 1999, “Introduction to the Generalized Method of Moments Estimation,” in Laszlo Matyas (ed.), *Generalized Method of Moments Estimation* . chap. 1, Cambridge University Press.

Hayashi, F., 2000, *Econometrics*, Princeton University Press.

Newey, W. K., and K. D. West, 1987, “A Simple Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix,” *Econometrica*, 55, 703–708.

Pindyck, R. S., and D. L. Rubinfeld, 1998, *Econometric Models and Economic Forecasts*, Irwin McGraw-Hill, Boston, Massachusetts, 4ed edn.

Verbeek, M., 2004, *A Guide to Modern Econometrics*, Wiley, Chichester, 2nd edn.

4 Least Squares

Reference: Greene (2000) 6
Additional references: Hayashi (2000) 1-2; Verbeek (2004) 1-4; Hamilton (1994) 8

4.1 Definition of the LS Estimator

4.1.1 LS with Summation Operators

Consider the linear model

$$y_t = x_t'\beta_0 + u_t, \tag{4.1}$$

where y_t and u_t are scalars, x_t a $k \times 1$ vector, and β_0 is a $k \times 1$ vector of the true coefficients. Least squares minimizes the sum of the squared fitted residuals

$$\sum_{t=1}^{T} e_t^2 = \sum_{t=1}^{T} \left(y_t - x_t'\beta\right)^2, \tag{4.2}$$

by choosing the vector β. The first order conditions are

$$\mathbf{0}_{kx1} = \sum_{t=1}^{T} x_t \left(y_t - x_t'\hat{\beta}_{LS}\right) \text{ or} \tag{4.3}$$

$$\sum_{t=1}^{T} x_t y_t = \sum_{t=1}^{T} x_t x_t'\hat{\beta}_{LS}, \tag{4.4}$$

which are the so called normal equations. These can be solved as

$$\hat{\beta}_{LS} = \left(\sum_{t=1}^{T} x_t x_t'\right)^{-1} \sum_{t=1}^{T} x_t y_t \tag{4.5}$$

$$= \left(\frac{1}{T}\sum_{t=1}^{T} x_t x_t'\right)^{-1} \frac{1}{T}\sum_{t=1}^{T} x_t y_t \tag{4.6}$$

Remark 4.1 *(Summation and vectors) Let z_t and x_t be the vectors*

$$z_t = \begin{bmatrix} z_{1t} \\ z_{2t} \end{bmatrix} \text{ and } x_t = \begin{bmatrix} x_{1t} \\ x_{2t} \\ x_{3t} \end{bmatrix},$$

then

$$\sum_{t=1}^{T} x_t z_t' = \sum_{t=1}^{T} \begin{bmatrix} x_{1t} \\ x_{2t} \\ x_{3t} \end{bmatrix} \begin{bmatrix} z_{1t} & z_{2t} \end{bmatrix} = \sum_{t=1}^{T} \begin{bmatrix} x_{1t}z_{1t} & x_{1t}z_{2t} \\ x_{2t}z_{1t} & x_{2t}z_{2t} \\ x_{3t}z_{1t} & x_{3t}z_{2t} \end{bmatrix} = \begin{bmatrix} \sum_{t=1}^{T} x_{1t}z_{1t} & \sum_{t=1}^{T} x_{1t}z_{2t} \\ \sum_{t=1}^{T} x_{2t}z_{1t} & \sum_{t=1}^{T} x_{2t}z_{2t} \\ \sum_{t=1}^{T} x_{3t}z_{1t} & \sum_{t=1}^{T} x_{3t}z_{2t} \end{bmatrix}.$$

4.1.2 LS in Matrix Form

Define the matrices

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{bmatrix}_{T\times 1}, \; u = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_T \end{bmatrix}_{T\times 1}, \; X = \begin{bmatrix} x_1' \\ x_2' \\ \vdots \\ x_T' \end{bmatrix}_{T\times k}, \text{ and } e = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_T \end{bmatrix}_{T\times 1}. \tag{4.7}$$

Write the model (4.1) as

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{bmatrix} = \begin{bmatrix} x_1' \\ x_2' \\ \vdots \\ x_T' \end{bmatrix} \beta_0 + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_T \end{bmatrix} \text{ or} \tag{4.8}$$

$$Y = X\beta_0 + u. \tag{4.9}$$

Remark 4.2 *Let x_t be a $k \times 1$ and z_t an $m \times 1$ vector. Define the matrices*

$$X = \begin{bmatrix} x_1' \\ x_2' \\ \vdots \\ x_T' \end{bmatrix}_{T\times k} \text{ and } Z = \begin{bmatrix} z_1' \\ z_2' \\ \vdots \\ z_T' \end{bmatrix}_{T\times m}.$$

We then have

$$\sum_{t=1}^{T} x_t z_t' = X'Z.$$

We can then rewrite the loss function (4.2) as $e'e$, the first order conditions (4.3) and (4.4) as (recall that $y_t = y_t'$ since it is a scalar)

$$\mathbf{0}_{kx1} = X'\left(Y - X\hat{\beta}_{LS}\right) \tag{4.10}$$

$$X'Y = X'X\hat{\beta}_{LS}, \tag{4.11}$$

and the solution (4.5) as

$$\hat{\beta}_{LS} = \left(X'X\right)^{-1} X'Y. \tag{4.12}$$

4.2 LS and R^2 *

The first order conditions in LS are

$$\sum_{t=1}^{T} x_t \hat{u}_t = \mathbf{0}, \text{ where } \hat{u}_t = y_t - \hat{y}_t, \text{ with } \hat{y}_t = x_t'\hat{\beta}. \tag{4.13}$$

This implies that the fitted residuals and fitted values are orthogonal, $\Sigma_{t=1}^{T} \hat{y}_t \hat{u}_t = \Sigma_{t=1}^{T} \hat{\beta}' x_t \hat{u}_t = 0$. If we let x_t include a constant, then (4.13) also implies that the fitted residuals have a zero mean, $\Sigma_{t=1}^{T} \hat{u}_t / T = 0$. We can then decompose the sample variance (denoted $\widehat{\text{Var}}$) of $y_t = \hat{y}_t + \hat{u}_t$ as

$$\widehat{\text{Var}}\left(y_t\right) = \widehat{\text{Var}}\left(\hat{y}_t\right) + \widehat{\text{Var}}\left(\hat{u}_t\right), \tag{4.14}$$

since $\hat{y}_t$ and $\hat{u}_t$ are uncorrelated in this case. (Note that $\text{Cov}(\hat{y}_t, \hat{u}_t) = \text{E}\hat{y}_t\hat{u}_t - \text{E}\hat{y}_t\text{E}\hat{u}_t$ so the orthogonality is not enough to allow the decomposition; we also need $\text{E}\hat{y}_t\text{E}\hat{u}_t = 0$—this holds for sample moments as well.)

We define R^2 as the fraction of $\widehat{\text{Var}}\left(y_t\right)$ that is explained by the model

$$R^2 = \frac{\widehat{\text{Var}}\left(\hat{y}_t\right)}{\widehat{\text{Var}}\left(y_t\right)} \tag{4.15}$$

$$= 1 - \frac{\widehat{\text{Var}}\left(\hat{u}_t\right)}{\widehat{\text{Var}}\left(y_t\right)}. \tag{4.16}$$

LS minimizes the sum of squared fitted errors, which is proportional to $\widehat{\text{Var}}\left(\hat{u}_t\right)$, so it maximizes R^2.

We can rewrite R^2 by noting that

$$\widehat{\text{Cov}}\left(y_t, \hat{y}_t\right) = \widehat{\text{Cov}}\left(\hat{y}_t + \hat{u}_t, \hat{y}_t\right) = \widehat{\text{Var}}\left(\hat{y}_t\right). \tag{4.17}$$

Use this to substitute for $\widehat{\text{Var}}\left(\hat{y}_t\right)$ in (4.15) and then multiply both sides with $\widehat{\text{Cov}}\left(y_t, \hat{y}_t\right) / \widehat{\text{Var}}\left(\hat{y}_t\right) = 1$ to get

$$R^2 = \frac{\widehat{\text{Cov}}\left(y_t, \hat{y}_t\right)^2}{\widehat{\text{Var}}\left(y_t\right)\widehat{\text{Var}}\left(\hat{y}_t\right)}$$

$$= \widehat{\text{Corr}}\left(y_t, \hat{y}_t\right)^2 \tag{4.18}$$

which shows that R^2 is the square of correlation coefficient of the actual and fitted value. Note that this interpretation of R^2 relies on the fact that $\widehat{\text{Cov}}\left(\hat{y}_t, \hat{u}_t\right) = 0$. From (4.14) this implies that the sample variance of the fitted variables is smaller than the sample variance of y_t. From (4.15) we see that this implies that $0 \leq R^2 \leq 1$.

To get a bit more intuition for what R^2 represents, suppose the estimated coefficients equal the true coefficients, so $\hat{y}_t = x_t'\beta_0$. In this case, $R^2 = \text{Corr}\left(x_t'\beta_0 + u_t, x_t'\beta_0\right)^2$, that is, the squared correlation of y_t with the systematic part of y_t. Clearly, if the model is perfect so $u_t = 0$, then $R^2 = 1$. On contrast, when there is no movements in the systematic part ($\beta_0 = \mathbf{0}$), then $R^2 = 0$.

Remark 4.3 *In a simple regression where $y_t = a + bx_t + u_t$, where x_t is a scalar, $R^2 = \widehat{Corr}\left(y_t, x_t\right)^2$. To see this, note that, in this case (4.18) can be written*

$$R^2 = \frac{\widehat{Cov}\left(y_t, \hat{b}x_t\right)^2}{\widehat{Var}\left(y_t\right)\widehat{Var}\left(\hat{b}x_t\right)} = \frac{\hat{b}^2\widehat{Cov}\left(y_t, x_t\right)^2}{\hat{b}^2\widehat{Var}\left(y_t\right)\widehat{Var}\left(x_t\right)},$$

so the $\hat{b}^2$ terms cancel.

Remark 4.4 *Now, consider the reverse regression $x_t = c + dy_t + v_t$. The LS estimator of the slope is $\hat{d}_{LS} = \widehat{Cov}\left(y_t, x_t\right) / \widehat{Var}\left(y_t\right)$. Recall that $\hat{b}_{LS} = \widehat{Cov}\left(y_t, x_t\right) / \widehat{Var}\left(x_t\right)$. We therefore have*

$$\hat{b}_{LS}\hat{d}_{LS} = \frac{\widehat{Cov}\left(y_t, x_t\right)^2}{\widehat{Var}\left(y_t\right)\widehat{Var}\left(x_t\right)} = R^2.$$

This shows that $\hat{d}_{LS} = 1/\hat{b}_{LS}$ if (and only if) $R^2 = 1$.

4.3 Finite Sample Properties of LS

Use the true model (4.1) to substitute for y_t in the definition of the LS estimator (4.6)

$$\begin{aligned}\hat{\beta}_{LS} &= \left(\frac{1}{T}\sum_{t=1}^{T} x_t x_t'\right)^{-1} \frac{1}{T}\sum_{t=1}^{T} x_t \left(x_t'\beta_0 + u_t\right)\\ &= \beta_0 + \left(\frac{1}{T}\sum_{t=1}^{T} x_t x_t'\right)^{-1} \frac{1}{T}\sum_{t=1}^{T} x_t u_t. \qquad (4.19)\end{aligned}$$

It is possible to show *unbiasedness of the LS estimator*, even if x_t stochastic and u_t is autocorrelated and heteroskedastic—provided $\mathrm{E}(u_t|x_{t-s}) = 0$ for all s. Let $\mathrm{E}\left(u_t|\{x_t\}_{t=1}^{T}\right)$ denote the expectation of u_t conditional on all values of x_{t-s}. Using iterated expectations on (4.19) then gives

$$\begin{aligned}\mathrm{E}\hat{\beta}_{LS} &= \beta_0 + \mathrm{E}_x\left[\left(\frac{1}{T}\sum_{t=1}^{T} x_t x_t'\right)^{-1} \frac{1}{T}\sum_{t=1}^{T} x_t \mathrm{E}\left(u_t|\{x_t\}_{t=1}^{T}\right)\right] \qquad (4.20)\\ &= \beta_0, \qquad (4.21)\end{aligned}$$

since $\mathrm{E}(u_t|x_{t-s}) = 0$ for all s. This is, for instance, the case when the regressors are deterministic. Notice that $\mathrm{E}(u_t|\, x_t) = 0$ is not enough for unbiasedness since (4.19) contains terms involving $x_{t-s}x_t u_t$ from the product of $(\frac{1}{T}\sum_{t=1}^{T} x_t x_t')^{-1}$ and $x_t u_t$.

Example 4.5 *(AR(1).) Consider estimating α in $y_t = \alpha y_{t-1} + u_t$. The LS estimator is*

$$\begin{aligned}\hat{\alpha}_{LS} &= \left(\frac{1}{T}\sum_{t=1}^{T} y_{t-1}^2\right)^{-1} \frac{1}{T}\sum_{t=1}^{T} y_{t-1} y_t\\ &= \alpha + \left(\frac{1}{T}\sum_{t=1}^{T} y_{t-1}^2\right)^{-1} \frac{1}{T}\sum_{t=1}^{T} y_{t-1} u_t.\end{aligned}$$

In this case, the assumption $E(u_t|x_{t-s}) = 0$ for all s (that is, $s = ..., -1, 0, 1, ...$) is false, since $x_{t+1} = y_t$ and u_t and y_t are correlated. We can therefore not use this way of proving that $\hat{\alpha}_{LS}$ is unbiased. In fact, it is not, and it can be shown that $\hat{\alpha}_{LS}$ is downward-biased if $\alpha > 0$, and that this bias gets quite severe as α gets close to unity.

The finite sample distribution of the LS estimator is typically unknown.

Even in the most restrictive case where u_t is iid $N\left(0, \sigma^2\right)$ and $\mathrm{E}(u_t|x_{t-s}) = 0$ for all s, we can only get that

$$\hat{\beta}_{LS}|\,\{x_t\}_{t=1}^{T} \sim N\left[\beta_0, \sigma^2\left(\frac{1}{T}\sum_{t=1}^{T} x_t x_t'\right)^{-1}\right]. \qquad (4.22)$$

This says that the estimator, *conditional* on the sample of regressors, is normally distributed. With deterministic x_t, this clearly means that $\hat{\beta}_{LS}$ is normally distributed in a small sample. The intuition is that the LS estimator with deterministic regressors is just a linear combination of the normally distributed y_t, so it must be normally distributed. However, if x_t is stochastic, then we have to take into account the distribution of $\{x_t\}_{t=1}^{T}$ to find the *unconditional* distribution of $\hat{\beta}_{LS}$. The principle is that

$$\mathrm{pdf}\left(\hat{\beta}\right) = \int_{-\infty}^{\infty} \mathrm{pdf}\left(\hat{\beta}, x\right) dx = \int_{-\infty}^{\infty} \mathrm{pdf}\left(\hat{\beta}\,|x\right) \mathrm{pdf}\,(x)\, dx,$$

so the distribution in (4.22) must be multiplied with the probability density function of $\{x_t\}_{t=1}^{T}$ and then integrated over $\{x_t\}_{t=1}^{T}$ to give the unconditional distribution (marginal) of $\hat{\beta}_{LS}$. This is typically not a normal distribution.

Another way to see the same problem is to note that $\hat{\beta}_{LS}$ in (4.19) is a product of two random variables, $(\Sigma_{t=1}^{T} x_t x_t'/T)^{-1}$ and $\Sigma_{t=1}^{T} x_t u_t/T$. Even if u_t happened to be normally distributed, there is no particular reason why $x_t u_t$ should be, and certainly no strong reason for why $(\Sigma_{t=1}^{T} x_t x_t'/T)^{-1}\Sigma_{t=1}^{T} x_t u_t/T$ should be.

4.4 Consistency of LS

Reference: Greene (2000) 9.3-5 and 11.2; Hamilton (1994) 8.2; Davidson (2000) 3

We now study if the LS estimator is consistent.

Remark 4.6 *Suppose the true parameter value is β_0. The estimator $\hat{\beta}_T$ (which, of course, depends on the sample size T) is said to be consistent if for every $\varepsilon > 0$ and $\delta > 0$ there exists N such that for $T \geq N$*

$$\Pr\left(\left\|\hat{\beta}_T - \beta_0\right\| > \delta\right) < \varepsilon.$$

($\|x\| = \sqrt{x'x}$, the Euclidean distance of x from zero.) We write this $\operatorname{plim}\hat{\beta}_T = \beta_0$ *or just* $\operatorname{plim}\hat{\beta} = \beta_0$*, or perhaps* $\hat{\beta} \to^p \beta_0$*. (For an estimator of a covariance matrix, the most convenient is to stack the unique elements in a vector and then apply the definition above.)*

Remark 4.7 *(Slutsky's theorem.) If $g(.)$ is a continuous function, then* $\operatorname{plim} g(z_T) = g(\operatorname{plim} z_T)$*. In contrast, note that $Eg(z_T)$ is generally not equal to $g(Ez_T)$, unless $g(.)$ is a linear function.*

Remark 4.8 *(Probability limit of product.) Let x_T and y_T be two functions of a sample of length T. If* $\operatorname{plim} x_T = a$ *and* $\operatorname{plim} y_T = b$*, then* $\operatorname{plim} x_T y_T = ab$*.*

Assume

$$\operatorname{plim} \frac{1}{T}\sum_{t=1}^{T} x_t x_t' = \Sigma_{xx} < \infty, \text{ and } \Sigma_{xx} \text{ invertible.} \tag{4.23}$$

The plim carries over to the inverse by Slutsky's theorem.[1] Use the facts above to write the probability limit of (4.19) as

$$\operatorname{plim}\hat{\beta}_{LS} = \beta_0 + \Sigma_{xx}^{-1} \operatorname{plim} \frac{1}{T}\sum_{t=1}^{T} x_t u_t. \tag{4.24}$$

To prove consistency of $\hat{\beta}_{LS}$ we therefore have to show that

$$\operatorname{plim} \frac{1}{T}\sum_{t=1}^{T} x_t u_t = \mathrm{E} x_t u_t = \mathrm{Cov}(x_t, u_t) = 0. \tag{4.25}$$

This is fairly easy to establish in special cases, for instance, when $w_t = x_t u_t$ is iid or when there is either heteroskedasticity or serial correlation. The case with both serial correlation and heteroskedasticity is just a bit more complicated. In other cases, it is clear that the covariance the residuals and the regressors are not all zero—for instance when some of the regressors are measured with error or when some of them are endogenous variables.

An example of a case where LS is not consistent is when the errors are autocorrelated *and* the regressors include lags of the dependent variable. For instance, suppose the error is a MA(1) process

$$u_t = \varepsilon_t + \theta_1 \varepsilon_{t-1}, \tag{4.26}$$

where ε_t is white noise and that the regression equation is an AR(1)

$$y_t = \rho y_{t-1} + u_t. \tag{4.27}$$

This is an ARMA(1,1) model and it is clear that the regressor and error in (4.27) are correlated, so LS is not a consistent estimator of an ARMA(1,1) model.

[1] This puts non-trivial restrictions on the data generating processes. For instance, if x_t include lagged values of y_t, then we typically require y_t to be stationary and ergodic, and that u_t is independent of x_{t-s} for $s \geq 0$.

4.5 Asymptotic Normality of LS

Reference: Greene (2000) 9.3-5 and 11.2; Hamilton (1994) 8.2; Davidson (2000) 3

Remark 4.9 *(Continuous mapping theorem.) Let the sequences of random matrices $\{x_T\}$ and $\{y_T\}$, and the non-random matrix $\{a_T\}$ be such that $x_T \overset{d}{\to} x$, $y_T \overset{p}{\to} y$, and $a_T \to a$ (a traditional limit). Let $g(x_T, y_T, a_T)$ be a continuous function. Then $g(x_T, y_T, a_T) \overset{d}{\to} g(x, y, a)$. Either of y_T and a_T could be irrelevant in g.*

Remark 4.10 *From the previous remark: if $x_T \overset{d}{\to} x$ (a random variable) and* $\operatorname{plim} Q_T = Q$ *(a constant matrix), then $Q_T x_T \overset{d}{\to} Qx$.*

Premultiply (4.19) by $\sqrt{T}$ and rearrange as

$$\sqrt{T}\left(\hat{\beta}_{LS} - \beta_0\right) = \left(\frac{1}{T}\sum_{t=1}^{T} x_t x_t'\right)^{-1} \frac{\sqrt{T}}{T}\sum_{t=1}^{T} x_t u_t. \tag{4.28}$$

If the first term on the right hand side converges in probability to a finite matrix (as assumed in (4.23)), and the vector of random variables $x_t u_t$ satisfies a central limit theorem, then

$$\sqrt{T}(\hat{\beta}_{LS} - \beta_0) \overset{d}{\to} N\left(0, \Sigma_{xx}^{-1} S_0 \Sigma_{xx}^{-1}\right), \text{ where} \tag{4.29}$$
$$\Sigma_{xx} = \operatorname{plim} \frac{1}{T}\sum_{t=1}^{T} x_t x_t' \text{ and } S_0 = \mathrm{Cov}\left(\frac{\sqrt{T}}{T}\sum_{t=1}^{T} x_t u_t\right).$$

The last matrix in the covariance matrix does not need to be transposed since it is symmetric (since Σ_{xx} is). This general expression is valid for both autocorrelated and heteroskedastic residuals—all such features are loaded into the S_0 matrix. Note that S_0 is the variance-covariance matrix of $\sqrt{T}$ times a sample average (of the vector of random variables $x_t u_t$), which can be complicated to specify and to estimate. In simple cases, we can derive what it is. To do so, we typically need to understand the properties of the residuals. Are they autocorrelated and/or heteroskedastic? In other cases we will have to use some kind of "non-parametric" approach to estimate it.

A common approach is to estimate Σ_{xx} by $\Sigma_{t=1}^{T} x_t x_t' / T$ and use the Newey-West estimator of S_0.

4.5.1 Special Case: Classical LS assumptions

Reference: Greene (2000) 9.4 or Hamilton (1994) 8.2.

We can recover the classical expression for the covariance, $\sigma^2 \Sigma_{xx}^{-1}$, if we assume that the regressors are stochastic, but require that x_t is independent of all u_{t+s} and that u_t is iid. It rules out, for instance, that u_t and x_{t-2} are correlated and also that the variance of u_t depends on x_t. Expand the expression for S_0 as Expand the expression for S_0 as

$$\begin{aligned} S_0 &= \mathrm{E}\left(\frac{\sqrt{T}}{T}\sum_{t=1}^{T} x_t u_t\right)\left(\frac{\sqrt{T}}{T}\sum_{t=1}^{T} u_t x_t'\right) \qquad (4.30) \\ &= \frac{1}{T}\mathrm{E}\left(... + x_{s-1}u_{s-1} + x_s u_s + ...\right)\left(... + u_{s-1}x_{s-1}' + u_s x_s' + ...\right). \end{aligned}$$

Note that

$$\begin{aligned} \mathrm{E}x_{t-s}u_{t-s}u_t x_t' &= \mathrm{E}x_{t-s}x_t' \mathrm{E}u_{t-s}u_t \text{ (since } u_t \text{ and } x_{t-s} \text{ independent)} \\ &= \begin{cases} 0 \text{ if } s \neq 0 \text{ (since } \mathrm{E}u_{t-s}u_t = 0 \text{ by iid } u_t) \\ \mathrm{E}x_t x_t' \mathrm{E}u_t u_t \text{ else.} \end{cases} \qquad (4.31) \end{aligned}$$

This means that all cross terms (involving different observations) drop out and that we can write

$$\begin{aligned} S_0 &= \frac{1}{T}\sum_{t=1}^{T} \mathrm{E}x_t x_t' \mathrm{E}u_t^2 \qquad (4.32) \\ &= \sigma^2 \frac{1}{T}\mathrm{E}\sum_{t=1}^{T} x_t x_t' \text{ (since } u_t \text{ is iid and } \sigma^2 = \mathrm{E}u_t^2) \qquad (4.33) \\ &= \sigma^2 \Sigma_{xx}. \qquad (4.34) \end{aligned}$$

Using this in (4.29) gives

$$\text{Asymptotic Cov}[\sqrt{T}(\hat{\beta}_{LS} - \beta_0)] = \Sigma_{xx}^{-1} S_0 \Sigma_{xx}^{-1} = \Sigma_{xx}^{-1}\sigma^2 \Sigma_{xx}\Sigma_{xx}^{-1} = \sigma^2 \Sigma_{xx}^{-1}.$$

4.5.2 Special Case: White's Heteroskedasticity

Reference: Greene (2000) 12.2 and Davidson and MacKinnon (1993) 16.2.

This section shows that the classical LS formula for the covariance matrix is valid even if the errors are heteroskedastic—provided the heteroskedasticity is independent of the regressors.

The only difference compared with the classical LS assumptions is that u_t is now allowed to be heteroskedastic, but this heteroskedasticity is not allowed to depend on the moments of x_t. This means that (4.32) holds, but (4.33) does not since $\mathrm{E}u_t^2$ is not the same for all t.

However, we can still simplify (4.32) a bit more. We assumed that $\mathrm{E}x_t x_t'$ and $\mathrm{E}u_t^2$ (which can both be time varying) are not related to each other, so we could perhaps multiply $\mathrm{E}x_t x_t'$ by $\Sigma_{t=1}^{T}\mathrm{E}u_t^2/T$ instead of by $\mathrm{E}u_t^2$. This is indeed true asymptotically—where any possible "small sample" relation between $\mathrm{E}x_t x_t'$ and $\mathrm{E}u_t^2$ must wash out due to the assumptions of independence (which are about population moments).

In large samples we therefore have

$$\begin{aligned} S_0 &= \left(\frac{1}{T}\sum_{t=1}^{T}\mathrm{E}u_t^2\right)\left(\frac{1}{T}\sum_{t=1}^{T}\mathrm{E}x_t x_t'\right) \\ &= \left(\frac{1}{T}\sum_{t=1}^{T}\mathrm{E}u_t^2\right)\left(\mathrm{E}\frac{1}{T}\sum_{t=1}^{T} x_t x_t'\right) \\ &= \omega^2 \Sigma_{xx}, \qquad (4.35) \end{aligned}$$

where ω^2 is a scalar. This is very similar to the classical LS case, except that ω^2 is the average variance of the residual rather than the constant variance. In practice, the estimator of ω^2 is the same as the estimator of σ^2, so we can actually apply the standard LS formulas in this case.

This is the motivation for why White's test for heteroskedasticity makes sense: if the heteroskedasticity is not correlated with the regressors, then the standard LS formula is correct (provided there is no autocorrelation).

4.6 Inference

Consider some estimator, $\hat{\beta}_{k\times 1}$, with an asymptotic normal distribution

$$\sqrt{T}(\hat{\beta}-\beta_0) \overset{d}{\rightarrow} N\left(0, V\right). \tag{4.36}$$

Suppose we want to test the null hypothesis that the s linear restrictions $R\beta_0 = r$ hold, where R is an $s \times k$ matrix and r is an $s \times 1$ vector. If the null hypothesis is true, then

$$\sqrt{T}(R\hat{\beta}-r) \overset{d}{\rightarrow} N(0, RVR'), \tag{4.37}$$

since the s linear combinations are linear combinations of random variables with an asymptotic normal distribution as in (4.37).

Remark 4.11 *If the $n \times 1$ vector $x \sim N(0, \Sigma)$, then $x'\Sigma^{-1}x \sim \chi^2_n$.*

Remark 4.12 *From the previous remark and Remark (4.9), it follows that if the $n \times 1$ vector $x \overset{d}{\rightarrow} N(0, \Sigma)$, then $x'\Sigma^{-1}x \overset{d}{\rightarrow} \chi^2_n$.*

From this remark, it follows that if the null hypothesis, $R\beta_0 = r$, is true, then Wald test statistics converges in distribution to a χ^2_s variable

$$T(R\hat{\beta}-r)'\left(RVR'\right)^{-1}(R\hat{\beta}-r) \overset{d}{\rightarrow} \chi^2_s. \tag{4.38}$$

Values of the test statistics above the $x\%$ critical value of the χ^2_s distribution mean that we reject the null hypothesis at the $x\%$ significance level.

When there is only one restriction ($s = 1$), then $\sqrt{T}(R\hat{\beta}-r)$ is a scalar, so the test can equally well be based on the fact that

$$\frac{\sqrt{T}(R\hat{\beta}-r)}{RVR'} \overset{d}{\rightarrow} N(0, 1).$$

In this case, we should reject the null hypothesis if the test statistics is either very low (negative) or very high (positive). In particular, let $\Phi()$ be the standard normal cumulative distribution function. We then reject the null hypothesis at the $x\%$ significance level if the test statistics is below x_L such that $\Phi(x_L) = (x/2)\%$ or above x_H such that $\Phi(x_H) = 1-(x/2)\%$ (that is with $(x/2)\%$ of the probability mass in each tail).

Example 4.13 *($TR^2/(1-R^2)$ as a test of the regression.) Recall from (4.15)-(4.16) that $R^2 = \widehat{Var}\left(\hat{y}_t\right)/\widehat{Var}\left(y_t\right) = 1-\widehat{Var}\left(\hat{u}_t\right)/\widehat{Var}\left(y_t\right)$, where $\hat{y}_t$ and $\hat{u}_t$ are the fitted value and residual respectively. We therefore get*

$$TR^2/(1-R^2) = T\widehat{Var}\left(\hat{y}_t\right)/\widehat{Var}\left(\hat{u}_t\right).$$

To simplify the algebra, assume that both y_t and x_t are demeaned and that no intercept is used. (We get the same results, but after more work, if we relax this assumption.) In this case, $\hat{y}_t = x_t'\hat{\beta}$, so we can rewrite the previous eqiuation as

$$TR^2/(1-R^2) = T\hat{\beta}'\Sigma_{xx}\hat{\beta}'/\widehat{Var}\left(\hat{u}_t\right).$$

This is identical to (4.38) when $R = I_k$ and $r = \mathbf{0}_{k\times 1}$ and the classical LS assumptions are fulfilled (so $V = \mathrm{Var}\left(\hat{u}_t\right)\Sigma_{xx}^{-1}$). The $TR^2/(1-R^2)$ is therefore a χ^2_k distributed statistics for testing if all the slope coefficients are zero.

Example 4.14 *(F version of the test.) There is also an $F_{k,T-k}$ version of the test in the previous example: $[R^2/k]/[(1-R^2)/(T-k)]$. Note that k times an $F_{k,T-k}$ variable converges to a χ^2_k variable as $T-k \rightarrow \infty$. This means that the χ^2_k form in the previous example can be seen as an asymptotic version of the (more common) F form.*

4.6.1 Tests of Non-Linear Restrictions*

To test non-linear restrictions, we can use the delta method which gives the asymptotic distribution of a function of a random variable.

Fact 4.15 ***Remark 4.16*** *(Delta method) Consider an estimator $\hat{\beta}_{k\times 1}$ which satisfies*

$$\sqrt{T}\left(\hat{\beta}-\beta_0\right) \overset{d}{\rightarrow} N\left(0, \Omega\right),$$

and suppose we want the asymptotic distribution of a transformation of β

$$\gamma_{q\times 1} = g\left(\beta\right),$$

where g (.) is has continuous first derivatives. The result is

$$\sqrt{T}\left[g\left(\hat{\beta}\right)-g\left(\beta_0\right)\right]\overset{d}{\to}N\left(0,\Psi_{q\times q}\right),\ \textit{where}$$
$$\Psi=\frac{\partial g\left(\beta_0\right)}{\partial\beta'}\Omega\frac{\partial g\left(\beta_0\right)'}{\partial\beta},\ \textit{where}\ \frac{\partial g\left(\beta_0\right)}{\partial\beta'}\ \textit{is}\ q\times k.$$

Proof. By the mean value theorem we have

$$g\left(\hat{\beta}\right)=g\left(\beta_0\right)+\frac{\partial g\left(\beta^*\right)}{\partial\beta'}\left(\hat{\beta}-\beta_0\right),$$

where

$$\frac{\partial g\left(\beta\right)}{\partial\beta'}=\begin{bmatrix}\frac{\partial g_1(\beta)}{\partial\beta_1} & \cdots & \frac{\partial g_1(\beta)}{\partial\beta_k}\\ \vdots & \ddots & \vdots\\ \frac{\partial g_q(\beta)}{\partial\beta_1} & \cdots & \frac{\partial g_q(\beta)}{\partial\beta_k}\end{bmatrix}_{q\times k},$$

and we evaluate it at β^* which is (weakly) between $\hat{\beta}$ and β_0. Premultiply by $\sqrt{T}$ and rearrange as

$$\sqrt{T}\left[g\left(\hat{\beta}\right)-g\left(\beta_0\right)\right]=\frac{\partial g\left(\beta^*\right)}{\partial\beta'}\sqrt{T}\left(\hat{\beta}-\beta_0\right).$$

If $\hat{\beta}$ is consistent (plim $\hat{\beta}=\beta_0$) and $\partial g\left(\beta^*\right)/\partial\beta'$ is continuous, then by Slutsky's theorem plim $\partial g\left(\beta^*\right)/\partial\beta'=\partial g\left(\beta_0\right)/\partial\beta'$, which is a constant. The result then follows from the continuous mapping theorem. ■

4.6.2 On F Tests*

F tests are sometimes used instead of chi–square tests. However, F tests rely on very special assumptions and typically converge to chi–square tests as the sample size increases. There are therefore few compelling theoretical reasons for why we should use F tests.[2] This section demonstrates that point.

Remark 4.17 *If $Y_1\sim\chi^2_{n_1}$, $Y_2\sim\chi^2_{n_2}$, and if Y_1 and Y_2 are independent, then $Z=(Y_1/n_1)/(Y_1/n_1)\sim F_{n_1,n_2}$. As $n_2\to\infty$, $n_1Z\overset{d}{\to}\chi^2_{n_1}$ (essentially because the denominator in Z is then equal to its expected value).*

[2] However, some simulation evidence suggests that F tests may have better small sample properties than chi-square test.

To use the F test to test s linear restrictions $R\beta_0=r$, we need to assume that the small sample distribution of the estimator is normal, $\sqrt{T}(\hat{\beta}-\beta_0)\sim N(0,\sigma^2W)$, where σ^2 is a scalar and W a known matrix. This would follow from an assumption that the residuals are normally distributed and that we either consider the distribution conditional on the regressors or that the regressors are deterministic. In this case $W=\Sigma_{xx}^{-1}$.

Consider the test statistics

$$F=T(R\hat{\beta}-r)'\left(R\hat{\sigma}^2WR'\right)^{-1}(R\hat{\beta}-r)/s.$$

This is similar to (4.38), expect that we use the estimated covariance matrix $\hat{\sigma}^2W$ instead of the true σ^2W (recall, W is assumed to be known) and that we have divided by the number of restrictions, s. Multiply and divide this expressions by σ^2

$$F=\frac{T(R\hat{\beta}-r)'\left(R\sigma^2WR'\right)^{-1}(R\hat{\beta}-r)/s}{\hat{\sigma}^2/\sigma^2}.$$

The numerator is an χ^2_s variable divided by its degrees of freedom, s. The denominator can be written $\hat{\sigma}^2/\sigma^2=\Sigma(\hat{u}_t/\sigma)^2/T$, where $\hat{u}_t$ are the fitted residuals. Since we just assumed that $u_t$are iid $N(0,\sigma^2)$, the denominator is an χ^2_T variable divided by its degrees of freedom, T. It can also be shown that the numerator and denominator are independent (essentially because the fitted residuals are orthogonal to the regressors), so F is an $F_{s,T}$ variable.

We need indeed very strong assumptions to justify the F distributions. Moreover, as $T\to\infty$, $sF\overset{d}{\to}\chi^2_n$ which is the Wald test—which do not need all these assumptions.

4.7 Diagnostic Tests of Autocorrelation, Heteroskedasticity, and Normality*

Reference: Greene (2000) 12.3, 13.5 and 9.7; Johnston and DiNardo (1997) 6; and Pindyck and Rubinfeld (1998) 6, Patterson (2000) 5

LS and IV are still consistent even if the residuals are autocorrelated, heteroskedastic, and/or non-normal, but the traditional expression for the variance of the parameter estimators is invalid. It is therefore important to investigate the properties of the residuals.

We would like to test the properties of the true residuals, u_t, but these are unobservable. We can instead use residuals from a consistent estimator as approximations, since the approximation error then goes to zero as the sample size increases. The residuals from an estimator are

$$\begin{aligned}\hat{u}_t &= y_t - x_t'\hat{\beta} \\ &= x_t'\left(\beta_0 - \hat{\beta}\right) + u_t. \qquad (4.39)\end{aligned}$$

If plim $\hat{\beta} = \beta_0$, then $\hat{u}_t$ converges in probability to the true residual ("pointwise consistency"). It therefore makes sense to use $\hat{u}_t$ to study the (approximate) properties of u_t. We want to understand if u_t are autocorrelated and/or heteroskedastic, since this affects the covariance matrix of the least squares estimator and also to what extent least squares is efficient. We might also be interested in studying if the residuals are normally distributed, since this also affects the efficiency of least squares (remember that LS is MLE is the residuals are normally distributed).

It is important that the fitted residuals used in the diagnostic tests are consistent. With poorly estimated residuals, we can easily find autocorrelation, heteroskedasticity, or non-normality even if the true residuals have none of these features.

4.7.1 Autocorrelation

Let $\hat{\rho}_s$ be the estimate of the sth autocorrelation coefficient of some variable, for instance, the fitted residuals. The sampling properties of $\hat{\rho}_s$ are complicated, but there are several useful large sample results for Gaussian processes (these results typically carry over to processes which are similar to the Gaussian—a homoskedastic process with finite 6th moment is typically enough). When the true autocorrelations are all zero (not ρ_0, of course), then for any i and j different from zero

$$\sqrt{T}\begin{bmatrix}\hat{\rho}_i \\ \hat{\rho}_j\end{bmatrix} \rightarrow^d N\left(\begin{bmatrix}0 \\ 0\end{bmatrix}, \begin{bmatrix}1 & 0 \\ 0 & 1\end{bmatrix}\right). \qquad (4.40)$$

This result can be used to construct tests for both single autocorrelations (t-test or χ^2 test) and several autocorrelations at once (χ^2 test).

Example 4.18 *(t-test) We want to test the hypothesis that $\rho_1 = 0$. Since the $N(0,1)$ distribution has 5% of the probability mass below -1.65 and another 5% above 1.65, we can reject the null hypothesis at the 10% level if $\sqrt{T}|\hat{\rho}_1| > 1.65$. With $T = 100$, we therefore need $|\hat{\rho}_1| > 1.65/\sqrt{100} = 0.165$ for rejection, and with $T = 1000$ we need $|\hat{\rho}_1| > 1.65/\sqrt{1000} \approx 0.0.53$.*

The *Box-Pierce test* follows directly from the result in (4.40), since it shows that $\sqrt{T}\hat{\rho}_i$ and $\sqrt{T}\hat{\rho}_j$ are iid N(0,1) variables. Therefore, the sum of the square of them is distributed as an χ^2 variable. The test statistics typically used is

$$Q_L = T\sum_{s=1}^{L}\hat{\rho}_s^2 \rightarrow^d \chi_L^2. \qquad (4.41)$$

Example 4.19 *(Box-Pierce) Let $\hat{\rho}_1 = 0.165$, and $T = 100$, so $Q_1 = 100 \times 0.165^2 = 2.72$. The 10% critical value of the χ_1^2 distribution is 2.71, so the null hypothesis of no autocorrelation is rejected.*

The choice of lag order in (4.41), L, should be guided by theoretical considerations, but it may also be wise to try different values. There is clearly a trade off: too few lags may miss a significant high-order autocorrelation, but too many lags can destroy the power of the test (as the test statistics is not affected much by increasing L, but the critical values increase).

Example 4.20 *(Residuals follow an AR(1)process) If $u_t = 0.9u_{t-1} + \varepsilon_t$, then the true autocorrelation coefficients are $\rho_j = 0.9^j$.*

A common test of the serial correlation of residuals from a regression is the *Durbin-Watson test*

$$d = 2\left(1 - \hat{\rho}_1\right), \qquad (4.42)$$

where the null hypothesis of no autocorrelation is

not rejected if $d > d^*_{upper}$
rejected if $d < d^*_{lower}$ (in favor of positive autocorrelation)
else inconclusive

where the upper and lower critical values can be found in tables. (Use $4 - d$ to let negative autocorrelation be the alternative hypothesis.) This test is typically not useful when lagged dependent variables enter the right hand side (d is biased towards showing no autocorrelation). Note that DW tests only for first-order autocorrelation.

Example 4.21 *(Durbin-Watson.) With $\hat{\rho}_1 = 0.2$ we get $d = 1.6$. For large samples, the 5% critical value is $d^*_{lower} \approx 1.6$, so $\hat{\rho}_1 > 0.2$ is typically considered as evidence of positive autocorrelation.*

The fitted residuals used in the autocorrelation tests must be consistent in order to interpret the result in terms of the properties of the true residuals. For instance, an excluded autocorrelated variable will probably give autocorrelated fitted residuals—and also make the coefficient estimator inconsistent (unless the excluded variable is uncorrelated with the regressors). Only when we know that the model is correctly specified can we interpret a finding of autocorrelated residuals as an indication of the properties of the true residuals.

4.7.2 Heteroskedasticity

Remark 4.22 *(Kronecker product.) If A and B are matrices, then*

$$A \otimes B = \begin{bmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{bmatrix}.$$

Example 4.23 *Let x_1 and x_2 be scalars. Then*

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \otimes \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} x_1 \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \\ x_2 \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \end{bmatrix} = \begin{bmatrix} x_1 x_1 \\ x_1 x_2 \\ x_2 x_1 \\ x_2 x_2 \end{bmatrix}.$$

White's test for heteroskedasticity tests the null hypothesis of homoskedasticity against the kind of heteroskedasticity which can be explained by the levels, squares, and cross products of the regressors. Let w_t be the unique elements in $x_t \otimes x_t$, where we have added a constant to x_t if there was not one from the start. Run a regression of the squared fitted LS residuals on w_t

$$\hat{u}_t^2 = w_t'\gamma + \varepsilon_t \tag{4.43}$$

and test if all elements (except the constant) in γ are zero (with a χ^2 or F test). The reason for this specification is that if u_t^2 is uncorrelated with $x_t \otimes x_t$, then the usual LS covariance matrix applies.

Breusch-Pagan's test is very similar, except that the vector w_t in (4.43) can be any vector which is thought of as useful for explaining the heteroskedasticity. The null hypothesis is that the variance is constant, which is tested against the alternative that the variance is some function of w_t.

The fitted residuals used in the heteroskedasticity tests must be consistent in order to interpret the result in terms of the properties of the true residuals. For instance, if some of the of elements in w_t belong to the regression equation, but are excluded, then fitted residuals will probably fail these tests.

4.7.3 Normality

We often make the assumption of normally distributed errors, for instance, in maximum likelihood estimation. This assumption can be tested by using the fitted errors. This works since moments estimated from the fitted errors are consistent estimators of the moments of the true errors. Define the degree of *skewness* and *excess kurtosis* for a variable z_t (could be the fitted residuals) as

$$\hat{\theta}_3 = \frac{1}{T}\sum_{t=1}^{T}(z_t - \bar{z})^3 / \hat{\sigma}^3, \tag{4.44}$$

$$\hat{\theta}_4 = \frac{1}{T}\sum_{t=1}^{T}(z_t - \bar{z})^4 / \hat{\sigma}^4 - 3, \tag{4.45}$$

where $\bar{z}$ is the sample mean and $\hat{\sigma}^2$ is the estimated variance.

Remark 4.24 *($\chi^2(n)$ distribution.) If x_i are independent $N(0, \sigma_i^2)$ variables, then $\Sigma_{i=1}^n x_i^2/\sigma_i^2 \sim \chi^2(n)$.*

In a normal distribution, the true values are zero and the test statistics $\hat{\theta}_3$ and $\hat{\theta}_4$ are themselves normally distributed with zero covariance and variances $6/T$ and $24/T$, respectively (straightforward, but tedious, to show). Therefore, under the null hypothesis of a normal distribution, $T\hat{\theta}_3^2/6$ and $T\hat{\theta}_4^2/24$ are independent and both asymptotically distributed as $\chi^2(1)$, so the sum is asymptotically a $\chi^2(2)$ variable

$$W = T\left(\hat{\theta}_3^2/6 + \hat{\theta}_4^2/24\right) \to^d \chi^2(2). \tag{4.46}$$

This is the *Jarque and Bera test of normality*.

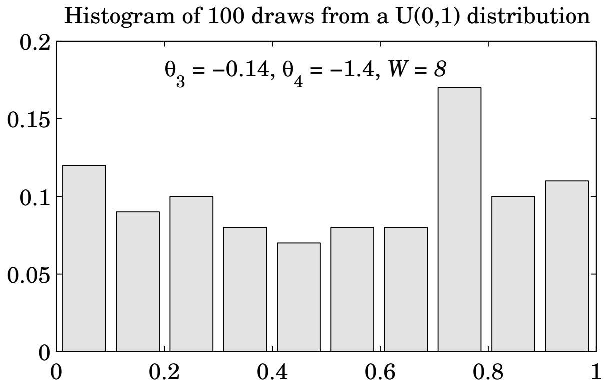


Figure 4.1: This figure shows a histogram from 100 draws of iid uniformly [0,1] distributed variables.

Bibliography

Davidson, J., 2000, *Econometric Theory*, Blackwell Publishers, Oxford.

Davidson, R., and J. G. MacKinnon, 1993, *Estimation and Inference in Econometrics*, Oxford University Press, Oxford.

Greene, W. H., 2000, *Econometric Analysis*, Prentice-Hall, Upper Saddle River, New Jersey, 4th edn.

Hamilton, J. D., 1994, *Time Series Analysis*, Princeton University Press, Princeton.

Hayashi, F., 2000, *Econometrics*, Princeton University Press.

Johnston, J., and J. DiNardo, 1997, *Econometric Methods*, McGraw-Hill, New York, 4th edn.

Patterson, K., 2000, *An Introduction to Applied Econometrics: A Time Series Approach*, MacMillan Press, London.

Pindyck, R. S., and D. L. Rubinfeld, 1998, *Econometric Models and Economic Forecasts*, Irwin McGraw-Hill, Boston, Massachusetts, 4ed edn.

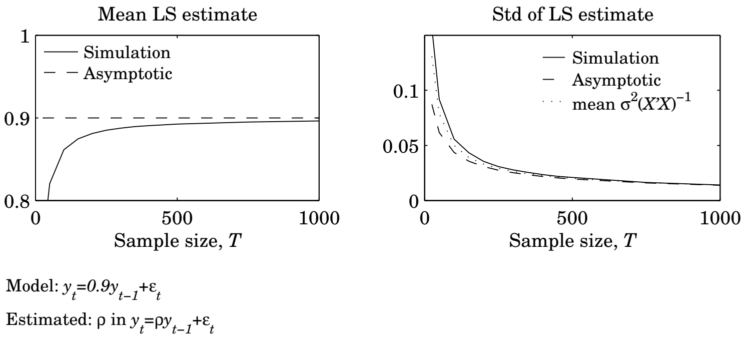


Figure 4.2: Distribution of LS estimator of autoregressive parameter.

Verbeek, M., 2004, *A Guide to Modern Econometrics*, Wiley, Chichester, 2nd edn.

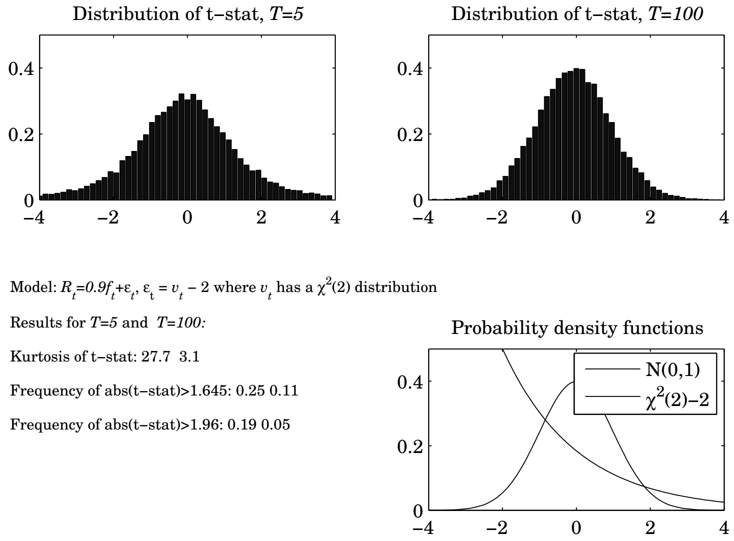


Figure 4.3: Distribution of LS estimator when residuals have a t_3 distribution.

5 Instrumental Variable Method

Reference: Greene (2000) 9.5 and 16.1-2
Additional references: Hayashi (2000) 3.1-4; Verbeek (2004) 5.1-4; Hamilton (1994) 8.2; and Pindyck and Rubinfeld (1998) 7

5.1 Consistency of Least Squares or Not?

Consider the linear model

$$y_t = x_t'\beta_0 + u_t, \tag{5.1}$$

where y_t and u_t are scalars, x_t a $k \times 1$ vector, and β_0 is a $k \times 1$ vector of the true coefficients. The least squares estimator is

$$\hat{\beta}_{LS} = \left(\frac{1}{T}\sum_{t=1}^{T} x_t x_t'\right)^{-1} \frac{1}{T}\sum_{t=1}^{T} x_t y_t \tag{5.2}$$

$$= \beta_0 + \left(\frac{1}{T}\sum_{t=1}^{T} x_t x_t'\right)^{-1} \frac{1}{T}\sum_{t=1}^{T} x_t u_t, \tag{5.3}$$

where we have used (5.1) to substitute for y_t. The probability limit is

$$\text{plim}\,\hat{\beta}_{LS} - \beta_0 = \left(\text{plim}\,\frac{1}{T}\sum_{t=1}^{T} x_t x_t'\right)^{-1} \text{plim}\,\frac{1}{T}\sum_{t=1}^{T} x_t u_t. \tag{5.4}$$

In many cases the law of large numbers applies to both terms on the right hand side. The first term is typically a matrix with finite elements and the second term is the covariance of the regressors and the true residuals. This covariance must be zero for LS to be consistent.

5.2 Reason 1 for IV: Measurement Errors

Reference: Greene (2000) 9.5.

Suppose the true model is

$$y_t^* = x_t^{*\prime}\beta_0 + u_t^*. \tag{5.5}$$

Data on y_t^* and x_t^* is not directly observable, so we instead run the regression

$$y_t = x_t'\beta + u_t, \tag{5.6}$$

where y_t and x_t are proxies for the correct variables (the ones that the model is true for). We can think of the difference as measurement errors

$$y_t = y_t^* + v_t^y \text{ and} \tag{5.7}$$

$$x_t = x_t^* + v_t^x, \tag{5.8}$$

where the errors are uncorrelated with the true values and the "true" residual u_t^*.

Use (5.7) and (5.8) in (5.5)

$$\begin{aligned} y_t - v_t^y &= \left(x_t - v_t^x\right)'\beta_0 + u_t^* \text{ or} \\ y_t &= x_t'\beta_0 + \varepsilon_t \text{ where } \varepsilon_t = -v_t^{x\prime}\beta_0 + v_t^y + u_t^*. \end{aligned} \tag{5.9}$$

Suppose that x_t^* is a measured with error. From (5.8) we see that v_t^x and x_t are correlated, so LS on (5.9) is inconsistent in this case. To make things even worse, measurement errors in only one of the variables typically affect all the coefficient estimates.

To illustrate the effect of the error, consider the case when x_t is a scalar. Then, the probability limit of the LS estimator of β in (5.9) is

$$\begin{aligned} \text{plim}\,\hat{\beta}_{LS} &= \text{Cov}\left(y_t, x_t\right)/\text{Var}\left(x_t\right) \\ &= \text{Cov}\left(x_t^*\beta_0 + u_t^*, x_t\right)/\text{Var}\left(x_t\right) \\ &= \text{Cov}\left(x_t\beta_0 - v_t^x\beta_0 + u_t^*, x_t\right)/\text{Var}\left(x_t\right) \\ &= \frac{\text{Cov}\left(x_t\beta_0, x_t\right) + \text{Cov}\left(-v_t^x\beta_0, x_t\right) + \text{Cov}\left(u_t^*, x_t\right)}{\text{Var}\left(x_t\right)} \\ &= \frac{\text{Var}\left(x_t\right)}{\text{Var}\left(x_t\right)}\beta_0 + \frac{\text{Cov}\left(-v_t^x\beta_0, x_t^* - v_t^x\right)}{\text{Var}\left(x_t\right)} \\ &= \beta_0 - \beta_0\text{Var}\left(v_t^x\right)/\text{Var}\left(x_t\right) \\ &= \beta_0\left[1 - \frac{\text{Var}\left(v_t^x\right)}{\text{Var}\left(x_t^*\right) + \text{Var}\left(v_t^x\right)}\right]. \end{aligned} \tag{5.10}$$

since x_t^* and v_t^x are uncorrelated with u_t^* and with each other. This shows that $\hat{\beta}_{LS}$ goes to zero as the measurement error becomes relatively more volatile compared with the true value. This makes a lot of sense, since when the measurement error is very large then the regressor x_t is dominated by noise that has nothing to do with the dependent variable.

Suppose instead that only y_t^* is measured with error. This not a big problem since this measurement error is uncorrelated with the regressor, so the consistency of least squares is not affected. In fact, a measurement error in the dependent variable is like increasing the variance in the residual.

5.3 Reason 2 for IV: Simultaneous Equations Bias (and Inconsistency)

Suppose economic theory tells you that the *structural form* of the m endogenous variables, y_t, and the k predetermined (exogenous) variables, z_t, is

$$Fy_t + Gz_t = u_t, \text{ where } u_t \text{ is iid with } \text{E}u_t = \mathbf{0} \text{ and } \text{Cov}\left(u_t\right) = \Sigma, \tag{5.11}$$

where F is $m \times m$, and G is $m \times k$. The disturbances are assumed to be uncorrelated with the predetermined variables, $\text{E}(z_t u_t') = \mathbf{0}$.

Suppose F is invertible. Solve for y_t to get the *reduced form*

$$y_t = -F^{-1}Gz_t + F^{-1}u_t \tag{5.12}$$

$$= \Pi z_t + \varepsilon_t, \text{ with } \text{Cov}\left(\varepsilon_t\right) = \Omega. \tag{5.13}$$

The reduced form coefficients, Π, can be consistently estimated by LS on each equation since the exogenous variables z_t are uncorrelated with the reduced form residuals (which are linear combinations of the structural residuals). The fitted residuals can then be used to get an estimate of the reduced form covariance matrix.

The jth line of the structural form (5.11) can be written

$$F_j y_t + G_j z_t = u_{jt}, \tag{5.14}$$

where F_j and G_j are the jth rows of F and G, respectively. Suppose the model is normalized so that the coefficient on y_{jt} is one (otherwise, divide (5.14) with this coefficient).

Then, rewrite (5.14) as

$$\begin{aligned} y_{jt} &= -G_{j1}\tilde{z}_t - F_{j1}\tilde{y}_t + u_{jt} \\ &= x_t'\beta + u_{jt}, \text{ where } x_t' = \left[\tilde{z}_t', \tilde{y}_t'\right], \end{aligned} \tag{5.15}$$

where $\tilde{z}_t$ and $\tilde{y}_t$ are the exogenous and endogenous variables that enter the jth equation, which we collect in the x_t vector to highlight that (5.15) looks like any other linear regression equation. The problem with (5.15), however, is that the residual is likely to be correlated with the regressors, so the LS estimator is inconsistent. The reason is that a shock to u_{jt} influences y_{jt}, which in turn will affect some other endogenous variables in the system (5.11). If any of these endogenous variable are in x_t in (5.15), then there is a correlation between the residual and (some of) the regressors.

Note that the concept of endogeneity discussed here only refers to *contemporaneous endogeneity* as captured by off-diagonal elements in F in (5.11). The vector of predetermined variables, z_t, could very well include lags of y_t without affecting the econometric endogeneity problem.

Example 5.1 *(Supply and Demand. Reference: GR 16, Hamilton 9.1.) Consider the simplest simultaneous equations model for supply and demand on a market. Supply is*

$$q_t = \gamma p_t + u_t^s,\ \gamma > 0,$$

and demand is

$$q_t = \beta p_t + \alpha A_t + u_t^d,\ \beta < 0,$$

where A_t is an observable demand shock (perhaps income). The structural form is therefore

$$\begin{bmatrix} 1 & -\gamma \\ 1 & -\beta \end{bmatrix}\begin{bmatrix} q_t \\ p_t \end{bmatrix} + \begin{bmatrix} 0 \\ -\alpha \end{bmatrix} A_t = \begin{bmatrix} u_t^s \\ u_t^d \end{bmatrix}.$$

The reduced form is

$$\begin{bmatrix} q_t \\ p_t \end{bmatrix} = \begin{bmatrix} \pi_{11} \\ \pi_{21} \end{bmatrix} A_t + \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix}.$$

If we knew the structural form, then we can solve for q_t and p_t to get the reduced form in terms of the structural parameters

$$\begin{bmatrix} q_t \\ p_t \end{bmatrix} = \begin{bmatrix} -\frac{\gamma}{\beta-\gamma}\alpha \\ -\frac{1}{\beta-\gamma}\alpha \end{bmatrix} A_t + \begin{bmatrix} \frac{\beta}{\beta-\gamma} & -\frac{\gamma}{\beta-\gamma} \\ \frac{1}{\beta-\gamma} & -\frac{1}{\beta-\gamma} \end{bmatrix}\begin{bmatrix} u_t^s \\ u_t^d \end{bmatrix}.$$

Example 5.2 *(Supply equation with LS.) Suppose we try to estimate the supply equation in Example 5.1 by LS, that is, we run the regression*

$$q_t = \theta p_t + \varepsilon_t.$$

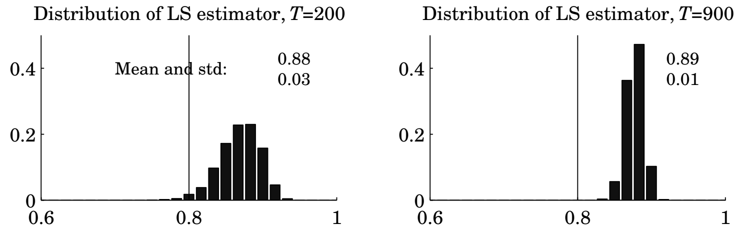
If data is generated by the model in Example 5.1, then the reduced form shows that p_t is correlated with u_t^s, so we cannot hope that LS will be consistent. In fact, when both q_t and p_t have zero means, then the probability limit of the LS estimator is

$$\begin{aligned} \operatorname{plim}\hat{\theta} &= \frac{Cov\left(q_t, p_t\right)}{Var\left(p_t\right)} \\ &= \frac{Cov\left(\frac{\gamma\alpha}{\gamma-\beta}A_t + \frac{\gamma}{\gamma-\beta}u_t^d - \frac{\beta}{\gamma-\beta}u_t^s, \frac{\alpha}{\gamma-\beta}A_t + \frac{1}{\gamma-\beta}u_t^d - \frac{1}{\gamma-\beta}u_t^d\right)}{Var\left(\frac{\alpha}{\gamma-\beta}A_t + \frac{1}{\gamma-\beta}u_t^d - \frac{1}{\gamma-\beta}u_t^s\right)}, \end{aligned}$$

where the second line follows from the reduced form. Suppose the supply and demand shocks are uncorrelated. In that case we get

$$\begin{aligned} \operatorname{plim}\hat{\theta} &= \frac{\frac{\gamma\alpha^2}{(\gamma-\beta)^2}Var\left(A_t\right) + \frac{\gamma}{(\gamma-\beta)^2}Var\left(u_t^d\right) + \frac{\beta}{(\gamma-\beta)^2}Var\left(u_t^s\right)}{\frac{\alpha^2}{(\gamma-\beta)^2}Var\left(A_t\right) + \frac{1}{(\gamma-\beta)^2}Var\left(u_t^d\right) + \frac{1}{(\gamma-\beta)^2}Var\left(u_t^s\right)} \\ &= \frac{\gamma\alpha^2 Var\left(A_t\right) + \gamma Var\left(u_t^d\right) + \beta Var\left(u_t^s\right)}{\alpha^2 Var\left(A_t\right) + Var\left(u_t^d\right) + Var\left(u_t^s\right)}. \end{aligned}$$

First, suppose the supply shocks are zero, $Var\left(u_t^s\right) = 0$, then $\operatorname{plim}\hat{\theta} = \gamma$, so we indeed estimate the supply elasticity, as we wanted. Think of a fixed supply curve, and a demand curve which moves around. These point of p_t and q_t should trace out the supply curve. It is clearly u_t^s that causes a simultaneous equations problem in estimating the supply curve: u_t^s affects both q_t and p_t and the latter is the regressor in the supply equation. With no movements in u_t^s there is no correlation between the shock and the regressor. Second, now suppose instead that the both demand shocks are zero (both $A_t = 0$ and $Var\left(u_t^d\right) = 0$). Then $\operatorname{plim}\hat{\theta} = \beta$, so the estimated value is not the supply, but the demand elasticity. Not good. This time, think of a fixed demand curve, and a supply curve which moves around.



True model: $y_t = \rho y_{t-1} + \varepsilon_t$, where $\varepsilon_t = \nu_t + \theta\nu_{t-1}$,

where ρ=0.8, θ=0.5 and ν_t is iid N(0,2)

Estimated model: $y_t = \rho y_{t-1} + u_t$

Figure 5.1: Distribution of LS estimator of autoregressive parameter

Example 5.3 *(A flat demand curve.) Suppose we change the demand curve in Example 5.1 to be infinitely elastic, but to still have demand shocks. For instance, the inverse demand curve could be* $p_t = \psi A_t + u_t^D$. *In this case, the supply and demand is no longer a simultaneous system of equations and both equations could be estimated consistently with LS. In fact, the system is recursive, which is easily seen by writing the system on vector form*

$$\begin{bmatrix} 1 & 0 \\ 1 & -\gamma \end{bmatrix}\begin{bmatrix} p_t \\ q_t \end{bmatrix} + \begin{bmatrix} -\psi \\ 0 \end{bmatrix} A_t = \begin{bmatrix} u_t^D \\ u_t^s \end{bmatrix}.$$

A supply shock, u_t^s, *affects the quantity, but this has no affect on the price (the regressor in the supply equation), so there is no correlation between the residual and regressor in the supply equation. A demand shock,* u_t^D, *affects the price and the quantity, but since quantity is not a regressor in the inverse demand function (only the exogenous* A_t *is) there is no correlation between the residual and the regressor in the inverse demand equation either.*

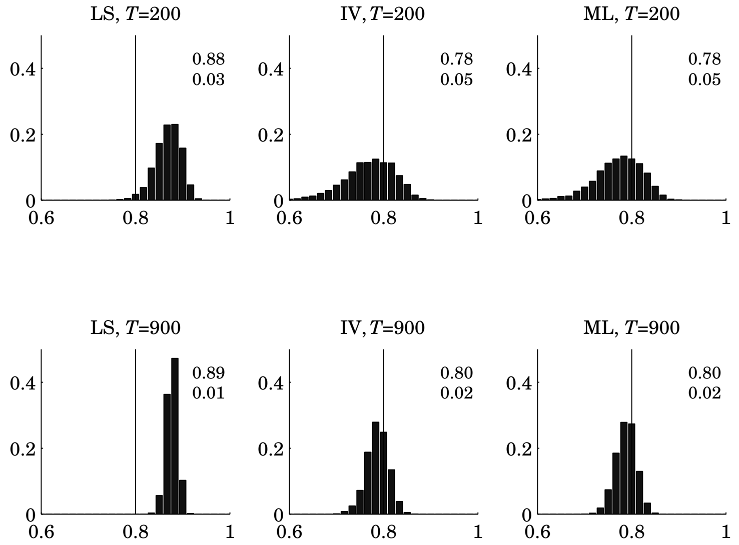


Figure 5.2: Distribution of LS, IV and ML estimators of autoregressive parameter. See Figure 5.1 for details.

5.4 Definition of the IV Estimator—Consistency of IV

Reference: Greene (2000) 9.5; Hamilton (1994) 8.2; and Pindyck and Rubinfeld (1998) 7.

Consider the linear model

$$y_t = x_t'\beta_0 + u_t, \tag{5.16}$$

where y_t is a scalar, x_t a $k \times 1$ vector, and β_0 is a vector of the true coefficients. If we suspect that x_t and u_t in (5.16) are correlated, then we may use the instrumental variables (IV) method. To do that, let z_t be a $k \times 1$ vector of instruments (as many instruments as regressors; we will later deal with the case when we have more instruments than regressors.) If x_t and u_t are not correlated, then setting $x_t = z_t$ gives the least squares (LS) method.

Recall that LS minimizes the variance of the fitted residuals, $\hat{u}_t = y_t - x_t'\hat{\beta}_{LS}$. The first order conditions for that optimization problem are

$$\mathbf{0}_{kx1} = \frac{1}{T}\sum_{t=1}^{T} x_t \left(y_t - x_t'\hat{\beta}_{LS}\right). \tag{5.17}$$

If x_t and u_t are correlated, then plim $\hat{\beta}_{LS} \neq \beta_0$. The reason is that the probability limit of the right hand side of (5.17) is $\text{Cov}(x_t, y_t - x_t'\hat{\beta}_{LS})$, which at $\hat{\beta}_{LS} = \beta_0$ is non-zero, so the first order conditions (in the limit) cannot be satisfied at the true parameter values. Note that since the LS estimator by construction forces the fitted residuals to be uncorrelated with the regressors, the properties of the LS residuals are of little help in deciding if to use LS or IV.

The idea of the IV method is to replace the first x_t in (5.17) with a vector (of similar size) of some instruments, z_t. The identifying assumption of the IV method is that the instruments are uncorrelated with the residuals (and, as we will see, correlated with the regressors)

$$\mathbf{0}_{kx1} = \text{E}z_t u_t \tag{5.18}$$

$$= \text{E}z_t \left(y_t - x_t'\beta_0\right). \tag{5.19}$$

The intuition is that the linear model (5.16) is assumed to be correctly specified: the residuals, u_t, represent factors which we cannot explain, so z_t should not contain any information about u_t.

The sample analogue to (5.19) defines the IV estimator of β as[1]

$$\mathbf{0}_{kx1} = \frac{1}{T}\sum_{t=1}^{T} z_t \left(y_t - x_t'\hat{\beta}_{IV}\right), \text{ or} \tag{5.20}$$

$$\hat{\beta}_{IV} = \left(\frac{1}{T}\sum_{t=1}^{T} z_t x_t'\right)^{-1} \frac{1}{T}\sum_{t=1}^{T} z_t y_t. \tag{5.21}$$

It is clearly necessay for $\Sigma z_t x_t'/T$ to have full rank to calculate the IV estimator.

Remark 5.4 *(Probability limit of product) For any random variables y_T and x_T where* plim $y_T = a$ *and* plim $x_T = b$ *(a and b are constants), we have* plim $y_T x_T = ab$.

[1] In matrix notation where z_t' is the t^{th} row of Z we have $\hat{\beta}_{IV} = (Z'X/T)^{-1}(Z'Y/T)$.

To see if the IV estimator is consistent, use (5.16) to substitute for y_t in (5.20) and take the probability limit

$$\text{plim}\,\frac{1}{T}\sum_{t=1}^{T} z_t x_t'\beta_0 + \text{plim}\,\frac{1}{T}\sum_{t=1}^{T} z_t u_t = \text{plim}\,\frac{1}{T}\sum_{t=1}^{T} z_t x_t'\hat{\beta}_{IV}. \tag{5.22}$$

Two things are required for consistency of the IV estimator, plim $\hat{\beta}_{IV} = \beta_0$. First, that plim $\Sigma z_t u_t/T = 0$. Provided a law of large numbers apply, this is condition (5.18). Second, that plim $\Sigma z_t x_t'/T$ has full rank. To see this, suppose plim $\Sigma z_t u_t/T = 0$ is satisfied. Then, (5.22) can be written

$$\left(\text{plim}\,\frac{1}{T}\sum_{t=1}^{T} z_t x_t'\right)\left(\beta_0 - \text{plim}\,\hat{\beta}_{IV}\right) = 0. \tag{5.23}$$

If plim $\Sigma z_t x_t'/T$ has reduced rank, then plim $\hat{\beta}_{IV}$ does not need to equal β_0 for (5.23) to be satisfied. In practical terms, the first order conditions (5.20) do then not define a unique value of the vector of estimates. If a law of large numbers applies, then plim $\Sigma z_t x_t'/T = \text{E}z_t x_t'$. If both z_t and x_t contain constants (or at least one of them has zero means), then a reduced rank of $\text{E}z_t x_t'$ would be a consequence of a reduced rank of the covariance matrix of the stochastic elements in z_t and x_t, for instance, that some of the instruments are uncorrelated with all the regressors. This shows that the instruments must indeed be correlated with the regressors for IV to be consistent (and to make sense).

Remark 5.5 *(Second moment matrix) Note that $Ezx' = EzEx' + Cov(z, x)$. If $Ez = \mathbf{0}$ and/or $Ex = \mathbf{0}$, then the second moment matrix is a covariance matrix. Alternatively, suppose both z and x contain constants normalized to unity: $z = [1, \tilde{z}']'$ and $x = [1, \tilde{x}']'$ where $\tilde{z}$ and $\tilde{x}$ are random vectors. We can then write*

$$\begin{aligned} Ezx' &= \begin{bmatrix} 1 \\ E\tilde{z} \end{bmatrix}\begin{bmatrix} 1 & E\tilde{x}' \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & Cov(\tilde{z}, \tilde{x}) \end{bmatrix} \\ &= \begin{bmatrix} 1 & E\tilde{x}' \\ E\tilde{z} & E\tilde{z}E\tilde{x}' + Cov(\tilde{z}, \tilde{x}) \end{bmatrix}. \end{aligned}$$

For simplicity, suppose $\tilde{z}$ and $\tilde{x}$ are scalars. Then Ezx' has reduced rank if $Cov(\tilde{z}, \tilde{x}) = 0$, since $Cov(\tilde{z}, \tilde{x})$ is then the determinant of Ezx'. This is true also when $\tilde{z}$ and $\tilde{x}$ are vectors.

Example 5.6 *(Supply equation with IV.) Suppose we try to estimate the supply equation in Example 5.1 by IV. The only available instrument is A_t, so (5.21) becomes*

$$\hat{\gamma}_{IV} = \left(\frac{1}{T}\sum_{t=1}^{T} A_t p_t\right)^{-1} \frac{1}{T}\sum_{t=1}^{T} A_t q_t,$$

so the probability limit is

$$\text{plim}\,\hat{\gamma}_{IV} = Cov\left(A_t, p_t\right)^{-1} Cov\left(A_t, q_t\right),$$

since all variables have zero means. From the reduced form in Example 5.1 we see that

$$Cov\left(A_t, p_t\right) = -\frac{1}{\beta - \gamma}\alpha Var\left(A_t\right) \text{ and } Cov\left(A_t, q_t\right) = -\frac{\gamma}{\beta - \gamma}\alpha Var\left(A_t\right),$$

so

$$\begin{aligned} \text{plim}\,\hat{\gamma}_{IV} &= \left[-\frac{1}{\beta - \gamma}\alpha Var\left(A_t\right)\right]^{-1}\left[-\frac{\gamma}{\beta - \gamma}\alpha Var\left(A_t\right)\right] \\ &= \gamma. \end{aligned}$$

This shows that $\hat{\gamma}_{IV}$ is consistent.

5.4.1 Asymptotic Normality of IV

Little is known about the finite sample distribution of the IV estimator, so we focus on the asymptotic distribution—assuming the IV estimator is consistent.

Remark 5.7 *If $x_T \xrightarrow{d} x$ (a random variable) and* plim $Q_T = Q$ *(a constant matrix), then $Q_T x_T \xrightarrow{d} Qx$.*

Use (5.16) to substitute for y_t in (5.20)

$$\hat{\beta}_{IV} = \beta_0 + \left(\frac{1}{T}\sum_{t=1}^{T} z_t x_t'\right)^{-1} \frac{1}{T}\sum_{t=1}^{T} z_t u_t. \quad (5.24)$$

Premultiply by $\sqrt{T}$ and rearrange as

$$\sqrt{T}(\hat{\beta}_{IV} - \beta_0) = \left(\frac{1}{T}\sum_{t=1}^{T} z_t x_t'\right)^{-1} \frac{\sqrt{T}}{T}\sum_{t=1}^{T} z_t u_t. \quad (5.25)$$

If the first term on the right hand side converges in probability to a finite matrix (as assumed in in proving consistency), and the vector of random variables $z_t u_t$ satisfies a central limit theorem, then

$$\begin{aligned} \sqrt{T}(\hat{\beta}_{IV} - \beta_0) &\xrightarrow{d} N\left(0, \Sigma_{zx}^{-1} S_0 \Sigma_{xz}^{-1}\right), \text{ where} \quad (5.26) \\ \Sigma_{zx} &= \frac{1}{T}\sum_{t=1}^{T} z_t x_t' \text{ and } S_0 = \text{Cov}\left(\frac{\sqrt{T}}{T}\sum_{t=1}^{T} z_t u_t\right). \end{aligned}$$

The last matrix in the covariance matrix follows from $(\Sigma_{zx}^{-1})' = (\Sigma_{zx}')^{-1} = \Sigma_{xz}^{-1}$. This general expression is valid for both autocorrelated and heteroskedastic residuals—all such features are loaded into the S_0 matrix. Note that S_0 is the variance-covariance matrix of $\sqrt{T}$ times a sample average (of the vector of random variables $x_t u_t$).

Example 5.8 *(Choice of instrument in IV, simplest case) Consider the simple regression*

$$y_t = \beta_1 x_t + u_t.$$

The asymptotic variance of the IV estimator is

$$AVar(\sqrt{T}(\hat{\beta}_{IV} - \beta_0)) = \text{Var}\left(\frac{\sqrt{T}}{T}\sum_{t=1}^{T} z_t u_t\right) / \text{Cov}\left(z_t, x_t\right)^2$$

If z_t and u_t is serially uncorrelated and independent of each other, then $\text{Var}(\Sigma_{t=1}^{T} z_t u_t / \sqrt{T}) = \text{Var}(z_t)\,\text{Var}(u_t)$. *We can then write*

$$AVar(\sqrt{T}(\hat{\beta}_{IV} - \beta_0)) = \text{Var}(u_t)\frac{\text{Var}(z_t)}{\text{Cov}\left(z_t, x_t\right)^2} = \frac{\text{Var}(u_t)}{\text{Var}(x_t) Corr\left(z_t, x_t\right)^2}.$$

An instrument with a weak correlation with the regressor gives an imprecise estimator. With a perfect correlation, then we get the precision of the LS estimator (which is precise, but perhaps not consistent).

5.4.2 2SLS

Suppose now that we have more instruments, z_t, than regressors, x_t. The IV method does not work since, there are then more equations than unknowns in (5.20). Instead, we can use the *2SLS* estimator. It has two steps. First, regress all elements in x_t on all elements in z_t with LS. Second, use the fitted values of x_t, denoted $\hat{x}_t$, as instruments in the IV method (use $\hat{x}_t$ in place of z_t in the equations above). In can be shown that this is the most efficient use of the information in z_t. The IV is clearly a special case of 2SLS (when z_t has the same number of elements as x_t).

It is immediate from (5.22) that 2SLS is consistent under the same condiditons as IV since $\hat{x}_t$ is a linear function of the instruments, so $\text{plim} \sum_{t=1}^{T} \hat{x}_t u_t / T = 0$, if all the instruments are uncorrelated with u_t.

The name, 2SLS, comes from the fact that we get exactly the same result if we replace the second step with the following: regress y_t on $\hat{x}_t$ with LS.

Example 5.9 *(Supply equation with 2SLS.). With only one instrument, A_t, this is the same as Example 5.6, but presented in another way. First, regress p_t on A_t*

$$p_t = \delta A_t + u_t \Rightarrow \text{plim}\,\hat{\delta}_{LS} = \frac{Cov\left(p_t, A_t\right)}{Var\left(A_t\right)} = -\frac{1}{\beta - \gamma}\alpha.$$

Construct the predicted values as

$$\hat{p}_t = \hat{\delta}_{LS} A_t.$$

Second, regress q_t on $\hat{p}_t$

$$q_t = \gamma \hat{p}_t + e_t, \textit{ with } \text{plim}\,\hat{\gamma}_{2SLS} = \text{plim}\,\frac{\widehat{Cov}\left(q_t, \hat{p}_t\right)}{\widehat{Var}\left(\hat{p}_t\right)}.$$

Use $\hat{p}_t = \hat{\delta}_{LS} A_t$ and Slutsky's theorem

$$\begin{aligned}
\text{plim}\,\hat{\gamma}_{2SLS} &= \frac{\text{plim}\,\widehat{Cov}\left(q_t, \hat{\delta}_{LS} A_t\right)}{\text{plim}\,\widehat{Var}\left(\hat{\delta}_{LS} A_t\right)} \\
&= \frac{Cov\left(q_t, A_t\right)\text{plim}\,\hat{\delta}_{LS}}{Var\left(A_t\right)\text{plim}\,\hat{\delta}_{LS}^2} \\
&= \frac{\left[-\frac{\gamma}{\beta-\gamma}\alpha Var\left(A_t\right)\right]\left[-\frac{1}{\beta-\gamma}\alpha\right]}{Var\left(A_t\right)\left[-\frac{1}{\beta-\gamma}\alpha\right]^2} \\
&= \gamma.
\end{aligned}$$

Note that the trick here is to suppress some the movements in p_t. Only those movements that depend on A_t (the observable shifts of the demand curve) are used. Movements in p_t which are due to the unobservable demand and supply shocks are disregarded in $\hat{p}_t$. We know from Example 5.2 that it is the supply shocks that make the LS estimate of the supply curve inconsistent. The IV method suppresses both them and the unobservable demand shock.

5.5 Hausman's Specification Test*

Reference: Greene (2000) 9.5

This test is constructed to test if an efficient estimator (like LS) gives (approximately) the same estimate as a consistent estimator (like IV). If not, the efficient estimator is most likely inconsistent. It is therefore a way to test for the presence of endogeneity and/or measurement errors.

Let $\hat{\beta}_e$ be an estimator that is consistent and asymptotically efficient when the null hypothesis, H_0, is true, but inconsistent when H_0 is false. Let $\hat{\beta}_c$ be an estimator that is consistent under both H_0 and the alternative hypothesis. When H_0 is true, the asymptotic distribution is such that

$$\text{Cov}\left(\hat{\beta}_e, \hat{\beta}_c\right) = \text{Var}\left(\hat{\beta}_e\right). \tag{5.27}$$

Proof. Consider the estimator $\lambda \hat{\beta}_c + (1-\lambda)\hat{\beta}_e$, which is clearly consistent under H_0 since both $\hat{\beta}_c$ and $\hat{\beta}_e$ are. The asymptotic variance of this estimator is

$$\lambda^2 \text{Var}\left(\hat{\beta}_c\right) + (1-\lambda)^2 \text{Var}\left(\hat{\beta}_e\right) + 2\lambda(1-\lambda)\,\text{Cov}\left(\hat{\beta}_c, \hat{\beta}_e\right),$$

which is minimized at $\lambda = 0$ (since $\hat{\beta}_e$ is asymptotically efficient). The first order condition with respect to λ

$$2\lambda \text{Var}\left(\hat{\beta}_c\right) - 2(1-\lambda)\,\text{Var}\left(\hat{\beta}_e\right) + 2(1-2\lambda)\,\text{Cov}\left(\hat{\beta}_c, \hat{\beta}_e\right) = 0$$

should therefore be zero at $\lambda = 0$ so

$$\text{Var}\left(\hat{\beta}_e\right) = \text{Cov}\left(\hat{\beta}_c, \hat{\beta}_e\right).$$

(See Davidson (2000) 8.1) ■

This means that we can write

$$\begin{aligned}\operatorname{Var}\left(\hat{\beta}_e-\hat{\beta}_c\right) &= \operatorname{Var}\left(\hat{\beta}_e\right)+\operatorname{Var}\left(\hat{\beta}_c\right)-2\operatorname{Cov}\left(\hat{\beta}_e,\hat{\beta}_c\right)\\ &= \operatorname{Var}\left(\hat{\beta}_c\right)-\operatorname{Var}\left(\hat{\beta}_e\right).\end{aligned} \tag{5.28}$$

We can use this to test, for instance, if the estimates from least squares ($\hat{\beta}_e$, since LS is efficient if errors are iid normally distributed) and instrumental variable method ($\hat{\beta}_c$, since consistent even if the true residuals are correlated with the regressors) are the same. In this case, H_0 is that the true residuals are uncorrelated with the regressors.

All we need for this test are the point estimates and consistent estimates of the variance matrices. Testing one of the coefficient can be done by a t test, and testing all the parameters by a χ^2 test

$$\left(\hat{\beta}_e-\hat{\beta}_c\right)'\operatorname{Var}\left(\hat{\beta}_e-\hat{\beta}_c\right)^{-1}\left(\hat{\beta}_e-\hat{\beta}_c\right)\sim\chi^2(j), \tag{5.29}$$

where j equals the number of regressors that are potentially endogenous or measured with error. Note that the covariance matrix in (5.28) and (5.29) is likely to have a reduced rank, so the inverse needs to be calculated as a generalized inverse.

5.6 Tests of Overidentifying Restrictions in 2SLS*

When we use 2SLS, then we can test if instruments affect the dependent variable only via their correlation with the regressors. If not, something is wrong with the model since some relevant variables are excluded from the regression.

Bibliography

Davidson, J., 2000, *Econometric Theory*, Blackwell Publishers, Oxford.

Greene, W. H., 2000, *Econometric Analysis*, Prentice-Hall, Upper Saddle River, New Jersey, 4th edn.

Hamilton, J. D., 1994, *Time Series Analysis*, Princeton University Press, Princeton.

Hayashi, F., 2000, *Econometrics*, Princeton University Press.

Pindyck, R. S., and D. L. Rubinfeld, 1998, *Econometric Models and Economic Forecasts*, Irwin McGraw-Hill, Boston, Massachusetts, 4ed edn.

Verbeek, M., 2004, *A Guide to Modern Econometrics*, Wiley, Chichester, 2nd edn.

6 Simulating the Finite Sample Properties

Reference: Greene (2000) 5.3

Additional references: Cochrane (2001) 15.2; Davidson and MacKinnon (1993) 21; Davison and Hinkley (1997); Efron and Tibshirani (1993) (bootstrapping, chap 9 in particular); and Berkowitz and Kilian (2000) (bootstrapping in time series models)

We know the small sample properties of regression coefficients in linear models with fixed regressors (X is non-stochastic) and iid normal error terms. Monte Carlo Simulations and bootstrapping are two common techniques used to understand the small sample properties when these conditions are not satisfied.

6.1 Monte Carlo Simulations in the Simplest Case

Monte Carlo simulations is essentially a way to generate many artificial (small) samples from a parameterized model and then estimating the statistics on each of those samples. The distribution of the statistics is then used as the small sample distribution of the estimator.

The following is an example of how Monte Carlo simulations could be done in the special case of a linear model for a scalar dependent variable

$$y_t = x_t'\beta + u_t, \tag{6.1}$$

where u_t is iid $N(0, \sigma^2)$ and x_t is stochastic but independent of $u_{t\pm s}$ for all s. This means that x_t cannot include lags of y_t.

Suppose we want to find the small sample distribution of a function of the estimate, $g(\hat{\beta})$. To do a Monte Carlo experiment, we need information on *(i)* β; *(ii)* the variance of u_t, σ^2; *(iii)* and a process for x_t.

The process for x_t is typically estimated from the data on x_t. For instance, we could estimate the VAR system $x_t = A_1 x_{t-1} + A_2 x_{t-2} + e_t$. An alternative is to take an actual sample of x_t and repeat it.

The values of β and σ^2 are often a mix of estimation results and theory. In some case, we simply take the point estimates. In other cases, we adjust the point estimates so that $g(\beta) = 0$ holds, that is, so you *simulate the model under the null hypothesis* in order to study the size of asymptotic tests and to find valid critical values for small samples. Alternatively, you may *simulate the model under an alternative hypothesis* in order to study the power of the test using either critical values from either the asymptotic distribution or from a (perhaps simulated) small sample distribution.

To make it a bit concrete, suppose you want to use these simulations to get a 5% critical value for testing the null hypothesis $g(\beta) = 0$. The Monte Carlo experiment follows these steps.

1. (a) Construct an artificial sample of the regressors (see above), $\{\tilde{x}_t\}_{t=1}^T$.

   (b) Draw random numbers $\{\tilde{u}_t\}_{t=1}^T$ and use those together with the artificial sample of $\tilde{x}_t$ to calculate an artificial sample $\{\tilde{y}_t\}_{t=1}^T$ by using (6.1). Calculate an estimate $\hat{\beta}$ and record it along with the value of $g(\hat{\beta})$ and perhaps also the test statistics of the hypothesis that $g(\beta) = 0$.

2. Repeat the previous steps N (3000, say) times. The more times you repeat, the better is the approximation of the small sample distribution.

3. Sort your simulated $\hat{\beta}$, $g(\hat{\beta})$, and the test statistics in ascending order. For a one-sided test (for instance, a chi-square test), take the $(0.95N)$th observations in these sorted vector as your 5% critical values. For a two-sided test (for instance, a t-test), take the $(0.025N)$th and $(0.975N)$th observations as the 5% critical values. You can also record how many times the 5% critical values from the asymptotic distribution would reject a true null hypothesis.

4. You may also want to plot a histogram of $\hat{\beta}$, $g(\hat{\beta})$, and the test statistics to see if there is a small sample bias, and how the distribution looks like. Is it close to normal? How wide is it?

5. See *Figure 6.1* for an example.

Remark 6.1 *(Generating $N(\mu, \Sigma)$ random numbers) Suppose you want to draw an $n \times 1$ vector ε_t of $N(\mu, \Sigma)$ variables. Use the Cholesky decomposition to calculate the lower triangular P such that $\Sigma = PP'$ (note that Gauss and MatLab returns P' instead of*

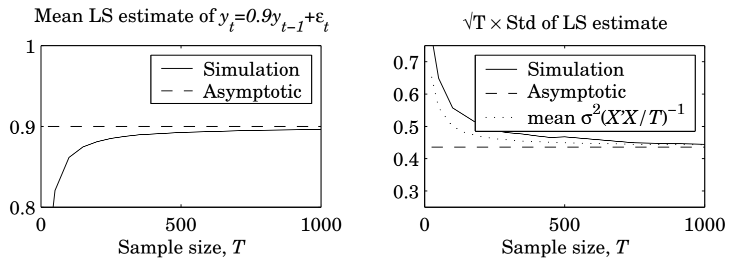


Figure 6.1: Results from a Monte Carlo experiment of LS estimation of the AR coefficient. Data generated by an AR(1) process, 5000 simulations.

P). Draw u_t from an $N(0, I)$ distribution (randn in MatLab, rndn in Gauss), and define $\varepsilon_t = \mu + Pu_t$. Note that $\text{Cov}(\varepsilon_t) = \text{E}\, Pu_tu_t'P' = PIP' = \Sigma$.

6.2 Monte Carlo Simulations in More Complicated Cases*

6.2.1 When x_t Includes Lags of y_t

If x_t contains lags of y_t, then we must set up the simulations so that feature is preserved in every artificial sample that we create. For instance, suppose x_t includes y_{t-1} and another vector z_t of variables which are independent of $u_{t\pm s}$ for all s. We can then generate an artificial sample as follows. First, create a sample $\{\tilde{z}_t\}_{t=1}^{T}$ by some time series model or by taking the observed sample itself (as we did with x_t in the simplest case). Second, observation t of $\{\tilde{x}_t, \tilde{y}_t\}$ is generated as

$$\tilde{x}_t = \begin{bmatrix} \tilde{y}_{t-1} \\ \tilde{z}_t \end{bmatrix} \text{ and } \tilde{y}_t = \tilde{x}_t'\beta + u_t, \tag{6.2}$$

which is repeated for $t = 1, ..., T$. We clearly need the initial value y_0 to start up the artificial sample, so one observation from the original sample is lost.

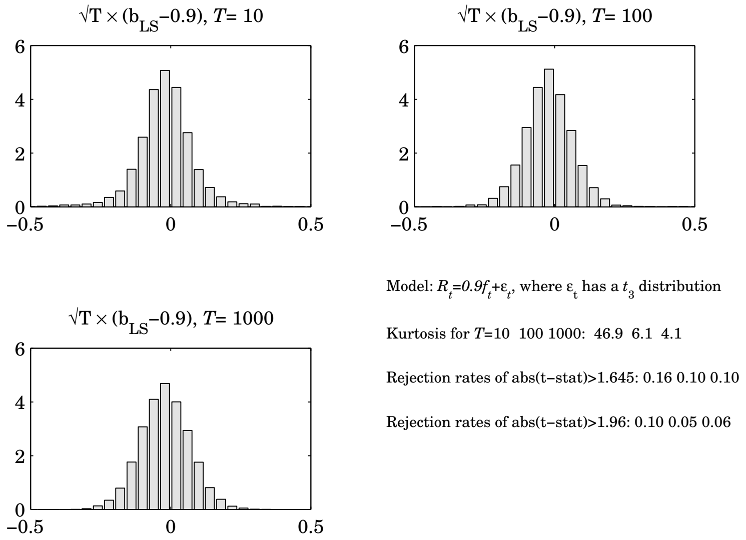


Figure 6.2: Results from a Monte Carlo experiment with thick-tailed errors. The regressor is iid normally distributed. The errors have a t_3-distribution, 5000 simulations.

6.2.2 More Complicated Errors

It is straightforward to sample the errors from other distributions than the normal, for instance, a uniform distribution. Equipped with uniformly distributed random numbers, you can always (numerically) invert the cumulative distribution function (cdf) of any distribution to generate random variables from any distribution by using the probability transformation method. See *Figure 6.2* for an example.

Remark 6.2 *Let $X \sim U(0,1)$ and consider the transformation $Y = F^{-1}(X)$, where $F^{-1}()$ is the inverse of a strictly increasing cdf F, then Y has the CDF $F()$. (Proof: follows from the lemma on change of variable in a density function.)*

Example 6.3 *The exponential cdf is $x = 1-\exp(-\theta y)$ with inverse $y = -\ln(1-x)/\theta$. Draw x from $U(0.1)$ and transform to y to get an exponentially distributed variable.*

It is more difficult to handle non-iid errors, for instance, heteroskedasticity and autocorrelation. We then need to model the error process and generate the errors from that model. For instance, if the errors are assumed to follow an AR(2) process, then we could estimate that process from the errors in (6.1) and then generate artificial samples of errors.

6.3 Bootstrapping in the Simplest Case

Bootstrapping is another way to do simulations, where we construct artificial samples by sampling from the actual data. The advantage of the bootstrap is then that we do not have to try to estimate the process of the errors and regressors as we must do in a Monte Carlo experiment. The real benefit of this is that we do not have to make any strong assumption about the distribution of the errors.

The bootstrap approach works particularly well when the errors are iid and independent of x_{t-s} for all s. This means that x_t cannot include lags of y_t. We here consider bootstrapping the linear model (6.1), for which we have point estimates (perhaps from LS) and fitted residuals. The procedure is similar to the Monte Carlo approach, except that the artificial sample is generated differently. In particular, Step 1 in the Monte Carlo simulation is replaced by the following:

1. Construct an artificial sample $\{\tilde{y}_t\}_{t=1}^{T}$ by

$$\tilde{y}_t = x_t'\beta + \tilde{u}_t, \tag{6.3}$$

where $\tilde{u}_t$ is drawn (with replacement) from the fitted residual and where β is the point estimate. Calculate an estimate $\hat{\beta}$ and record it along with the value of $g(\hat{\beta})$ and perhaps also the test statistics of the hypothesis that $g(\beta) = 0$.

6.4 Bootstrapping in More Complicated Cases*

6.4.1 Case 2: Errors are iid but Correlated With x_{t+s}

When x_t contains lagged values of y_t, then we have to modify the approach in (6.3) since $\tilde{u}_t$ can become correlated with x_t. For instance, if x_t includes y_{t-1} and we happen to sample $\tilde{u}_t = \hat{u}_{t-1}$, then we get a non-zero correlation. The easiest way to handle this is as in the Monte Carlo simulations: replace any y_{t-1} in x_t by $\tilde{y}_{t-1}$, that is, the corresponding observation in the artificial sample.

6.4.2 Case 3: Errors are Heteroskedastic but Uncorrelated with of $x_{t\pm s}$

Case 1 and 2 both draw errors randomly—based on the assumption that the errors are iid. Suppose instead that the errors are heteroskedastic, but still serially uncorrelated. We know that if the heteroskedastcity is related to the regressors, then the traditional LS covariance matrix is not correct (this is the case that White's test for heteroskedasticity tries to identify). It would then be wrong it pair x_t with just any $\hat{u}_s$ since that destroys the relation between x_t and the variance of u_t.

An alternative way of bootstrapping can then be used: generate the artificial sample by drawing (with replacement) *pairs* (y_s, x_s), that is, we let the artificial pair in t be $(\tilde{y}_t, \tilde{x}_t) = (x_s'\hat{\beta}_0 + \hat{u}_s, x_s)$ for some random draw of s so we are always pairing the residual, $\hat{u}_s$, with the contemporaneous regressors, x_s. Note that is we are always sampling with replacement—otherwise the approach of drawing pairs would be just re-create the original data set. For instance, if the data set contains 3 observations, then artificial sample could be

$$\begin{bmatrix} (\tilde{y}_1, \tilde{x}_1) \\ (\tilde{y}_2, \tilde{x}_2) \\ (\tilde{y}_3, \tilde{x}_3) \end{bmatrix} = \begin{bmatrix} (x_2'\hat{\beta}_0 + \hat{u}_2, x_2) \\ (x_3'\hat{\beta}_0 + \hat{u}_3, x_3) \\ (x_3'\hat{\beta}_0 + \hat{u}_3, x_3) \end{bmatrix}$$

In contrast, when we sample (with replacement) $\hat{u}_s$, as we did above, then an artificial sample could be

$$\begin{bmatrix} (\tilde{y}_1, \tilde{x}_1) \\ (\tilde{y}_2, \tilde{x}_2) \\ (\tilde{y}_3, \tilde{x}_3) \end{bmatrix} = \begin{bmatrix} (x_1'\hat{\beta}_0 + \hat{u}_2, x_1) \\ (x_2'\hat{\beta}_0 + \hat{u}_1, x_2) \\ (x_3'\hat{\beta}_0 + \hat{u}_2, x_3) \end{bmatrix}.$$

Davidson and MacKinnon (1993) argue that bootstrapping the pairs (y_s, x_s) makes little sense when x_s contains lags of y_s, since there is no way to construct lags of y_s in the bootstrap. However, what is important for the estimation is sample averages of various functions of the dependent and independent variable within a period—not how the line up over time (provided the assumption of no autocorrelation of the residuals is true).

6.4.3 Other Approaches

There are many other ways to do bootstrapping. For instance, we could sample the regressors and residuals independently of each other and construct an artificial sample of the dependent variable $\tilde{y}_t = \tilde{x}_t'\hat{\beta} + \tilde{u}_t$. This clearly makes sense if the residuals and regressors are independent of each other and errors are iid. In that case, the advantage of this approach is that we do not keep the regressors fixed.

6.4.4 Serially Dependent Errors

It is quite hard to handle the case when the errors are serially dependent, since we must the sample in such a way that we do not destroy the autocorrelation structure of the data. A common approach is to fit a model for the residuals, for instance, an AR(1), and then bootstrap the (hopefully iid) innovations to that process.

Another approach amounts to resampling of blocks of data. For instance, suppose the sample has 10 observations, and we decide to create blocks of 3 observations. The first block is $(\hat{u}_1, \hat{u}_2, \hat{u}_3)$, the second block is $(\hat{u}_2, \hat{u}_3, \hat{u}_4)$, and so forth until the last block, $(\hat{u}_8, \hat{u}_9, \hat{u}_{10})$. If we need a sample of length 3τ, say, then we simply draw τ of those block randomly (with replacement) and stack them to form a longer series. To handle end point effects (so that all data points have the same probability to be drawn), we also create blocks by "wrapping" the data around a circle. In practice, this means that we add a the following blocks: $(\hat{u}_{10}, \hat{u}_1, \hat{u}_2)$ and $(\hat{u}_9, \hat{u}_{10}, \hat{u}_1)$. An alternative approach is to have non-overlapping blocks. See Berkowitz and Kilian (2000) for some other recent methods.

Bibliography

Berkowitz, J., and L. Kilian, 2000, "Recent Developments in Bootstrapping Time Series," *Econometric-Reviews*, 19, 1–48.

Cochrane, J. H., 2001, *Asset Pricing*, Princeton University Press, Princeton, New Jersey.

Davidson, R., and J. G. MacKinnon, 1993, *Estimation and Inference in Econometrics*, Oxford University Press, Oxford.

Davison, A. C., and D. V. Hinkley, 1997, *Bootstrap Methods and Their Applications*, Cambridge University Press.

Efron, B., and R. J. Tibshirani, 1993, *An Introduction to the Bootstrap*, Chapman and Hall, New York.

Greene, W. H., 2000, *Econometric Analysis*, Prentice-Hall, Upper Saddle River, New Jersey, 4th edn.

7 GMM

References: Greene (2000) 4.7 and 11.5-6

Additional references: Hayashi (2000) 3-4; Verbeek (2004) 5; Hamilton (1994) 14; Ogaki (1993), Johnston and DiNardo (1997) 10; Harris and Matyas (1999); Pindyck and Rubinfeld (1998) Appendix 10.1; Cochrane (2001) 10-11

7.1 Method of Moments

Let $m(x_t)$ be a $k \times 1$ vector valued continuous function of a stationary process, and let the probability limit of the mean of $m(.)$ be a function $\gamma(.)$ of a $k \times 1$ vector β of parameters. We want to estimate β. The method of moments (MM, not yet generalized to GMM) estimator is obtained by replacing the probability limit with the sample mean and solving the system of k equations

$$\frac{1}{T}\sum_{t=1}^{T} m(x_t) - \gamma(\beta) = \mathbf{0}_{k\times 1} \tag{7.1}$$

for the parameters β.

It is clear that this is a consistent estimator of β if γ is continuous. (Proof: the sample mean is a consistent estimator of $\gamma(.)$, and by Slutsky's theorem $\operatorname{plim}\gamma(\hat{\beta}) = \gamma(\operatorname{plim}\hat{\beta})$ if γ is a continuous function.)

Example 7.1 *(Moment conditions for variances and covariance) Suppose the series x_t and y_t have zero means. The following moment conditions define the traditional variance and covariance estimators*

$$\frac{1}{T}\sum\nolimits_{t=1}^{T} x_t^2 - \sigma_{xx} = 0$$
$$\frac{1}{T}\sum\nolimits_{t=1}^{T} y_t^2 - \sigma_{yy} = 0$$
$$\frac{1}{T}\sum\nolimits_{t=1}^{T} x_t y_t - \sigma_{xy} = 0.$$

It does not matter if the parameterers are estimated separately or jointly. In contrast, if we want the correlation, ρ_{xy}, instead of the covariance, then we change the last moment condition to

$$\frac{1}{T}\sum\nolimits_{t=1}^{T} x_t y_t - \rho_{xy}\sqrt{\sigma_{xx}}\sqrt{\sigma_{yy}} = 0,$$

which must be estimated jointly with the first two conditions.

Example 7.2 *(MM for an MA(1).) For an MA(1), $y_t = \epsilon_t + \theta\epsilon_{t-1}$, we have*

$$\begin{aligned} Ey_t^2 &= E(\epsilon_t + \theta\epsilon_{t-1})^2 &&= \sigma_\epsilon^2\left(1+\theta^2\right) \\ E(y_t y_{t-1}) &= E\left[(\epsilon_t + \theta\epsilon_{t-1})(\epsilon_{t-1} + \theta\epsilon_{t-2})\right] &&= \sigma_\epsilon^2\theta. \end{aligned}$$

The moment conditions could therefore be

$$\begin{bmatrix} \frac{1}{T}\sum_{t=1}^{T} y_t^2 - \sigma_\epsilon^2\left(1+\theta^2\right) \\ \frac{1}{T}\sum_{t=1}^{T} y_t y_{t-1} - \sigma_\epsilon^2\theta \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

which allows us to estimate θ and σ^2.

7.2 Generalized Method of Moments

GMM extends MM by allowing for more orthogonality conditions than parameters. This could, for instance, increase efficiency and/or provide new aspects which can be tested.

Many (most) traditional estimation methods, like LS, IV, and MLE are special cases of GMM. This means that the properties of GMM are very general, and therefore fairly difficult to prove.

7.3 Moment Conditions in GMM

Suppose we have q (unconditional) moment conditions,

$$\begin{aligned} \mathrm{E}m(w_t,\beta_0) &= \begin{bmatrix} \mathrm{E}m_1(w_t,\beta_0) \\ \vdots \\ \mathrm{E}m_q(w_t,\beta_0) \end{bmatrix} \\ &= \mathbf{0}_{q\times 1}, \end{aligned} \tag{7.2}$$

from which we want to estimate the $k \times 1$ ($k \leq q$) vector of parameters, β. The true values are β_0. We assume that w_t is a stationary and ergodic (vector) process (otherwise the sample means does not converge to anything meaningful as the sample size increases). The sample averages, or "sample moment conditions," evaluated at some value of β, are

$$\bar{m}(\beta) = \frac{1}{T}\sum_{t=1}^{T} m(w_t, \beta). \tag{7.3}$$

The sample average $\bar{m}(\beta)$ is a vector of functions of random variables, so they are random variables themselves and depend on the sample used. It will later be interesting to calculate the variance of $\bar{m}(\beta)$. Note that $\bar{m}(\beta_1)$ and $\bar{m}(\beta_2)$ are sample means obtained by using two different parameter vectors, but on the same sample of data.

Example 7.3 *(Moments conditions for IV/2SLS.) Consider the linear model $y_t = x_t'\beta_0 + u_t$, where x_t and β are $k \times 1$ vectors. Let z_t be a $q \times 1$ vector, with $q \geq k$. The moment conditions and their sample analogues are*

$$\mathbf{0}_{q\times 1} = Ez_t u_t = E[z_t(y_t - x_t'\beta_0)], \text{ and } \bar{m}(\beta) = \frac{1}{T}\sum_{t=1}^{T} z_t(y_t - x_t'\beta),$$

(or $Z'(Y - X\beta)/T$ in matrix form). Let $q = k$ to get IV; let $z_t = x_t$ to get LS.

Example 7.4 *(Moments conditions for MLE.) The maximum likelihood estimator maximizes the log likelihood function, $\frac{1}{T}\Sigma_{t=1}^{T}\ln L(w_t;\beta)$, which requires $\frac{1}{T}\Sigma_{t=1}^{T}\partial \ln L(w_t;\beta)/\partial\beta = 0$. A key regularity condition for the MLE is that $E\partial \ln L(w_t;\beta_0)/\partial\beta = 0$, which is just like a GMM moment condition.*

7.3.1 Digression: From Conditional to Unconditional Moment Conditions

Suppose we are instead given *conditional* moment restrictions

$$\mathrm{E}\left[u(x_t, \beta_0)|z_t\right] = \mathbf{0}_{m\times 1}, \tag{7.4}$$

where z_t is a vector of conditioning (predetermined) variables. We want to transform this to unconditional moment conditions.

Remark 7.5 *($E(u|z) = 0$ versus $Euz = 0$.) For any random variables u and z,*

$$Cov\left(z, u\right) = Cov\left[z, E\left(u|z\right)\right].$$

The condition $E(u|z) = 0$ then implies $Cov(z,u) = 0$. Recall that $Cov(z,u) = Ezu - EzEu$, and that $E(u|z) = 0$ implies that $Eu = 0$ (by iterated expectations). We therefore get that

$$E\left(u|z\right) = 0 \Rightarrow \begin{bmatrix} Cov\left(z,u\right) = 0 \\ Eu = 0 \end{bmatrix} \Rightarrow Euz = 0.$$

Example 7.6 *(Euler equation for optimal consumption.) The standard Euler equation for optimal consumption choice which with isoelastic utility $U\left(C_t\right) = C_t^{1-\gamma}/\left(1-\gamma\right)$ is*

$$E\left[R_{t+1}\beta\left(\frac{C_{t+1}}{C_t}\right)^{-\gamma} - 1\middle|\Omega_t\right] = 0,$$

where R_{t+1} is a gross return on an investment and Ω_t is the information set in t. Let $z_t \in \Omega_t$, for instance asset returns or consumption t or earlier. The Euler equation then implies

$$E\left[R_{t+1}\beta\left(\frac{C_{t+1}}{C_t}\right)^{-\gamma} z_t - z_t\right] = 0.$$

Let $z_t = (z_{1t}, ..., z_{nt})'$, and define the new (unconditional) moment conditions as

$$m(w_t, \beta) = u(x_t, \beta) \otimes z_t = \begin{bmatrix} u_1(x_t,\beta)z_{1t} \\ u_1(x_t,\beta)z_{2t} \\ \vdots \\ u_1(x_t,\beta)z_{nt} \\ u_2(x_t,\beta)z_{1t} \\ \vdots \\ u_m(x_t,\beta)z_{nt} \end{bmatrix}_{q\times 1}, \tag{7.5}$$

which by (7.4) must have an expected value of zero, that is

$$\mathrm{E}m(w_t, \beta_0) = \mathbf{0}_{q\times 1}. \tag{7.6}$$

This a set of unconditional moment conditions—just as in (7.2). The sample moment conditions (7.3) are therefore valid also in the conditional case, although we have to specify $m(w_t, \beta)$ as in (7.5).

Note that the choice of instruments is often arbitrary: it often amounts to using only a subset of the information variables. GMM is often said to be close to economic theory, but it should be admitted that economic theory sometimes tells us fairly little about which instruments, z_t, to use.

Example 7.7 *(Euler equation for optimal consumption, continued) The orthogonality conditions from the consumption Euler equations in Example 7.6 are highly non-linear, and theory tells us very little about how the prediction errors are distributed. GMM has the advantage of using the theoretical predictions (moment conditions) with a minimum of distributional assumptions. The drawback is that it is sometimes hard to tell exactly which features of the (underlying) distribution that are tested.*

7.4 The Optimization Problem in GMM

7.4.1 The Loss Function

The GMM estimator $\hat{\beta}$ minimizes the weighted quadratic form

$$J = \begin{bmatrix} \bar{m}_1(\beta) \\ \vdots \\ \vdots \\ \bar{m}_q(\beta) \end{bmatrix}' \begin{bmatrix} W_{11} & \cdots & \cdots & W_{1q} \\ \vdots & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ W_{1q} & \cdots & \cdots & W_{qq} \end{bmatrix} \begin{bmatrix} \bar{m}_1(\beta) \\ \vdots \\ \vdots \\ \bar{m}_q(\beta) \end{bmatrix} \tag{7.7}$$

$$= \bar{m}(\beta)' W \bar{m}(\beta), \tag{7.8}$$

where $\bar{m}(\beta)$ is the sample average of $m(w_t, \beta)$ given by (7.3), and where W is some $q \times q$ symmetric positive definite weighting matrix. (We will soon discuss a good choice of weighting matrix.) There are k parameters in β to estimate, and we have q moment conditions in $\bar{m}(\beta)$. We therefore have $q - k$ *overidentifying moment restrictions*.

With $q = k$ the model is exactly identified (as many equations as unknowns), and it should be possible to set all q sample moment conditions to zero by a choosing the $k = q$ parameters. It is clear that the choice of the weighting matrix has no effect in this case since $\bar{m}(\hat{\beta}) = 0$ at the point estimates $\hat{\beta}$.

Example 7.8 *(Simple linear regression.) Consider the model*

$$y_t = x_t \beta_0 + u_t, \tag{7.9}$$

where y_t and x_t are zero mean scalars. The moment condition and loss function are

$$\bar{m}(\beta) = \frac{1}{T}\sum_{t=1}^{T} x_t(y_t - x_t\beta) \text{ and}$$

$$J = W\left[\frac{1}{T}\sum_{t=1}^{T} x_t(y_t - x_t\beta)\right]^2,$$

so the scalar W is clearly irrelevant in this case.

Example 7.9 *(IV/2SLS method continued.) From Example 7.3, we note that the loss function for the IV/2SLS method is*

$$\bar{m}(\beta)' W \bar{m}(\beta) = \left[\frac{1}{T}\sum_{t=1}^{T} z_t(y_t - x_t'\beta)\right]' W \left[\frac{1}{T}\sum_{t=1}^{T} z_t(y_t - x_t'\beta)\right].$$

When $q = k$, then the model is exactly identified, so the estimator could actually be found by setting all moment conditions to zero. We then get the IV estimator

$$\mathbf{0} = \frac{1}{T}\sum_{t=1}^{T} z_t(y_t - x_t'\hat{\beta}_{IV}) \text{ or}$$

$$\hat{\beta}_{IV} = \left(\frac{1}{T}\sum_{t=1}^{T} z_t x_t'\right)^{-1} \frac{1}{T}\sum_{t=1}^{T} z_t y_t$$

$$= \hat{\Sigma}_{zx}^{-1}\hat{\Sigma}_{zy},$$

where $\hat{\Sigma}_{zx} = \Sigma_{t=1}^{T} z_t x_t'/T$ and similarly for the other second moment matrices. Let $z_t = x_t$ to get LS

$$\hat{\beta}_{LS} = \hat{\Sigma}_{xx}^{-1}\hat{\Sigma}_{xy}.$$

7.4.2 First Order Conditions

Remark 7.10 *(Matrix differentiation of non-linear functions.) Let the vector $y_{n\times 1}$ be a function of the vector $x_{m\times 1}$*

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = f(x) = \begin{bmatrix} f_1(x) \\ \vdots \\ f_n(x) \end{bmatrix}.$$

Then, $\partial y/\partial x'$ is an $n \times m$ matrix

$$\frac{\partial y}{\partial x'} = \begin{bmatrix} \frac{\partial f_1(x)}{\partial x'} \\ \vdots \\ \frac{\partial f_1(x)}{\partial x'} \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1(x)}{\partial x_1} & \cdots & \frac{\partial f_1(x)}{\partial x_m} \\ \vdots & & \vdots \\ \frac{\partial f_n(x)}{\partial x_1} & \cdots & \frac{\partial f_n(x)}{\partial x_m} \end{bmatrix}.$$

(Note that the notation implies that the derivatives of the first element in y, denoted y_1, with respect to each of the elements in x' are found in the first row of $\partial y/\partial x'$. A rule to help memorizing the format of $\partial y/\partial x'$: y is a column vector and x' is a row vector.)

Remark 7.11 *When $y = Ax$ where A is an $n \times m$ matrix, then $f_i(x)$ in Remark 7.10 is a linear function. We then get $\partial y/\partial x' = \partial(Ax)/\partial x' = A$.*

Remark 7.12 *As a special case of the previous remark $y = z'x$ where both z and x are vectors. Then $\partial(z'x)/\partial x' = z'$ (since z' plays the role of A).*

Remark 7.13 *(Matrix differentiation of quadratic forms.) Let $x_{n\times 1}$, $f(x)_{m\times 1}$, and $A_{m\times m}$ symmetric. Then*

$$\frac{\partial f(x)' A f(x)}{\partial x} = 2\left(\frac{\partial f(x)}{\partial x'}\right)' A f(x).$$

Remark 7.14 *If $f(x) = x$, then $\partial f(x)/\partial x' = I$, so $\partial(x'Ax)/\partial x = 2Ax$.*

The k first order conditions for minimizing the GMM loss function in (7.8) with respect to the k parameters are that the partial derivatives with respect to β equal zero at the estimate, $\hat{\beta}$,

$$\begin{aligned} \mathbf{0}_{k\times 1} &= \frac{\partial \bar{m}(\hat{\beta})' W \bar{m}(\hat{\beta})}{\partial \beta} \\ &= \begin{bmatrix} \frac{\partial \bar{m}_1(\hat{\beta})}{\partial \beta_1} & \cdots & \frac{\partial \bar{m}_1(\hat{\beta})}{\partial \beta_k} \\ \vdots & & \vdots \\ \vdots & & \vdots \\ \frac{\partial \bar{m}_q(\hat{\beta})}{\partial \beta_1} & \cdots & \frac{\partial \bar{m}_q(\hat{\beta})}{\partial \beta_k} \end{bmatrix}' \begin{bmatrix} W_{11} & \cdots & \cdots & W_{1q} \\ \vdots & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ W_{1q} & \cdots & \cdots & W_{qq} \end{bmatrix} \begin{bmatrix} \bar{m}_1(\hat{\beta}) \\ \vdots \\ \vdots \\ \bar{m}_q(\hat{\beta}) \end{bmatrix} \text{ (with } \hat{\beta}_{k\times 1}\text{)}, \qquad (7.10) \\ &= \underbrace{\left(\frac{\partial \bar{m}(\hat{\beta})}{\partial \beta'}\right)'}_{k\times q} \underbrace{W}_{q\times q} \underbrace{\bar{m}(\hat{\beta})}_{q\times 1}. \qquad (7.11) \end{aligned}$$

We can solve for the GMM estimator, $\hat{\beta}$, from (7.11). This set of equations must often be solved by numerical methods, except in linear models (the moment conditions are linear functions of the parameters) where we can find analytical solutions by matrix inversion.

Example 7.15 *(First order conditions of simple linear regression.) The first order conditions of the loss function in Example 7.8 is*

$$\begin{aligned} 0 &= \frac{d}{d\beta} W \left[\frac{1}{T}\sum_{t=1}^{T} x_t(y_t - x_t\hat{\beta})\right]^2 \\ &= \left[-\frac{1}{T}\sum_{t=1}^{T} x_t^2\right] W \left[\frac{1}{T}\sum_{t=1}^{T} x_t(y_t - x_t\hat{\beta})\right], \text{ or} \\ \hat{\beta} &= \left(\frac{1}{T}\sum_{t=1}^{T} x_t^2\right)^{-1} \frac{1}{T}\sum_{t=1}^{T} x_t y_t. \end{aligned}$$

Example 7.16 *(First order conditions of IV/2SLS.) The first order conditions corresponding to (7.11) of the loss function in Example 7.9 (when $q \geq k$) are*

$$\begin{aligned} \mathbf{0}_{k\times 1} &= \left[\frac{\partial \bar{m}(\hat{\beta})}{\partial \beta'}\right]' W \bar{m}(\hat{\beta}) \\ &= \left[\frac{\partial}{\partial \beta'}\frac{1}{T}\sum_{t=1}^{T} z_t(y_t - x_t'\hat{\beta})\right]' W \frac{1}{T}\sum_{t=1}^{T} z_t(y_t - x_t'\hat{\beta}) \\ &= \left[-\frac{1}{T}\sum_{t=1}^{T} z_t x_t'\right]' W \frac{1}{T}\sum_{t=1}^{T} z_t(y_t - x_t'\hat{\beta}) \\ &= -\hat{\Sigma}_{xz} W (\hat{\Sigma}_{zy} - \hat{\Sigma}_{zx}\hat{\beta}). \end{aligned}$$

We can solve for $\hat{\beta}$ from the first order conditions as

$$\hat{\beta}_{2SLS} = \left(\hat{\Sigma}_{xz} W \hat{\Sigma}_{zx}\right)^{-1} \hat{\Sigma}_{xz} W \hat{\Sigma}_{zy}.$$

When $q = k$, then the first order conditions can be premultiplied with $(\hat{\Sigma}_{xz} W)^{-1}$, since $\hat{\Sigma}_{xz} W$ is an invertible $k \times k$ matrix in this case, to give

$$\mathbf{0}_{k\times 1} = \hat{\Sigma}_{zy} - \hat{\Sigma}_{zx}\hat{\beta}, \text{ so } \hat{\beta}_{IV} = \hat{\Sigma}_{zx}^{-1}\hat{\Sigma}_{zy}.$$

This shows that the first order conditions are just the same as the sample moment condi-

tions, which can be made to hold exactly since there are as many parameters as there are equations.

7.5 Asymptotic Properties of GMM

We know very little about the general small sample properties, including bias, of GMM. We therefore have to rely either on simulations (Monte Carlo or bootstrap) or on the asymptotic results. This section is about the latter.

GMM estimates are typically consistent and normally distributed, even if the series $m(w_t, \beta)$ in the moment conditions (7.3) are serially correlated and heteroskedastic—provided w_t is a stationary and ergodic process. The reason is essentially that the estimators are (at least as a first order approximation) linear combinations of sample means which typically are consistent (LLN) and normally distributed (CLT). More about that later. The proofs are hard, since the GMM is such a broad class of estimators. This section discusses, in an informal way, how we can arrive at those results.

7.5.1 Consistency

Sample moments are typically consistent, so $\text{plim}\, m(\beta) = \text{E}\, m(w_t, \beta)$. This must hold at any parameter vector in the relevant space (for instance, those inducing stationarity and variances which are strictly positive). Then, if the moment conditions (7.2) are true only at the true parameter vector, β_0, (otherwise the parameters are "unidentified") and that they are continuous in β, then GMM is consistent. The idea is thus that GMM asymptotically solves

$$\begin{aligned}\mathbf{0}_{q\times 1} &= \text{plim}\, \bar{m}(\hat{\beta}) \\ &= \text{E}\, m(w_t, \hat{\beta}),\end{aligned}$$

which only holds at $\hat{\beta} = \beta_0$. Note that this is an application of Slutsky's theorem.

Remark 7.17 *(Slutsky's theorem.) If $\{x_T\}$ is a sequence of random matrices such that* $\text{plim}\, x_T = x$ *and $g(x_T)$ a continuous function, then* $\text{plim}\, g(x_T) = g(x)$.

Example 7.18 *(Consistency of 2SLS.) By using $y_t = x_t'\beta_0 + u_t$, the first order conditions in Example 7.16 can be rewritten*

$$\begin{aligned}\mathbf{0}_{k\times 1} &= \hat{\Sigma}_{xz} W \frac{1}{T}\sum_{t=1}^{T} z_t (y_t - x_t'\hat{\beta}) \\ &= \hat{\Sigma}_{xz} W \frac{1}{T}\sum_{t=1}^{T} z_t \left[u_t + x_t'\left(\beta_0 - \hat{\beta}\right)\right] \\ &= \hat{\Sigma}_{xz} W \hat{\Sigma}_{zu} + \hat{\Sigma}_{xz} W \hat{\Sigma}_{zx}\left(\beta_0 - \hat{\beta}\right).\end{aligned}$$

Take the probability limit

$$\mathbf{0}_{k\times 1} = \text{plim}\, \hat{\Sigma}_{xz} W \,\text{plim}\, \hat{\Sigma}_{zu} + \text{plim}\, \hat{\Sigma}_{xz} W \,\text{plim}\, \hat{\Sigma}_{zx}\left(\beta_0 - \text{plim}\, \hat{\beta}\right).$$

In most cases, $\text{plim}\, \hat{\Sigma}_{xz}$ *is some matrix of constants, and* $\text{plim}\, \hat{\Sigma}_{zu} = \text{E}\, z_t u_t = \mathbf{0}_{q\times 1}$. *It then follows that* $\text{plim}\, \hat{\beta} = \beta_0$. *Note that the whole argument relies on that the moment condition, $\text{E}\, z_t u_t = \mathbf{0}_{q\times 1}$, is true. If it is not, then the estimator is inconsistent. For instance, when the instruments are invalid (correlated with the residuals) or when we use LS ($z_t = x_t$) when there are measurement errors or in a system of simultaneous equations.*

7.5.2 Asymptotic Normality

To give the asymptotic distribution of $\sqrt{T}(\hat{\beta} - \beta_0)$, we need to define three things. (As usual, we also need to scale with $\sqrt{T}$ to get a non-trivial asymptotic distribution; the asymptotic distribution of $\hat{\beta} - \beta_0$ is a spike at zero.) *First*, let S_0 (a $q \times q$ matrix) denote the asymptotic covariance matrix (as sample size goes to infinity) of $\sqrt{T}$ times the sample moment conditions evaluated at the true parameters

$$\begin{aligned}S_0 &= \text{ACov}\left[\sqrt{T}\bar{m}(\beta_0)\right] && (7.12) \\ &= \text{ACov}\left[\frac{1}{\sqrt{T}}\sum_{t=1}^{T} m(w_t, \beta_0)\right], && (7.13)\end{aligned}$$

where we use the definition of $\bar{m}(\beta_0)$ in (7.3). (To estimate S_0 it is important to recognize that it is a scaled sample average.) Let $R(s)$ be the $q \times q$ covariance (matrix) of the vector

$m(w_t, \beta_0)$ with the vector $m(w_{t-2}, \beta_0)$

$$\begin{aligned} R(s) &= \text{Cov}\left[m(w_t, \beta_0), m(w_{t-s}, \beta_0)\right] \\ &= \text{E}\, m(w_t, \beta_0) m(w_{t-s}, \beta_0)'. \end{aligned} \tag{7.14}$$

Then, it is well known that

$$\text{ACov}\left[\sqrt{T}\bar{m}\left(\beta_0\right)\right] = \sum_{s=-\infty}^{\infty} R(s). \tag{7.15}$$

In practice, we often estimate this by using the Newey-West estimator (or something similar).

Second, let D_0 (a $q \times k$ matrix) denote the probability limit of the gradient of the sample moment conditions with respect to the parameters, evaluated at the true parameters

$$D_0 = \text{plim}\, \frac{\partial \bar{m}(\beta_0)}{\partial \beta'}, \text{ where} \tag{7.16}$$

$$\frac{\partial \bar{m}(\beta_0)}{\partial \beta'} = \begin{bmatrix} \frac{\partial \bar{m}_1(\beta)}{\partial \beta_1} & \cdots & \frac{\partial \bar{m}_1(\beta)}{\partial \beta_k} \\ \vdots & & \vdots \\ \vdots & & \vdots \\ \frac{\partial \bar{m}_q(\beta)}{\partial \beta_1} & \cdots & \frac{\partial \bar{m}_q(\beta)}{\partial \beta_k} \end{bmatrix} \text{ at the true } \beta \text{ vector.} \tag{7.17}$$

Note that a similar gradient, but evaluated at $\hat{\beta}$, also shows up in the first order conditions (7.11). *Third*, let the weighting matrix be the inverse of the covariance matrix of the moment conditions (once again evaluated at the true parameters)

$$W = S_0^{-1}. \tag{7.18}$$

It can be shown that this choice of weighting matrix gives the asymptotically most efficient estimator for a *given* set of orthogonality conditions. For instance, in 2SLS, this means a given set of instruments and (7.18) then shows only how to use these instruments in the most efficient way. Of course, another set of instruments might be better (in the sense of giving a smaller Cov($\hat{\beta}$)).

With the definitions in (7.12) and (7.16) and the choice of weighting matrix in (7.18) and the added assumption that the rank of D_0 equals k (number of parameters) then we can show (under fairly general conditions) that

$$\sqrt{T}(\hat{\beta} - \beta_0) \xrightarrow{d} N(\mathbf{0}_{k\times 1}, V), \text{ where } V = \left(D_0' S_0^{-1} D_0\right)^{-1}. \tag{7.19}$$

This holds also when the model is exactly identified, so we really do not use any weighting matrix.

To prove this note the following.

Remark 7.19 *(Continuous mapping theorem.) Let the sequences of random matrices* $\{x_T\}$ *and* $\{y_T\}$*, and the non-random matrix* $\{a_T\}$ *be such that* $x_T \xrightarrow{d} x$*,* $y_T \xrightarrow{p} y$*, and* $a_T \to a$ *(a traditional limit). Let* $g(x_T, y_T, a_T)$ *be a continuous function. Then* $g(x_T, y_T, a_T) \xrightarrow{d} g(x, y, a)$*. Either of* y_T *and* a_T *could be irrelevant in* g*. (See Mittelhammer (1996) 5.3.)*

Example 7.20 *For instance, the sequences in Remark 7.19 could be* $x_T = \sqrt{T}\Sigma_{t=}^T w_t/T$*, the scaled sample average of a random variable* w_t*;* $y_T = \Sigma_{t=}^T w_t^2/T$*, the sample second moment; and* $a_T = \Sigma_{t=1}^T 0.7^t$*.*

Remark 7.21 *From the previous remark: if* $x_T \xrightarrow{d} x$ *(a random variable) and* plim $Q_T = Q$ *(a constant matrix), then* $Q_T x_T \xrightarrow{d} Qx$*.*

Proof. (The asymptotic distribution (7.19). Sketch of proof.) This proof is essentially an application of the delta rule. By the mean-value theorem the sample moment condition evaluated at the GMM estimate, $\hat{\beta}$, is

$$\bar{m}(\hat{\beta}) = \bar{m}(\beta_0) + \frac{\partial \bar{m}(\beta_1)}{\partial \beta'}(\hat{\beta} - \beta_0) \tag{7.20}$$

for some values β_1 between $\hat{\beta}$ and β_0. (This point is different for different elements in $\bar{m}$.) Premultiply with $[\partial \bar{m}(\hat{\beta})/\partial \beta']' W$. By the first order condition (7.11), the left hand side is then zero, so we have

$$\mathbf{0}_{k\times 1} = \left(\frac{\partial \bar{m}(\hat{\beta})}{\partial \beta'}\right)' W \bar{m}(\beta_0) + \left(\frac{\partial \bar{m}(\hat{\beta})}{\partial \beta'}\right)' W \frac{\partial \bar{m}(\beta_1)}{\partial \beta'}(\hat{\beta} - \beta_0). \tag{7.21}$$

Multiply with $\sqrt{T}$ and solve as

$$\sqrt{T}\left(\hat{\beta}-\beta_0\right) = \underbrace{-\left[\left(\frac{\partial\bar{m}(\hat{\beta})}{\partial\beta'}\right)' W \frac{\partial\bar{m}(\beta_1)}{\partial\beta'}\right]^{-1}\left(\frac{\partial\bar{m}(\hat{\beta})}{\partial\beta'}\right)' W}_{\Gamma}\sqrt{T}\bar{m}(\beta_0). \tag{7.22}$$

If

$$\operatorname{plim}\frac{\partial\bar{m}(\hat{\beta})}{\partial\beta'} = \frac{\partial\bar{m}(\beta_0)}{\partial\beta'} = D_0, \text{ then } \operatorname{plim}\frac{\partial\bar{m}(\beta_1)}{\partial\beta'} = D_0,$$

since β_1 is between β_0 and $\hat{\beta}$. Then

$$\operatorname{plim}\Gamma = -\left(D_0'WD_0\right)^{-1}D_0'W. \tag{7.23}$$

The last term in (7.22), $\sqrt{T}\bar{m}(\beta_0)$, is $\sqrt{T}$ times a vector of sample averages, so by a CLT it converges in distribution to N(0, S_0), where S_0 is defined as in (7.12). By the rules of limiting distributions (see Remark 7.19) we then have that

$$\sqrt{T}\left(\hat{\beta}-\beta_0\right) \xrightarrow{d} \operatorname{plim}\Gamma \times \text{ something that is } N\left(0, S_0\right), \text{ that is,}$$
$$\sqrt{T}\left(\hat{\beta}-\beta_0\right) \xrightarrow{d} N\left[\mathbf{0}_{k\times 1}, (\operatorname{plim}\Gamma)S_0(\operatorname{plim}\Gamma')\right].$$

The covariance matrix is then

$$\begin{aligned}\text{ACov}[\sqrt{T}(\hat{\beta}-\beta_0)] &= (\operatorname{plim}\Gamma)S_0(\operatorname{plim}\Gamma') \\ &= \left(D_0'WD_0\right)^{-1}D_0'WS_0[\left(D_0'WD_0\right)^{-1}D_0'W]' \qquad (7.24)\\ &= \left(D_0'WD_0\right)^{-1}D_0'WS_0W'D_0\left(D_0'WD_0\right)^{-1}. \qquad (7.25)\end{aligned}$$

If $W = W' = S_0^{-1}$, then this expression simplifies to (7.19). (See, for instance, Hamilton (1994) 14 (appendix) for more details.) ■

It is straightforward to show that the difference between the covariance matrix in (7.25) and $\left(D_0'S_0^{-1}D_0\right)^{-1}$ (as in (7.19)) is a positive semi-definite matrix: any linear combination of the parameters has a smaller variance if $W = S_0^{-1}$ is used as the weighting matrix.

All the expressions for the asymptotic distribution are supposed to be evaluated at the true parameter vector β_0, which is unknown. However, D_0 in (7.16) can be estimated by $\partial\bar{m}(\hat{\beta})/\partial\beta'$, where we use the point estimate instead of the true value of the parameter vector. In practice, this means plugging in the point estimates into the sample moment conditions and calculate the derivatives with respect to parameters (for instance, by a numerical method).

Similarly, S_0 in (7.13) can be estimated by, for instance, Newey-West's estimator of Cov$[\sqrt{T}\bar{m}(\hat{\beta})]$, once again using the point estimates in the moment conditions.

Example 7.22 *(Covariance matrix of 2SLS.) Define*

$$S_0 = ACov\left[\sqrt{T}\bar{m}\left(\beta_0\right)\right] = ACov\left(\frac{\sqrt{T}}{T}\sum_{t=1}^{T}z_t u_t\right)$$
$$D_0 = \operatorname{plim}\frac{\partial\bar{m}(\beta_0)}{\partial\beta'} = \operatorname{plim}\left(-\frac{1}{T}\sum_{t=1}^{T}z_t x_t'\right) = -\Sigma_{zx}.$$

This gives the asymptotic covariance matrix of $\sqrt{T}(\hat{\beta}-\beta_0)$

$$V = \left(D_0'S_0^{-1}D_0\right)^{-1} = \left(\Sigma_{zx}'S_0^{-1}\Sigma_{zx}\right)^{-1}.$$

7.6 Summary of GMM

$$\begin{aligned}\text{Economic model}: &\ \mathrm{E}m(w_t,\beta_0) = \mathbf{0}_{q\times 1}, \beta \text{ is } k\times 1\\ \text{Sample moment conditions}: &\ \bar{m}(\beta) = \frac{1}{T}\sum_{t=1}^{T}m(w_t,\beta)\\ \text{Loss function}: &\ J = \bar{m}(\beta)'W\bar{m}(\beta)\\ \text{First order conditions}: &\ \mathbf{0}_{k\times 1} = \frac{\partial\bar{m}(\hat{\beta})'W\bar{m}(\hat{\beta})}{\partial\beta} = \left(\frac{\partial\bar{m}(\hat{\beta})}{\partial\beta'}\right)'W\bar{m}(\hat{\beta})\\ \text{Consistency}: &\ \hat{\beta} \text{ is typically consistent if } \mathrm{E}m(w_t,\beta_0) = \mathbf{0}\\ \text{Define}: &\ S_0 = \text{Cov}\left[\sqrt{T}\bar{m}\left(\beta_0\right)\right] \text{ and } D_0 = \operatorname{plim}\frac{\partial\bar{m}(\beta_0)}{\partial\beta'}\\ \text{Choose:} &\ W = S_0^{-1}\\ \text{Asymptotic distribution}: &\ \sqrt{T}(\hat{\beta}-\beta_0) \xrightarrow{d} N(\mathbf{0}_{k\times 1}, V), \text{ where } V = \left(D_0'S_0^{-1}D_0\right)^{-1}\end{aligned}$$

7.7 Efficient GMM and Its Feasible Implementation

The efficient GMM (remember: for a *given* set of moment conditions) requires that we use $W = S_0^{-1}$, which is tricky since S_0 should be calculated by using the true (unknown) parameter vector. However, the following *two-stage procedure* usually works fine:

- First, estimate model with some (symmetric and positive definite) weighting matrix. The identity matrix is typically a good choice for models where the moment conditions are of the same order of magnitude (if not, consider changing the moment conditions). This gives consistent estimates of the parameters β. Then a consistent estimate $\hat{S}$ can be calculated (for instance, with Newey-West).

- Use the consistent $\hat{S}$ from the first step to define a new weighting matrix as $W = \hat{S}^{-1}$. The algorithm is run again to give asymptotically efficient estimates of β.

- Iterate at least once more. (You may want to consider iterating until the point estimates converge.)

Example 7.23 *(Implementation of 2SLS.) Under the classical 2SLS assumptions, there is no need for iterating since the efficient weighting matrix is $\Sigma_{zz}^{-1}/\sigma^2$. Only σ^2 depends on the estimated parameters, but this scaling factor of the loss function does not affect $\hat{\beta}_{2SLS}$.*

One word of warning: if the number of parameters in the covariance matrix $\hat{S}$ is large compared to the number of data points, then $\hat{S}$ tends to be unstable (fluctuates a lot between the steps in the iterations described above) and sometimes also close to singular. The *saturation ratio* is sometimes used as an indicator of this problem. It is defined as the number of data points of the moment conditions (qT) divided by the number of estimated parameters (the k parameters in $\hat{\beta}$ and the unique $q(q+1)/2$ parameters in $\hat{S}$ if it is estimated with Newey-West). A value less than 10 is often taken to be an indicator of problems. A possible solution is then to impose restrictions on S, for instance, that the autocorrelation is a simple AR(1) and then estimate S using these restrictions (in which case you cannot use Newey-West, or course).

7.8 Testing in GMM

The result in (7.19) can be used to do *Wald tests of the parameter vector*. For instance, suppose we want to test the s linear restrictions that $R\beta_0 = r$ (R is $s \times k$ and r is $s \times 1$) then it must be the case that under null hypothesis

$$\sqrt{T}(R\hat{\beta} - r) \xrightarrow{d} N(\mathbf{0}_{s\times 1}, RVR'). \tag{7.26}$$

Remark 7.24 *(Distribution of quadratic forms.) If the $n \times 1$ vector $x \sim N(0, \Sigma)$, then $x'\Sigma^{-1}x \sim \chi^2_n$.*

From this remark and the continuous mapping theorem in Remark (7.19) it follows that, under the null hypothesis that $R\beta_0 = r$, the Wald test statistics is distributed as a χ^2_s variable

$$T(R\hat{\beta} - r)'\left(RVR'\right)^{-1}(R\hat{\beta} - r) \xrightarrow{d} \chi^2_s. \tag{7.27}$$

We might also want to *test the overidentifying restrictions*. The first order conditions (7.11) imply that k linear combinations of the q moment conditions are set to zero by solving for $\hat{\beta}$. Therefore, we have $q - k$ remaining overidentifying restrictions which should also be close to zero if the model is correct (fits data). Under the null hypothesis that the moment conditions hold (so the overidentifying restrictions hold), we know that $\sqrt{T}\bar{m}(\beta_0)$ is a (scaled) sample average and therefore has (by a CLT) an asymptotic normal distribution. It has a zero mean (the null hypothesis) and the covariance matrix in (7.12). In short,

$$\sqrt{T}\bar{m}(\beta_0) \xrightarrow{d} N\left(\mathbf{0}_{q\times 1}, S_0\right). \tag{7.28}$$

If would then perhaps be natural to expect that the quadratic form $T\bar{m}(\hat{\beta})'S_0^{-1}\bar{m}(\hat{\beta})$ should be converge in distribution to a χ^2_q variable. That is not correct, however, since $\hat{\beta}$ chosen is such a way that k linear combinations of the first order conditions always (in every sample) are zero. There are, in effect, only $q-k$ nondegenerate random variables in the quadratic form (see Davidson and MacKinnon (1993) 17.6 for a detailed discussion). The correct result is therefore that if we have used optimal weight matrix is used, $W = S_0^{-1}$, then

$$T\bar{m}(\hat{\beta})'S_0^{-1}\bar{m}(\hat{\beta}) \xrightarrow{d} \chi^2_{q-k}, \text{ if } W = S_0^{-1}. \tag{7.29}$$

The left hand side equals T times of value of the loss function (7.8) evaluated at the point estimates, so we could equivalently write what is often called the *J test*

$$TJ(\hat{\beta}) \sim \chi^2_{q-k}, \text{ if } W = S_0^{-1}. \tag{7.30}$$

This also illustrates that with no overidentifying restrictions (as many moment conditions as parameters) there are, of course, no restrictions to test. Indeed, the loss function value is then always zero at the point estimates.

Example 7.25 *(Test of overidentifying assumptions in 2SLS.) In contrast to the IV method, 2SLS allows us to test overidentifying restrictions (we have more moment conditions than parameters, that is, more instruments than regressors). This is a test of whether the residuals are indeed uncorrelated with all the instruments. If not, the model should be rejected. It can be shown that test (7.30) is (asymptotically, at least) the same as the traditional (Sargan (1964), see Davidson (2000) 8.4) test of the overidentifying restrictions in 2SLS. In the latter, the fitted residuals are regressed on the instruments; TR^2 from that regression is χ^2 distributed with as many degrees of freedom as the number of overidentifying restrictions.*

Example 7.26 *(Results from GMM on CCAPM; continuing Example 7.6.) The instruments could be anything known at t or earlier could be used as instruments. Actually, Hansen and Singleton (1982) and Hansen and Singleton (1983) use lagged $R_{i,t+1}c_{t+1}/c_t$ as instruments, and estimate γ to be 0.68 to 0.95, using monthly data. However, $TJ_T(\hat{\beta})$ is large and the model can usually be rejected at the 5% significance level. The rejection is most clear when multiple asset returns are used. If T-bills and stocks are tested at the same time, then the rejection would probably be overwhelming.*

Another test is to compare a restricted and a less restricted model, where we have used the optimal weighting matrix for the less restricted model in estimating both the less restricted and more restricted model (the weighting matrix is treated as a fixed matrix in the latter case). It can be shown that the test of the s restrictions (the "D test", similar in flavour to an LR test), is

$$T[J(\hat{\beta}^{restricted}) - J(\hat{\beta}^{less\ restricted})] \sim \chi^2_s, \text{ if } W = S_0^{-1}. \tag{7.31}$$

The weighting matrix is typically based on the unrestricted model. Note that (7.30) is a special case, since the model with allows q non-zero parameters (as many as the moment conditions) always attains $J = 0$, and that by imposing $s = q - k$ restrictions we get a restricted model.

7.9 GMM with Sub-Optimal Weighting Matrix*

When the optimal weighting matrix is not used, that is, when (7.18) does not hold, then the asymptotic covariance matrix of the parameters is given by (7.25) instead of the result in (7.19). That is,

$$\sqrt{T}(\hat{\beta} - \beta_0) \xrightarrow{d} N(\mathbf{0}_{k\times 1}, V_2), \text{ where } V_2 = \left(D_0'WD_0\right)^{-1} D_0'WS_0W'D_0 \left(D_0'WD_0\right)^{-1}. \tag{7.32}$$

The consistency property is not affected.

The test of the overidentifying restrictions (7.29) and (7.30) are not longer valid. Instead, the result is that

$$\sqrt{T}\bar{m}(\hat{\beta}) \rightarrow^d N\left(\mathbf{0}_{q\times 1}, \Psi_2\right), \text{ with} \tag{7.33}$$

$$\Psi_2 = [I - D_0\left(D_0'WD_0\right)^{-1} D_0'W]S_0[I - D_0\left(D_0'WD_0\right)^{-1} D_0'W]'. \tag{7.34}$$

This covariance matrix has rank $q - k$ (the number of overidentifying restriction). This distribution can be used to test hypotheses about the moments, for instance, that a particular moment condition is zero.

Proof. (Sketch of proof of (7.33)-(7.34)) Use (7.22) in (7.20) to get

$$\begin{aligned}\sqrt{T}\bar{m}(\hat{\beta}) &= \sqrt{T}\bar{m}(\beta_0) + \sqrt{T}\frac{\partial\bar{m}(\beta_1)}{\partial\beta'}\Gamma\bar{m}(\beta_0)\\ &= \left[I + \frac{\partial\bar{m}(\beta_1)}{\partial\beta'}\Gamma\right]\sqrt{T}\bar{m}(\beta_0).\end{aligned}$$

The term in brackets has a probability limit, which by (7.23) equals $I - D_0\left(D_0'WD_0\right)^{-1} D_0'W$. Since $\sqrt{T}\bar{m}(\beta_0) \rightarrow^d N\left(\mathbf{0}_{q\times 1}, S_0\right)$ we get (7.33). ■

Remark 7.27 *If the $n \times 1$ vector $X \sim N(0, \Sigma)$, where Σ has rank $r \leq n$ then $Y = X'\Sigma^+X \sim \chi^2_r$ where Σ^+ is the pseudo inverse of Σ.*

Remark 7.28 *The symmetric Σ can be decomposed as $\Sigma = Z\Lambda Z'$ where Z are the orthogonal eigenvectors ($Z'Z = I$) and Λ is a diagonal matrix with the eigenvalues* *along the main diagonal. The pseudo inverse can then be calculated as* $\Sigma^{+} = Z\Lambda^{+}Z'$*, where*

$$\Lambda^{+} = \begin{bmatrix} \Lambda_{11}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix},$$

with the reciprocals of the non-zero eigen values along the principal diagonal of Λ_{11}^{-1}.

This remark and (7.34) implies that the test of overidentifying restrictions (Hansen's J statistics) analogous to (7.29) is

$$T\bar{m}(\hat{\beta})'\Psi_2^{+}\bar{m}(\hat{\beta}) \overset{d}{\to} \chi^2_{q-k}. \tag{7.35}$$

It requires calculation of a generalized inverse (denoted by superscript $^{+}$), but this is fairly straightforward since Ψ_2 is a symmetric matrix. It can be shown (a bit tricky) that this simplifies to (7.29) when the optimal weighting matrix is used.

7.10 GMM without a Loss Function*

Suppose we sidestep the whole optimization issue and instead specify k linear combinations (as many as there are parameters) of the q moment conditions directly. That is, instead of the first order conditions (7.11) we postulate that the estimator should solve

$$\mathbf{0}_{k\times 1} = \underbrace{A}_{k\times q}\underbrace{\bar{m}(\hat{\beta})}_{q\times 1} \text{ } (\hat{\beta} \text{ is } k\times 1). \tag{7.36}$$

The matrix A is chosen by the researcher and it must have rank k (lower rank means that we effectively have too few moment conditions to estimate the k parameters in β). If A is random, then it should have a finite probability limit A_0 (also with rank k). One simple case when this approach makes sense is when we want to use a subset of the moment conditions to estimate the parameters (some columns in A are then filled with zeros), but we want to study the distribution of all the moment conditions.

By comparing (7.11) and (7.36) we see that A plays the same role as $[\partial\bar{m}(\hat{\beta})/\partial\beta']'W$, but with the difference that A is chosen and not allowed to depend on the parameters. In the asymptotic distribution, it is the probability limit of these matrices that matter, so we can actually substitute A_0 for $D_0'W$ in the proof of the asymptotic distribution. The

covariance matrix in (7.32) then becomes

$$\begin{aligned} V_3 &= (A_0D_0)^{-1}A_0S_0[(A_0D_0)^{-1}A_0]' \\ &= (A_0D_0)^{-1}A_0S_0A_0'[(A_0D_0)^{-1}]', \end{aligned} \tag{7.37}$$

which can be used to test hypotheses about the parameters.

Similarly,the asymptotic distribution of the moment conditions is

$$\sqrt{T}\bar{m}(\hat{\beta}) \to^{d} N\left(\mathbf{0}_{q\times 1}, \Psi_3\right)\text{, with} \tag{7.38}$$

$$\Psi_3 = [I - D_0(A_0D_0)^{-1}A_0]S_0[I - D_0(A_0D_0)^{-1}A_0]', \tag{7.39}$$

where Ψ_3 has reduced rank. As before, this covariance matrix can be used to construct both t type and χ^2 tests of the moment conditions. For instance, the test of overidentifying restrictions (Hansen's J statistics)

$$T\bar{m}(\hat{\beta})'\Psi_3^{+}\bar{m}(\hat{\beta}) \overset{d}{\to} \chi^2_{q-k}, \tag{7.40}$$

where Ψ_3^{+} is a generalized inverse of Ψ_3.

7.11 Simulated Moments Estimator*

Reference: Ingram and Lee (1991)

It sometimes happens that it is not possible to calculate the theoretical moments in GMM explicitly. For instance, suppose we want to match the variance of the model with the variance of data

$$\mathrm{E}\, m(w_t, \beta_0) = 0\text{, where} \tag{7.41}$$

$$m(w_t, \beta) = (w_t - \mu)^2 - \text{Var\_in\_model}(\beta)\,, \tag{7.42}$$

but the model is so non-linear that we cannot find a closed form expression for Var_of_model(β_0). Similary, we could match a covariance of

The SME involves *(i)* drawing a set of random numbers for the stochastic shocks in the model; *(ii)* for a given set of parameter values generate a model simulation with T_{sim} observations, calculating the moments and using those instead of Var_of_model(β_0) (or similarly for other moments), which is then used to evaluate the loss function J_T. This is repeated for various sets of parameter values until we find the one which minimizes J_T.

Basically all GMM results go through, but the covariance matrix should be scaled up with $1 + T/T_{sim}$, where T is the sample length. Note that the same sequence of random numbers should be reused over and over again (as the parameter values are changed).

Example 7.29 *Suppose w_t has two elements, x_t and y_t, and that we want to match both variances and also the covariance. For simplicity, suppose both series have zero means. Then we can formulate the moment conditions*

$$m(x_t, y_t, \beta) = \begin{bmatrix} x_t^2 - Var(x)\_in\_model(\beta) \\ y_t^2 - Var(y)\_in\_model(\beta) \\ x_t y_t - Cov(x,y)\_in\_model(\beta) \end{bmatrix}. \tag{7.43}$$

Bibliography

Cochrane, J. H., 2001, *Asset Pricing*, Princeton University Press, Princeton, New Jersey.

Davidson, J., 2000, *Econometric Theory*, Blackwell Publishers, Oxford.

Davidson, R., and J. G. MacKinnon, 1993, *Estimation and Inference in Econometrics*, Oxford University Press, Oxford.

Greene, W. H., 2000, *Econometric Analysis*, Prentice-Hall, Upper Saddle River, New Jersey, 4th edn.

Hamilton, J. D., 1994, *Time Series Analysis*, Princeton University Press, Princeton.

Hansen, L., and K. Singleton, 1982, "Generalized Instrumental Variables Estimation of Nonlinear Rational Expectations Models," *Econometrica*, 50, 1269–1288.

Hansen, L., and K. Singleton, 1983, "Stochastic Consumption, Risk Aversion and the Temporal Behavior of Asset Returns," *Journal of Political Economy*, 91, 249–268.

Harris, D., and L. Matyas, 1999, "Introduction to the Generalized Method of Moments Estimation," in Laszlo Matyas (ed.), *Generalized Method of Moments Estimation* . chap. 1, Cambridge University Press.

Hayashi, F., 2000, *Econometrics*, Princeton University Press.

Ingram, B.-F., and B.-S. Lee, 1991, "'Simulation Estimation of Time-Series Models," *Journal of Econometrics*, 47, 197–205.

Johnston, J., and J. DiNardo, 1997, *Econometric Methods*, McGraw-Hill, New York, 4th edn.

Mittelhammer, R. C., 1996, *Mathematical Statistics for Economics and Business*, Springer-Verlag, New York.

Ogaki, M., 1993, "Generalized Method of Moments: Econometric Applications," in G. S. Maddala, C. R. Rao, and H. D. Vinod (ed.), *Handbook of Statistics*, vol. 11, . chap. 17, pp. 455–487, Elsevier.

Pindyck, R. S., and D. L. Rubinfeld, 1998, *Econometric Models and Economic Forecasts*, Irwin McGraw-Hill, Boston, Massachusetts, 4ed edn.

Verbeek, M., 2004, *A Guide to Modern Econometrics*, Wiley, Chichester, 2nd edn.

8 Examples and Applications of GMM

8.1 GMM and Classical Econometrics: Examples

8.1.1 The LS Estimator (General)

The model is

$$y_t = x_t'\beta_0 + u_t, \tag{8.1}$$

where β is a $k \times 1$ vector.

The k moment conditions are

$$\bar{m}(\beta) = \frac{1}{T}\sum_{t=1}^{T} x_t(y_t - x_t'\beta) = \frac{1}{T}\sum_{t=1}^{T} x_t y_t - \frac{1}{T}\sum_{t=1}^{T} x_t x_t'\beta. \tag{8.2}$$

The point estimates are found by setting all moment conditions to zero (the model is exactly identified), $\bar{m}(\beta) = \mathbf{0}_{k\times 1}$, which gives

$$\hat{\beta} = \left(\frac{1}{T}\sum_{t=1}^{T} x_t x_t'\right)^{-1} \frac{1}{T}\sum_{t=1}^{T} x_t y_t \beta. \tag{8.3}$$

If we define

$$S_0 = \text{ACov}\left[\sqrt{T}\bar{m}(\beta_0)\right] = \text{ACov}\left(\frac{\sqrt{T}}{T}\sum_{t=1}^{T} x_t u_t\right) \tag{8.4}$$

$$D_0 = \text{plim}\,\frac{\partial \bar{m}(\beta_0)}{\partial \beta'} = \text{plim}\left(-\frac{1}{T}\sum_{t=1}^{T} x_t x_t'\right) = -\Sigma_{xx}. \tag{8.5}$$

then the asymptotic covariance matrix of $\sqrt{T}(\hat{\beta} - \beta_0)$

$$V_{LS} = \left(D_0' S_0^{-1} D_0\right)^{-1} = \left(\Sigma_{xx}' S_0^{-1} \Sigma_{xx}\right)^{-1} = \Sigma_{xx}^{-1} S_0 \Sigma_{xx}^{-1}. \tag{8.6}$$

We can then either try to estimate S_0 by Newey-West, or make further assumptions to simplify S_0 (see below).

8.1.2 The IV/2SLS Estimator (General)

The model is (8.1), but we use an IV/2SLS method. The q moment conditions (with $q \geq k$) are

$$\bar{m}(\beta) = \frac{1}{T}\sum_{t=1}^{T} z_t(y_t - x_t'\beta) = \frac{1}{T}\sum_{t=1}^{T} z_t y_t - \frac{1}{T}\sum_{t=1}^{T} z_t x_t'\beta. \tag{8.7}$$

The loss function is (for some positive definite weighting matrix W, not necessarily the optimal)

$$\bar{m}(\beta)'W\bar{m}(\beta) = \left[\frac{1}{T}\sum_{t=1}^{T} z_t(y_t - x_t'\beta)\right]' W \left[\frac{1}{T}\sum_{t=1}^{T} z_t(y_t - x_t'\beta)\right], \tag{8.8}$$

and the k first order conditions, $(\partial \bar{m}(\hat{\beta})/\partial \beta')' W \bar{m}(\hat{\beta}) = 0$, are

$$\begin{aligned}
\mathbf{0}_{k\times 1} &= \left[\frac{\partial}{\partial \beta'}\frac{1}{T}\sum_{t=1}^{T} z_t(y_t - x_t'\hat{\beta})\right]' W \frac{1}{T}\sum_{t=1}^{T} z_t(y_t - x_t'\hat{\beta}) \\
&= \left[-\frac{1}{T}\sum_{t=1}^{T} z_t x_t'\right]' W \frac{1}{T}\sum_{t=1}^{T} z_t(y_t - x_t'\hat{\beta}) \\
&= -\hat{\Sigma}_{xz} W (\hat{\Sigma}_{zy} - \hat{\Sigma}_{zx}\hat{\beta}).
\end{aligned} \tag{8.9}$$

We solve for $\hat{\beta}$ as

$$\hat{\beta} = \left(\hat{\Sigma}_{xz} W \hat{\Sigma}_{zx}\right)^{-1} \hat{\Sigma}_{xz} W \hat{\Sigma}_{zy}. \tag{8.10}$$

Define

$$S_0 = \text{ACov}\left[\sqrt{T}\bar{m}(\beta_0)\right] = \text{ACov}\left(\frac{\sqrt{T}}{T}\sum_{t=1}^{T} z_t u_t\right) \tag{8.11}$$

$$D_0 = \text{plim}\,\frac{\partial \bar{m}(\beta_0)}{\partial \beta'} = \text{plim}\left(-\frac{1}{T}\sum_{t=1}^{T} z_t x_t'\right) = -\Sigma_{zx}. \tag{8.12}$$

This gives the asymptotic covariance matrix of $\sqrt{T}(\hat{\beta} - \beta_0)$

$$V = \left(D_0' S_0^{-1} D_0\right)^{-1} = \left(\Sigma_{zx}' S_0^{-1} \Sigma_{zx}\right)^{-1}. \tag{8.13}$$

When the model is exactly identified ($q = k$), then we can make some simplifications since $\hat{\Sigma}_{xz}$ is then invertible. This is the case of the classical IV estimator. We get

$$\hat{\beta} = \hat{\Sigma}_{zx}^{-1}\hat{\Sigma}_{zy} \text{ and } V = \Sigma_{zx}^{-1}S_0\left(\Sigma_{zx}'\right)^{-1} \text{ if } q = k. \tag{8.14}$$

(Use the rule $(ABC)^{-1} = C^{-1}B^{-1}A^{-1}$ to show this.)

8.1.3 Classical LS Assumptions

Reference: Greene (2000) 9.4 and Hamilton (1994) 8.2.

This section returns to the LS estimator in Section (8.1.1) in order to highlight the classical LS assumptions that give the variance matrix $\sigma^2\Sigma_{xx}^{-1}$.

We allow the regressors to be stochastic, but require that x_t is independent of all u_{t+s} and that u_t is iid. It rules out, for instance, that u_t and x_{t-2} are correlated and also that the variance of u_t depends on x_t. Expand the expression for S_0 as

$$\begin{aligned} S_0 &= \mathrm{E}\left(\frac{\sqrt{T}}{T}\sum_{t=1}^{T}x_tu_t\right)\left(\frac{\sqrt{T}}{T}\sum_{t=1}^{T}u_tx_t'\right) \qquad (8.15)\\ &= \frac{1}{T}\mathrm{E}\left(... + x_{s-1}u_{s-1} + x_su_s + ...\right)\left(... + u_{s-1}x_{s-1}' + u_sx_s' + ...\right). \end{aligned}$$

Note that

$$\begin{aligned} \mathrm{E}x_{t-s}u_{t-s}u_tx_t' &= \mathrm{E}x_{t-s}x_t'\mathrm{E}u_{t-s}u_t \text{ (since } u_t \text{ and } x_{t-s} \text{ independent)}\\ &= \begin{cases} 0 \text{ if } s \neq 0 \text{ (since } \mathrm{E}u_{s-1}u_s = 0 \text{ by iid } u_t)\\ \mathrm{E}x_tx_t'\mathrm{E}u_tu_t \text{ else.} \end{cases} \end{aligned} \tag{8.16}$$

This means that all cross terms (involving different observations) drop out and that we can write

$$\begin{aligned} S_0 &= \frac{1}{T}\sum_{t=1}^{T}\mathrm{E}x_tx_t'\mathrm{E}u_t^2 && (8.17)\\ &= \sigma^2\frac{1}{T}\mathrm{E}\sum_{t=1}^{T}x_tx_t' \text{ (since } u_t \text{ is iid and } \sigma^2 = \mathrm{E}u_t^2) && (8.18)\\ &= \sigma^2\Sigma_{xx}. && (8.19) \end{aligned}$$

Using this in (8.6) gives

$$V = \sigma^2\Sigma_{xx}^{-1}. \tag{8.20}$$

8.1.4 Almost Classical LS Assumptions: White's Heteroskedasticity.

Reference: Greene (2000) 12.2 and Davidson and MacKinnon (1993) 16.2.

The only difference compared with the classical LS assumptions is that u_t is now allowed to be heteroskedastic, but this heteroskedasticity is not allowed to depend on the moments of x_t. This means that (8.17) holds, but (8.18) does not since $\mathrm{E}u_t^2$ is not the same for all t.

However, we can still simplify (8.17) a bit more. We assumed that $\mathrm{E}x_tx_t'$ and $\mathrm{E}u_t^2$ (which can both be time varying) are not related to each other, so we could perhaps multiply $\mathrm{E}x_tx_t'$ by $\Sigma_{t=1}^{T}\mathrm{E}u_t^2/T$ instead of by $\mathrm{E}u_t^2$. This is indeed true asymptotically—where any possible "small sample" relation between $\mathrm{E}x_tx_t'$ and $\mathrm{E}u_t^2$ must wash out due to the assumptions of independence (which are about population moments).

In large samples we therefore have

$$\begin{aligned} S_0 &= \left(\frac{1}{T}\sum_{t=1}^{T}\mathrm{E}u_t^2\right)\left(\frac{1}{T}\sum_{t=1}^{T}\mathrm{E}x_tx_t'\right)\\ &= \left(\frac{1}{T}\sum_{t=1}^{T}\mathrm{E}u_t^2\right)\left(\mathrm{E}\frac{1}{T}\sum_{t=1}^{T}x_tx_t'\right)\\ &= \omega^2\Sigma_{xx}, \qquad (8.21) \end{aligned}$$

where ω^2 is a scalar. This is very similar to the classical LS case, except that ω^2 is the average variance of the residual rather than the constant variance. In practice, the estimator of ω^2 is the same as the estimator of σ^2, so we can actually apply the standard LS formulas in this case.

This is the motivation for why White's test for heteroskedasticity makes sense: if the heteroskedasticity is not correlated with the regressors, then the standard LS formula is correct (provided there is no autocorrelation).

8.1.5 Estimating the Mean of a Process

Suppose u_t is heteroskedastic, but not autocorrelated. In the regression $y_t = \alpha + u_t$, $x_t = z_t = 1$. This is a special case of the previous example, since $\mathrm{E}u_t^2$ is certainly unrelated to $\mathrm{E}x_t x_t' = 1$ (since it is a constant). Therefore, the LS covariance matrix is the correct variance of the sample mean as an estimator of the mean, even if u_t are heteroskedastic (provided there is no autocorrelation).

8.1.6 The Classical 2SLS Assumptions*

Reference: Hamilton (1994) 9.2.

The classical 2SLS case assumes that z_t is independent of all u_{t+s} and that u_t is iid. The covariance matrix of the moment conditions are

$$S_0 = \mathrm{E}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} z_t u_t\right)\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} u_t z_t'\right), \tag{8.22}$$

so by following the same steps in (8.16)-(8.19) we get $S_0 = \sigma^2 \Sigma_{zz}$.The optimal weighting matrix is therefore $W = \Sigma_{zz}^{-1}/\sigma^2$ (or $(Z'Z/T)^{-1}/\sigma^2$ in matrix form). We use this result in (8.10) to get

$$\hat{\beta}_{2SLS} = \left(\hat{\Sigma}_{xz}\hat{\Sigma}_{zz}^{-1}\hat{\Sigma}_{zx}\right)^{-1}\hat{\Sigma}_{xz}\hat{\Sigma}_{zz}^{-1}\hat{\Sigma}_{zy}, \tag{8.23}$$

which is the classical 2SLS estimator.

Since this GMM is efficient (for a given set of moment conditions), we have established that 2SLS uses its given set of instruments in the efficient way—provided the classical 2SLS assumptions are correct. Also, using the weighting matrix in (8.13) gives

$$V = \left(\Sigma_{xz}\frac{1}{\sigma^2}\Sigma_{zz}^{-1}\Sigma_{zx}\right)^{-1}. \tag{8.24}$$

8.2 Identification of Systems of Simultaneous Equations

Reference: Greene (2000) 16.1-3

This section shows how the GMM moment conditions can be used to understand if the parameters in a system of simultaneous equations are identified or not.

The structural model (form) is

$$Fy_t + Gz_t = u_t, \tag{8.25}$$

where y_t is a vector of endogenous variables, z_t a vector of predetermined (exogenous) variables, F is a square matrix, and G is another matrix.[1] We can write the jth equation of the structural form (8.25) as

$$y_{jt} = x_t'\beta + u_{jt}, \tag{8.26}$$

where x_t contains the endogenous and exogenous variables that enter the jth equation with non-zero coefficients, that is, subsets of y_t and z_t.

We want to estimate β in (8.26). Least squares is inconsistent if some of the regressors are endogenous variables (in terms of (8.25), this means that the jth row in F contains at least one additional non-zero element apart from coefficient on y_{jt}). Instead, we use IV/2SLS. By assumption, the structural model summarizes all relevant information for the endogenous variables y_t. This implies that the only useful instruments are the variables in z_t. (A valid instrument is uncorrelated with the residuals, but correlated with the regressors.) The moment conditions for the jth equation are then

$$\mathrm{E}z_t\left(y_{jt} - x_t'\beta\right) = \mathbf{0} \text{ with sample moment conditions } \frac{1}{T}\sum_{t=1}^{T} z_t\left(y_{jt} - x_t'\beta\right) = \mathbf{0}. \tag{8.27}$$

If there are as many moment conditions as there are elements in β, then this equation is *exactly identified*, so the sample moment conditions can be inverted to give the Instrumental variables (IV) estimator of β. If there are more moment conditions than elements in β, then this equation is *overidentified* and we must devise some method for weighting the different moment conditions. This is the 2SLS method. Finally, when there are fewer moment conditions than elements in β, then this equation is *unidentified*, and we cannot hope to estimate the structural parameters of it.

We can partition the vector of regressors in (8.26) as $x_t' = [\tilde{z}_t', \tilde{y}_t']$, where y_{1t} and z_{1t} are the subsets of z_t and y_t respectively, that enter the right hand side of (8.26). Partition z_t conformably $z_t' = [\tilde{z}_t', z_t^{*\prime}]$, where z_t^* are the exogenous variables that do not enter

[1] By premultiplying with F^{-1} and rearranging we get the reduced form $y_t = \Pi z_t + \varepsilon_t$, with $\Pi = -F^{-1}$ and $\mathrm{Cov}(\varepsilon_t) = F^{-1}\mathrm{Cov}(u_t)(F^{-1})'$.

(8.26). We can then rewrite the moment conditions in (8.27) as

$$\mathrm{E}\begin{bmatrix} \tilde{z}_t \\ z_t^* \end{bmatrix}\left(y_{jt} - \begin{bmatrix} \tilde{z}_t \\ \tilde{y}_t \end{bmatrix}'\beta\right) = \mathbf{0}. \tag{8.28}$$

$$\begin{aligned} y_{jt} &= -G_j\tilde{z}_t - F_j\tilde{y}_t + u_{jt} \\ &= x_t'\beta + u_{jt}, \text{ where } x_t' = \left[\tilde{z}_t', \tilde{y}_t'\right], \end{aligned} \tag{8.29}$$

This shows that we need at least as many elements in z_t^* as in $\tilde{y}_t$ to have this equations identified, which confirms the old-fashioned rule of thumb: *there must be at least as many excluded exogenous variables (z_t^*) as included endogenous variables ($\tilde{y}_t$) to have the equation identified.*

This section has discussed identification of structural parameters when 2SLS/IV, one equation at a time, is used. There are other ways to obtain identification, for instance, by imposing restrictions on the covariance matrix. See, for instance, Greene (2000) 16.1-3 for details.

Example 8.1 *(Supply and Demand. Reference: GR 16, Hamilton 9.1.) Consider the simplest simultaneous equations model for supply and demand on a market. Supply is*

$$q_t = \gamma p_t + u_t^s,\ \gamma > 0,$$

and demand is

$$q_t = \beta p_t + \alpha A_t + u_t^d,\ \beta < 0,$$

where A_t is an observable exogenous demand shock (perhaps income). The only meaningful instrument is A_t. From the supply equation we then get the moment condition

$$EA_t\left(q_t - \gamma p_t\right) = 0,$$

which gives one equation in one unknown, γ. The supply equation is therefore exactly identified. In contrast, the demand equation is unidentified, since there is only one (meaningful) moment condition

$$EA_t\left(q_t - \beta p_t - \alpha A_t\right) = 0,$$

but two unknowns (β and α).

Example 8.2 *(Supply and Demand: overidentification.) If we change the demand equation in Example 8.1 to*

$$q_t = \beta p_t + \alpha A_t + bB_t + u_t^d,\ \beta < 0.$$

There are now two moment conditions for the supply curve (since there are two useful instruments)

$$E\begin{bmatrix} A_t\left(q_t - \gamma p_t\right) \\ B_t\left(q_t - \gamma p_t\right) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

but still only one parameter: the supply curve is now overidentified. The demand curve is still underidentified (two instruments and three parameters).

8.3 Testing for Autocorrelation

This section discusses how GMM can be used to test if a series is autocorrelated. The analysis focuses on first-order autocorrelation, but it is straightforward to extend it to higher-order autocorrelation.

Consider a scalar random variable x_t with a zero mean (it is easy to extend the analysis to allow for a non-zero mean). Consider the moment conditions

$$m_t(\beta) = \begin{bmatrix} x_t^2 - \sigma^2 \\ x_t x_{t-1} - \rho\sigma^2 \end{bmatrix}, \text{ so } \bar{m}(\beta) = \frac{1}{T}\sum_{t=1}^{T}\begin{bmatrix} x_t^2 - \sigma^2 \\ x_t x_{t-1} - \rho\sigma^2 \end{bmatrix}, \text{ with } \beta = \begin{bmatrix} \sigma^2 \\ \rho \end{bmatrix}. \tag{8.30}$$

σ^2 is the variance and ρ the first-order autocorrelation so $\rho\sigma^2$ is the first-order autocovariance. We want to test if $\rho = 0$. We could proceed along two different routes: estimate ρ and test if it is different from zero or set ρ to zero and then test overidentifying restrictions. We analyze how these two approaches work when the null hypothesis of $\rho = 0$ is true.

8.3.1 Estimating the Autocorrelation Coefficient

We estimate both σ^2 and ρ by using the moment conditions (8.30) and then test if $\rho = 0$. To do that we need to calculate the asymptotic variance of $\hat{\rho}$ (there is little hope of being able to calculate the small sample variance, so we have to settle for the asymptotic variance as an approximation).

We have an exactly identified system so the weight matrix does not matter—we can then proceed as if we had used the optimal weighting matrix (all those results apply).

To find the asymptotic covariance matrix of the parameters estimators, we need the probability limit of the Jacobian of the moments and the covariance matrix of the moments—evaluated at the true parameter values. Let $\bar{m}_i(\beta_0)$ denote the ith element of the $\bar{m}(\beta)$ vector—evaluated at the true parameter values. The probability of the Jacobian is

$$D_0 = \text{plim} \begin{bmatrix} \partial \bar{m}_1(\beta_0)/\partial\sigma^2 & \partial \bar{m}_1(\beta_0)/\partial\rho \\ \partial \bar{m}_2(\beta_0)/\partial\sigma^2 & \partial \bar{m}_2(\beta_0)/\partial\rho \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ -\rho & -\sigma^2 \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 0 & -\sigma^2 \end{bmatrix}, \tag{8.31}$$

since $\rho = 0$ (the true value). Note that we differentiate with respect to σ^2, not σ, since we treat σ^2 as a parameter.

The covariance matrix is more complicated. The definition is

$$S_0 = \mathrm{E}\left[\frac{\sqrt{T}}{T}\sum_{t=1}^{T} m_t(\beta_0)\right]\left[\frac{\sqrt{T}}{T}\sum_{t=1}^{T} m_t(\beta_0)\right]'.$$

Assume that there is no autocorrelation in $m_t(\beta_0)$. We can then simplify as

$$S_0 = \mathrm{E}\, m_t(\beta_0) m_t(\beta_0)'.$$

This assumption is stronger than assuming that $\rho = 0$, but we make it here in order to illustrate the asymptotic distribution. To get anywhere, we assume that x_t is iid $N(0, \sigma^2)$. In this case (and with $\rho = 0$ imposed) we get

$$\begin{aligned} S_0 &= \mathrm{E}\begin{bmatrix} x_t^2 - \sigma^2 \\ x_t x_{t-1} \end{bmatrix}\begin{bmatrix} x_t^2 - \sigma^2 \\ x_t x_{t-1} \end{bmatrix}' = \mathrm{E}\begin{bmatrix} (x_t^2 - \sigma^2)^2 & (x_t^2 - \sigma^2)x_t x_{t-1} \\ (x_t^2 - \sigma^2)x_t x_{t-1} & (x_t x_{t-1})^2 \end{bmatrix} \\ &= \begin{bmatrix} \mathrm{E}\, x_t^4 - 2\sigma^2\,\mathrm{E}\, x_t^2 + \sigma^4 & 0 \\ 0 & \mathrm{E}\, x_t^2 x_{t-1}^2 \end{bmatrix} = \begin{bmatrix} 2\sigma^4 & 0 \\ 0 & \sigma^4 \end{bmatrix}. \end{aligned} \tag{8.32}$$

To make the simplification in the second line we use the facts that $\mathrm{E}\, x_t^4 = 3\sigma^4$ if $x_t \sim N(0, \sigma^2)$, and that the normality and the iid properties of x_t together imply $\mathrm{E}\, x_t^2 x_{t-1}^2 = \mathrm{E}\, x_t^2\, \mathrm{E}\, x_{t-1}^2$ and $\mathrm{E}\, x_t^3 x_{t-1} = \mathrm{E}\, \sigma^2 x_t x_{t-1} = 0$.

By combining (8.31) and (8.32) we get that

$$\begin{aligned} \mathrm{ACov}\left(\sqrt{T}\begin{bmatrix} \hat{\sigma}^2 \\ \hat{\rho} \end{bmatrix}\right) &= \left(D_0' S_0^{-1} D_0\right)^{-1} \\ &= \left(\begin{bmatrix} -1 & 0 \\ 0 & -\sigma^2 \end{bmatrix}'\begin{bmatrix} 2\sigma^4 & 0 \\ 0 & \sigma^4 \end{bmatrix}^{-1}\begin{bmatrix} -1 & 0 \\ 0 & -\sigma^2 \end{bmatrix}\right)^{-1} \\ &= \begin{bmatrix} 2\sigma^4 & 0 \\ 0 & 1 \end{bmatrix}. \end{aligned} \tag{8.33}$$

This shows the standard expression for the uncertainty of the variance and that the $\sqrt{T}\hat{\rho}$. Since GMM estimators typically have an asymptotic distribution we have $\sqrt{T}\hat{\rho} \to^d N(0, 1)$, so we can test the null hypothesis of no first-order autocorrelation by the test statistics

$$T\hat{\rho}^2 \sim \chi_1^2. \tag{8.34}$$

This is the same as the *Box-Ljung test for first-order autocorrelation*.

This analysis shows that we are able to arrive at simple expressions for the sampling uncertainty of the variance and the autocorrelation—provided we are willing to make strong assumptions about the data generating process. In particular, ewe assumed that data was iid $N(0, \sigma^2)$. One of the strong points of GMM is that we could perform similar tests without making strong assumptions—provided we use a correct estimator of the asymptotic covariance matrix S_0 (for instance, Newey-West).

8.3.2 Testing the Overidentifying Restriction of No Autocorrelation*

We can estimate σ^2 alone and then test if both moment condition are satisfied at $\rho = 0$. There are several ways of doing that, but the perhaps most straightforward is skip the loss function approach to GMM and instead specify the "first order conditions" directly as

$$\begin{aligned} 0 &= A\bar{m} \\ &= \begin{bmatrix} 1 & 0 \end{bmatrix}\frac{1}{T}\sum_{t=1}^{T}\begin{bmatrix} x_t^2 - \sigma^2 \\ x_t x_{t-1} \end{bmatrix}, \end{aligned} \tag{8.35}$$

which sets $\hat{\sigma}^2$ equal to the sample variance.

The only parameter in this estimation problem is σ^2, so the matrix of derivatives becomes

$$D_0 = \operatorname{plim}\begin{bmatrix}\partial\bar{m}_1(\beta_0)/\partial\sigma^2\\ \partial\bar{m}_2(\beta_0)/\partial\sigma^2\end{bmatrix} = \begin{bmatrix}-1\\0\end{bmatrix}. \tag{8.36}$$

By using this result, the A matrix in (8.36) and the S_0 matrix in (8.32,) it is straighforward to calculate the asymptotic covariance matrix the moment conditions. In general, we have

$$\mathrm{ACov}[\sqrt{T}\bar{m}(\hat{\beta})] = [I - D_0\,(A_0 D_0)^{-1}\,A_0]S_0[I - D_0\,(A_0 D_0)^{-1}\,A_0]'. \tag{8.37}$$

The term in brackets is here (since $A_0 = A$ since it is a matrix with constants)

$$\underbrace{\begin{bmatrix}1&0\\0&1\end{bmatrix}}_{I_2} - \underbrace{\begin{bmatrix}-1\\0\end{bmatrix}}_{D_0}\left(\underbrace{\begin{bmatrix}1&0\end{bmatrix}}_{A_0}\underbrace{\begin{bmatrix}-1\\0\end{bmatrix}}_{D_0}\right)^{-1}\underbrace{\begin{bmatrix}1&0\end{bmatrix}}_{A_0} = \begin{bmatrix}0&0\\0&1\end{bmatrix}. \tag{8.38}$$

We therefore get

$$\mathrm{ACov}[\sqrt{T}\bar{m}(\hat{\beta})] = \begin{bmatrix}0&0\\0&1\end{bmatrix}\begin{bmatrix}2\sigma^4&0\\0&\sigma^4\end{bmatrix}\begin{bmatrix}0&0\\0&1\end{bmatrix}' = \begin{bmatrix}0&0\\0&\sigma^4\end{bmatrix}. \tag{8.39}$$

Note that the first moment condition has no sampling variance at the estimated parameters, since the choice of $\hat{\sigma}^2$ always sets the first moment condition equal to zero.

The test of the overidentifying restriction that the second moment restriction is also zero is

$$T\bar{m}'\left(\mathrm{ACov}[\sqrt{T}\bar{m}(\hat{\beta})]\right)^{+}\bar{m} \sim \chi^2_1, \tag{8.40}$$

where we have to use a generalized inverse if the covariance matrix is singular (which it is in (8.39)).

In this case, we get the test statistics (note the generalized inverse)

$$T\begin{bmatrix}0\\ \Sigma_{t=1}^{T}x_t x_{t-1}/T\end{bmatrix}'\begin{bmatrix}0&0\\0&1/\sigma^4\end{bmatrix}\begin{bmatrix}0\\ \Sigma_{t=1}^{T}x_t x_{t-1}/T\end{bmatrix} = T\frac{\left[\Sigma_{t=1}^{T}x_t x_{t-1}/T\right]^2}{\sigma^4}, \tag{8.41}$$

which is the T times the square of the sample covariance divided by σ^4. A sample correlation, $\hat{\rho}$, would satisfy $\Sigma_{t=1}^{T}x_t x_{t-1}/T = \hat{\rho}\hat{\sigma}^2$, which we can use to rewrite (8.41) as $T\hat{\rho}^2\hat{\sigma}^4/\sigma^4$. By approximating σ^4 by $\hat{\sigma}^4$ we get the same test statistics as in (8.34).

8.4 Estimating and Testing a Normal Distribution

8.4.1 Estimating the Mean and Variance

This section discusses how the GMM framework can be used to test if a variable is normally distributed. The analysis cold easily be changed in order to test other distributions as well.

Suppose we have a sample of the scalar random variable x_t and that we want to test if the series is normally distributed. We analyze the asymptotic distribution under the null hypothesis that x_t is $N(\mu,\sigma^2)$.

We specify four moment conditions

$$m_t = \begin{bmatrix}x_t-\mu\\(x_t-\mu)^2-\sigma^2\\(x_t-\mu)^3\\(x_t-\mu)^4-3\sigma^4\end{bmatrix} \text{ so } \bar{m} = \frac{1}{T}\sum_{t=1}^{T}\begin{bmatrix}x_t-\mu\\(x_t-\mu)^2-\sigma^2\\(x_t-\mu)^3\\(x_t-\mu)^4-3\sigma^4\end{bmatrix} \tag{8.42}$$

Note that $\mathrm{E}\,m_t = \mathbf{0}_{4\times 1}$ if x_t is normally distributed.

Let $\bar{m}_i(\beta_0)$ denote the ith element of the $\bar{m}(\beta)$ vector—evaluated at the true parameter values. The probability of the Jacobian is

$$\begin{aligned} D_0 &= \operatorname{plim}\begin{bmatrix}\partial\bar{m}_1(\beta_0)/\partial\mu & \partial\bar{m}_1(\beta_0)/\partial\sigma^2\\ \partial\bar{m}_2(\beta_0)/\partial\mu & \partial\bar{m}_2(\beta_0)/\partial\sigma^2\\ \partial\bar{m}_3(\beta_0)/\partial\mu & \partial\bar{m}_3(\beta_0)/\partial\sigma^2\\ \partial\bar{m}_4(\beta_0)/\partial\mu & \partial\bar{m}_4(\beta_0)/\partial\sigma^2\end{bmatrix}\\ &= \operatorname{plim}\frac{1}{T}\sum_{t=1}^{T}\begin{bmatrix}-1&0\\-2(x_t-\mu)&-1\\-3(x_t-\mu)^2&0\\-4(x_t-\mu)^3&-6\sigma^2\end{bmatrix} = \begin{bmatrix}-1&0\\0&-1\\-3\sigma^2&0\\0&-6\sigma^2\end{bmatrix}. \end{aligned} \tag{8.43}$$

(Recall that we treat σ^2, not σ, as a parameter.)

The covariance matrix of the scaled moment conditions (at the true parameter values) is

$$S_0 = \mathrm{E}\left[\frac{\sqrt{T}}{T}\sum_{t=1}^{T}m_t(\beta_0)\right]\left[\frac{\sqrt{T}}{T}\sum_{t=1}^{T}m_t(\beta_0)\right]', \tag{8.44}$$

which can be a very messy expression. Assume that there is no autocorrelation in $m_t(\beta_0)$, which would certainly be true if x_t is iid. We can then simplify as

$$S_0 = \mathrm{E}\, m_t(\beta_0) m_t(\beta_0)', \tag{8.45}$$

which is the form we use here for illustration. We therefore have (provided $m_t(\beta_0)$ is not autocorrelated)

$$S_0 = \mathrm{E}\left(\begin{bmatrix} x_t-\mu \\ (x_t-\mu)^2-\sigma^2 \\ (x_t-\mu)^3 \\ (x_t-\mu)^4-3\sigma^4 \end{bmatrix}\right)\left(\begin{bmatrix} x_t-\mu \\ (x_t-\mu)^2-\sigma^2 \\ (x_t-\mu)^3 \\ (x_t-\mu)^4-3\sigma^4 \end{bmatrix}\right)' = \begin{bmatrix} \sigma^2 & 0 & 3\sigma^4 & 0 \\ 0 & 2\sigma^4 & 0 & 12\sigma^6 \\ 3\sigma^4 & 0 & 15\sigma^6 & 0 \\ 0 & 12\sigma^6 & 0 & 96\sigma^8 \end{bmatrix}. \tag{8.46}$$

It is straightforward to derive this result once we have the information in the following remark.

Remark 8.3 *If $X \sim N(\mu,\sigma^2)$, then the first few moments around the mean of a are* $\mathrm{E}(X-\mu)=0$, $\mathrm{E}(X-\mu)^2=\sigma^2$, $\mathrm{E}(X-\mu)^3=0$ *(all odd moments are zero),* $\mathrm{E}(X-\mu)^4=3\sigma^4$, $\mathrm{E}(X-\mu)^6=15\sigma^6$, *and* $\mathrm{E}(X-\mu)^8=105\sigma^8$.

Suppose we use the efficient weighting matrix. The asymptotic covariance matrix of the estimated mean and variance is then $((D_0' S_0^{-1} D_0)^{-1})$

$$\left(\begin{bmatrix} -1 & 0 \\ 0 & -1 \\ -3\sigma^2 & 0 \\ 0 & -6\sigma^2 \end{bmatrix}' \begin{bmatrix} \sigma^2 & 0 & 3\sigma^4 & 0 \\ 0 & 2\sigma^4 & 0 & 12\sigma^6 \\ 3\sigma^4 & 0 & 15\sigma^6 & 0 \\ 0 & 12\sigma^6 & 0 & 96\sigma^8 \end{bmatrix}^{-1} \begin{bmatrix} -1 & 0 \\ 0 & -1 \\ -3\sigma^2 & 0 \\ 0 & -6\sigma^2 \end{bmatrix}\right)^{-1} = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix}^{-1}$$
$$= \begin{bmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{bmatrix}. \tag{8.47}$$

This is the same as the result from maximum likelihood estimation which use the sample mean and sample variance as the estimators. The extra moment conditions (overidentifying restrictions) does not produce any more efficient estimators—for the simple reason that the first two moments completely characterizes the normal distribution.

8.4.2 Testing Normality*

The payoff from the overidentifying restrictions is that we can test if the series is actually normally distributed. There are several ways of doing that, but the perhaps most straightforward is skip the loss function approach to GMM and instead specify the "first order conditions" directly as

$$\begin{aligned} 0 &= A\bar{m} \\ &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \frac{1}{T}\sum_{t=1}^{T} \begin{bmatrix} x_t-\mu \\ (x_t-\mu)^2-\sigma^2 \\ (x_t-\mu)^3 \\ (x_t-\mu)^4-3\sigma^4 \end{bmatrix}. \end{aligned} \tag{8.48}$$

The asymptotic covariance matrix the moment conditions is as in (8.37). In this case, the matrix with brackets is

$$\underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}}_{I_4} - \underbrace{\begin{bmatrix} -1 & 0 \\ 0 & -1 \\ -3\sigma^2 & 0 \\ 0 & -6\sigma^2 \end{bmatrix}}_{D_0} \left(\underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}}_{A_0} \underbrace{\begin{bmatrix} -1 & 0 \\ 0 & -1 \\ -3\sigma^2 & 0 \\ 0 & -6\sigma^2 \end{bmatrix}}_{D_0} \right)^{-1} \underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}}_{A_0}$$
$$= \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -3\sigma^2 & 0 & 1 & 0 \\ 0 & -6\sigma^2 & 0 & 1 \end{bmatrix} \tag{8.49}$$

We therefore get

$$\text{ACov}[\sqrt{T}\bar{m}(\hat{\beta})] = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -3\sigma^2 & 0 & 1 & 0 \\ 0 & -6\sigma^2 & 0 & 1 \end{bmatrix} \begin{bmatrix} \sigma^2 & 0 & 3\sigma^4 & 0 \\ 0 & 2\sigma^4 & 0 & 12\sigma^6 \\ 3\sigma^4 & 0 & 15\sigma^6 & 0 \\ 0 & 12\sigma^6 & 0 & 96\sigma^8 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -3\sigma^2 & 0 & 1 & 0 \\ 0 & -6\sigma^2 & 0 & 1 \end{bmatrix}'$$

$$= \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 6\sigma^6 & 0 \\ 0 & 0 & 0 & 24\sigma^8 \end{bmatrix} \qquad (8.50)$$

We now form the test statistics for the overidentifying restrictions as in (8.40). In this case, it is (note the generalized inverse)

$$T \begin{bmatrix} 0 \\ 0 \\ \Sigma_{t=1}^{T}(x_t-\mu)^3/T \\ \Sigma_{t=1}^{T}[(x_t-\mu)^4-3\sigma^4]/T \end{bmatrix}' \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1/(6\sigma^6) & 0 \\ 0 & 0 & 0 & 1/(24\sigma^8) \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ \Sigma_{t=1}^{T}(x_t-\mu)^3/T \\ \Sigma_{t=1}^{T}[(x_t-\mu)^4-3\sigma^4]/T \end{bmatrix}$$

$$= \frac{T}{6}\frac{\left[\Sigma_{t=1}^{T}(x_t-\mu)^3/T\right]^2}{\sigma^6} + \frac{T}{24}\frac{\left\{\Sigma_{t=1}^{T}[(x_t-\mu)^4-3\sigma^4]/T\right\}^2}{\sigma^8}. \qquad (8.51)$$

When we approximate σ by $\hat{\sigma}$ then this is the same as the *Jarque and Bera test of normality*.

The analysis shows (once again) that we can arrive at simple closed form results by making strong assumptions about the data generating process. In particular, we assumed that the moment conditions were serially uncorrelated. The GMM test, with a modified estimator of the covariance matrix S_0, can typically be much more general.

8.5 Testing the Implications of an RBC Model

Reference: Christiano and Eichenbaum (1992)

This section shows how the GMM framework can be used to test if an RBC model fits data.

Christiano and Eichenbaum (1992) try to test if the RBC model predictions correspond are significantly different from correlations and variances of data. The first step is to define a vector of parameters and some second moments

$$\Psi = \left[\delta, ..., \sigma_\lambda, \frac{\sigma_{c^p}}{\sigma_y}, ..., \text{Corr}\left(\frac{y}{n}, n\right)\right], \qquad (8.52)$$

and estimate it with GMM using moment conditions. One of the moment condition is that the sample average of the labor share in value added equals the coefficient on labor in a Cobb-Douglas production function, another is that just the definitions of a standard deviation, and so forth.

The distribution of the estimator for Ψ is asymptotically normal. Note that the covariance matrix of the moments is calculated similarly to the Newey-West estimator.

The second step is to note that the RBC model generates second moments as a function $h\,(.)$ of the model parameters $\{\delta, ..., \sigma_\lambda\}$, which are in Ψ, that is, the model generated second moments can be thought of as $h\,(\Psi)$.

The third step is to test if the non-linear restrictions of the model (the model mapping from parameters to second moments) are satisfied. That is, the restriction that the model second moments are as in data

$$H\,(\Psi) = h\,(\Psi) - \left[\frac{\sigma_{c^p}}{\sigma_y}, ..., \text{Corr}\left(\frac{y}{n}, n\right)\right] = 0, \qquad (8.53)$$

is tested with a Wald test. (Note that this is much like the $R\beta = 0$ constraints in the linear case.) From the delta-method we get

$$\sqrt{T}H(\hat{\Psi}) \overset{d}{\to} N\left(0, \frac{\partial H}{\partial \Psi'}\text{Cov}(\hat{\Psi})\frac{\partial H'}{\partial \Psi}\right). \qquad (8.54)$$

Forming the quadratic form

$$TH(\hat{\Psi})'\left(\frac{\partial H}{\partial \Psi'}\text{Cov}(\hat{\Psi})\frac{\partial H'}{\partial \Psi}\right)^{-1} H(\hat{\Psi}), \qquad (8.55)$$

will as usual give a χ^2 distributed test statistic with as many degrees of freedoms as restrictions (the number of functions in (8.53)).

8.6 IV on a System of Equations*

Suppose we have two equations

$$\begin{aligned} y_{1t} &= x_{1t}'\beta_1 + u_{1t} \\ y_{2t} &= x_{2t}'\beta_2 + u_{2t}, \end{aligned}$$

and two sets of instruments, z_{1t} and z_{2t} with the same dimensions as x_{1t} and x_{2t}, respectively. The sample moment conditions are

$$\bar{m}(\beta_1, \beta_2) = \frac{1}{T}\sum_{t=1}^{T} \left[\begin{array}{c} z_{1t}\left(y_{1t} - x_{1t}'\beta_1\right) \\ z_{2t}\left(y_{2t} - x_{2t}'\beta_2\right) \end{array} \right],$$

Let $\beta = (\beta_1', \beta_2')'$. Then

$$\begin{aligned} \frac{\partial \bar{m}(\beta_1, \beta_2)}{\partial \beta'} &= \left[\begin{array}{cc} \frac{\partial}{\partial \beta_1'} \frac{1}{T}\sum_{t=1}^{T} z_{1t}\left(y_{1t} - x_{1t}'\beta_1\right) & \frac{\partial}{\partial \beta_2'} \frac{1}{T}\sum_{t=1}^{T} z_{1t}\left(y_{1t} - x_{1t}'\beta_1\right) \\ \frac{\partial}{\partial \beta_1'} \frac{1}{T}\sum_{t=1}^{T} z_{2t}\left(y_{2t} - x_{2t}'\beta_2\right) & \frac{\partial}{\partial \beta_2'} \frac{1}{T}\sum_{t=1}^{T} z_{2t}\left(y_{2t} - x_{2t}'\beta_2\right) \end{array} \right] \\ &= \left[\begin{array}{cc} \frac{1}{T}\sum_{t=1}^{T} z_{1t}x_{1t}' & \mathbf{0} \\ \mathbf{0} & \frac{1}{T}\sum_{t=1}^{T} z_{2t}x_{2t}' \end{array} \right]. \end{aligned}$$

This is invertible so we can premultiply the first order condition with the inverse of $[\partial \bar{m}(\beta)/\partial \beta']' A$ and get $\bar{m}(\beta) = \mathbf{0}_{k\times 1}$. We can solve this system for β_1 and β_2 as

$$\begin{aligned} \left[\begin{array}{c} \beta_1 \\ \beta_2 \end{array} \right] &= \left[\begin{array}{cc} \frac{1}{T}\sum_{t=1}^{T} z_{1t}x_{1t}' & \mathbf{0} \\ \mathbf{0} & \frac{1}{T}\sum_{t=1}^{T} z_{2t}x_{2t}' \end{array} \right]^{-1} \left[\begin{array}{c} \frac{1}{T}\sum_{t=1}^{T} z_{1t}y_{1t} \\ \frac{1}{T}\sum_{t=1}^{T} z_{2t}y_{2t} \end{array} \right] \\ &= \left[\begin{array}{cc} \left(\frac{1}{T}\sum_{t=1}^{T} z_{1t}x_{1t}'\right)^{-1} & \mathbf{0} \\ \mathbf{0} & \left(\frac{1}{T}\sum_{t=1}^{T} z_{2t}x_{2t}'\right)^{-1} \end{array} \right] \left[\begin{array}{c} \frac{1}{T}\sum_{t=1}^{T} z_{1t}y_{1t} \\ \frac{1}{T}\sum_{t=1}^{T} z_{2t}y_{2t} \end{array} \right]. \end{aligned}$$

This is IV on each equation separately, which follows from having an exactly identified system.

Bibliography

Christiano, L. J., and M. Eichenbaum, 1992, "Current Real-Business-Cycle Theories and Aggregate Labor-Market Fluctuations," *American Economic Review*, 82, 430–450.

Davidson, R., and J. G. MacKinnon, 1993, *Estimation and Inference in Econometrics*, Oxford University Press, Oxford.

Greene, W. H., 2000, *Econometric Analysis*, Prentice-Hall, Upper Saddle River, New Jersey, 4th edn.

Hamilton, J. D., 1994, *Time Series Analysis*, Princeton University Press, Princeton.

11 Vector Autoregression (VAR)

Reference: Hamilton (1994) 10-11; Greene (2000) 17.5; Johnston and DiNardo (1997) 9.1-9.2 and Appendix 9.2; and Pindyck and Rubinfeld (1998) 9.2 and 13.5.

Let y_t be an $n \times 1$ vector of variables. The VAR(p) is

$$y_t = \mu + A_1 y_{t-1} + ... + A_p y_{t-p} + \varepsilon_t, \varepsilon_t \text{ is white noise, } \mathrm{Cov}(\varepsilon_t) = \Omega. \tag{11.1}$$

Example 11.1 *(VAR(2) of 2×1 vector.) Let $y_t = [\ x_t \quad z_t\ ]'$. Then*

$$\begin{bmatrix} x_t \\ z_t \end{bmatrix} = \begin{bmatrix} A_{1,11} & A_{1,12} \\ A_{1,21} & A_{1,22} \end{bmatrix} \begin{bmatrix} x_{t-1} \\ z_{t-1} \end{bmatrix} + \begin{bmatrix} A_{2,11} & A_{2,12} \\ A_{2,21} & A_{2,22} \end{bmatrix} \begin{bmatrix} x_{t-2} \\ z_{t-2} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{bmatrix}. \tag{11.2}$$

Issues:

- Variable selection
- Lag length
- Estimation
- Purpose: data description (Granger-causality, impulse response, forecast error variance decomposition), forecasting, policy analysis (Lucas critique)?

11.1 Canonical Form

A VAR(p) can be rewritten as a VAR(1). For instance, a VAR(2) can be written as

$$\begin{bmatrix} y_t \\ y_{t-1} \end{bmatrix} = \begin{bmatrix} \mu \\ 0 \end{bmatrix} + \begin{bmatrix} A_1 & A_2 \\ I & 0 \end{bmatrix} \begin{bmatrix} y_{t-1} \\ y_{t-2} \end{bmatrix} + \begin{bmatrix} \varepsilon_t \\ 0 \end{bmatrix} \text{ or} \tag{11.3}$$

$$y_t^* = \mu^* + A y_{t-1}^* + \varepsilon_t^*. \tag{11.4}$$

Example 11.2 *(Canonical form of a univariate AR(2).)*

$$\begin{bmatrix} y_t \\ y_{t-1} \end{bmatrix} = \begin{bmatrix} \mu \\ 0 \end{bmatrix} + \begin{bmatrix} a_1 & a_2 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} y_{t-1} \\ y_{t-2} \end{bmatrix} + \begin{bmatrix} \varepsilon_t \\ 0 \end{bmatrix}.$$

Example 11.3 *(Canonical for of VAR(2) of 2×1 vector.) Continuing on the previous example, we get*

$$\begin{bmatrix} x_t \\ z_t \\ x_{t-1} \\ z_{t-1} \end{bmatrix} = \begin{bmatrix} A_{1,11} & A_{1,11} & A_{2,11} & A_{2,12} \\ A_{1,21} & A_{1,22} & A_{2,21} & A_{2,22} \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_{t-1} \\ z_{t-1} \\ x_{t-2} \\ z_{t-2} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \\ 0 \\ 0 \end{bmatrix}.$$

11.2 Moving Average Form and Stability

Consider a VAR(1), or a VAR(1) representation of a VAR(p) or an AR(p)

$$y_t^* = A y_{t-1}^* + \varepsilon_t^*. \tag{11.5}$$

Solve recursively backwards (substitute for $y_{t-s}^* = A y_{t-s-1}^* + \varepsilon_{t-s}^*$, $s = 1, 2,...$) to get the vector moving average representation (VMA), or impulse response function

$$\begin{aligned} y_t^* &= A\left(A y_{t-2}^* + \varepsilon_{t-1}^*\right) + \varepsilon_t^* \\ &= A^2 y_{t-2}^* + A\varepsilon_{t-1}^* + \varepsilon_t^* \\ &= A^2\left(A y_{t-3}^* + \varepsilon_{t-2}^*\right) + A\varepsilon_{t-1}^* + \varepsilon_t^* \\ &= A^3 y_{t-3}^* + A^2\varepsilon_{t-2}^* + A\varepsilon_{t-1}^* + \varepsilon_t^* \\ &\vdots \\ &= A^{K+1} y_{t-K-1}^* + \sum_{s=0}^{K} A^s \varepsilon_{t-s}^*. \end{aligned} \tag{11.6}$$

Remark 11.4 *(Spectral decomposition.) The n eigenvalues (λ_i) and associated eigenvectors (z_i) of the $n \times n$ matrix A satisfies*

$$(A - \lambda_i I_n)\, z_i = \mathbf{0}_{n\times 1}.$$

If the eigenvectors are linearly independent, then

$$A = Z\Lambda Z^{-1}, \text{ where } \Lambda = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix} \text{ and } Z = \begin{bmatrix} z_1 & z_2 & \cdots & z_n \end{bmatrix}.$$

Note that we therefore get

$$A^2 = AA = Z\Lambda Z^{-1} Z\Lambda Z^{-1} = Z\Lambda\Lambda Z^{-1} = Z\Lambda^2 Z^{-1} \Rightarrow A^q = Z\Lambda^q Z^{-1}.$$

Remark 11.5 *(Modulus of complex number.) If $\lambda = a + bi$, where $i = \sqrt{-1}$, then $|\lambda| = |a + bi| = \sqrt{a^2 + b^2}$.*

We want $\lim_{K\to\infty} A^{K+1} y^*_{t-K-1} = 0$ (stable VAR) to get a moving average representation of y_t (where the influence of the starting values vanishes asymptotically). We note from the spectral decompositions that $A^{K+1} = Z\Lambda^{K+1}Z^{-1}$, where Z is the matrix of eigenvectors and Λ a diagonal matrix with eigenvalues. Clearly, $\lim_{K\to\infty} A^{K+1} y^*_{t-K-1} = 0$ is satisfied if the eigenvalues of A are all less than one in modulus.

Example 11.6 *(AR(1).) For the univariate AR(1) $y_t = ay_{t-1} + \varepsilon_t$, the characteristic equation is $(a - \lambda)\, z = 0$, which is only satisfied if the eigenvalue is $\lambda = a$. The AR(1) is therefore stable (and stationarity) if $-1 < a < 1$.*

If we have a stable VAR, then (11.6) can be written

$$\begin{aligned} y^*_t &= \sum_{s=0}^{\infty} A^s \varepsilon^*_{t-s} \qquad (11.7) \\ &= \varepsilon^*_t + A\varepsilon^*_{t-1} + A^2\varepsilon^*_{t-2} + ... \end{aligned}$$

We may pick out the first n equations from (11.7) (to extract the "original" variables from the canonical form) and write them as

$$y_t = \varepsilon_t + C_1\varepsilon_{t-1} + C_2\varepsilon_{t-2} + ..., \qquad (11.8)$$

which is the vector moving average, VMA, form of the VAR.

Example 11.7 *(AR(2), Example (11.2) continued.) Let $\mu = 0$ in 11.2 and note that the VMA of the canonical form is*

$$\begin{bmatrix} y_t \\ y_{t-1} \end{bmatrix} = \begin{bmatrix} \varepsilon_t \\ 0 \end{bmatrix} + \begin{bmatrix} a_1 & a_2 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \varepsilon_{t-1} \\ 0 \end{bmatrix} + \begin{bmatrix} a_1^2 + a_2 & a_1 a_2 \\ a_1 & a_2 \end{bmatrix} \begin{bmatrix} \varepsilon_{t-2} \\ 0 \end{bmatrix} + ...$$

The MA of y_t is therefore

$$y_t = \varepsilon_t + a_1\varepsilon_{t-1} + \left(a_1^2 + a_2\right)\varepsilon_{t-2} + ...$$

Note that

$$\frac{\partial y_t}{\partial \varepsilon'_{t-s}} = C_s \text{ or } \frac{\partial \mathrm{E}_t y_{t+s}}{\partial \varepsilon'_t} = C_s, \text{ with } C_0 = I \qquad (11.9)$$

so the *impulse response function* is given by $\{I, C_1, C_2, ...\}$. Note that it is typically only meaningful to discuss impulse responses to uncorrelated shocks with economic interpretations. The idea behind structural VARs (discussed below) is to impose enough restrictions to achieve this.

Example 11.8 *(Impulse response function for AR(1).) Let $y_t = \rho y_{t-1} + \varepsilon_t$. The MA representation is $y_t = \sum_{s=0}^{t} \rho^s \varepsilon_{t-s}$, so $\partial y_t / \partial \varepsilon_{t-s} = \partial E_t y_{t+s} / \partial \varepsilon_t = \rho^s$. Stability requires $|\rho| < 1$, so the effect of the initial value eventually dies off ($\lim_{s\to\infty} \partial y_t / \partial \varepsilon_{t-s} = 0$).*

Example 11.9 *(Numerical VAR(1) of 2×1 vector.) Consider the VAR(1)*

$$\begin{bmatrix} x_t \\ z_t \end{bmatrix} = \begin{bmatrix} 0.5 & 0.2 \\ 0.1 & -0.3 \end{bmatrix} \begin{bmatrix} x_{t-1} \\ z_{t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{bmatrix}.$$

The eigenvalues are approximately 0.52 and -0.32, so this is a stable VAR. The VMA is

$$\begin{bmatrix} x_t \\ z_t \end{bmatrix} = \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{bmatrix} + \begin{bmatrix} 0.5 & 0.2 \\ 0.1 & -0.3 \end{bmatrix} \begin{bmatrix} \varepsilon_{1,t-1} \\ \varepsilon_{2,t-1} \end{bmatrix} + \begin{bmatrix} 0.27 & 0.04 \\ 0.02 & 0.11 \end{bmatrix} \begin{bmatrix} \varepsilon_{1,t-2} \\ \varepsilon_{2,t-2} \end{bmatrix} + ...$$

11.3 Estimation

The MLE, conditional on the initial observations, of the VAR is the same as OLS estimates of each equation separately. The MLE of the ij^{th} element in $\mathrm{Cov}(\varepsilon_t)$ is given by $\sum_{t=1}^{T} \hat{v}_{it}\hat{v}_{jt}/T$, where $\hat{v}_{it}$ and $\hat{v}_{jt}$ are the OLS residuals.

Note that the VAR system is a system of "seemingly unrelated regressions," with the same regressors in each equation. The OLS on each equation is therefore the GLS, which coincides with MLE if the errors are normally distributed.

11.4 Granger Causality

Main message: Granger-causality might be useful, but it is not the same as causality.

Definition: if z cannot help forecast x, then z does not Granger-cause x; the MSE of the forecast $\mathrm{E}(x_t | x_{t-s}, z_{t-s}, s > 0)$ equals the MSE of the forecast $\mathrm{E}(x_t | x_{t-s}, s > 0)$.

Test: Redefine the dimensions of x_t and z_t in (11.2): let x_t be $n_1 \times 1$ and z_t is $n_2 \times 1$. If the $n_1 \times n_2$ matrices $A_{1,12} = \mathbf{0}$ and $A_{2,12} = \mathbf{0}$, then z fail to Granger-cause x. (In general, we would require $A_{s,12} = \mathbf{0}$ for $s = 1, ..., p$.) This carries over to the MA representation in (11.8), so $C_{s,12} = \mathbf{0}$.

These restrictions can be tested with an F-test. The easiest case is when x is a scalar, since we then simply have a set of linear restrictions on a single OLS regression.

Example 11.10 *(RBC and nominal neutrality.) Suppose we have an RBC model which says that money has no effect on the real variables (for instance, output, capital stock, and the productivity level). Money stock should not Granger-cause real variables.*

Example 11.11 *(Granger causality and causality.) Do Christmas cards cause Christmas?*

Example 11.12 *(Granger causality and causality II, from Hamilton 11.) Consider the price P_t of an asset paying dividends D_t. Suppose the expected return ($E_t(P_{t+1} + D_{t+1})/P_t$) is a constant, R. The price then satisfies P_t =$E_t \sum_{s=1}^{\infty} R^{-s} D_{t+s}$. Suppose $D_t = u_t + \delta u_{t-1} + v_t$, so $E_t D_{t+1} = \delta u_t$ and $E_t D_{t+s} = 0$ for $s > 1$. This gives $P_t = \delta u_t / R$, and $D_t = u_t + v_t + R P_{t-1}$, so the VAR is*

$$\begin{bmatrix} P_t \\ D_t \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ R & 0 \end{bmatrix} \begin{bmatrix} P_{t-1} \\ D_{t-1} \end{bmatrix} + \begin{bmatrix} \delta u_t / R \\ u_t + v_t \end{bmatrix},$$

where P Granger-causes D. Of course, the true causality is from D to P. Problem: forward looking behavior.

Example 11.13 *(Money and output, Sims (1972).) Sims found that output, y does not Granger-cause money, m, but that m Granger causes y. His interpretation was that money supply is exogenous (set by the Fed) and that money has real effects. Notice how he used a combination of two Granger causality test to make an economic interpretation.*

Example 11.14 *(Granger causality and omitted information.*) Consider the VAR*

$$\begin{bmatrix} y_{1t} \\ y_{2t} \\ y_{3t} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & 0 \\ 0 & a_{22} & 0 \\ 0 & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} y_{1t-1} \\ y_{2t-1} \\ y_{3t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \\ \varepsilon_{3t} \end{bmatrix}$$

Notice that y_{2t} and y_{3t} do not depend on y_{1t-1}, so the latter should not be able to Granger-cause y_{3t}. However, suppose we forget to use y_{2t} in the regression and then ask if y_{1t} Granger causes y_{3t}. The answer might very well be yes since y_{1t-1} contains information about y_{2t-1} which does affect y_{3t}. (If you let y_{1t} be money, y_{2t} be the (autocorrelated) Solow residual, and y_{3t} be output, then this is a short version of the comment in King (1986) comment on Bernanke (1986) (see below) on why money may appear to Granger-cause output). Also note that adding a nominal interest rate to Sims (see above) money-output VAR showed that money cannot be taken to be exogenous.

11.5 Forecasts Forecast Error Variance

The error forecast of the s period ahead forecast is

$$y_{t+s} - \mathrm{E}_t y_{t+s} = \varepsilon_{t+s} + C_1 \varepsilon_{t+s-1} + ... + C_{s-1} \varepsilon_{t+1}, \tag{11.10}$$

so the covariance matrix of the (s periods ahead) forecasting errors is

$$\mathrm{E}\left(y_{t+s} - \mathrm{E}_t y_{t+s}\right)\left(y_{t+s} - \mathrm{E}_t y_{t+s}\right)' = \Omega + C_1 \Omega C_1' + ... + C_{s-1} \Omega C_{s-1}'. \tag{11.11}$$

For a VAR(1), $C_s = A^s$, so we have

$$y_{t+s} - \mathrm{E}_t y_{t+s} = \varepsilon_{t+s} + A \varepsilon_{t+s-1} + ... + A^s \varepsilon_{t+1}, \text{ and} \tag{11.12}$$

$$\mathrm{E}\left(y_{t+s} - \mathrm{E}_t y_{t+s}\right)\left(y_{t+s} - \mathrm{E}_t y_{t+s}\right)' = \Omega + A \Omega A' + ... + A^{s-1} \Omega (A^{s-1})'. \tag{11.13}$$

Note that $\lim_{s \to \infty} \mathrm{E}_t y_{t+s} = 0$, that is, the forecast goes to the unconditional mean (which is zero here, since there are no constants - you could think of y_t as a deviation from the mean). Consequently, the forecast error becomes the VMA representation (11.8). Similarly, the forecast error variance goes to the unconditional variance.

Example 11.15 *(Unconditional variance of VAR(1).) Letting $s \to \infty$ in (11.13) gives*

$$
\begin{aligned}
Ey_t y_t' &= \sum_{s=0}^{\infty} A^s \Omega \left(A^s\right)' \\
&= \Omega + [A\Omega A' + A^2\Omega(A^2)' + ...] \\
&= \Omega + A\left(\Omega + A\Omega A' + ...\right) A' \\
&= \Omega + A(Ey_t y_t')A',
\end{aligned}
$$

which suggests that we can calculate $Ey_t y_t'$ by an iteration (backwards in time) $\Phi_t = \Omega + A\Phi_{t+1}A'$, starting from $\Phi_T = I$, until convergence.

11.6 Forecast Error Variance Decompositions*

If the shocks are uncorrelated, then it is often useful to calculate the fraction of Var($y_{i,t+s}$−E$_t y_{i,t+s}$) due to the j^{th} shock, the *forecast error variance decomposition*. Suppose the covariance matrix of the shocks, here Ω, is a diagonal $n \times n$ matrix with the variances ω_{ii} along the diagonal. Let c_{qi} be the ith column of C_q. We then have

$$
C_q \Omega C_q' = \sum_{i=1}^{n} \omega_{ii} c_{qi} \left(c_{qi}\right)'. \tag{11.14}
$$

Example 11.16 *(Illustration of (11.14) with $n = 2$.) Suppose*

$$
C_q = \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix} \text{ and } \Omega = \begin{bmatrix} \omega_{11} & 0 \\ 0 & \omega_{22} \end{bmatrix},
$$

then

$$
C_q \Omega C_q' = \begin{bmatrix} \omega_{11}c_{11}^2 + \omega_{22}c_{12}^2 & \omega_{11}c_{11}c_{21} + \omega_{22}c_{12}c_{22} \\ \omega_{11}c_{11}c_{21} + \omega_{22}c_{12}c_{22} & \omega_{11}c_{21}^2 + \omega_{22}c_{22}^2 \end{bmatrix},
$$

which should be compared with

$$
\begin{aligned}
&\omega_{11} \begin{bmatrix} c_{11} \\ c_{21} \end{bmatrix} \begin{bmatrix} c_{11} \\ c_{21} \end{bmatrix}' + \omega_{22} \begin{bmatrix} c_{12} \\ c_{22} \end{bmatrix} \begin{bmatrix} c_{12} \\ c_{22} \end{bmatrix}' \\
&= \omega_{11} \begin{bmatrix} c_{11}^2 & c_{11}c_{21} \\ c_{11}c_{21} & c_{21}^2 \end{bmatrix} + \omega_{22} \begin{bmatrix} c_{12}^2 & c_{12}c_{22} \\ c_{12}c_{22} & c_{22}^2 \end{bmatrix}.
\end{aligned}
$$

Applying this on (11.11) gives

$$
\begin{aligned}
\mathrm{E}\left(y_{t+s} - \mathrm{E}_t y_{t+s}\right)\left(y_{t+s} - \mathrm{E}_t y_{t+s}\right)' &= \sum_{i=1}^{n} \omega_{ii} I + \sum_{i=1}^{n} \omega_{ii} c_{1i} \left(c_{1i}\right)' + ... + \sum_{i=1}^{n} \omega_{ii} c_{s-1i} \left(c_{s-1i}\right)' \\
&= \sum_{i=1}^{n} \omega_{ii} \left[I + c_{1i}\left(c_{1i}\right)' + ... + c_{s-1i}\left(c_{s-1i}\right)'\right],
\end{aligned} \tag{11.15}
$$

which shows how the covariance matrix for the s-period forecast errors can be decomposed into its n components.

11.7 Structural VARs

11.7.1 Structural and Reduced Forms

We are usually not interested in the impulse response function (11.8) or the variance decomposition (11.11) with respect to ε_t, but with respect to some structural shocks, u_t, which have clearer interpretations (technology, monetary policy shock, etc.).

Suppose the *structural form* of the model is

$$
Fy_t = \alpha + B_1 y_{t-1} + ... + B_p y_{t-p} + u_t, u_t \text{ is white noise, Cov}(u_t) = D. \tag{11.16}
$$

This could, for instance, be an economic model derived from theory.[1]

Provided F^{-1} exists, it is possible to write the time series process as

$$
\begin{aligned}
y_t &= F^{-1}\alpha + F^{-1}B_1 y_{t-1} + ... + F^{-1}B_p y_{t-p} + F^{-1}u_t & (11.17)\\
&= \mu + A_1 y_{t-1} + ... + A_p y_{t-p} + \varepsilon_t, \text{ Cov}\left(\varepsilon_t\right) = \Omega, & (11.18)
\end{aligned}
$$

[1] This is a "structural model" in a traditional, Cowles commission, sense. This might be different from what modern macroeconomists would call structural.

where

$$\mu = F^{-1}\alpha,\ A_s = F^{-1}B_s, \text{ and } \varepsilon_t = F^{-1}u_t \text{ so } \Omega = F^{-1}D\left(F^{-1}\right)'. \tag{11.19}$$

Equation (11.18) is a VAR model, so a VAR can be thought of as a *reduced form* of the structural model (11.16).

The key to understanding the relation between the structural model and the VAR is the F matrix, which controls how the endogenous variables, y_t, are linked to each other *contemporaneously*. In fact, *identification of a VAR* amounts to choosing an F matrix. Once that is done, impulse responses and forecast error variance decompositions can be made with respect to the structural shocks. For instance, the impulse response function of the VAR, (11.8), can be rewritten in terms of $u_t = F\varepsilon_t$ (from (11.19))

$$\begin{aligned} y_t &= \varepsilon_t + C_1\varepsilon_{t-1} + C_2\varepsilon_{t-2} + ... \\ &= F^{-1}F\varepsilon_t + C_1F^{-1}F\varepsilon_{t-1} + C_2F^{-1}F\varepsilon_{t-2} + ... \\ &= F^{-1}u_t + C_1F^{-1}u_{t-1} + C_2F^{-1}u_{t-2} + ... \end{aligned} \tag{11.20}$$

Remark 11.17 *The easiest way to calculate this representation is by first finding F^{-1} (see below), then writing (11.18) as*

$$y_t = \mu + A_1y_{t-1} + ... + A_py_{t-p} + F^{-1}u_t. \tag{11.21}$$

To calculate the impulse responses to the first element in u_t, set $y_{t-1}, ..., y_{t-p}$ equal to the long-run average, $(I - A_1 - ... - Ap)^{-1}\mu$, make the first element in u_t unity and all other elements zero. Calculate the response by iterating forward on (11.21), but putting all elements in $u_{t+1}, u_{t+2}, ...$ to zero. This procedure can be repeated for the other elements of u_t.

We would typically pick F such that the elements in u_t are uncorrelated with each other, so they have a clear interpretation.

The VAR form can be estimated directly from data. Is it then possible to recover the structural parameters in (11.16) from the estimated VAR (11.18)? Not without restrictions on the structural parameters in F, B_s, α, and D. To see why, note that in the structural form (11.16) we have $(p+1)\,n^2$ parameters in $\{F, B_1, \ldots, B_p\}$, n parameters in α, and $n(n+1)/2$ unique parameters in D (it is symmetric). In the VAR (11.18) we have fewer

parameters: pn^2 in $\{A_1, \ldots, A_p\}$, n parameters in in μ, and $n(n+1)/2$ unique parameters in Ω. This means that we have to impose at least n^2 restrictions on the structural parameters $\{F, B_1, \ldots, B_p, \alpha, D\}$ to identify all of them. This means, of course, that many different structural models have can have exactly the same reduced form.

Example 11.18 *(Structural form of the 2×1 case.) Suppose the structural form of the previous example is*

$$\begin{bmatrix} F_{11} & F_{12} \\ F_{21} & F_{22} \end{bmatrix}\begin{bmatrix} x_t \\ z_t \end{bmatrix} = \begin{bmatrix} B_{1,11} & B_{1,12} \\ B_{1,21} & B_{1,22} \end{bmatrix}\begin{bmatrix} x_{t-1} \\ z_{t-1} \end{bmatrix} + \begin{bmatrix} B_{2,11} & B_{2,12} \\ B_{2,21} & B_{2,22} \end{bmatrix}\begin{bmatrix} x_{t-2} \\ z_{t-2} \end{bmatrix} + \begin{bmatrix} u_{1,t} \\ u_{2,t} \end{bmatrix}.$$

This structural form has $3 \times 4 + 3$ unique parameters. The VAR in (11.2) has $2 \times 4 + 3$. We need at least 4 restrictions on $\{F, B_1, B_2, D\}$ to identify them from $\{A_1, A_2, \Omega\}$.

11.7.2 "Triangular" Identification 1: Triangular F with $F_{ii} = 1$ and Diagonal D

Reference: Sims (1980).

The perhaps most common way to achieve identification of the structural parameters is to restrict the contemporaneous response of the different endogenous variables, y_t, to the different structural shocks, u_t. Within in this class of restrictions, the triangular identification is the most popular: assume that F is lower triangular ($n(n+1)/2$ restrictions) with diagonal element equal to unity, and that D is diagonal ($n(n-1)/2$ restrictions), which gives n^2 restrictions (exact identification).

A lower triangular F matrix is very restrictive. It means that the first variable can react to lags and the first shock, the second variable to lags and the first two shocks, etc. This is a recursive simultaneous equations model, and we obviously need to be careful with how we order the variables. The assumptions that $F_{ii} = 1$ is just a normalization.

A *diagonal D matrix* seems to be something that we would often like to have in a structural form in order to interpret the shocks as, for instance, demand and supply shocks. The diagonal elements of D are the variances of the structural shocks.

Example 11.19 *(Lower triangular F: going from structural form to VAR.) Suppose the structural form is*

$$\begin{bmatrix} 1 & 0 \\ -\alpha & 1 \end{bmatrix}\begin{bmatrix} x_t \\ z_t \end{bmatrix} = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}\begin{bmatrix} x_{t-1} \\ z_{t-1} \end{bmatrix} + \begin{bmatrix} u_{1,t} \\ u_{2,t} \end{bmatrix}.$$

This is a recursive system where x_t does not not depend on the contemporaneous z_t, and therefore not on the contemporaneous u_{2t} (see first equation). However, z_t does depend on x_t (second equation). The VAR (reduced form) is obtained by premultiplying by F^{-1}

$$\begin{bmatrix} x_t \\ z_t \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \alpha & 1 \end{bmatrix} \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} \begin{bmatrix} x_{t-1} \\ z_{t-1} \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ \alpha & 1 \end{bmatrix} \begin{bmatrix} u_{1,t} \\ u_{2,t} \end{bmatrix}$$
$$= \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} x_{t-1} \\ z_{t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{bmatrix}.$$

This means that $\varepsilon_{1t} = u_{1t}$, so the first VAR shock equals the first structural shock. In contrast, $\varepsilon_{2,t} = \alpha u_{1,t} + u_{2,t}$, so the second VAR shock is a linear combination of the first two shocks. The covariance matrix of the VAR shocks is therefore

$$Cov \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{bmatrix} = \begin{bmatrix} Var(u_{1t}) & \alpha Var(u_{1t}) \\ \alpha Var(u_{1t}) & \alpha^2 Var(u_{1t}) + Var(u_{2t}) \end{bmatrix}.$$

This set of identifying restrictions can be implemented by estimating the structural form with LS—equation by equation. The reason is that this is just the old fashioned fully recursive system of simultaneous equations. See, for instance, Greene (2000) 16.3.

11.7.3 "Triangular" Identification 2: Triangular F and $D = I$

The identifying restrictions in Section 11.7.2 is actually the same as assuming that F is triangular and that $D = I$. In this latter case, the restriction on the diagonal elements of F has been moved to the diagonal elements of D. This is just a change of normalization (that the structural shocks have unit variance). It happens that this alternative normalization is fairly convenient when we want to estimate the VAR first and then recover the structural parameters from the VAR estimates.

Example 11.20 *(Change of normalization in Example 11.19) Suppose the structural shocks in Example 11.19 have the covariance matrix*

$$D = Cov \begin{bmatrix} u_{1,t} \\ u_{2,t} \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}.$$

Premultiply the structural form in Example 11.19 by

$$\begin{bmatrix} 1/\sigma_1 & 0 \\ 0 & 1/\sigma_2 \end{bmatrix}$$

to get

$$\begin{bmatrix} 1/\sigma_1 & 0 \\ -\alpha/\sigma_2 & 1/\sigma_2 \end{bmatrix} \begin{bmatrix} x_t \\ z_t \end{bmatrix} = \begin{bmatrix} B_{11}/\sigma_1 & B_{12}/\sigma_1 \\ B_{21}/\sigma_2 & B_{22}/\sigma_2 \end{bmatrix} \begin{bmatrix} x_{t-1} \\ z_{t-1} \end{bmatrix} + \begin{bmatrix} u_{1,t}/\sigma_1 \\ u_{2,t}/\sigma_2 \end{bmatrix}.$$

This structural form has a triangular F matrix (with diagonal elements that can be different from unity), and a covariance matrix equal to an identity matrix.

The reason why this alternative normalization is convenient is that it allows us to use the widely available Cholesky decomposition.

Remark 11.21 *(Cholesky decomposition) Let Ω be an $n \times n$ symmetric positive definite matrix. The Cholesky decomposition gives the unique lower triangular P such that $\Omega = PP'$ (some software returns an upper triangular matrix, that is, Q in $\Omega = Q'Q$ instead).*

Remark 11.22 *Note the following two important features of the Cholesky decomposition. First, each column of P is only identified up to a sign transformation; they can be reversed at will. Second, the diagonal elements in P are typically not unity.*

Remark 11.23 *(Changing sign of column and inverting.) Suppose the square matrix A_2 is the same as A_1 except that the i^{th} and j^{th} columns have the reverse signs. Then A_2^{-1} is the same as A_1^{-1} except that the i^{th} and j^{th} rows have the reverse sign.*

This set of identifying restrictions can be implemented by estimating the VAR with LS and then take the following steps.

- *Step 1.* From (11.19) $\Omega = F^{-1} I \left(F^{-1}\right)'$ (recall $D = I$ is assumed), so a Cholesky decomposition recovers F^{-1} (lower triangular F gives a similar structure of F^{-1}, and vice versa, so this works). The signs of each column of F^{-1} can be chosen freely, for instance, so that a productivity shock gets a positive, rather than negative, effect on output. Invert F^{-1} to get F.

- *Step 2.* Invert the expressions in (11.19) to calculate the structural parameters from the VAR parameters as $\alpha = F\mu$, and $B_s = FA_s$.

Example 11.24 *(Identification of the 2×1 case.) Suppose the structural form of the previous example is*

$$\begin{bmatrix} F_{11} & 0 \\ F_{21} & F_{22} \end{bmatrix} \begin{bmatrix} x_t \\ z_t \end{bmatrix} = \begin{bmatrix} B_{1,11} & B_{1,12} \\ B_{1,21} & B_{1,22} \end{bmatrix} \begin{bmatrix} x_{t-1} \\ z_{t-1} \end{bmatrix} + \begin{bmatrix} B_{2,11} & B_{2,12} \\ B_{2,21} & B_{2,22} \end{bmatrix} \begin{bmatrix} x_{t-2} \\ z_{t-2} \end{bmatrix} + \begin{bmatrix} u_{1,t} \\ u_{2,t} \end{bmatrix},$$

$$\textit{with } D = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Step 1 above solves

$$\begin{bmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{12} & \Omega_{22} \end{bmatrix} = \begin{bmatrix} F_{11} & 0 \\ F_{21} & F_{22} \end{bmatrix}^{-1} \left(\begin{bmatrix} F_{11} & 0 \\ F_{21} & F_{22} \end{bmatrix}^{-1} \right)'$$

$$= \begin{bmatrix} \frac{1}{F_{11}^2} & -\frac{F_{21}}{F_{11}^2 F_{22}} \\ -\frac{F_{21}}{F_{11}^2 F_{22}} & \frac{F_{21}^2 + F_{11}^2}{F_{11}^2 F_{22}^2} \end{bmatrix}$$

for the three unknowns F_{11}, F_{21}, and F_{22} in terms of the known Ω_{11}, Ω_{12}, and Ω_{22}. Note that the identifying restrictions are that $D = I$ (three restrictions) and $F_{12} = 0$ (one restriction). (This system is just four nonlinear equations in three unknown - one of the equations for Ω_{12} is redundant. You do not need the Cholesky decomposition to solve it, since it could be solved with any numerical solver of non-linear equations—but why make life even more miserable?)

A practical consequence of this normalization is that the impulse response of shock i equal to unity is exactly the same as the impulse response of shock i equal to Std(u_{it}) in the normalization in Section 11.7.2.

11.7.4 Other Identification Schemes*

Reference: Bernanke (1986).

Not all economic models can be written in this recursive form. However, there are often cross-restrictions between different elements in F or between elements in F and D, or some other type of restrictions on F which may allow us to identify the system.

Suppose we have (estimated) the parameters of the VAR (11.18), and that we want to impose $D =$Cov$(u_t) = I$. From (11.19) we then have $(D = I)$

$$\Omega = F^{-1} \left(F^{-1}\right)'. \tag{11.22}$$

As before we need $n(n-1)/2$ restrictions on F, but this time we don't want to impose the restriction that all elements in F above the principal diagonal are zero. Given these restrictions (whatever they are), we can solve for the remaining elements in B, typically with a numerical method for solving systems of non-linear equations.

11.7.5 What if the VAR Shocks are Uncorrelated ($\Omega = I$)?*

Suppose we estimate a VAR and find that the covariance matrix of the estimated residuals is (almost) an identity matrix (or diagonal). Does this mean that the identification is superfluous? No, not in general. Yes, if we also want to impose the restrictions that F is triangular.

There are many ways to reshuffle the shocks and still get orthogonal shocks. Recall that the structural shocks are linear functions of the VAR shocks, $u_t = F\varepsilon_t$, and that we assume that Cov$(\varepsilon_t) = \Omega = I$ and we want Cov$(u_t) = I$, that, is from (11.19) we then have $(D = I)$

$$FF' = I. \tag{11.23}$$

There are many such F matrices: the class of those matrices even have a name: orthogonal matrices (all columns in F are orthonormal). However, there is only one lower triangular F which satisfies (11.23) (the one returned by a Cholesky decomposition, which is I).

Suppose you know that F is lower triangular (and you intend to use this as the identifying assumption), but that your estimated Ω is (almost, at least) diagonal. The logic then requires that F is not only lower triangular, but also diagonal. This means that $u_t = \varepsilon_t$ (up to a scaling factor). Therefore, a finding that the VAR shocks are uncorrelated combined with the identifying restriction that F is triangular implies that the structural and reduced form shocks are proportional. We can draw no such conclusion if the identifying assumption is something else than lower triangularity.

Example 11.25 *(Rotation of vectors ("Givens rotations").) Consider the transformation of the vector ε into the vector u, $u = G'\varepsilon$, where $G = I_n$ except that $G_{ik} = c$, $G_{ik} = s$, $G_{ki} = -s$, and $G_{kk} = c$. If we let $c = \cos\theta$ and $s = \sin\theta$ for some angle θ, then*

$G'G = I$. To see this, consider the simple example where $i = 2$ and $k = 3$

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & c & s \\ 0 & -s & c \end{bmatrix}' \begin{bmatrix} 1 & 0 & 0 \\ 0 & c & s \\ 0 & -s & c \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & c^2+s^2 & 0 \\ 0 & 0 & c^2+s^2 \end{bmatrix},$$

which is an identity matrix since $\cos^2\theta + \sin^2\theta = 1$. The transformation $u = G'\varepsilon$ gives

$$\begin{aligned} u_t &= \varepsilon_t \, for \, t \neq i, k \\ u_i &= \varepsilon_i c - \varepsilon_k s \\ u_k &= \varepsilon_i s + \varepsilon_k c. \end{aligned}$$

The effect of this transformation is to rotate the i^{th} and k^{th} vectors counterclockwise through an angle of θ. (Try it in two dimensions.) There is an infinite number of such transformations (apply a sequence of such transformations with different i and k, change θ, etc.).

Example 11.26 *(Givens rotations and the F matrix.) We could take F in (11.23) to be (the transpose) of any such sequence of givens rotations. For instance, if G_1 and G_2 are givens rotations, then $F = G_1'$ or $F = G_2'$ or $F = G_1'G_2'$ are all valid.*

11.7.6 Identification via Long-Run Restrictions - No Cointegration*

Suppose we have estimated a VAR system (11.1) for the first differences of some variables $y_t = \Delta x_t$, and that we have calculated the impulse response function as in (11.8), which we rewrite as

$$\begin{aligned} \Delta x_t &= \varepsilon_t + C_1\varepsilon_{t-1} + C_2\varepsilon_{t-2} + ... \\ &= C\,(\mathrm{L})\,\varepsilon_t, \text{ with } \mathrm{Cov}(\varepsilon_t) = \Omega. \end{aligned} \tag{11.24}$$

To find the MA of the level of x_t, we solve recursively

$$\begin{aligned} x_t &= C\,(\mathrm{L})\,\varepsilon_t + x_{t-1} \\ &= C\,(\mathrm{L})\,\varepsilon_t + C\,(\mathrm{L})\,\varepsilon_{t-1} + x_{t-2} \\ &\vdots \\ &= C\,(\mathrm{L})\,(\varepsilon_t + \varepsilon_{t-1} + \varepsilon_{t-2} + ...) \\ &= \varepsilon_t + (C_1 + I)\,\varepsilon_{t-1} + (C_2 + C_1 + I)\,\varepsilon_{t-2} + ... \\ &= C^+\,(\mathrm{L})\,\varepsilon_t, \text{ where } C_s^+ = \sum_{j=0}^{s} C_s \text{ with } C_0 = I. \end{aligned} \tag{11.25}$$

As before the structural shocks, u_t, are

$$u_t = F\varepsilon_t \text{ with } \mathrm{Cov}(u_t) = D.$$

The VMA in term of the structural shocks is therefore

$$x_t = C^+\,(\mathrm{L})\,F^{-1}u_t, \text{ where } C_s^+ = \sum_{j=0}^{s} C_s \text{ with } C_0 = I. \tag{11.26}$$

The C^+ (L) polynomial is known from the estimation, so we need to identify F in order to use this equation for impulse response function and variance decompositions with respect to the structural shocks.

As before we assume that $D = I$, so

$$\Omega = F^{-1}D\left(F^{-1}\right)' \tag{11.27}$$

in (11.19) gives $n(n+1)/2$ restrictions.

We now add restrictions on the long run impulse responses. From (11.26) we have

$$\begin{aligned} \lim_{s\to\infty} \frac{\partial x_{t+s}}{\partial u_t'} &= \lim_{s\to\infty} C_s^+ F^{-1} \\ &= C(1)F^{-1}, \end{aligned} \tag{11.28}$$

where $C(1) = \sum_{j=0}^{\infty} C_s$. We impose $n(n-1)/2$ restrictions on these long run responses. Together we have n^2 restrictions, which allows to identify all elements in F.

In general, (11.27) and (11.28) is a set of non-linear equations which have to solved for the elements in F. However, it is common to assume that (11.28) is a lower triangular matrix. We can then use the following "trick" to find F. Since $\varepsilon_t = F^{-1}u_t$

$$\begin{aligned} \mathrm{E}C(1)\varepsilon_t\varepsilon_t' C(1)' &= \mathrm{E}C(1)F^{-1}u_t u_t' \left(F^{-1}\right)' C(1)' \\ C(1)\Omega C(1)' &= C(1)F^{-1}\left(F^{-1}\right)' C(1)'. \end{aligned} \tag{11.29}$$

We can therefore solve for a lower triangular matrix

$$\Lambda = C(1)F^{-1} \tag{11.30}$$

by calculating the Cholesky decomposition of the left hand side of (11.29) (which is available from the VAR estimate). Finally, we solve for F^{-1} from (11.30).

Example 11.27 *(The 2×1 case.) Suppose the structural form is*

$$\begin{bmatrix} F_{11} & F_{12} \\ F_{21} & F_{22} \end{bmatrix} \begin{bmatrix} \Delta x_t \\ \Delta z_t \end{bmatrix} = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} \begin{bmatrix} \Delta x_{t-1} \\ \Delta z_{t-1} \end{bmatrix} + \begin{bmatrix} u_{1,t} \\ u_{2,t} \end{bmatrix}.$$

and we have an estimate of the reduced form

$$\begin{bmatrix} \Delta x_t \\ \Delta z_t \end{bmatrix} = A \begin{bmatrix} \Delta x_{t-1} \\ \Delta z_{t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{bmatrix}, \text{ with Cov}\left(\begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{bmatrix}\right) = \Omega.$$

The VMA form (as in (11.24))

$$\begin{bmatrix} \Delta x_t \\ \Delta z_t \end{bmatrix} = \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{bmatrix} + A \begin{bmatrix} \varepsilon_{1,t-1} \\ \varepsilon_{2,t-1} \end{bmatrix} + A^2 \begin{bmatrix} \varepsilon_{1,t-2} \\ \varepsilon_{2,t-2} \end{bmatrix} + ...$$

and for the level (as in (11.25))

$$\begin{bmatrix} x_t \\ z_t \end{bmatrix} = \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{bmatrix} + (A + I) \begin{bmatrix} \varepsilon_{1,t-1} \\ \varepsilon_{2,t-1} \end{bmatrix} + \left(A^2 + A + I\right) \begin{bmatrix} \varepsilon_{1,t-2} \\ \varepsilon_{2,t-2} \end{bmatrix} + ...$$

or since $\varepsilon_t = F^{-1}u_t$

$$\begin{bmatrix} x_t \\ z_t \end{bmatrix} = F^{-1} \begin{bmatrix} u_{1,t} \\ u_{2,t} \end{bmatrix} + (A + I)\, F^{-1} \begin{bmatrix} u_{1,t-1} \\ u_{2,t-1} \end{bmatrix} + \left(A^2 + A + I\right) F^{-1} \begin{bmatrix} u_{1,t-2} \\ u_{2,t-2} \end{bmatrix} + ...$$

There are 8+3 parameters in the structural form and 4+3 parameters in the VAR, so we need four restrictions. Assume that Cov$(u_t) = I$ (three restrictions) and that the long run response of $u_{1,t-s}$ on x_t is zero, that is,

$$\begin{aligned} \begin{bmatrix} \textit{unrestricted} & \textit{0} \\ \textit{unrestricted} & \textit{unrestricted} \end{bmatrix} &= \left(I + A + A^2 + ...\right) \begin{bmatrix} F_{11} & F_{12} \\ F_{21} & F_{22} \end{bmatrix}^{-1} \\ &= (I - A)^{-1} \begin{bmatrix} F_{11} & F_{12} \\ F_{21} & F_{22} \end{bmatrix}^{-1} \\ &= \begin{bmatrix} 1 - A_{11} & -A_{12} \\ -A_{21} & 1 - A_{22} \end{bmatrix}^{-1} \begin{bmatrix} F_{11} & F_{12} \\ F_{21} & F_{22} \end{bmatrix}^{-1}. \end{aligned}$$

The upper right element of the right hand side is

$$\frac{-F_{12} + F_{12}A_{22} + A_{12}F_{11}}{(1 - A_{22} - A_{11} + A_{11}A_{22} - A_{12}A_{21})\,(F_{11}F_{22} - F_{12}F_{21})}$$

which is one restriction on the elements in F. The other three are given by $F^{-1}\left(F^{-1}\right)' = \Omega$, that is,

$$\begin{bmatrix} \frac{F_{22}^2+F_{12}^2}{(F_{11}F_{22}-F_{12}F_{21})^2} & -\frac{F_{22}F_{21}+F_{12}F_{11}}{(F_{11}F_{22}-F_{12}F_{21})^2} \\ -\frac{F_{22}F_{21}+F_{12}F_{11}}{(F_{11}F_{22}-F_{12}F_{21})^2} & \frac{F_{21}^2+F_{11}^2}{(F_{11}F_{22}-F_{12}F_{21})^2} \end{bmatrix} = \begin{bmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{12} & \Omega_{22} \end{bmatrix}.$$

11.8 Cointegration, Common Trends, and Identification via Long-Run Restrictions*

These notes are a reading guide to Mellander, Vredin, and Warne (1992), which is well beyond the first year course in econometrics. See also Englund, Vredin, and Warne (1994). (I have not yet double checked this section.)

11.8.1 Common Trends Representation and Cointegration

The common trends representation of the n variables in y_t is

$$y_t = y_0 + \Upsilon\tau_t + \Phi\,(\mathrm{L}) \begin{bmatrix} \varphi_t \\ \psi_t \end{bmatrix}, \text{ with Cov}\left(\begin{bmatrix} \varphi_t \\ \psi_t \end{bmatrix}\right) = I_n \tag{11.31}$$

$$\tau_t = \tau_{t-1} + \varphi_t, \tag{11.32}$$

where Φ (L) is a stable matrix polynomial in the lag operator. We see that the $k \times 1$ vector φ_t has permanent effects on (at least some elements in) y_t, while the $r \times 1$ $(r = n - k)$ ψ_t does not.

The last component in (11.31) is stationary, but τ_t is a $k \times 1$ vector of random walks, so the $n \times k$ matrix Υ makes y_t share the non-stationary components: there are k *common trends*. If $k < n$, then we could find (at least) r linear combinations of y_t, $\alpha' y_t$ where α' is an $r \times n$ matrix of *cointegrating vectors*, which are such that the trends cancel each other ($\alpha' \Upsilon = \mathbf{0}$).

Remark 11.28 *(Lag operator.) We have the following rules:* (i) $L^k x_t = x_{t-k}$*;* (ii) *if* $\Phi(L) = a + bL^{-m} + cL^n$*, then* $\Phi(L)(x_t + y_t) = a(x_t + y_t) + b(x_{t+m} + y_{t+m}) + c(x_{t-n} + y_{t-n})$ *and* $\Phi(1) = a + b + c$*.*

Example 11.29 *(Söderlind and Vredin (1996)). Suppose we have*

$$y_t = \begin{bmatrix} \ln Y_t \text{ (output)} \\ \ln P_t \text{ (price level)} \\ \ln M_t \text{ (money stock)} \\ \ln R_t \text{ (gross interest rate)} \end{bmatrix}, \Upsilon = \begin{bmatrix} 0 & 1 \\ 1 & -1 \\ 1 & 0 \\ 0 & 0 \end{bmatrix}, \text{ and } \tau_t = \begin{bmatrix} \text{money supply trend} \\ \text{productivity trend} \end{bmatrix},$$

then we see that $\ln R_t$ *and* $\ln Y_t + \ln P_t - \ln M_t$ *(that is, log velocity) are stationary, so*

$$\alpha' = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 1 & -1 & 0 \end{bmatrix}$$

are (or rather, span the space of) cointegrating vectors. We also see that $\alpha' \Upsilon = \mathbf{0}_{2\times 2}$*.*

11.8.2 VAR Representation

The VAR representation is as in (11.1). In practice, we often estimate the parameters in A_s^*, α, the $n \times r$ matrix γ, and Ω =Cov(ε_t) in the vector "error correction form"

$$\Delta y_t = A_1^* \Delta y_t + ... + A_{p-1}^* \Delta y_{t-p+1} + \gamma \alpha' y_{t-1} + \varepsilon_t, \text{ with Cov}(\varepsilon_t) = \Omega. \quad (11.33)$$

This can easily be rewritten on the VAR form (11.1) or on the vector MA representation for Δy_t

$$\begin{aligned} \Delta y_t &= \varepsilon_t + C_1 \varepsilon_{t-1} + C_2 \varepsilon_{t-2} + ... &(11.34)\\ &= C(\text{L})\,\varepsilon_t. &(11.35) \end{aligned}$$

To find the MA of the level of y_t, we recurse on (11.35)

$$\begin{aligned} y_t &= C(\text{L})\,\varepsilon_t + y_{t-1} \\ &= C(\text{L})\,\varepsilon_t + C(\text{L})\,\varepsilon_{t-1} + y_{t-2} \\ &\vdots \\ &= C(\text{L})(\varepsilon_t + \varepsilon_{t-1} + \varepsilon_{t-2} + ... + \varepsilon_0) + y_0. \quad (11.36) \end{aligned}$$

We now try to write (11.36) in a form which resembles the common trends representation (11.31)-(11.32) as much as possible.

11.8.3 Multivariate Beveridge-Nelson decomposition

We want to split a vector of non-stationary series into some random walks and the rest (which is stationary). Rewrite (11.36) by adding and subtracting $C(1)(\varepsilon_t + \varepsilon_{t-1} + ...)$

$$y_t = C(1)(\varepsilon_t + \varepsilon_{t-1} + \varepsilon_{t-2} + ... + \varepsilon_0) + [C(\text{L}) - C(1)](\varepsilon_t + \varepsilon_{t-1} + \varepsilon_{t-2} + ... + \varepsilon_0). \quad (11.37)$$

Suppose $\varepsilon_s = 0$ for $s < 0$ and consider the second term in (11.37). It can be written

$$\begin{aligned} &\left[I + C_1\text{L} + C_2\text{L}^2 + ... - C(1)\right](\varepsilon_t + \varepsilon_{t-1} + \varepsilon_{t-2} + ... + \varepsilon_0) \\ &= \text{/*since } C(1) = I + C_1 + C_2 + ...\text{*/} \\ &[-C_1 - C_2 - C_3 - ...]\,\varepsilon_t + [-C_2 - C_3 - ...]\,\varepsilon_{t-1} + [-C_3 - ...]\,\varepsilon_{t-2}. \quad (11.38) \end{aligned}$$

Now define the random walks

$$\begin{aligned} \xi_t &= \xi_{t-1} + \varepsilon_t, \quad (11.39) \\ &= \varepsilon_t + \varepsilon_{t-1} + \varepsilon_{t-2} + ... + \varepsilon_0. \end{aligned}$$

Use (11.38) and (11.39) to rewrite (11.37) as

$$y_t = C(1)\,\xi_t + C^*(\mathrm{L})\,\varepsilon_t, \text{ where} \tag{11.40}$$

$$C_s^* = -\sum_{j=s+1}^{\infty} C_j. \tag{11.41}$$

11.8.4 Identification of the Common Trends Shocks

Rewrite (11.31)-(11.32) and (11.39)-(11.40) as

$$y_t = C(1)\sum_{s=0}^{t}\varepsilon_t + C^*(\mathrm{L})\,\varepsilon_t, \text{ with } \mathrm{Cov}(\varepsilon_t) = \Omega, \text{ and} \tag{11.42}$$

$$= \begin{bmatrix} \Upsilon & \mathbf{0}_{n\times r} \end{bmatrix}\begin{bmatrix} \sum_{s=0}^{t}\varphi_t \\ \psi_t \end{bmatrix} + \Phi(\mathrm{L})\begin{bmatrix} \varphi_t \\ \psi_t \end{bmatrix}, \text{ with Cov}\left(\begin{bmatrix} \varphi_t \\ \psi_t \end{bmatrix}\right) = I_n. \tag{11.43}$$

Since both ε_t and $\begin{bmatrix} \varphi_t' & \psi_t' \end{bmatrix}'$ are white noise, we notice that the response of y_{t+s} to either must be the same, that is,

$$\left(C(1) + C_s^*\right)\varepsilon_t = \left(\begin{bmatrix} \Upsilon & \mathbf{0}_{n\times r} \end{bmatrix} + \Phi_s\right)\begin{bmatrix} \varphi_t \\ \psi_t \end{bmatrix} \text{ for all } t \text{ and } s \geq 0. \tag{11.44}$$

This means that the VAR shocks are linear combinations of the structural shocks (as in the standard setup without cointegration)

$$\begin{bmatrix} \varphi_t \\ \psi_t \end{bmatrix} = F\varepsilon_t$$
$$= \begin{bmatrix} F_k \\ F_r \end{bmatrix}\varepsilon_t. \tag{11.45}$$

Combining (11.44) and (11.45) gives that

$$C(1) + C_s^* = \Upsilon F_k + \Phi_s \begin{bmatrix} F_k \\ F_r \end{bmatrix} \tag{11.46}$$

must hold for all $s \geq 0$. In particular, it must hold for $s \to \infty$ where both C_s^* and Φ_s vanishes

$$C(1) = \Upsilon F_k. \tag{11.47}$$

The identification therefore amounts to finding the n^2 coefficients in F, exactly as in the usual case without cointegration. Once that is done, we can calculate the impulse responses and variance decompositions with respect to the structural shocks by using $\varepsilon_t = F^{-1}\begin{bmatrix} \varphi_t' & \psi_t' \end{bmatrix}'$ in (11.42).[2] As before, assumptions about the covariance matrix of the structural shocks are not enough to achieve identification. In this case, we typically rely on the information about long-run behavior (as opposed to short-run correlations) to supply the remaining restrictions.

- *Step 1.* From (11.31) we see that $\alpha'\Upsilon = \mathbf{0}_{r\times k}$ must hold for $\alpha' y_t$ to be stationary. Given an (estimate of) α, this gives rk equations from which we can identify rk elements in Υ. (It will soon be clear why it is useful to know Υ).

- *Step 2.* From (11.44) we have $\Upsilon\varphi_t = C(1)\,\varepsilon_t$ as $s \to \infty$. The variances of both sides must be equal

  $$\mathrm{E}\Upsilon\varphi_t\varphi_t'\Upsilon' = \mathrm{E}C(1)\,\varepsilon_t\varepsilon_t' C(1)', \text{ or}$$
  $$\Upsilon\Upsilon' = C(1)\,\Omega C(1)', \tag{11.48}$$

  which gives $k(k+1)/2$ restrictions on Υ (the number of unique elements in the symmetric $\Upsilon\Upsilon'$). (However, each column of Υ is only identified up to a sign transformation: neither step 1 or 2 is affected by multiplying each element in column j of Υ by -1.)

- *Step 3.* Υ has nk elements, so we still need $nk - rk - k(k+1)/2 = k(k-1)/2$ further restrictions on Υ to identify all elements. They could be, for instance, that money supply shocks have no long run effect on output (some $\Upsilon_{ij} = 0$). We now know Υ.

[2] Equivalently, we can use (11.47) and (11.46) to calculate Υ and Φ_s (for all s) and then calculate the impulse response function from (11.43).

- *Step 4.* Combining $\operatorname{Cov}\left(\begin{bmatrix} \varphi_t \\ \psi_t \end{bmatrix}\right) = I_n$ with (11.45) gives

$$\begin{bmatrix} I_k & 0 \\ 0 & I_r \end{bmatrix} = \begin{bmatrix} F_k \\ F_r \end{bmatrix} \Omega \begin{bmatrix} F_k \\ F_r \end{bmatrix}', \tag{11.49}$$

  which gives $n\left(n+1\right)/2$ restrictions.

  - *Step 4a.* Premultiply (11.47) with Υ' and solve for F_k

$$F_k = \left(\Upsilon'\Upsilon\right)^{-1} \Upsilon' C(1). \tag{11.50}$$

    (This means that $\mathrm{E}\varphi_t\varphi_t' = F_k \Omega F_k' = \left(\Upsilon'\Upsilon\right)^{-1} \Upsilon' C(1) \Omega C\left(1\right)' \Upsilon \left(\Upsilon'\Upsilon\right)^{-1}$. From (11.48) we see that this indeed is I_k as required by (11.49).) We still need to identify F_r.

  - *Step 4b.* From (11.49), $\mathrm{E}\varphi_t\psi_t' = \mathbf{0}_{k\times r}$, we get $F_k \Omega F_r' = \mathbf{0}_{k\times r}$, which gives kr restrictions on the rn elements in F_r. Similarly, from $\mathrm{E}\psi_t\psi_t' = I_r$, we get $F_r \Omega F_r' = I_r$, which gives $r\left(r+1\right)/2$ additional restrictions on F_r. We still need $r\left(r-1\right)/2$ restrictions. Exactly how they look does not matter for the impulse response function of φ_t (as long as $\mathrm{E}\varphi_t\psi_t^{'} = \mathbf{0}$). Note that restrictions on F_r are restrictions on $\partial y_t / \partial \psi_t'$, that is, on the contemporaneous response. This is exactly as in the standard case without cointegration.

A summary of identifying assumptions used by different authors is found in Englund, Vredin, and Warne (1994).

Bibliography

Bernanke, B., 1986, “Alternative Explanations of the Money-Income Correlation,” *Carnegie-Rochester Series on Public Policy*, 25, 49–100.

Englund, P., A. Vredin, and A. Warne, 1994, “Macroeconomic Shocks in an Open Economy - A Common Trends Representation of Swedish Data 1871-1990,” in Villy Bergström, and Anders Vredin (ed.), *Measuring and Interpreting Business Cycles* . pp. 125–233, Claredon Press.

Greene, W. H., 2000, *Econometric Analysis*, Prentice-Hall, Upper Saddle River, New Jersey, 4th edn.

Hamilton, J. D., 1994, *Time Series Analysis*, Princeton University Press, Princeton.

Johnston, J., and J. DiNardo, 1997, *Econometric Methods*, McGraw-Hill, New York, 4th edn.

King, R. G., 1986, “Money and Business Cycles: Comments on Bernanke and Related Literature,” *Carnegie-Rochester Series on Public Policy*, 25, 101–116.

Mellander, E., A. Vredin, and A. Warne, 1992, “Stochastic Trends and Economic Fluctuations in a Small Open Economy,” *Journal of Applied Econometrics*, 7, 369–394.

Pindyck, R. S., and D. L. Rubinfeld, 1998, *Econometric Models and Economic Forecasts*, Irwin McGraw-Hill, Boston, Massachusetts, 4ed edn.

Sims, C. A., 1980, “Macroeconomics and Reality,” *Econometrica*, 48, 1–48.

Söderlind, P., and A. Vredin, 1996, “Applied Cointegration Analysis in the Mirror of Macroeconomic Theory,” *Journal of Applied Econometrics*, 11, 363–382.

12 Kalman filter

12.1 Conditional Expectations in a Multivariate Normal Distribution

Reference: Harvey (1989), Lütkepohl (1993), and Hamilton (1994)

Suppose $Z_{m\times 1}$ and $X_{n\times 1}$ are jointly normally distributed

$$\begin{bmatrix} Z \\ X \end{bmatrix} = N\left(\begin{bmatrix} \bar{Z} \\ \bar{X} \end{bmatrix}, \begin{bmatrix} \Sigma_{zz} & \Sigma_{zx} \\ \Sigma_{xz} & \Sigma_{xx} \end{bmatrix}\right). \tag{12.1}$$

The distribution of the random variable Z conditional on that $X = x$ is also normal with mean (expectation of the random variable Z conditional on that the random variable X has the value x)

$$\underbrace{\mathrm{E}\,(Z|x)}_{m\times 1} = \underbrace{\bar{Z}}_{m\times 1} + \underbrace{\Sigma_{zx}}_{m\times n}\underbrace{\Sigma_{xx}^{-1}}_{n\times n}\underbrace{\left(x - \bar{X}\right)}_{n\times 1}, \tag{12.2}$$

and variance (variance of Z conditional on that $X = x$)

$$\begin{aligned} \mathrm{Var}\,(Z|x) &= \mathrm{E}\left\{[Z - \mathrm{E}\,(Z|x)]^2 \middle| x\right\} \\ &= \Sigma_{zz} - \Sigma_{zx}\Sigma_{xx}^{-1}\Sigma_{xz}. \end{aligned} \tag{12.3}$$

The conditional variance is the variance of the prediction error $Z - \mathrm{E}(Z|x)$.

Both $\mathrm{E}(Z|x)$ and $\mathrm{Var}(Z|x)$ are in general stochastic variables, but for the multivariate normal distribution $\mathrm{Var}(Z|x)$ is constant. Note that $\mathrm{Var}(Z|x)$ is less than Σ_{zz} (in a matrix sense) if x contains any relevant information (so Σ_{zx} is not zero, that is, $\mathrm{E}(z|x)$ is not a constant).

It can also be useful to know that $\mathrm{Var}(Z)$ =E[Var $(Z|X)$] + Var[E $(Z|X)$] (the X is now random), which here becomes $\Sigma_{zz} - \Sigma_{zx}\Sigma_{xx}^{-1}\Sigma_{xz} + \Sigma_{zx}\Sigma_{xx}^{-1}\mathrm{Var}(X)\ \Sigma_{xx}^{-1}\Sigma_{xZ} = \Sigma_{zz}$.

12.2 Kalman Recursions

12.2.1 State space form

The measurement equation is

$$y_t = Z\alpha_t + \epsilon_t,\ \text{with Var}\,(\epsilon_t) = H, \tag{12.4}$$

where y_t and ϵ_t are $n\times 1$ vectors, and Z an $n\times m$ matrix. (12.4) expresses some observable variables y_t in terms of some (partly) unobservable state variables α_t. The transition equation for the states is

$$\alpha_t = T\alpha_{t-1} + u_t,\ \text{with Var}\,(u_t) = Q, \tag{12.5}$$

where α_t and u_t are $m \times 1$ vectors, and T an $m \times m$ matrix. This system is time invariant since all coefficients are constant. It is assumed that all errors are normally distributed, and that $\mathrm{E}(\epsilon_t u_{t-s}) = 0$ for all s.

Example 12.1 *(AR(2).) The process $x_t = \rho_1 x_{t-1} + \rho_2 x_{t-2} + e_t$ can be rewritten as*

$$\underbrace{x_t}_{y_t} = \underbrace{\begin{bmatrix} 1 & 0 \end{bmatrix}}_{Z}\underbrace{\begin{bmatrix} x_t \\ x_{t-1} \end{bmatrix}}_{\alpha_t} + \underbrace{0}_{\epsilon_t},$$

$$\underbrace{\begin{bmatrix} x_t \\ x_{t-1} \end{bmatrix}}_{\alpha_t} = \underbrace{\begin{bmatrix} \rho_1 & \rho_2 \\ 1 & 0 \end{bmatrix}}_{T}\underbrace{\begin{bmatrix} x_{t-1} \\ x_{t-2} \end{bmatrix}}_{\alpha_{t-1}} + \underbrace{\begin{bmatrix} e_t \\ 0 \end{bmatrix}}_{u_t},$$

with $H = 0$, and $Q = \begin{bmatrix} Var\,(e_t) & 0 \\ 0 & 0 \end{bmatrix}$. In this case $n = 1$, $m = 2$.

12.2.2 Prediction equations: $\mathbf{E}(\alpha_t|\mathbf{I}_{t-1})$

Suppose we have an estimate of the state in $t - 1$ based on the information set in $t - 1$, denoted by $\hat{\alpha}_{t-1}$, and that this estimate has the variance

$$P_{t-1} = \mathrm{E}\left[\left(\hat{\alpha}_{t-1} - \alpha_{t-1}\right)\left(\hat{\alpha}_{t-1} - \alpha_{t-1}\right)'\right]. \tag{12.6}$$

Now we want an estimate of α_t based $\hat{\alpha}_{t-1}$. From (12.5) the obvious estimate, denoted by $\alpha_{t|t-1}$, is

$$\hat{\alpha}_{t|t-1} = T\hat{\alpha}_{t-1}. \tag{12.7}$$

The variance of the prediction error is

$$\begin{aligned} P_{t|t-1} &= \mathrm{E}\left[\left(\alpha_t - \hat{\alpha}_{t|t-1}\right)\left(\alpha_t - \hat{\alpha}_{t|t-1}\right)'\right] \\ &= \mathrm{E}\left\{\left[T\alpha_{t-1} + u_t - T\hat{\alpha}_{t-1}\right]\left[T\alpha_{t-1} + u_t - T\hat{\alpha}_{t-1}\right]'\right\} \\ &= \mathrm{E}\left\{\left[T\left(\hat{\alpha}_{t-1} - \alpha_{t-1}\right) - u_t\right]\left[T\left(\hat{\alpha}_{t-1} - \alpha_{t-1}\right) - u_t\right]'\right\} \\ &= T\mathrm{E}\left[\left(\hat{\alpha}_{t-1} - \alpha_{t-1}\right)\left(\hat{\alpha}_{t-1} - \alpha_{t-1}\right)'\right]T' + \mathrm{E}u_t u_t' \\ &= TP_{t-1}T' + Q, \end{aligned} \tag{12.8}$$

where we have used (12.5), (12.6), and the fact that u_t is uncorrelated with $\hat{\alpha}_{t-1} - \alpha_{t-1}$.

Example 12.2 *(AR(2) continued.) By substitution we get*

$$\hat{\alpha}_{t|t-1} = \begin{bmatrix} \hat{x}_{t|t-1} \\ \hat{x}_{t-1|t-1} \end{bmatrix} = \begin{bmatrix} \rho_1 & \rho_2 \\ 1 & 0 \end{bmatrix}\begin{bmatrix} \hat{x}_{t-1|t-1} \\ \hat{x}_{t-2|t-1} \end{bmatrix}, \textit{ and}$$

$$P_{t|t-1} = \begin{bmatrix} \rho_1 & \rho_2 \\ 1 & 0 \end{bmatrix} P_{t-1} \begin{bmatrix} \rho_1 & 1 \\ \rho_2 & 0 \end{bmatrix} + \begin{bmatrix} Var(\epsilon_t) & 0 \\ 0 & 0 \end{bmatrix}$$

If we treat x_{-1} and x_0 as given, then $P_0 = \mathbf{0}_{2\times 2}$ which would give $P_{1|0} = \begin{bmatrix} Var(\epsilon_t) & 0 \\ 0 & 0 \end{bmatrix}$.

12.2.3 Updating equations: $\mathbf{E}(\alpha_t|\mathbf{I}_{t-1}) \to \mathbf{E}(\alpha_t|\mathbf{I}_t)$

The best estimate of y_t, given $\hat{a}_{t|t-1}$, follows directly from (12.4)

$$\hat{y}_{t|t-1} = Z\hat{\alpha}_{t|t-1}, \tag{12.9}$$

with prediction error

$$v_t = y_t - \hat{y}_{t|t-1} = Z\left(\alpha_t - \hat{\alpha}_{t|t-1}\right) + \epsilon_t. \tag{12.10}$$

The variance of the prediction error is

$$\begin{aligned} F_t &= \mathrm{E}\left(v_t v_t'\right) \\ &= \mathrm{E}\left\{\left[Z\left(\alpha_t - \hat{\alpha}_{t|t-1}\right) + \epsilon_t\right]\left[Z\left(\alpha_t - \hat{\alpha}_{t|t-1}\right) + \epsilon_t\right]'\right\} \\ &= Z\mathrm{E}\left[\left(\alpha_t - \hat{\alpha}_{t|t-1}\right)\left(\alpha_t - \hat{\alpha}_{t|t-1}\right)'\right]Z' + \mathrm{E}\epsilon_t\epsilon_t' \\ &= ZP_{t|t-1}Z' + H, \end{aligned} \tag{12.11}$$

where we have used the definition of $P_{t|t-1}$ in (12.8), and of H in 12.4. Similarly, the covariance of the prediction errors for y_t and for α_t is

$$\begin{aligned} \mathrm{Cov}\left(\alpha_t - \hat{\alpha}_{t|t-1}, y_t - \hat{y}_{t|t-1}\right) &= \mathrm{E}\left(\alpha_t - \hat{\alpha}_{t|t-1}\right)\left(y_t - \hat{y}_{t|t-1}\right) \\ &= \mathrm{E}\left\{\left(\alpha_t - \hat{\alpha}_{t|t-1}\right)\left[Z\left(\alpha_t - \hat{\alpha}_{t|t-1}\right) + \epsilon_t\right]'\right\} \\ &= \mathrm{E}\left[\left(\alpha_t - \hat{\alpha}_{t|t-1}\right)\left(\alpha_t - \hat{\alpha}_{t|t-1}\right)'\right]Z' \\ &= P_{t|t-1}Z'. \end{aligned} \tag{12.12}$$

Suppose that y_t is observed and that we want to update our estimate of α_t from $\hat{\alpha}_{t|t-1}$ to $\hat{\alpha}_t$, where we want to incorporate the new information conveyed by y_t.

Example 12.3 *(AR(2) continued.) We get*

$$\hat{y}_{t|t-1} = Z\hat{\alpha}_{t|t-1} = \begin{bmatrix} 1 & 0 \end{bmatrix}\begin{bmatrix} \hat{x}_{t|t-1} \\ \hat{x}_{t-1|t-1} \end{bmatrix} = \hat{x}_{t|t-1} = \rho_1\hat{x}_{t-1|t-1} + \rho_2\hat{x}_{t-2|t-1}, \textit{ and}$$

$$F_t = \begin{bmatrix} 1 & 0 \end{bmatrix}\left\{\begin{bmatrix} \rho_1 & \rho_2 \\ 1 & 0 \end{bmatrix} P_{t-1}\begin{bmatrix} \rho_1 & 1 \\ \rho_2 & 0 \end{bmatrix} + \begin{bmatrix} Var(\epsilon_t) & 0 \\ 0 & 0 \end{bmatrix}\right\}\begin{bmatrix} 1 & 0 \end{bmatrix}'.$$

If $P_0 = \mathbf{0}_{2\times 2}$ as before, then $F_1 = P_1 = \begin{bmatrix} Var(\epsilon_t) & 0 \\ 0 & 0 \end{bmatrix}$.

By applying the rules (12.2) and (12.3) we note that the expectation of α_t (like z in (12.2)) conditional on y_t (like x in (12.2)) is (note that y_t is observed so we can use it to guess α_t)

$$\underbrace{\hat{\alpha}_t}_{\mathrm{E}(z|x)} = \underbrace{\hat{\alpha}_{t|t-1}}_{\mathrm{E}z} + \underbrace{P_{t|t-1}Z'}_{\Sigma_{zx}}\left(\underbrace{ZP_{t|t-1}Z' + H}_{\Sigma_{xx} = F_t}\right)^{-1}\left(y_t - \underbrace{Z\hat{\alpha}_{t|t-1}}_{\mathrm{E}x}\right) \tag{12.13}$$

with variance

$$\underbrace{P_t}_{\text{Var}(z|x)} = \underbrace{P_{t|t-1}}_{\Sigma_{zz}} - \underbrace{P'_{t|t-1}Z'}_{\Sigma_{zx}}\underbrace{\left(ZP_{t|t-1}Z' + H\right)^{-1}}_{\Sigma_{xx}^{-1}}\underbrace{ZP_{t|t-1}}_{\Sigma_{xz}}, \tag{12.14}$$

where $\hat{\alpha}_{t|t-1}$ ("Ez") is from (12.7), $P_{t|t-1}Z'$ ("Σ_{zx}") from (12.12), $ZP_{t|t-1}Z' + H$ ("Σ_{xx}") from (12.11), and $Z\hat{\alpha}_{t|t-1}$ ("Ex") from (12.9).

(12.13) uses the new information in y_t, that is, the observed prediction error, in order to update the estimate of α_t from $\hat{\alpha}_{t|t-1}$ to $\hat{\alpha}_t$.

Proof. The last term in (12.14) follows from the expected value of the square of the last term in (12.13)

$$P_{t|t-1}Z'\left(ZP_{t|t-1}Z' + H\right)^{-1} \mathrm{E}\left(y_t - Z\alpha_{t|t-1}\right)\left(y_t - Z\alpha_{t|t-1}\right)'\left(ZP_{t|t-1}Z' + H\right)^{-1} ZP_{t|t-1}, \tag{12.15}$$

where we have exploited the symmetry of covariance matrices. Note that $y_t - Z\alpha_{t|t-1} = y_t - \hat{y}_{t|t-1}$, so the middle term in the previous expression is

$$\mathrm{E}\left(y_t - Z\alpha_{t|t-1}\right)\left(y_t - Z\alpha_{t|t-1}\right)' = ZP_{t|t-1}Z' + H. \tag{12.16}$$

Using this gives the last term in (12.14). ■

12.2.4 The Kalman Algorithm

The Kalman algorithm calculates optimal predictions of α_t in a recursive way. You can also calculate the prediction errors v_t in (12.10) as a by-prodct, which turns out to be useful in estimation.

1. Pick starting values for P_0 and α_0. Let $t = 1$.

2. Calculate (12.7), (12.8), (12.13), and (12.14) in that order. This gives values for $\hat{\alpha}_t$ and P_t. If you want v_t for estimation purposes, calculate also (12.10) and (12.11). Increase t with one step.

3. Iterate on 2 until $t = T$.

One choice of starting values that work in stationary models is to set P_0 to the unconditional covariance matrix of α_t, and α_0 to the unconditional mean. This is the matrix P

to which (12.8) converges: $P = TPT' + Q$. (The easiest way to calculate this is simply to start with $P = I$ and iterate until convergence.)

In non-stationary model we could set

$$P_0 = 1000 * I_m, \text{ and } \alpha_0 = \mathbf{0}_{m\times 1}, \tag{12.17}$$

in which case the first m observations of $\hat{\alpha}_t$ and v_t should be disregarded.

12.2.5 MLE based on the Kalman filter

For any (conditionally) Gaussian time series model for the observable y_t the log likelihood for an observation is

$$\ln L_t = -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln|F_t| - \frac{1}{2}v_t' F_t^{-1} v_t. \tag{12.18}$$

In case the starting conditions are as in (12.17), the overall log likelihood function is

$$\ln L = \begin{cases} \sum_{t=1}^{T} \ln L_t \text{ in stationary models} \\ \sum_{t=m+1}^{T} \ln L_t \text{ in non-stationary models.} \end{cases} \tag{12.19}$$

12.2.6 Inference and Diagnostics

We can, of course, use all the asymptotic MLE theory, like likelihood ratio tests etc. For diagnostoic tests, we will most often want to study the *normalized* residuals

$$\tilde{v}_{it} = v_{it}/\sqrt{\text{element } ii \text{ in } F_t}, \, i = 1, ..., n,$$

since element ii in F_t is the standard deviation of the scalar residual v_{it}. Typical tests are CUSUMQ tests for structural breaks, various tests for serial correlation, heteroskedasticity, and normality.

Bibliography

Hamilton, J. D., 1994, *Time Series Analysis*, Princeton University Press, Princeton.

Harvey, A. C., 1989, *Forecasting, Structural Time Series Models and the Kalman Filter*, Cambridge University Press.

Lütkepohl, H., 1993, *Introduction to Multiple Time Series*, Springer-Verlag, 2nd edn.

13 Outliers and Robust Estimators

13.1 Influential Observations and Standardized Residuals

Reference: Greene (2000) 6.9; Rousseeuw and Leroy (1987)

Consider the linear model

$$y_t = x_t'\beta_0 + u_t, \tag{13.1}$$

where x_t is $k \times 1$. The LS estimator

$$\hat{\beta} = \left(\sum_{t=1}^{T} x_t x_t'\right)^{-1} \sum_{t=1}^{T} x_t y_t, \tag{13.2}$$

which is the solution to

$$\min_{\beta} \sum_{t=1}^{T} \left(y_t - x_t'\beta\right)^2. \tag{13.3}$$

The fitted values and residuals are

$$\hat{y}_t = x_t'\hat{\beta}, \text{ and } \hat{u}_t = y_t - \hat{y}_t. \tag{13.4}$$

Suppose we were to reestimate β on the whole sample, except observation s. This would give us an estimate $\hat{\beta}^{(s)}$. The fitted values and residual are then

$$\hat{y}_t^{(s)} = x_t'\hat{\beta}^{(s)}, \text{and } \hat{u}_t^{(s)} = y_t - \hat{y}_t^{(s)}. \tag{13.5}$$

A common way to study the sensitivity of the results with respect to excluding observations is to plot $\hat{\beta}^{(s)} - \hat{\beta}$, and $\hat{y}_s^{(s)} - \hat{y}_s$. Note that we here plot the fitted value of y_s using the coefficients estimated by excluding observation s from the sample. Extreme values prompt a closer look at data (errors in data?) and perhaps also a more robust estimation method than LS, which is very sensitive to outliers.

Another useful way to spot outliers is to study the *standardized residuals*, $\hat{u}_s/\hat{\sigma}$ and $\hat{u}_s^{(s)}/\hat{\sigma}^{(s)}$, where $\hat{\sigma}$ and $\hat{\sigma}^{(s)}$ are standard deviations estimated from the whole sample and excluding observation s, respectively. Values below -2 or above 2 warrant attention (recall that $\Pr(x > 1.96) \approx 0.025$ in a $N(0,1)$ distribution).

Sometimes the residuals are instead standardized by taking into account the uncertainty of the estimated coefficients. Note that

$$\begin{aligned} \hat{u}_t^{(s)} &= y_t - x_t'\hat{\beta}^{(s)} \\ &= u_t + x_t'\left(\beta - \hat{\beta}^{(s)}\right), \end{aligned} \tag{13.6}$$

since $y_t = x_t'\beta + u_t$. The variance of $\hat{u}_t$ is therefore the variance of the sum on the right hand side of this expression. When we use the variance of u_t as we did above to standardize the residuals, then we disregard the variance of $\hat{\beta}^{(s)}$. In general, we have

$$\operatorname{Var}\left(\hat{u}_t^{(s)}\right) = \operatorname{Var}(u_t) + x_t'\operatorname{Var}\left(\beta - \hat{\beta}^{(s)}\right)x_t + 2\operatorname{Cov}\left[u_t, x_t'\left(\beta - \hat{\beta}^{(s)}\right)\right]. \tag{13.7}$$

When $t = s$, which is the case we care about, the covariance term drops out since $\hat{\beta}^{(s)}$ cannot be correlated with u_s since period s is not used in the estimation (this statement assumes that shocks are not autocorrelated). The first term is then estimated as the usual variance of the residuals (recall that period s is not used) and the second term is the estimated covariance matrix of the parameter vector (once again excluding period s) pre- and postmultiplied by x_s.

Example 13.1 *(Errors are iid independent of the regressors.) In this case the variance of the parameter vector is estimated as $\hat{\sigma}^2(\Sigma x_t x_t')^{-1}$ (excluding period s), so we have*

$$Var\left(\hat{u}_t^{(s)}\right) = \hat{\sigma}^2\left(1 + x_s'(\Sigma x_t x_t')^{-1}x_s\right).$$

13.2 Recursive Residuals*

Reference: Greene (2000) 7.8

Recursive residuals are a version of the technique discussed in Section 13.1. They are used when data is a time series. Suppose we have a sample $t = 1, ..., T$,.and that $t = 1, ..., s$ are used to estimate a first estimate, $\hat{\beta}^{[s]}$ (not to be confused with $\hat{\beta}^{(s)}$ used in Section 13.1). We then make a one-period ahead forecast and record the fitted value and the forecast error

$$\hat{y}_{s+1}^{[s]} = x_{s+1}'\hat{\beta}^{[s]}, \text{ and } \hat{u}_{s+1}^{[s]} = y_{s+1} - \hat{y}_{s+1}^{[s]}. \tag{13.8}$$

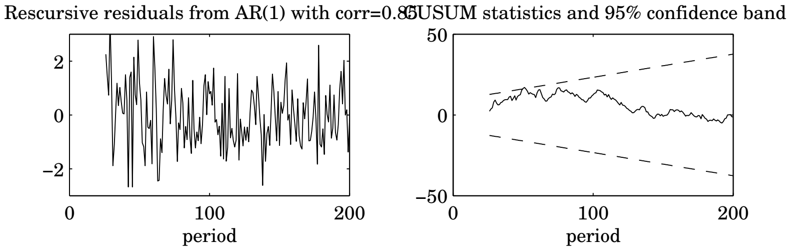


Figure 13.1: This figure shows recursive residuals and CUSUM statistics, when data are simulated from $y_t = 0.85y_{t-1} + u_t$, with $\operatorname{Var}(u_t) = 1$.

This is repeated for the rest of the sample by extending the sample used in the estimation by one period, making a one-period ahead forecast, and then repeating until we reach the end of the sample.

A first diagnosis can be made by examining the standardized residuals, $\hat{u}_{s+1}^{[s]}/\hat{\sigma}^{[s]}$, where $\hat{\sigma}^{[s]}$ can be estimated as in (13.7) with a zero covariance term, since u_{s+1} is not correlated with data for earlier periods (used in calculating $\hat{\beta}^{[s]}$), provided errors are not autocorrelated. As before, standardized residuals outside ± 2 indicates problems: outliers or structural breaks (if the residuals are persistently outside ± 2).

The *CUSUM test* uses these standardized residuals to form a sequence of test statistics. A (persistent) jump in the statistics is a good indicator of a structural break. Suppose we use r observations to form the first estimate of β, so we calculate $\hat{\beta}^{[s]}$ and $\hat{u}_{s+1}^{[s]}/\hat{\sigma}^{[s]}$ for $s = r, ..., T$. Define the cumulative sums of standardized residuals

$$W_t = \sum_{s=r}^{t} \hat{u}_{s+1}^{[s]}/\hat{\sigma}^{[s]}, t = r, ..., T. \tag{13.9}$$

Under the null hypothesis that no structural breaks occurs, that is, that the true β is the same for the whole sample, W_t has a zero mean and a variance equal to the number of elements in the sum, $t - r + 1$. This follows from the fact that the standardized residuals all have zero mean and unit variance and are uncorrelated with each other. Typically, W_t is plotted along with a 95% confidence interval, which can be shown to be $\pm\left(a\sqrt{T-r} + 2a\,(t-r)/\sqrt{T-r}\right)$ with $a = 0.948$. The hypothesis of no structural break is rejected if the W_t is outside this band for at least one observation. (The derivation of this confidence band is somewhat tricky, but it incorporates the fact that W_t and W_{t+1}

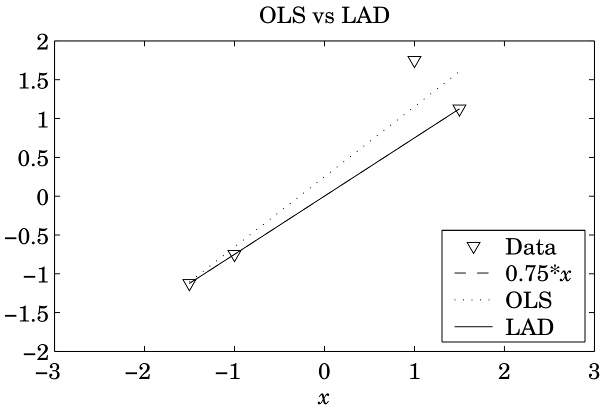


Figure 13.2: This figure shows an example of how LS and LAD can differ. In this case $y_t = 0.75x_t + u_t$, but only one of the errors has a non-zero value.

are very correlated.)

13.3 Robust Estimation

Reference: Greene (2000) 9.8.1; Rousseeuw and Leroy (1987); Donald and Maddala (1993); and Judge, Griffiths, Lütkepohl, and Lee (1985) 20.4.

The idea of robust estimation is to give less weight to extreme observations than in Least Squares. When the errors are normally distributed, then there should be very few extreme observations, so LS makes a lot of sense (and is indeed the MLE). When the errors have distributions with fatter tails (like the Laplace or two-tailed exponential distribution, $f(u) = \exp(-|u|/\sigma)/2\sigma$), then LS is no longer optimal and can be fairly sensitive to outliers. The ideal way to proceed would be to apply MLE, but the true distribution is often unknown. Instead, one of the "robust estimators" discussed below is often used.

Let $\hat{u}_t = y_t - x_t'\hat{\beta}$. Then, the least absolute deviations (LAD), least median squares (LMS), and least trimmed squares (LTS) estimators solve

$$\hat{\beta}_{LAD} = \arg\min_{\beta} \sum_{t=1}^{T} |\hat{u}_t| \tag{13.10}$$

$$\hat{\beta}_{LMS} = \arg\min_{\beta} \left[\text{median}\left(\hat{u}_t^2\right)\right] \tag{13.11}$$

$$\hat{\beta}_{LTS} = \arg\min_{\beta} \sum_{i=1}^{h} \hat{u}_i^2,\ \hat{u}_1^2 \leq \hat{u}_2^2 \leq \ldots \text{ and } h \leq T. \tag{13.12}$$

Note that the LTS estimator in (13.12) minimizes of the sum of the h smallest squared residuals.

These estimators involve non-linearities, so they are more computationally intensive than LS. In some cases, however, a simple iteration may work.

Example 13.2 *(Algorithm for LAD.) The LAD estimator can be written*

$$\hat{\beta}_{LAD} = \arg\min_{\beta} \sum_{t=1}^{T} w_t \hat{u}_t^2,\ w_t = 1/\left|\hat{u}_t\right|,$$

so it is a weighted least squares where both y_t and x_t are multiplied by $1/|\hat{u}_t|$. It can be shown that iterating on LS with the weights given by $1/|\hat{u}_t|$, where the residuals are from the previous iteration, converges very quickly to the LAD estimator.

It can be noted that LAD is actually the MLE for the Laplace distribution discussed above.

13.4 Multicollinearity*

Reference: Greene (2000) 6.7

When the variables in the x_t vector are very highly correlated (they are "multicollinear") then data cannot tell, with the desired precision, if the movements in y_t was due to movements in x_{it} or x_{jt}. This means that the point estimates might fluctuate wildly over subsamples and it is often the case that individual coefficients are insignificant even though the R^2 is high and the joint significance of the coefficients is also high. The estimators are still consistent and asymptotically normally distributed, just very imprecise.

A common indicator for multicollinearity is to standardize each element in x_t by subtracting the sample mean and then dividing by its standard deviation

$$\tilde{x}_{it} = (x_{it} - \bar{x}_{it}) / \text{std}(x_{it}). \tag{13.13}$$

(Another common procedure is to use $\tilde{x}_{it} = x_{it}/(\Sigma_{t=1}^{T} x_{it}^2/T)^{1/2}$.)

Then calculate the eigenvalues, λ_j, of the second moment matrix of $\tilde{x}_t$

$$A = \frac{1}{T}\sum_{t=1}^{T} \tilde{x}_t \tilde{x}_t'. \tag{13.14}$$

The *condition number* of a matrix is the ratio of the largest (in magnitude) of the eigenvalues to the smallest

$$c = |\lambda|_{\max} / |\lambda|_{\min}. \tag{13.15}$$

(Some authors take $c^{1/2}$ to be the condition number; others still define it in terms of the "singular values" of a matrix.) If the regressors are uncorrelated, then the condition value of A is one. This follows from the fact that A is a (sample) covariance matrix. If it is diagonal, then the eigenvalues are equal to diagonal elements, which are all unity since the standardization in (13.13) makes all variables have unit variances. Values of c above several hundreds typically indicate serious problems.

Bibliography

Donald, S. G., and G. S. Maddala, 1993, "Identifying Outliers and Influential Observations in Econometric Models," in G. S. Maddala, C. R. Rao, and H. D. Vinod (ed.), *Handbook of Statistics, Vol 11* . pp. 663–701, Elsevier Science Publishers B.V.

Greene, W. H., 2000, *Econometric Analysis*, Prentice-Hall, Upper Saddle River, New Jersey, 4th edn.

Judge, G. G., W. E. Griffiths, H. Lütkepohl, and T.-C. Lee, 1985, *The Theory and Practice of Econometrics*, John Wiley and Sons, New York, 2nd edn.

Rousseeuw, P. J., and A. M. Leroy, 1987, *Robust Regression and Outlier Detection*, John Wiley and Sons, New York.

14 Generalized Least Squares

Reference: Greene (2000) 11.3-4

Additional references: Hayashi (2000) 1.6; Johnston and DiNardo (1997) 5.4; Verbeek (2004) 4

14.1 Introduction

Instead of using LS in the presence of autocorrelation/heteroskedasticity (and, of course, adjusting the variance-covariance matrix), we may apply the generalized least squares method. It can often improve efficiency.

The linear model $y_t = x_t'\beta_0 + u_t$ written on matrix form (GLS is one of the cases in econometrics where matrix notation really pays off) is

$$y = X\beta_0 + u, \text{ where} \tag{14.1}$$

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{bmatrix}, X = \begin{bmatrix} x_1' \\ x_2' \\ \vdots \\ x_T' \end{bmatrix}, \text{and } u = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_T \end{bmatrix}.$$

Suppose that the covariance matrix of the residuals (across time) is

$$\begin{aligned} \mathrm{E}uu' &= \begin{bmatrix} \mathrm{E}u_1u_1 & \mathrm{E}u_1u_2 & \cdots & \mathrm{E}u_1u_T \\ \mathrm{E}u_2u_1 & \mathrm{E}u_2u_2 & & \mathrm{E}u_2u_T \\ \vdots & & \ddots & \vdots \\ \mathrm{E}u_Tu_1 & \mathrm{E}u_Tu_2 & & \mathrm{E}u_Tu_T \end{bmatrix} \\ &= \Omega_{T\times T}. \end{aligned} \tag{14.2}$$

This allows for both heteroskedasticity (different elements along the main diagonal) and autocorrelation (non-zero off-diagonal elements). LS is still consistent even if Ω is not proportional to an identity matrix, but it is not efficient. Generalized least squares (GLS) is. The trick of GLS is to transform the variables and the do LS.

14.2 GLS as Maximum Likelihood

Remark 14.1 *If the $n \times 1$ vector x has a multivariate normal distribution with mean vector μ and covariance matrix Ω, then the joint probability density function is* $(2\pi)^{-n/2} |\Omega|^{-1/2} \exp[-(x-\mu)'\Omega^{-1}(x-\mu)/2]$.

If the $T\times 1$ vector u is $N(\mathbf{0}, \Omega)$, then the joint pdf of u is $(2\pi)^{-n/2} |\Omega|^{-1/2} \exp[-u'\Omega^{-1}u/2]$. Change variable from u to $y - X\beta$ (the Jacobian of this transformation equals one), and take logs to get the (scalar) log likelihood function

$$\ln L = -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln|\Omega| - \frac{1}{2}(y - X\beta)'\Omega^{-1}(y - X\beta). \tag{14.3}$$

To simplify things, suppose we know Ω. It is then clear that we maximize the likelihood function by minimizing the last term, which is a weighted sum of squared errors.

In the *classical LS* case, $\Omega = \sigma^2 I$, so the last term in (14.3) is proportional to the unweighted sum of squared errors. The LS is therefore the MLE when the errors are iid normally distributed.

When errors are *heteroskedastic*, but not autocorrelated, then Ω has the form

$$\Omega = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \sigma_T^2 \end{bmatrix}. \tag{14.4}$$

In this case, we can decompose Ω^{-1} as

$$\Omega^{-1} = P'P, \text{ where } P = \begin{bmatrix} 1/\sigma_1 & 0 & \cdots & 0 \\ 0 & 1/\sigma_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & 1/\sigma_T \end{bmatrix}. \tag{14.5}$$

The last term in (14.3) can then be written

$$\begin{aligned} -\frac{1}{2}(y - X\beta)'\Omega^{-1}(y - X\beta) &= -\frac{1}{2}(y - X\beta)'P'P(y - X\beta) \\ &= -\frac{1}{2}(Py - PX\beta)'(Py - PX\beta). \end{aligned} \tag{14.6}$$

This very practical result says that if we define $y_t^* = y_t/\sigma_t$ and $x_t^* = x_t/\sigma_t$, then we get ML estimates of β running an LS regression of y_t^* on x_t^*. (One of the elements in x_t could be a constant—also this one should be transformed). This is the generalized least squares (GLS).

Remark 14.2 *Let A be an $n\times n$ symmetric positive definite matrix. It can be decomposed as $A = PP'$. There are many such P matrices, but only one which is lower triangular P (see next remark).*

Remark 14.3 *Let A be an $n\times n$ symmetric positive definite matrix. The Cholesky decomposition gives the unique lower triangular P_1 such that $A = P_1P_1'$ or an upper triangular matrix P_2 such that $A = P_2'P_2$ (clearly $P_2 = P_1'$). Note that P_1 and P_2 must be invertible (since A is).*

When errors are *autocorrelated* (with or without heteroskedasticity), then it is typically harder to find a straightforward analytical decomposition of Ω^{-1}. We therefore move directly to the general case. Since the covariance matrix is symmetric and positive definite, Ω^{-1} is too. We therefore decompose it as

$$\Omega^{-1} = P'P. \tag{14.7}$$

The Cholesky decomposition is often a convenient tool, but other decompositions can also be used. We can then apply (14.6) also in this case—the only difference is that P is typically more complicated than in the case without autocorrelation. In particular, the transformed variables Py and PX cannot be done line by line (y_t^* is a function of y_t, y_{t-1}, and perhaps more).

Example 14.4 *(AR(1) errors, see Davidson and MacKinnon (1993) 10.6.) Let $u_t = au_{t-1} + \varepsilon_t$ where ε_t is iid. We have $Var(u_t) = \sigma^2/\left(1 - a^2\right)$, and $Corr(u_t, u_{t-s}) = a^s$.*

For $T = 4$, the covariance matrix of the errors is

$$\begin{aligned}\Omega &= Cov\left(\begin{bmatrix} u_1 & u_2 & u_3 & u_4 \end{bmatrix}'\right)\\ &= \frac{\sigma^2}{1-a^2}\begin{bmatrix} 1 & a & a^2 & a^3 \\ a & 1 & a & a^2 \\ a^2 & a & 1 & a \\ a^3 & a^2 & a & 1 \end{bmatrix}.\end{aligned}$$

The inverse is

$$\Omega^{-1} = \frac{1}{\sigma^2}\begin{bmatrix} 1 & -a & 0 & 0 \\ -a & 1+a^2 & -a & 0 \\ 0 & -a & 1+a^2 & -a \\ 0 & 0 & -a & 1 \end{bmatrix},$$

and note that we can decompose it as

$$\Omega^{-1} = \underbrace{\frac{1}{\sigma}\begin{bmatrix} \sqrt{1-a^2} & 0 & 0 & 0 \\ -a & 1 & 0 & 0 \\ 0 & -a & 1 & 0 \\ 0 & 0 & -a & 1 \end{bmatrix}'}_{P'}\underbrace{\frac{1}{\sigma}\begin{bmatrix} \sqrt{1-a^2} & 0 & 0 & 0 \\ -a & 1 & 0 & 0 \\ 0 & -a & 1 & 0 \\ 0 & 0 & -a & 1 \end{bmatrix}}_{P}.$$

This is not a Cholesky decomposition, but certainly a valid decomposition (in case of doubt, do the multiplication). Premultiply the system

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} x_1' \\ x_2' \\ x_3' \\ x_4' \end{bmatrix}\beta_0 + \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{bmatrix}$$

by P to get

$$\frac{1}{\sigma}\begin{bmatrix} \sqrt{(1-a^2)}y_1 \\ y_2 - ay_1 \\ y_3 - ay_2 \\ y_4 - ay_3 \end{bmatrix} = \frac{1}{\sigma}\begin{bmatrix} \sqrt{(1-a^2)}x_1' \\ x_2' - ax_1' \\ x_3' - ax_2' \\ x_4' - ax_3' \end{bmatrix}\beta_0 + \frac{1}{\sigma}\begin{bmatrix} \sqrt{(1-a^2)}u_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \end{bmatrix}.$$

Note that all the residuals are uncorrelated in this formulation. Apart from the first observation, they are also identically distributed. The importance of the first observation becomes smaller as the sample size increases—in the limit, GLS is efficient.

14.3 GLS as a Transformed LS

When the errors are not normally distributed, then the MLE approach in the previous section is not valid. But we can still note that GLS has the same properties as LS has with iid non-normally distributed errors. In particular, the Gauss-Markov theorem applies, so the GLS is most efficient within the class of linear (in y_t) and unbiased estimators (assuming, of course, that GLS and LS really are unbiased, which typically requires that u_t is uncorrelated with x_{t-s} for all s). This follows from that the transformed system

$$\begin{aligned} Py &= PX\beta_0 + Pu \\ y^* &= X^*\beta_0 + u^*, \end{aligned} \tag{14.8}$$

have iid errors, u^*. So see this, note that

$$\begin{aligned} \mathrm{E}u^*u^{*\prime} &= \mathrm{E}Puu'P' \\ &= P\mathrm{E}uu'P'. \end{aligned} \tag{14.9}$$

Recall that $\mathrm{E}uu' = \Omega$, $P'P = \Omega^{-1}$ and that P' is invertible. Multiply both sides by P'

$$\begin{aligned} P'\mathrm{E}u^*u^{*\prime} &= P'P\mathrm{E}uu'P' \\ &= \Omega^{-1}\Omega P' \\ &= P', \text{ so } \mathrm{E}u^*u^{*\prime} = I. \end{aligned} \tag{14.10}$$

14.4 Feasible GLS

In practice, we usually do not know Ω. *Feasible GLS* (FGSL) is typically implemented by first estimating the model (14.1) with LS, then calculating a consistent estimate of Ω, and finally using GLS as if Ω was known with certainty. Very little is known about the finite sample properties of FGLS, but (the large sample properties) consistency, asymptotic normality, and asymptotic efficiency (assuming normally distributed errors) can often be established. Evidence from simulations suggests that the FGLS estimator can be a lot worse than LS if the estimate of Ω is bad.

To use maximum likelihood when Ω is unknown requires that we make assumptions about the structure of Ω (in terms of a small number of parameters), and more generally about the distribution of the residuals. We must typically use numerical methods to maximize the likelihood function.

Example 14.5 *(MLE and AR(1) errors.) If u_t in Example 14.4 are normally distributed, then we can use the Ω^{-1} in (14.3) to express the likelihood function in terms of the unknown parameters: β, σ, and a. Maximizing this likelihood function requires a numerical optimization routine.*

Bibliography

Davidson, R., and J. G. MacKinnon, 1993, *Estimation and Inference in Econometrics*, Oxford University Press, Oxford.

Greene, W. H., 2000, *Econometric Analysis*, Prentice-Hall, Upper Saddle River, New Jersey, 4th edn.

Hayashi, F., 2000, *Econometrics*, Princeton University Press.

Johnston, J., and J. DiNardo, 1997, *Econometric Methods*, McGraw-Hill, New York, 4th edn.

Verbeek, M., 2004, *A Guide to Modern Econometrics*, Wiley, Chichester, 2nd edn.

14.5 Nonparametric Regression

Reference: Campbell, Lo, and MacKinlay (1997) 12.3; Härdle (1990); Pagan and Ullah (1999); Mittelhammer, Judge, and Miller (2000) 21

14.5.1 Introduction

Nonparametric regressions are used when we are unwilling to impose a parametric form on the regression equation—and we have a lot of data.

Let the scalars y_t and x_t be related as

$$y_t = b(x_t) + \varepsilon_t, \tag{14.1}$$

where ε_t is uncorrelated over time and where $\mathrm{E}\,\varepsilon_t = 0$ and $\mathrm{E}(\varepsilon_t|x_t) = 0$. The function $b()$ is unknown and possibly non-linear.

Suppose the sample had 3 observations (say, $t = 3$, 27, and 99) with exactly the same value of x_t, say 1.9. A natural way of estimating $b(x)$ at $x = 1.9$ would then be to average over these 3 observations as we can expect average of the error terms to be close to zero (iid and zero mean).

Unfortunately, we seldom have repeated observations of this type. Instead, we may try to approximate the value of $b(x)$ (x is a single value, 1.9, say) by averaging over observations where x_t is close to x. The general form of this type of estimator is

$$\hat{b}(x) = \frac{\sum_{t=1}^{T} w(x_t - x) y_t}{\sum_{t=1}^{T} w(x_t - x)}, \tag{14.2}$$

where $w(x_t - x)/\Sigma_{t=1}^{T} w(x_t - x)$ is the weight on observation t. Note that the denominator makes the weights sum to unity. The basic assumption behind (14.2) is that the $b(x)$ function is smooth so local (around x) averaging makes sense.

As an example of a $w(.)$ function, it could give equal weight to the k values of x_t which are closest to x and zero weight to all other observations (this is the "k-nearest neighbor" estimator, see Härdle (1990) 3.2). As another example, the weight function could be defined so that it trades off the expected squared errors, $\mathrm{E}[y_t - \hat{b}(x)]^2$, and the expected squared acceleration, $\mathrm{E}[d^2\hat{b}(x)/dx^2]^2$. This defines a cubic spline (often used in macroeconomics when $x_t = t$, and is then called the Hodrick-Prescott filter).

Remark 14.6 *(Easy way to calculate the "nearest neighbor" estimator, univariate case) Create a matrix Z where row t is (y_t, x_t). Sort the rows of Z according to the second column (x). Calculate an equally weighted centered moving average of the first column (y).*

14.5.2 Kernel Regression

A *Kernel regression* uses a pdf as the weight function, $w(x_t - x) = K[(x_t - x)/h]$, where the choice of h (also called bandwidth) allows us to easily vary the relative weights of different observations. The pdf of $N(0,1)$ is often used for $K()$. This weighting function is positive, so all observations get a positive weight, but the weights are highest for observations close to x and then taper off in a bell-shaped way. A low value of h means that the weights taper off fast. See Figure 14.1 for an example.

With the $N(0,1)$ kernel, we get the following estimator of $b(x)$ at a point x

$$\hat{b}(x) = \frac{\sum_{t=1}^{T} K\left(\frac{x_t - x}{h}\right) y_t}{\sum_{t=1}^{T} K\left(\frac{x_t - x}{h}\right)}, \text{ where } K(u) = \frac{\exp\left(-u^2/2\right)}{\sqrt{2\pi}}. \tag{14.3}$$

Remark 14.7 *(Kernel as a pdf of $N(x, h^2)$) If $K(z)$ is the pdf of an $N(0,1)$ variable, then $K[(x_t - x)/h]/h$ is the same as using an $N(x, h^2)$ pdf of x_t. Clearly, the $1/h$ term would cancel in (14.3).*

In practice we have to estimate $\hat{b}(x)$ at a finite number of points x. This could, for instance, be 100 evenly spread points in the interval between the minimum and the maximum values observed in the sample. See Figure 14.2 for an illustration. Special corrections might be needed if there are a lot of observations stacked close to the boundary of the support of x (see Härdle (1990) 4.4).

Example 14.8 *(Kernel regression) Suppose the sample has three data points $[x_1, x_2, x_3] = [1.5, 2, 2.5]$ and $[y_1, y_2, y_3] = [5, 4, 3.5]$. Consider the estimation of $b(x)$ at $x = 1.9$.*

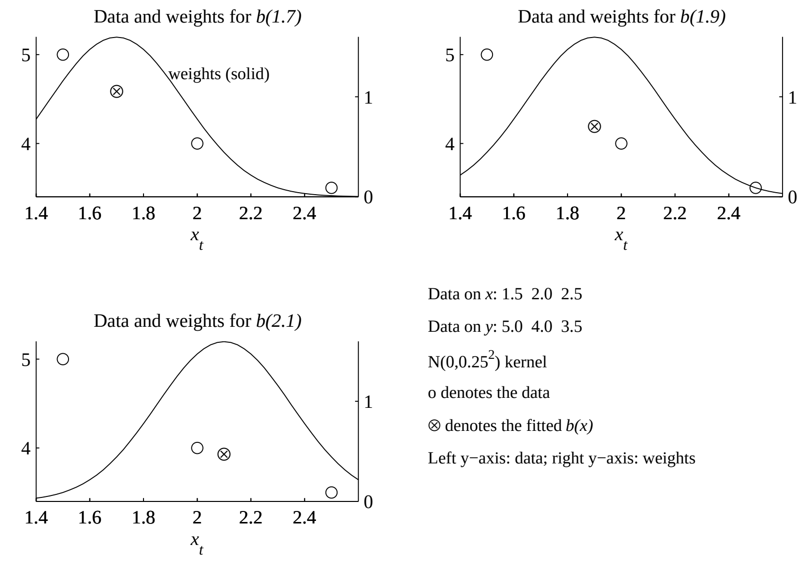


Figure 14.1: Example of kernel regression with three data points

With $h = 1$, the numerator in (14.3) is

$$\begin{aligned}\sum_{t=1}^{T} K(x_t - x) y_t &= \left(e^{-(1.5-1.9)^2/2} \times 5 + e^{-(2-1.9)^2/2} \times 4 + e^{-(2.5-1.9)^2/2} \times 3.5\right)/\sqrt{2\pi} \\ &\approx (0.92 \times 5 + 1.0 \times 4 + 0.84 \times 3.5)/\sqrt{2\pi} \\ &= 11.52/\sqrt{2\pi}.\end{aligned}$$

The denominator is

$$\begin{aligned}\sum_{t=1}^{T} K(x_t - x) &= \left(e^{-(1.5-1.9)^2/2} + e^{-(2-1.9)^2/2} + e^{-(2.5-1.9)^2/2}\right)/\sqrt{2\pi} \\ &\approx 2.75/\sqrt{2\pi}.\end{aligned}$$

The estimate at $x = 1.9$ is therefore

$$\hat{b}(1.9) \approx 11.52/2.75 \approx 4.19.$$

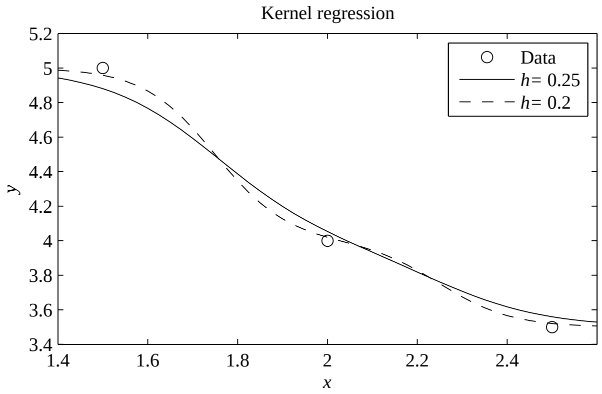


Figure 14.2: Example of kernel regression with three data points

Kernel regressions are typically consistent, provided longer samples are accompanied by smaller values of h, so the weighting function becomes more and more local as the sample size increases. It can be shown (see Härdle (1990) 3.1 and Pagan and Ullah (1999) 3.3–4) that under the assumption that x_t is iid, the mean squared error, variance and bias of the estimator at the value x are approximately (for general kernel functions)

$$\begin{aligned}
\text{MSE}(x) &= \text{Var}\left[\hat{b}(x)\right] + \left\{\text{Bias}[\hat{b}(x)]\right\}^2, \text{ with} \\
\text{Var}\left[\hat{b}(x)\right] &= \frac{1}{Th}\frac{\sigma^2(x)}{f(x)} \times \textstyle\int_{-\infty}^{\infty} K(u)^2 du \\
\text{Bias}[\hat{b}(x)] &= h^2 \times \left[\frac{1}{2}\frac{d^2 b(x)}{dx^2} + \frac{df(x)}{dx}\frac{1}{f(x)}\frac{db(x)}{dx}\right] \times \textstyle\int_{-\infty}^{\infty} K(u)u^2 du. \qquad (14.4)
\end{aligned}$$

In these expressions, $\sigma^2(x)$ is the variance of the residuals in (14.1) and $f(x)$ the marginal density of x. The remaining terms are functions of either the true regression function or the kernel.

With a $N(0,1)$ kernel these expressions can be simplified since

$$\textstyle\int_{-\infty}^{\infty} K(u)^2 du = \frac{1}{2\sqrt{\pi}} \text{ and } \int_{-\infty}^{\infty} K(u)u^2 du = 1, \text{ if } N(0,1) \text{ kernel.} \qquad (14.5)$$

In this case (14.4) becomes

$$\begin{aligned}
\text{Var}\left[\hat{b}(x)\right] &= \frac{1}{Th}\frac{\sigma^2(x)}{f(x)} \times \frac{1}{2\sqrt{\pi}} \\
\text{Bias}[\hat{b}(x)] &= h^2 \times \left[\frac{1}{2}\frac{d^2 b(x)}{dx^2} + \frac{df(x)}{dx}\frac{1}{f(x)}\frac{db(x)}{dx}\right]. \qquad (14.6)
\end{aligned}$$

A smaller h increases the variance (we effectively use fewer data points to estimate $b(x)$) but decreases the bias of the estimator (it becomes more local to x). If h decreases less than proportionally with the sample size (so hT in the denominator of the first term increases with T), then the variance goes to zero and the estimator is consistent (since the bias in the second term decreases as h does).

The variance is a function of the variance of the residuals and the "peakedness" of the kernel, but not of the $b(x)$ function. The more concentrated the kernel is ($\int K(u)^2 du$ large) around x (for a given h), the less information is used in forming the average around x, and the uncertainty is therefore larger—which is similar to using a small h. A low density of the regressors ($f(x)$ low) means that we have little data at x which drives up the uncertainty of the estimator.

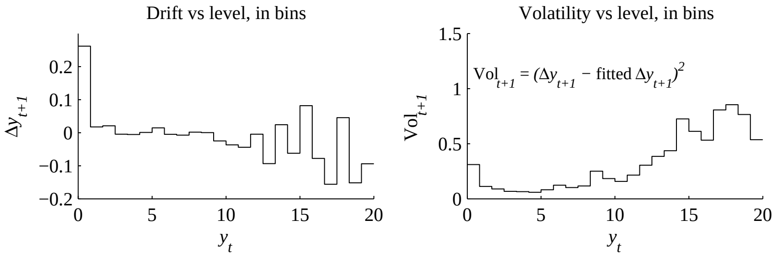
The bias increases (in magnitude) with the curvature of the $b(x)$ function (that is, $(d^2 b(x)/dx^2)^2$). This makes sense, since rapid changes of the slope of $b(x)$ make it hard to get $b(x)$ right by averaging at nearby x values. It also increases with the variance of the kernel since a large kernel variance is similar to a large h.

It is clear that the choice of h has a major importance on the estimation results. A lower value of h means a more "local" averaging, which has the potential of picking up sharp changes in the regression function—at the cost of being more affected by randomness. See Figures 14.3–14.4 for an example.

A good (but computationally intensive) approach to choose h is by the leave-one-out *cross-validation* technique. This approach would, for instance, choose h to minimize the expected (or average) prediction error

$$\text{EPE}(h) = \textstyle\sum_{t=1}^{T} \left[y_t - \hat{b}_{-t}(x_t, h)\right]^2 / T, \qquad (14.7)$$

where $\hat{b}_{-t}(x_t, h)$ is the fitted value at x_t when we use a regression function estimated on a sample that excludes observation t, and a bandwidth h. This means that each prediction is out-of-sample. To calculate (14.7) we clearly need to make T estimations (for each



Daily federal funds rates 1954−2006

Figure 14.3: Federal funds rate

x_t)—and then repeat this for different values of h to find the minimum.

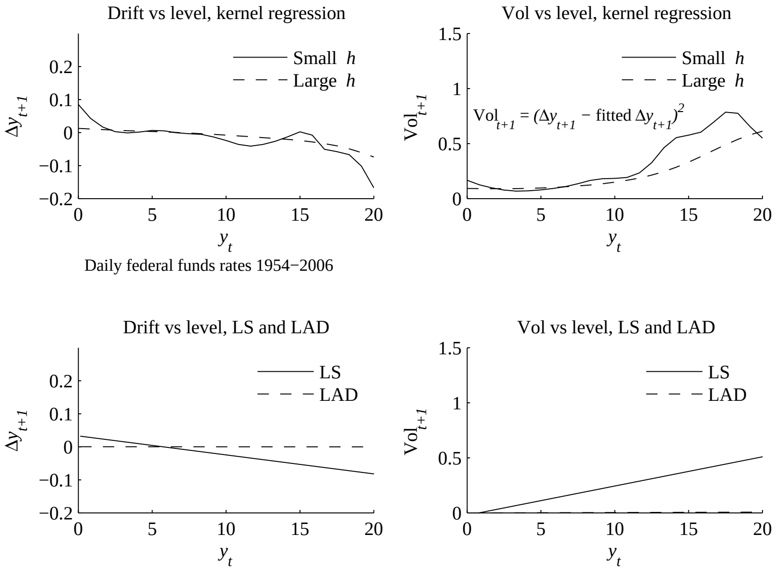
Remark 14.9 *(Speed and fast Fourier transforms) The calculation of the kernel estimator can often be speeded up by the use of a fast Fourier transform.*

If the observations are independent, then it can be shown (see Härdle (1990) 4.2, Pagan and Ullah (1999) 3.3–6, and also (14.6)) that, with a Gaussian kernel, the estimator at point x is asymptotically normally distributed

$$\sqrt{Th}\left[\hat{b}(x) - \mathrm{E}\,\hat{b}(x)\right] \to^d N\left[0, \frac{1}{2\sqrt{\pi}}\frac{\sigma^2(x)}{f(x)}\right], \tag{14.8}$$

where $\sigma^2(x)$ is the variance of the residuals in (14.1) and $f(x)$ the marginal density of x. (A similar expression holds for other choices of the kernel.) This expression assumes that the asymptotic bias is zero, which is guaranteed if h is decreased (as T increases) slightly faster than $T^{-1/5}$. In practice, to implement this expression $\sigma^2(x)$ is estimated as a kernel regression (14.3) but with $[y_t - \hat{b}(x_t)]^2$ substituted for y_t , $f(x)$ is estimated by a kernel density estimate as in (14.24) and it is assumed that the asymptotic bias is zero ($\mathrm{E}\,\hat{b}(x) = b(x)$). Notice that the estimation of $\sigma^2(x)$ is quite computationally intensive since it requires estimating $\hat{b}(x)$ at every point $x = x_t$ in the sample.

See Figure 14.5 for an example—and Figure 14.7 for the reason why the confidence band varies across x values.



Daily federal funds rates 1954−2006

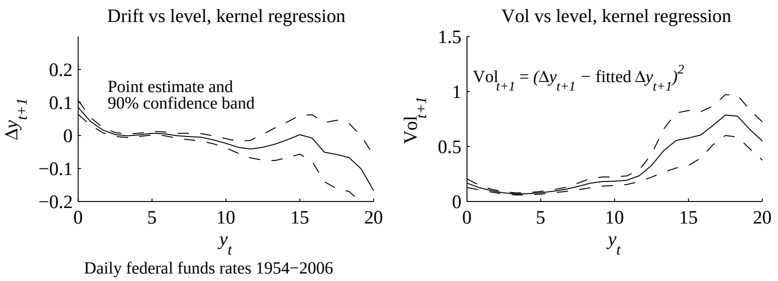
Figure 14.4: Federal funds rate

14.5.3 Multivariate Kernel Regression

Suppose that y_t depends on two variables (x_t and z_t)

$$y_t = b(x_t, z_t) + \varepsilon_t, \tag{14.9}$$

where ε_t is uncorrelated over time and where $\mathrm{E}\,\varepsilon_t = 0$ and $\mathrm{E}(\varepsilon_t|x_t, z_t) = 0$. This makes the estimation problem much harder since there are typically few observations in every bivariate bin (rectangle) of x and z. For instance, with as little as a 20 intervals of each of x and z, we get 400 bins, so we need a large sample to have a reasonable number of observations in every bin.



Daily federal funds rates 1954−2006

Figure 14.5: Federal funds rate

In any case, the most common way to implement the kernel regressor is to let

$$\hat{b}(x,z) = \frac{\sum_{t=1}^{T} K_x\left(\frac{x_t - x}{h_x}\right) K_z\left(\frac{z_t - z}{h_z}\right) y_t}{\sum_{t=1}^{T} K_x\left(\frac{x_t - x}{h_x}\right) K_z\left(\frac{z_t - z}{h_z}\right)}, \tag{14.10}$$

where $K_x(u)$ and $K_z(v)$ are two kernels like in (14.3) and where we may allow h_x and h_y to be different (and depend on the variance of x_t and y_t). In this case, the weight of the observation (x_t, z_t) is proportional to $K_x\left(\frac{x_t - x}{h_x}\right) K_z\left(\frac{z_t - z}{h_z}\right)$, which is high if both x_t and y_t are close to x and y respectively.

14.5.4 Semiparametric Estimation

A possible way out of the curse of dimensionality of the multivariate kernel regression is to specify a partially linear model

$$y_t = z_t'\beta + b(x_t) + \varepsilon_t, \tag{14.11}$$

where ε_t is uncorrelated over time and where $\mathrm{E}\,\varepsilon_t = 0$ and $\mathrm{E}(\varepsilon_t|x_t, z_t) = 0$. This model is linear in z_t, but possibly non-linear in x_t since the function $b(x_t)$ is unknown.

To construct an estimator, start by taking expectations of (14.11) conditional on x_t

$$\mathrm{E}(y_t|x_t) = \mathrm{E}(z_t|x_t)'\beta + b(x_t). \tag{14.12}$$

Subtract from (14.11) to get

$$y_t - \mathrm{E}(y_t|x_t) = [z_t - \mathrm{E}(z_t|x_t)]'\beta + \varepsilon_t. \tag{14.13}$$

The “double residual” method (see Pagan and Ullah (1999) 5.2) has several steps. First, estimate $\mathrm{E}(y_t|x_t)$ by a kernel regression of y_t on x_t, $\hat{b}_y(x)$, and $\mathrm{E}(z_t|x_t)$ by a similar kernel regression of z_t on x_t, $\hat{b}_z(x)$. Second, use these estimates in (14.13)

$$y_t - \hat{b}_y(x_t) = [z_t - \hat{b}_z(x_t)]'\beta + \varepsilon_t \tag{14.14}$$

and estimate β by least squares. Third, use these estimates in (14.12) to estimate $b(x_t)$ as

$$\hat{b}(x_t) = \hat{b}_y(x_t) - \hat{b}_z(x_t)'\hat{\beta}. \tag{14.15}$$

It can be shown that (under the assumption that y_t, z_t and x_t are iid)

$$\sqrt{T}(\hat{\beta} - \beta) \to^d N\left[0, \mathrm{Var}(\varepsilon_t)\,\mathrm{Cov}(z_t|x_t)^{-1}\right]. \tag{14.16}$$

We can consistently estimate $\mathrm{Var}(\varepsilon_t)$ by the sample variance of the fitted residuals in (14.11)—plugging in the estimated β and $b(x_t)$: and we can also consistently estimate $\mathrm{Cov}(z_t|x_t)$ by the sample variance of $z_t - \hat{b}_z(x_t)$. Clearly, this result is as if we knew the non-parametric parts of the problem (which relies on the consistency of their estimators), so it is only an asymptotic results. By using this result, together with the known properties of the kernel regressions (see, for instance, 14.6), it should be possible to apply the delta method to (14.15) to construct the asymptotic variance of $\hat{b}(x_t)$ (that is, at a given point x_t).

14.6 Estimating and Testing Distributions

Reference: Harvey (1989) 260, Davidson and MacKinnon (1993) 267, Silverman (1986); Mittelhammer (1996), DeGroot (1986)

14.6.1 Parametric Tests of Normal Distribution

The skewness, kurtosis and Bera-Jarque test for normality are useful diagnostic tools. For an iid normally distributed variable, $x_t \sim$iid $N\left(\mu, \sigma^2\right)$, they are

$$\begin{array}{lcll} & & \text{Test statistic} & \text{Distribution} \\ \text{skewness} & = & \frac{1}{T}\sum_{t=1}^{T}\left(\frac{x_t-\mu}{\sigma}\right)^3 & N\left(0, \frac{6}{T}\right) \\ \text{kurtosis} & = & \frac{1}{T}\sum_{t=1}^{T}\left(\frac{x_t-\mu}{\sigma}\right)^4 & N\left(3, \frac{24}{T}\right) \\ \text{Bera-Jarque} & = & \frac{T}{6}\text{skewness}^2 + \frac{T}{24}\left(\text{kurtosis} - 3\right)^2 & \chi_2^2. \end{array} \tag{14.17}$$

This is implemented by using the estimated mean and standard deviation. The distributions stated on the right hand side of (14.17) are under the null hypothesis that x_t is iid $N\left(\mu, \sigma^2\right)$.

The intuition for the χ_2^2 distribution of the Bera-Jarque test is that both the skewness and kurtosis are, if properly scaled, $N(0,1)$ variables. It can also be shown that they, under the null hypothesis, are uncorrelated. The Bera-Jarque test statistic is therefore a sum of the square of two uncorrelated $N(0,1)$ variables, which has a χ_2^2 distribution.

The Bera-Jarque test can also be implemented as a test of overidentifying restrictions in GMM. The moment conditions

$$g(\mu, \sigma^2) = \frac{1}{T}\sum_{t=1}^{T}\begin{bmatrix} x_t - \mu \\ (x_t-\mu)^2 - \sigma^2 \\ (x_t-\mu)^3 \\ (x_t-\mu)^4 - 3\sigma^4 \end{bmatrix}, \tag{14.18}$$

should all be zero if x_t is $N(\mu, \sigma^2)$. We can estimate the two parameters, μ and σ^2, by using the first two moment conditions only, and then test if all four moment conditions are satisfied. It can be shown that this is the same as the Bera-Jarque test if x_t is indeed iid $N(\mu, \sigma^2)$.

14.6.2 Nonparametric Tests of General Distributions

The *Kolmogorov-Smirnov* test is designed to test if an empirical distribution function, EDF(x), conforms with a theoretical cdf, $F(x)$. The empirical distribution function is defined as the fraction of observations which are less or equal to x, that is,

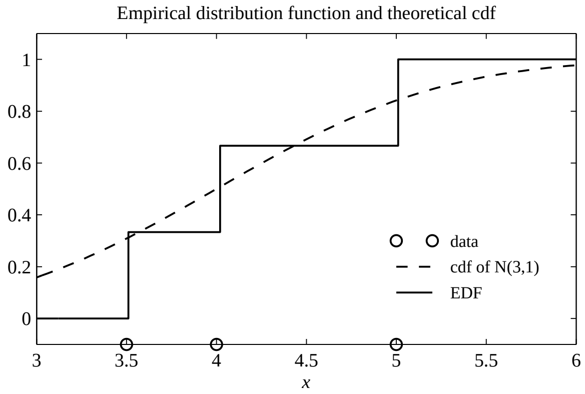


Figure 14.6: Example of empirical distribution function

$$\text{EDF}(x) = \frac{1}{T}\sum_{t=1}^{T}\delta(x_t \leq x), \text{ where } \delta(q) = \begin{cases} 1 \text{ if } q \text{ is true} \\ 0 \text{ else.} \end{cases} \tag{14.19}$$

The EDF(x_t) and $F(x_t)$ are often plotted against the sorted (in ascending order) sample $\{x_t\}_{t=1}^{T}$. See Figure 14.6.

Example 14.10 *(EDF) Suppose we have a sample with three data points:* $[x_1, x_2, x_3] = [5, 3.5, 4]$. *The empirical distribution function is then as in Figure 14.6.*

Define the absolute value of the maximum distance

$$D_T = \max_{x_t}\left|\text{EDF}(x_t) - F(x_t)\right|. \tag{14.20}$$

Example 14.11 *(Kolmogorov-Smirnov test statistic) Figure 14.6 also shows the cumulative distribution function (cdf) of a normally distributed variable. The test statistic (14.20) is then the largest difference (in absolute terms) of the EDF and the cdf—among the observed values of x_t.*

We reject the null hypothesis that $\text{EDF}(x) = F(x)$ if $\sqrt{T} D_t > c$, where c is a critical value which can be calculated from

$$\lim_{T\to\infty} \Pr\left(\sqrt{T} D_T \leq c\right) = 1 - 2\sum_{i=1}^{\infty} (-1)^{i-1} e^{-2i^2c^2}. \tag{14.21}$$

It can be approximated by replacing ∞ with a large number (for instance, 100). For instance, $c = 1.35$ provides a 5% critical value. There is a corresponding test for comparing two empirical cdfs.

Pearson's χ^2 *test* does the same thing as the K-S test but for a discrete distribution. Suppose you have K categories with N_i values in category i. The theoretical distribution predicts that the fraction p_i should be in category i, with $\sum_{i=1}^{K} p_i = 1$. Then

$$\sum_{i=1}^{K} \frac{(N_i - Tp_i)^2}{Tp_i} \sim \chi^2_{K-1}. \tag{14.22}$$

There is a corresponding test for comparing two empirical distributions.

14.6.3 Kernel Density Estimation

Reference: Silverman (1986)

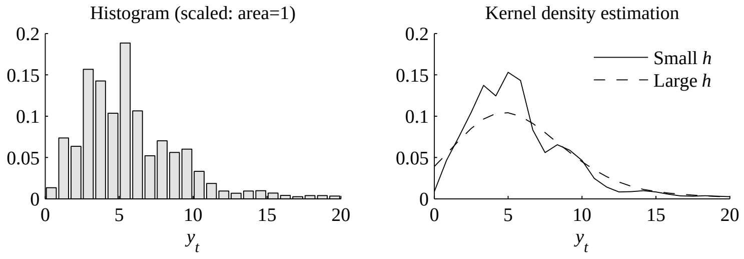
A histogram is just a count of the relative number of observations that fall in (pre-specified) non-overlapping intervals. If we also divide by the width of the interval, then the area under the histogram is unity, so the scaled histogram can be interpreted as a density function. For instance, if the intervals ("bins") are a wide, then the scaled histogram can be defined as

$$g(x|x \text{ is in bin}_i) = \frac{1}{T}\sum_{t=1}^{T} \frac{1}{a}\delta(x_t \text{ is in bin}_i), \text{ where } \delta(q) = \begin{cases} 1 \text{ if } q \text{ is true} \\ 0 \text{ else.} \end{cases} \tag{14.23}$$

Note that the area under $g(x)$ indeed integrates to unity.

We can gain efficiency by using a more sophisticated estimator. In particular, using a pdf instead of the binary function is often both convenient and more efficient. The $N(0, h^2)$ is often used. The kernel density estimator of the pdf at some point x is then

$$\hat{f}(x) = \frac{1}{Th}\sum\nolimits_{t=1}^{T} K\left(\frac{x_t - x}{h}\right), \text{ where } K(u) = \frac{\exp\left(-u^2/2\right)}{\sqrt{2\pi}}. \tag{14.24}$$



Daily federal funds rates 1954−2006

K−S (against N(μ,σ²)): sqrt(T)D = 15.32
Skewness: 1.251
kurtosis: 5.235
Bera−Jarque: 8996

Figure 14.7: Federal funds rate

The value $h = \text{Std}(x_t)1.06T^{-1/5}$ is sometimes recommended, since it can be shown to be the optimal choice (in MSE sense) if data is normally distributed and the $N(0, 1)$ kernel is used. Clearly, using $K\left[(x_t - x)/h\right]/h$ is the same as using pdf of $N(x, h^2)$. The bandwidth h could be chosen by a leave-one-out cross-validation technique.

The results on bias and variance in (14.4) are approximately true also for the kernel density estimation if we interpret $b(x)$ as the pdf of x. In particular, it can be shown that (with iid data and a Gaussian kernel) the asymptotic distribution is

$$\sqrt{Th}\left[\hat{f}(x) - \mathrm{E}\,\hat{f}(x)\right] \to^d N\left[0, \frac{1}{2\sqrt{\pi}} f(x)\right], \tag{14.25}$$

The easiest way to handle a bounded support of x is to transform the variable into one with an unbounded support, estimate the pdf for this variable, and then use the "change of variable" technique to transform to the pdf of the original variable.

We can also estimate multivariate pdfs. Let x_t be a $d \times 1$ matrix and $\hat{\Omega}$ be the estimated covariance matrix of x_t. We can then estimate the pdf at a point x by using a multivariate

Gaussian kernel as

$$\hat{f}(x) = \frac{1}{Th^d(2\pi)^{d/2}|\hat{\Omega}|^{1/2}} \sum\nolimits_{t=1}^{T} \exp[-(x-x_t)'\hat{\Omega}^{-1}(x-x_t)/(2h^2)]. \qquad (14.26)$$

The value $h = 0.96T^{-1/(d+4)}$ is sometimes recommended.

Kernel Density Estimation and LAD

Reference: Amemiya (1985) 4.6

Let $\hat{u}_t(b)$ be the residuals in a regression model for choice b of the coefficients, $\hat{u}_t(b) = y_t - x_t'b$. The least absolute deviations (LAD) estimator solve

$$\hat{\beta}_{LAD} = \arg\min_b \sum_{t=1}^{T} |\hat{u}_t(b)| \qquad (14.27)$$

This estimator involve non-linearities, but a a simple iteration works nicely. If we assume that the median of the true residual, u_t, is zero, then we (typically) have

$$\sqrt{T}(\hat{\beta}_{LAD} - \beta_0) \to^d N\left[0, f(0)^{-2}\Sigma_{xx}^{-1}/4\right], \text{ where } \Sigma_{xx} = \text{plim} \sum\nolimits_{t=1}^{T} x_t x_t'/T, \qquad (14.28)$$

where $f(0)$ is the value of the pdf of u_t at zero. Unless we know this density function (or else we would probably have used MLE instead of LAD), we need to estimate it—for instance with a kernel density method.

Example 14.12 *($N(0,\sigma^2)$) When $u_t \sim N(0,\sigma^2)$, then $f(0) = 1/\sqrt{2\pi\sigma^2}$, so the covariance matrix in (14.28) becomes $\pi\sigma^2\Sigma_{xx}^{-1}/2$. This is $\pi/2$ times larger than when using LS.*

Remark 14.13 *(Algorithm for LAD) The LAD estimator can be written*

$$\hat{\beta}_{LAD} = \arg\min_\beta \sum_{t=1}^{T} w_t \hat{u}_t(b)^2, \; w_t = 1/\left|\hat{u}_t(b)\right|,$$

so it is a weighted least squares where both y_t and x_t are multiplied by $1/\left|\hat{u}_t(b)\right|$. It can be shown that iterating on LS with the weights given by $1/\left|\hat{u}_t(b)\right|$, where the residuals are from the previous iteration, converges very quickly to the LAD estimator.

Bibliography

Amemiya, T., 1985, *Advanced Econometrics*, Harvard University Press, Cambridge, Massachusetts.

Campbell, J. Y., A. W. Lo, and A. C. MacKinlay, 1997, *The Econometrics of Financial Markets*, Princeton University Press, Princeton, New Jersey.

Davidson, R., and J. G. MacKinnon, 1993, *Estimation and Inference in Econometrics*, Oxford University Press, Oxford.

DeGroot, M. H., 1986, *Probability and Statistics*, Addison-Wesley, Reading, Massachusetts.

Härdle, W., 1990, *Applied Nonparametric Regression*, Cambridge University Press, Cambridge.

Harvey, A. C., 1989, *Forecasting, Structural Time Series Models and the Kalman Filter*, Cambridge University Press.

Mittelhammer, R. C., 1996, *Mathematical Statistics for Economics and Business*, Springer-Verlag, New York.

Mittelhammer, R. C., G. J. Judge, and D. J. Miller, 2000, *Econometric Foundations*, Cambridge University Press, Cambridge.

Pagan, A., and A. Ullah, 1999, *Nonparametric Econometrics*, Cambridge University Press.

Silverman, B. W., 1986, *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.

21 Some Statistics

This section summarizes some useful facts about statistics. Heuristic proofs are given in a few cases.

Some references: Mittelhammer (1996), DeGroot (1986), Greene (2000), Davidson (2000), Johnson, Kotz, and Balakrishnan (1994).

21.1 Distributions and Moment Generating Functions

Most of the stochastic variables we encounter in econometrics are continuous. For a continuous random variable X, the range is uncountably infinite and the probability that $X \leq x$ is $\Pr(X \leq x) = \int_{-\infty}^{x} f(q)dq$ where $f(q)$ is the continuous probability density function of X. Note that X is a random variable, x is a number (1.23 or so), and q is just a dummy argument in the integral.

Fact 21.1 *(cdf and pdf) The cumulative distribution function of the random variable X is $F(x) = \Pr(X \leq x) = \int_{-\infty}^{x} f(q)dq$. Clearly, $f(x) = dF(x)/dx$. Note that x is just a number, not random variable.*

Fact 21.2 *(Moment generating function of X) The moment generating function of the random variable X is $mgf(t) = \mathrm{E}\, e^{tX}$. The rth moment is the rth derivative of $mgf(t)$ evaluated at $t = 0$: $\mathrm{E}\, X^r = d mgf(0)/dt^r$. If a moment generating function exists (that is, $\mathrm{E}\, e^{tX} < \infty$ for some small interval $t \in (-h, h)$), then it is unique.*

Fact 21.3 *(Moment generating function of a function of X) If X has the moment generating function $mgf_X(t) = \mathrm{E}\, e^{tX}$, then $g(X)$ has the moment generating function $\mathrm{E}\, e^{tg(X)}$. The affine function $a + bX$ (a and b are constants) has the moment generating function $mgf_{g(X)}(t) = \mathrm{E}\, e^{t(a+bX)} = e^{ta}\, \mathrm{E}\, e^{tbX} = e^{ta} mgf_X(bt)$. By setting $b = 1$ and $a = -\mathrm{E}\, X$ we obtain a mgf for central moments (variance, skewness, kurtosis, etc), $mgf_{(X-\mathrm{E}\, X)}(t) = e^{-t\,\mathrm{E}\, X} mgf_X(t)$.*

Example 21.4 *When $X \sim N(\mu, \sigma^2)$, then $mgf_X(t) = \exp\left(\mu t + \sigma^2 t^2/2\right)$. Let $Z = (X-\mu)/\sigma$ so $a = -\mu/\sigma$ and $b = 1/\sigma$. This gives $mgf_Z(t) = \exp(-\mu t/\sigma) mgf_X(t/\sigma) = \exp\left(t^2/2\right)$. (Of course, this result can also be obtained by directly setting $\mu = 0$ and $\sigma = 1$ in mgf_X.)*

Fact 21.5 *(Change of variable, univariate case, monotonic function) Suppose X has the probability density function $f_X(c)$ and cumulative distribution function $F_X(c)$. Let $Y = g(X)$ be a continuously differentiable function with $dg/dX \neq 0$ (so $g(X)$ is monotonic) for all c such that $f_X(c) > 0$. Then the cdf of Y is*

$$F_Y(c) = \Pr[Y \leq c] = \Pr[g(X) \leq c] = \Pr[X \leq g^{-1}(c)] = F_X[g^{-1}(c)],$$

where g^{-1} is the inverse function of g such that $g^{-1}(Y) = X$. We also have that the pdf of Y is

$$f_Y(c) = f_X[g^{-1}(c)] \left| \frac{dg^{-1}(c)}{dc} \right|.$$

Proof. Differentiate $F_Y(c)$, that is, $F_X[g^{-1}(c)]$ with respect to c. ■

Example 21.6 *Let $X \sim U(0, 1)$ and $Y = g(X) = F^{-1}(X)$ where $F(c)$ is a strictly increasing cdf. We then get*

$$f_Y(c) = \frac{dF(c)}{dc}.$$

The variable Y then has the pdf $dF(c)/dc$ and the cdf $F(c)$.

Example 21.7 *Let $Y = \exp(X)$, so the inverse function is $X = \ln Y$ with derivative $1/Y$. Then, $f_Y(c) = f_X(\ln c)/\exp(c)$.*

Fact 21.8 *(Distribution of truncated a random variable) Let the probability distribution and density functions of X be $F(x)$ and $f(x)$, respectively. The corresponding functions, conditional on $a < X \leq b$ are $F(x)/[F(b) - F(a)]$ and $f(x)/[F(b) - F(a)]$.*

21.2 Joint and Conditional Distributions and Moments

21.2.1 Joint and Conditional Distributions

Fact 21.9 *(Joint and marginal cdf) Let X and Y be (possibly vectors of) random variables and let x and y be two numbers. The joint cumulative distribution function of X and Y is $H(x, y) = \Pr(X \leq x, Y \leq y) = \int_{-\infty}^{x}\int_{-\infty}^{y} h(q_x, q_y)dq_y dq_x$, where $h(x, y) = \partial^2 F(x, y)/\partial x \partial y$ is the joint probability density function.*

Fact 21.10 *(Joint and marginal pdf) The marginal cdf of X is obtained by integrating out Y: $F(x) = \Pr(X \leq x, Y \text{ anything}) = \int_{-\infty}^{x} \left[\int_{-\infty}^{\infty} h(q_x, q_y) dq_y\right] dq_x$. This shows that the marginal pdf of x is $f(x) = dF(x)/dx = \int_{-\infty}^{\infty} h(q_x, q_y) dq_y$.*

Fact 21.11 *(Conditional distribution) The pdf of Y conditional on $X = x$ (a number) is $g(y|x) = h(x, y)/f(x)$. This is clearly proportional to the joint pdf (at the given value x).*

Fact 21.12 *(Change of variable, multivariate case, monotonic function) The result in Fact 21.5 still holds if X and Y are both $n \times 1$ vectors, but the derivative are now $\partial g^{-1}(c)/\partial dc'$ which is an $n \times n$ matrix. If g_i^{-1} is the ith function in the vector g^{-1} then*

$$\frac{\partial g^{-1}(c)}{\partial dc'} = \begin{bmatrix} \frac{\partial g_1^{-1}(c)}{\partial c_1} & \cdots & \frac{\partial g_1^{-1}(c)}{\partial c_n} \\ \vdots & & \vdots \\ \frac{\partial g_n^{-1}(c)}{\partial c_1} & \cdots & \frac{\partial g_n^{-1}(c)}{\partial c_m} \end{bmatrix}.$$

21.2.2 Moments of Joint Distributions

Fact 21.13 *(Caucy-Schwartz)* $(\mathrm{E}\, XY)^2 \leq \mathrm{E}(X^2)\, \mathrm{E}(Y^2)$.

Proof. $0 \leq \mathrm{E}[(aX+Y)^2] = a^2\, \mathrm{E}(X^2) + 2a\, \mathrm{E}(XY) + \mathrm{E}(Y^2)$. Set $a = -\mathrm{E}(XY)/\, \mathrm{E}(X^2)$ to get

$$0 \leq -\frac{[\mathrm{E}(XY)]^2}{\mathrm{E}(X^2)} + \mathrm{E}(Y^2), \text{ that is, } \frac{[\mathrm{E}(XY)]^2}{\mathrm{E}(X^2)} \leq \mathrm{E}(Y^2).$$

■

Fact 21.14 *$(-1 \leq Corr(X, y) \leq 1)$. Let Y and X in Fact 21.13 be zero mean variables (or variables minus their means). We then get $[\mathrm{Cov}(X, Y)]^2 \leq \mathrm{Var}(X)\, \mathrm{Var}(Y)$, that is, $-1 \leq \mathrm{Cov}(X, Y)/[Std(X)Std(Y)] \leq 1$.*

21.2.3 Conditional Moments

Fact 21.15 *(Conditional moments) $\mathrm{E}\,(Y|x) = \int y g(y|x) dy$ and $\mathrm{Var}\,(Y|x) = \int [y - \mathrm{E}\,(Y|x)] g(y|x) dy$.*

Fact 21.16 *(Conditional moments as random variables) Before we observe X, the conditional moments are random variables—since X is. We denote these random variables by $\mathrm{E}\,(Y|X)$, $\mathrm{Var}\,(Y|X)$, etc.*

Fact 21.17 *(Law of iterated expectations) $\mathrm{E}\, Y = \mathrm{E}[\mathrm{E}\,(Y|X)]$. Note that $\mathrm{E}\,(Y|X)$ is a random variable since it is a function of the random variable X. It is not a function of Y, however. The outer expectation is therefore an expectation with respect to X only.*

Proof. $\mathrm{E}[\mathrm{E}\,(Y|X)] = \int \left[\int y g(y|x) dy\right] f(x) dx = \int\int y g(y|x) f(x) dy dx = \int\int y h(y, x) dy dx = \mathrm{E}\, Y$. ■

Fact 21.18 *(Conditional vs. unconditional variance)* $\mathrm{Var}\,(Y) = \mathrm{Var}\,[\mathrm{E}\,(Y|X)] + \mathrm{E}\,[\mathrm{Var}\,(Y|X)]$.

Fact 21.19 *(Properties of Conditional Expectations) (a) $Y = \mathrm{E}\,(Y|X) + U$ where U and $\mathrm{E}\,(Y|X)$ are uncorrelated:* $\mathrm{Cov}\,(X, Y) = \mathrm{Cov}\,[X, \mathrm{E}\,(Y|X) + U] = \mathrm{Cov}\,[X, \mathrm{E}\,(Y|X)]$. *It follows that (b)* $\mathrm{Cov}[Y, \mathrm{E}\,(Y|X)] = \mathrm{Var}[\mathrm{E}\,(Y|X)]$*; and (c)* $\mathrm{Var}\,(Y) = \mathrm{Var}\,[\mathrm{E}\,(Y|X)] + \mathrm{Var}\,(U)$. *Property (c) is the same as Fact 21.18, where* $\mathrm{Var}\,(U) = \mathrm{E}\,[\mathrm{Var}\,(Y|X)]$.

Proof. $\mathrm{Cov}\,(X, Y) = \int\int x(y - \mathrm{E}\, y) h(x, y) dy dx = \int x \left[\int (y - \mathrm{E}\, y) g(y|x) dy\right] f(x) dx$, but the term in brackets is $\mathrm{E}\,(Y|X) - \mathrm{E}\, Y$. ■

Fact 21.20 *(Conditional expectation and unconditional orthogonality)* $\mathrm{E}\,(Y|Z) = 0 \Rightarrow \mathrm{E}\, YZ = 0$.

Proof. Note from Fact 21.19 that $\mathrm{E}(Y|X) = 0$ implies $\mathrm{Cov}\,(X, Y) = 0$ so $\mathrm{E}\, XY = \mathrm{E}\, X\, \mathrm{E}\, Y$ (recall that $\mathrm{Cov}\,(X, Y) = \mathrm{E}\, XY - \mathrm{E}\, X\, \mathrm{E}\, Y$.) Note also that $\mathrm{E}\,(Y|X) = 0$ implies that $\mathrm{E}\, Y = 0$ (by iterated expectations). We therefore get

$$\mathrm{E}\,(Y|X) = 0 \Rightarrow \begin{bmatrix} \mathrm{Cov}\,(X, Y) = 0 \\ \mathrm{E}\, Y = 0 \end{bmatrix} \Rightarrow \mathrm{E}\, YX = 0.$$

■

21.2.4 Regression Function and Linear Projection

Fact 21.21 *(Regression function) Suppose we use information in some variables X to predict Y. The choice of the forecasting function $\hat{Y} = k(X) = \mathrm{E}\,(Y|X)$ minimizes $\mathrm{E}[Y - k(X)]^2$. The conditional expectation $\mathrm{E}\,(Y|X)$ is also called the regression function of Y on X. See Facts 21.19 and 21.20 for some properties of conditional expectations.*

Fact 21.22 *(Linear projection) Suppose we want to forecast the scalar Y using the $k \times 1$ vector X and that we restrict the forecasting rule to be linear $\hat{Y} = X'\beta$. This rule is a linear projection, denoted $P(Y|X)$, if β satisfies the orthogonality conditions* $\mathrm{E}[X(Y - X'\beta)] = \mathbf{0}_{k\times 1}$, *that is, if* $\beta = (\mathrm{E}\, XX')^{-1}\,\mathrm{E}\, XY$. *A linear projection minimizes* $\mathrm{E}[Y - k(X)]^2$ *within the class of linear $k(X)$ functions.*

Fact 21.23 *(Properties of linear projections) (a) The orthogonality conditions in Fact 21.22 mean that*

$$Y = X'\beta + \varepsilon,$$

where $\mathrm{E}(X\varepsilon) = \mathbf{0}_{k\times 1}$. *This implies that* $\mathrm{E}[P(Y|X)\varepsilon] = 0$, *so the forecast and forecast error are orthogonal. (b) The orthogonality conditions also imply that* $\mathrm{E}[XY] = \mathrm{E}[XP(Y|X)]$. *(c) When X contains a constant, so* $\mathrm{E}\,\varepsilon = 0$, *then (a) and (b) carry over to covariances:* $\mathrm{Cov}[P(Y|X), \varepsilon] = 0$ *and* $\mathrm{Cov}[X, Y] = \mathrm{Cov}[XP, (Y|X)]$.

Example 21.24 *($P(1|X)$) When $Y_t = 1$, then* $\beta = (\mathrm{E}\, XX')^{-1}\,\mathrm{E}\, X$. *For instance, suppose $X = [x_{1t}, x_{t2}]'$. Then*

$$\beta = \begin{bmatrix} \mathrm{E}\, x_{1t}^2 & \mathrm{E}\, x_{1t}x_{2t} \\ \mathrm{E}\, x_{2t}x_{1t} & \mathrm{E}\, x_{2t}^2 \end{bmatrix}^{-1} \begin{bmatrix} \mathrm{E}\, x_{1t} \\ \mathrm{E}\, x_{2t} \end{bmatrix}.$$

If $x_{1t} = 1$ in all periods, then this simplifies to $\beta = [1, 0]'$.

Remark 21.25 *Some authors prefer to take the transpose of the forecasting rule, that is, to use $\hat{Y} = \beta' X$. Clearly, since XX' is symmetric, we get* $\beta' = \mathrm{E}(YX')(\mathrm{E}\, XX')^{-1}$.

Fact 21.26 *(Linear projection with a constant in X) If X contains a constant, then $P(aY + b|X) = aP(Y|X) + b$.*

Fact 21.27 *(Linear projection versus regression function) Both the linear regression and the regression function (see Fact 21.21) minimize* $\mathrm{E}[Y - k(X)]^2$, *but the linear projection imposes the restriction that $k(X)$ is linear, whereas the regression function does not impose any restrictions. In the special case when Y and X have a joint normal distribution, then the linear projection is the regression function.*

Fact 21.28 *(Linear projection and OLS) The linear projection is about population moments, but OLS is its sample analogue.*

21.3 Convergence in Probability, Mean Square, and Distribution

Fact 21.29 *(Convergence in probability) The sequence of random variables $\{X_T\}$ converges in probability to the random variable X if (and only if) for all $\varepsilon > 0$*

$$\lim_{T\to\infty} \Pr(|X_T - X| < \varepsilon) = 1.$$

We denote this $X_T \overset{p}{\to} X$ or $\operatorname{plim} X_T = X$ *(X is the probability limit of X_T). Note: (a) X can be a constant instead of a random variable; (b) if X_T and X are matrices, then $X_T \overset{p}{\to} X$ if the previous condition holds for every element in the matrices.*

Example 21.30 *Suppose $X_T = 0$ with probability $(T-1)/T$ and $X_T = T$ with probability $1/T$. Note that* $\lim_{T\to\infty} \Pr(|X_T - 0| = 0) = \lim_{T\to\infty}(T-1)/T = 1$, *so* $\lim_{T\to\infty} \Pr(|X_T - 0| = \varepsilon) = 1$ *for any $\varepsilon > 0$. Note also that* $\mathrm{E}\, X_T = 0 \times (T-1)/T + T \times 1/T = 1$, *so X_T is biased.*

Fact 21.31 *(Convergence in mean square) The sequence of random variables $\{X_T\}$ converges in mean square to the random variable X if (and only if)*

$$\lim_{T\to\infty} \mathrm{E}(X_T - X)^2 = 0.$$

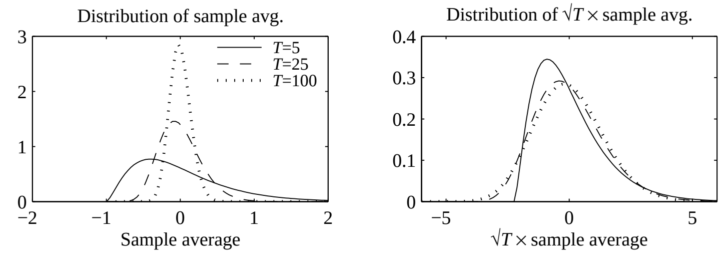
We denote this $X_T \overset{m}{\to} X$. Note: (a) X can be a constant instead of a random variable; (b) if X_T and X are matrices, then $X_T \overset{m}{\to} X$ if the previous condition holds for every element in the matrices.

Fact 21.32 *(Convergence in mean square to a constant) If X in Fact 21.31 is a constant, then then $X_T \overset{m}{\to} X$ if (and only if)*

$$\lim_{T\to\infty} (\mathrm{E}\, X_T - X)^2 = 0 \textit{ and } \lim_{T\to\infty} \mathrm{Var}(X_T{}^2) = 0.$$

This means that both the variance and the squared bias go to zero as $T \to \infty$.

Proof. $\mathrm{E}(X_T - X)^2 = \mathrm{E}\, X_T^2 - 2X\,\mathrm{E}\, X_T + X^2$. Add and subtract $(\mathrm{E}\, X_T)^2$ and recall that $\mathrm{Var}(X_T) = \mathrm{E}\, X_T^2 - (\mathrm{E}\, X_T)^2$. This gives $\mathrm{E}(X_T - X)^2 = \mathrm{Var}(X_T) - 2X\,\mathrm{E}\, X_T + X^2 + (\mathrm{E}\, X_T)^2 = \mathrm{Var}(X_T) + (\mathrm{E}\, X_T - X)^2$. ■



Sample average of $z_t - 1$ where z_t has a $\chi^2(1)$ distribution

Figure 21.1: Sampling distributions

Fact 21.33 *(Convergence in distribution) Consider the sequence of random variables* $\{X_T\}$ *with the associated sequence of cumulative distribution functions* $\{F_T\}$. *If* $\lim_{T\to\infty} F_T = F$ *(at all points), then* F *is the limiting cdf of* X_T. *If there is a random variable* X *with cdf* F, *then* X_T *converges in distribution to* X: $X_T \overset{d}{\to} X$. *Instead of comparing cdfs, the comparison can equally well be made in terms of the probability density functions or the moment generating functions.*

Fact 21.34 *(Relation between the different types of convergence) We have* $X_T \overset{m}{\to} X \Rightarrow X_T \overset{p}{\to} X \Rightarrow X_T \overset{d}{\to} X$. *The reverse implications are not generally true.*

Example 21.35 *Consider the random variable in Example 21.30. The expected value is* $\mathrm{E}\,X_T = 0(T-1)/T + T/T = 1$. *This means that the squared bias does not go to zero, so* X_T *does not converge in mean square to zero.*

Fact 21.36 *(Slutsky's theorem) If* $\{X_T\}$ *is a sequence of random matrices such that* $\operatorname{plim} X_T = X$ *and* $g(X_T)$ *a continuous function, then* $\operatorname{plim} g(X_T) = g(X)$.

Fact 21.37 *(Continuous mapping theorem) Let the sequences of random matrices* $\{X_T\}$ *and* $\{Y_T\}$, *and the non-random matrix* $\{a_T\}$ *be such that* $X_T \overset{d}{\to} X$, $Y_T \overset{p}{\to} Y$, *and* $a_T \to a$ *(a traditional limit). Let* $g(X_T, Y_T, a_T)$ *be a continuous function. Then* $g(X_T, Y_T, a_T) \overset{d}{\to} g(X, Y, a)$.

21.4 Laws of Large Numbers and Central Limit Theorems

Fact 21.38 *(Khinchine's theorem) Let* X_t *be independently and identically distributed (iid) with* $\mathrm{E}\,X_t = \mu < \infty$. *Then* $\Sigma_{t=1}^T X_t/T \overset{p}{\to} \mu$.

Fact 21.39 *(Chebyshev's theorem) If* $\mathrm{E}\,X_t = 0$ *and* $\lim_{T\to\infty} \mathrm{Var}(\Sigma_{t=1}^T X_t/T) = 0$, *then* $\Sigma_{t=1}^T X_t/T \overset{p}{\to} 0$.

Fact 21.40 *(The Lindeberg-Lévy theorem) Let* X_t *be independently and identically distributed (iid) with* $\mathrm{E}\,X_t = 0$ *and* $\mathrm{Var}(X_t) < \infty$. *Then* $\frac{1}{\sqrt{T}} \Sigma_{t=1}^T X_t/\sigma \overset{d}{\to} N(0,1)$.

21.5 Stationarity

Fact 21.41 *(Covariance stationarity)* X_t *is covariance stationary if*

$$\begin{gathered} \mathrm{E}\,X_t = \mu \text{ is independent of } t, \\ \mathrm{Cov}\,(X_{t-s}, X_t) = \gamma_s \text{ depends only on } s, \text{ and} \\ \text{both } \mu \text{ and } \gamma_s \text{ are finite.} \end{gathered}$$

Fact 21.42 *(Strict stationarity)* X_t *is strictly stationary if, for all* s, *the joint distribution of* $X_t, X_{t+1}, ..., X_{t+s}$ *does not depend on* t.

Fact 21.43 *(Strict stationarity versus covariance stationarity) In general, strict stationarity does not imply covariance stationarity or vice versa. However, strict stationary with finite first two moments implies covariance stationarity.*

21.6 Martingales

Fact 21.44 *(Martingale) Let* Ω_t *be a set of information in* t, *for instance* $Y_t, Y_{t-1}, ...$ *If* $\mathrm{E}\,|Y_t| < \infty$ *and* $\mathrm{E}(Y_{t+1}|\Omega_t) = Y_t$, *then* Y_t *is a martingale.*

Fact 21.45 *(Martingale difference) If* Y_t *is a martingale, then* $X_t = Y_t - Y_{t-1}$ *is a martingale difference:* X_t *has* $\mathrm{E}\,|X_t| < \infty$ *and* $\mathrm{E}(X_{t+1}|\Omega_t) = 0$.

Fact 21.46 *(Innovations as a martingale difference sequence) The forecast error* $X_{t+1} = Y_{t+1} - \mathrm{E}(Y_{t+1}|\Omega_t)$ *is a martingale difference.*

Fact 21.47 *(Properties of martingales) (a) If Y_t is a martingale, then* $\mathrm{E}(Y_{t+s}|\Omega_t) = Y_t$ *for* $s \geq 1$. *(b) If X_t is a martingale difference, then* $\mathrm{E}(X_{t+s}|\Omega_t) = 0$ *for* $s \geq 1$.

Proof. (a) Note that $\mathrm{E}(Y_{t+2}|\Omega_{t+1}) = Y_{t+1}$ and take expectations conditional on Ω_t: $\mathrm{E}[\mathrm{E}(Y_{t+2}|\Omega_{t+1})|\Omega_t] = \mathrm{E}(Y_{t+1}|\Omega_t) = Y_t$. By iterated expectations, the first term equals $\mathrm{E}(Y_{t+2}|\Omega_t)$. Repeat this for $t+3$, $t+4$, etc. (b) Essentially the same proof. ■

Fact 21.48 *(Properties of martingale differences) If X_t is a martingale difference and g_{t-1} is a function of Ω_{t-1}, then $X_t g_{t-1}$ is also a martingale difference.*

Proof. $\mathrm{E}(X_{t+1}g_t|\Omega_t) = \mathrm{E}(X_{t+1}|\Omega_t)g_t$ since g_t is a function of Ω_t. ■

Fact 21.49 *(Martingales, serial independence, and no autocorrelation) (a) X_t is serially uncorrelated if* $\mathrm{Cov}(X_t, X_{t+s}) = 0$ *for all* $s \neq 0$. *This means that a linear projection of X_{t+s} on $X_t, X_{t-1},...$ is a constant, so it cannot help predict X_{t+s}. (b) X_t is a martingale difference with respect to its history if* $\mathrm{E}(X_{t+s}|X_t, X_{t-1}, ...) = 0$ *for all* $s \geq 1$. *This means that no function of $X_t, X_{t-1}, ...$ can help predict X_{t+s}. (c) X_t is serially independent if* $\mathrm{pdf}(X_{t+s}|X_t, X_{t-1}, ...) = \mathrm{pdf}(X_{t+s})$. *This means than no function of $X_t, X_{t-1}, ...$ can help predict any function of X_{t+s}.*

Fact 21.50 *(WLN for martingale difference) If X_t is a martingale difference, then* $\mathrm{plim}\ \Sigma_{t=1}^T X_t/T = 0$ *if either (a) X_t is strictly stationary and* $\mathrm{E}\,|x_t| < 0$ *or (b)* $\mathrm{E}\,|x_t|^{1+\delta} < \infty$ *for* $\delta > 0$ *and all t. (See Davidson (2000) 6.2)*

Fact 21.51 *(CLT for martingale difference) Let X_t be a martingale difference. If* $\mathrm{plim}\ \Sigma_{t=1}^T (X_t^2 - \mathrm{E}\,X_t^2)/T = 0$ *and either*

(a) X_t is strictly stationary or
(b) $\max_{t\in[1,T]} \frac{(\mathrm{E}\,|X_t|^{2+\delta})^{1/(2+\delta)}}{\Sigma_{t=1}^T \mathrm{E}\,X_t^2/T} < \infty$ *for* $\delta > 0$ *and all* $T > 1$,

then $(\Sigma_{t=1}^T X_t/\sqrt{T})/(\Sigma_{t=1}^T \mathrm{E}\,X_t^2/T)^{1/2} \xrightarrow{d} N(0,1)$. *(See Davidson (2000) 6.2)*

21.7 Special Distributions

21.7.1 The Normal Distribution

Fact 21.52 *(Univariate normal distribution) If* $X \sim N(\mu, \sigma^2)$, *then the probability density function of X, $f(x)$ is*

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}.$$

The moment generating function is $mgf_X(t) = \exp\left(\mu t + \sigma^2 t^2/2\right)$ *and the moment generating function around the mean is* $mgf_{(X-\mu)}(t) = \exp\left(\sigma^2 t^2/2\right)$.

Example 21.53 *The first few moments around the mean are* $\mathrm{E}(X-\mu) = 0$, $\mathrm{E}(X-\mu)^2 = \sigma^2$, $\mathrm{E}(X-\mu)^3 = 0$ *(all odd moments are zero),* $\mathrm{E}(X-\mu)^4 = 3\sigma^4$, $\mathrm{E}(X-\mu)^6 = 15\sigma^6$, *and* $\mathrm{E}(X-\mu)^8 = 105\sigma^8$.

Fact 21.54 *(Standard normal distribution) If* $X \sim N(0,1)$, *then the moment generating function is* $mgf_X(t) = \exp\left(t^2/2\right)$. *Since the mean is zero, $m(t)$ gives central moments. The first few are* $\mathrm{E}\,X = 0$, $\mathrm{E}\,X^2 = 1$, $\mathrm{E}\,X^3 = 0$ *(all odd moments are zero), and* $EX^4 = 3$. *The distribution function,* $\Pr(X \leq a) = \Phi(a) = 1/2 + 1/2\,\mathrm{erf}(a/\sqrt{2})$, *where* $\mathrm{erf}()$ *is the error function. The complementary error function is* $\mathrm{erfc}(z) = 1 - \mathrm{erf}(z)$. *Since the distribution is symmetric around zero, we have* $\Phi(-a) = \Pr(X \leq -a) = \Pr(X \geq a) = 1 - \Phi(a)$. *Clearly,* $1 - \Phi(a) = \Phi(-a) = 1/2\,\mathrm{erfc}(a/\sqrt{2})$.

Fact 21.55 *(Multivariate normal distribution) If X is an $n \times 1$ vector of random variables with a multivariate normal distribution, with a mean vector μ and variance-covariance matrix Σ, $N(\mu, \Sigma)$, then the density function is*

$$f(x) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right].$$

Fact 21.56 *(Conditional normal distribution) Suppose $Z_{m\times 1}$ and $X_{n\times 1}$ are jointly normally distributed*

$$\begin{bmatrix} Z \\ X \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_Z \\ \mu_X \end{bmatrix}, \begin{bmatrix} \Sigma_{ZZ} & \Sigma_{ZX} \\ \Sigma_{XZ} & \Sigma_{XX} \end{bmatrix}\right).$$

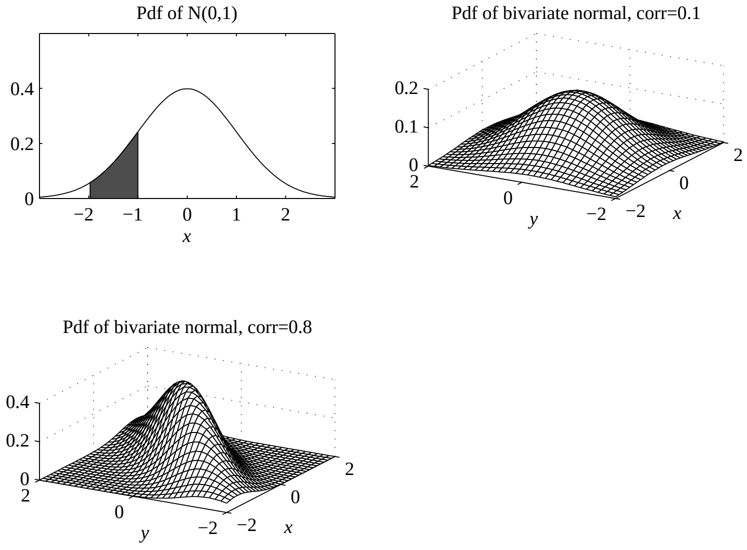


Figure 21.2: Normal distributions

The distribution of the random variable Z conditional on that $X = x$ (a number) is also normal with mean

$$\mathrm{E}\,(Z|x) = \mu_Z + \Sigma_{ZX}\Sigma_{XX}^{-1}\,(x - \mu_X)\,,$$

and variance (variance of Z conditional on that $X = x$, that is, the variance of the prediction error $Z - \mathrm{E}\,(Z|x)$)

$$\mathrm{Var}\,(Z|x) = \Sigma_{ZZ} - \Sigma_{ZX}\Sigma_{XX}^{-1}\Sigma_{XZ}.$$

Note that the conditional variance is constant in the multivariate normal distribution ($\mathrm{Var}\,(Z|X)$ *is not a random variable in this case). Note also that* $\mathrm{Var}\,(Z|x)$ *is less than* $\mathrm{Var}(Z) = \Sigma_{ZZ}$ *(in a matrix sense) if X contains any relevant information (so Σ_{ZX} is not zero, that is,* $\mathrm{E}\,(Z|x)$ *is not the same for all x).*

Fact 21.57 *(Stein's lemma) If Y has normal distribution and $h()$ is a differentiable function such that* $\mathrm{E}\,|h'(Y)| < \infty$*, then* $\mathrm{Cov}[Y, h(Y)] = \mathrm{Var}(Y)\,\mathrm{E}\,h'(Y)$.

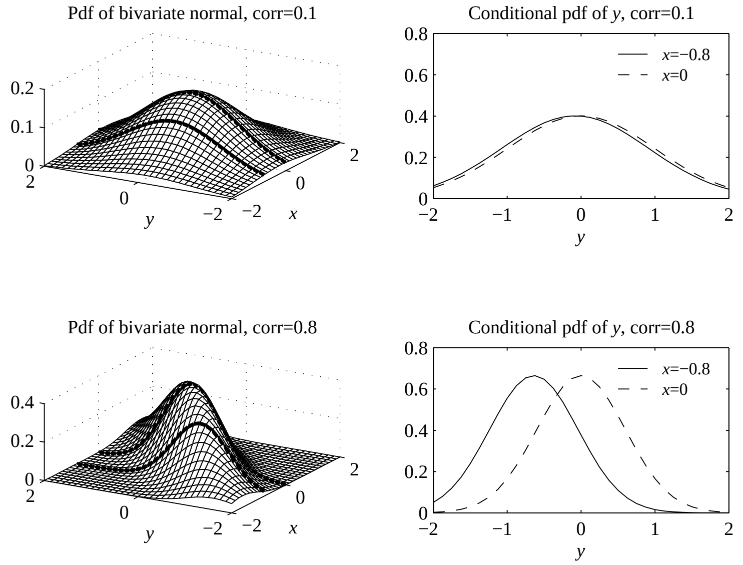


Figure 21.3: Density functions of normal distributions

Proof. $\mathrm{E}[(Y-\mu)h(Y)] = \int_{-\infty}^{\infty}(Y-\mu)h(Y)\phi(Y;\mu,\sigma^2)dY$, where $\phi(Y;\mu,\sigma^2)$ is the pdf of $N(\mu,\sigma^2)$. Note that $d\phi(Y;\mu,\sigma^2)/dY = -\phi(Y;\mu,\sigma^2)(Y-\mu)/\sigma^2$, so the integral can be rewritten as $-\sigma^2\int_{-\infty}^{\infty}h(Y)d\phi(Y;\mu,\sigma^2)$. Integration by parts ("$\int u dv = uv - \int v du$") gives $-\sigma^2\left[h(Y)\phi(Y;\mu,\sigma^2)\big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty}\phi(Y;\mu,\sigma^2)h'(Y)dY\right] = \sigma^2\,\mathrm{E}\,h'(Y)$. ■

Fact 21.58 *(Stein's lemma 2) It follows from Fact 21.57 that if X and Y have a bivariate normal distribution and $h()$ is a differentiable function such that* $\mathrm{E}\,|h'(Y)| < \infty$*, then* $\mathrm{Cov}[X, h(Y)] = \mathrm{Cov}(X,Y)\,\mathrm{E}\,h'(Y)$.

Example 21.59 *(a) With $h(Y) = \exp(Y)$ we get* $\mathrm{Cov}[X, \exp(Y)] = \mathrm{Cov}(X,Y)\,\mathrm{E}\exp(Y)$*; (b) with $h(Y) = Y^2$ we get* $\mathrm{Cov}[X, Y^2] = \mathrm{Cov}(X,Y)2\,\mathrm{E}\,Y$ *so with* $\mathrm{E}\,Y = 0$ *we get a zero covariance.*

Fact 21.60 *(Stein's lemma 3) Fact 21.58 still holds if the joint distribution of X and Y is* *a mixture of n bivariate normal distributions, provided the mean and variance of Y is the same in each of the n components.*

Fact 21.61 *(Truncated normal distribution) Let $X \sim N(\mu, \sigma^2)$, and consider truncating the distribution so that we want moments conditional on $a < X \leq b$. Define $a_0 = (a-\mu)/\sigma$ and $b_0 = (b-\mu)/\sigma$. Then,*

$$\begin{aligned}\mathrm{E}(X|a < X \leq b) &= \mu - \sigma\frac{\phi(b_0)-\phi(a_0)}{\Phi(b_0)-\Phi(a_0)} \text{ and} \\ \mathrm{Var}(X|a < X \leq b) &= \sigma^2\left\{1-\frac{b_0\phi(b_0)-a_0\phi(a_0)}{\Phi(b_0)-\Phi(a_0)}-\left[\frac{\phi(b_0)-\phi(a_0)}{\Phi(b_0)-\Phi(a_0)}\right]^2\right\}.\end{aligned}$$

Fact 21.62 *(Lower truncation) In Fact 21.61, let $b \to \infty$, so we only have the truncation $a < X$. Then, we have*

$$\begin{aligned}\mathrm{E}(X|a < X) &= \mu + \sigma\frac{\phi(a_0)}{1-\Phi(a_0)} \text{ and} \\ \mathrm{Var}(X|a < X) &= \sigma^2\left\{1+\frac{a_0\phi(a_0)}{1-\Phi(a_0)}-\left[\frac{\phi(a_0)}{1-\Phi(a_0)}\right]^2\right\}.\end{aligned}$$

(The latter follows from $\lim_{b\to\infty} b_0\phi(b_0) = 0$.)

Example 21.63 *Suppose $X \sim N(0, \sigma^2)$ and we want to calculate $\mathrm{E}\,|x|$. This is the same as $\mathrm{E}(X|X > 0) = 2\sigma\phi(0)$.*

Fact 21.64 *(Upper truncation) In Fact 21.61, let $a \to -\infty$, so we only have the truncation $X \leq b$. Then, we have*

$$\begin{aligned}\mathrm{E}(X|X \leq b) &= \mu - \sigma\frac{\phi(b_0)}{\Phi(b_0)} \text{ and} \\ \mathrm{Var}(X|X \leq b) &= \sigma^2\left\{1-\frac{b_0\phi(b_0)}{\Phi(b_0)}-\left[\frac{\phi(b_0)}{\Phi(b_0)}\right]^2\right\}.\end{aligned}$$

(The latter follows from $\lim_{a\to-\infty} a_0\phi(a_0) = 0$.)

Fact 21.65 *(Delta method) Consider an estimator $\hat{\beta}_{k\times 1}$ which satisfies*

$$\sqrt{T}\left(\hat{\beta}-\beta_0\right) \xrightarrow{d} N(0, \Omega),$$

and suppose we want the asymptotic distribution of a transformation of β

$$\gamma_{q\times 1} = g(\beta),$$

where $g(.)$ is has continuous first derivatives. The result is

$$\begin{aligned}\sqrt{T}\left[g\left(\hat{\beta}\right)-g(\beta_0)\right] &\xrightarrow{d} N\left(0, \Psi_{q\times q}\right), \text{ where} \\ \Psi &= \frac{\partial g(\beta_0)}{\partial\beta'}\Omega\frac{\partial g(\beta_0)'}{\partial\beta}, \text{ where } \frac{\partial g(\beta_0)}{\partial\beta'} \text{ is } q\times k.\end{aligned}$$

Proof. By the mean value theorem we have

$$g\left(\hat{\beta}\right) = g(\beta_0) + \frac{\partial g(\beta^*)}{\partial\beta'}\left(\hat{\beta}-\beta_0\right),$$

where

$$\frac{\partial g(\beta)}{\partial\beta'} = \begin{bmatrix}\frac{\partial g_1(\beta)}{\partial\beta_1} & \cdots & \frac{\partial g_1(\beta)}{\partial\beta_k} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_q(\beta)}{\partial\beta_1} & \cdots & \frac{\partial g_q(\beta)}{\partial\beta_k}\end{bmatrix}_{q\times k},$$

and we evaluate it at β^* which is (weakly) between $\hat{\beta}$ and β_0. Premultiply by $\sqrt{T}$ and rearrange as

$$\sqrt{T}\left[g\left(\hat{\beta}\right)-g(\beta_0)\right] = \frac{\partial g(\beta^*)}{\partial\beta'}\sqrt{T}\left(\hat{\beta}-\beta_0\right).$$

If $\hat{\beta}$ is consistent ($\operatorname{plim}\hat{\beta} = \beta_0$) and $\partial g(\beta^*)/\partial\beta'$ is continuous, then by Slutsky's theorem $\operatorname{plim}\partial g(\beta^*)/\partial\beta' = \partial g(\beta_0)/\partial\beta'$, which is a constant. The result then follows from the continuous mapping theorem. ■

21.7.2 The Lognormal Distribution

Fact 21.66 *(Univariate lognormal distribution) If $x \sim N(\mu, \sigma^2)$ and $y = \exp(x)$ then the probability density function of y, $f(y)$ is*

$$f(y) = \frac{1}{y\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2}(\frac{\ln y-\mu}{\sigma})^2},\ y > 0.$$

The rth moment of y is $\mathrm{E}\,y^r = \exp(r\mu + r^2\sigma^2/2)$.

Example 21.67 *The first two moments are $\mathrm{E}\,y = \exp\left(\mu+\sigma^2/2\right)$ and $\mathrm{E}\,y^2 = \exp(2\mu +$*

$2\sigma^2)$. *We therefore get* $\mathrm{Var}(y) = \exp\left(2\mu + \sigma^2\right)\left[\exp\left(\sigma^2\right) - 1\right]$ *and* $\mathrm{Std}\,(y)\,/\,\mathrm{E}\,y = \sqrt{\exp(\sigma^2) - 1}$.

Fact 21.68 *(Moments of a truncated lognormal distribution) If* $x \sim N(\mu, \sigma^2)$ *and* $y = \exp(x)$ *then* $\mathrm{E}(y^r|y > a) = \mathrm{E}(y^r)\Phi(r\sigma - a_0)/\Phi(-a_0)$*, where* $a_0 = (\ln a - \mu)\,/\sigma$*. Note that the denominator is* $\Pr(y > a) = \Phi(-a_0)$*. In contrast,* $\mathrm{E}(y^r|y \leq b) = \mathrm{E}(y^r)\Phi(-r\sigma + b_0)/\Phi(b_0)$*, where* $b_0 = (\ln b - \mu)\,/\sigma$*. The denominator is* $\Pr(y \leq b) = \Phi(b_0)$*.*

Example 21.69 *The first two moments of the truncated (from below) lognormal distribution are* $\mathrm{E}(y|y > a) = \exp\left(\mu + \sigma^2/2\right)\Phi(\sigma - a_0)/\Phi(-a_0)$ *and* $\mathrm{E}(y^2|y > a) = \exp\left(2\mu + 2\sigma^2\right)\Phi(2\sigma - a_0)/\Phi(-a_0)$*.*

Example 21.70 *The first two moments of the truncated (from above) lognormal distribution are* $\mathrm{E}(y|y \leq b) = \exp\left(\mu + \sigma^2/2\right)\Phi(-\sigma + b_0)/\Phi(b_0)$ *and* $\mathrm{E}(y^2|y \leq b) = \exp\left(2\mu + 2\sigma^2\right)\Phi(-2\sigma + b_0)/\Phi(b_0)$*.*

Fact 21.71 *(Bivariate lognormal distribution). Let* x_1 *and* x_2 *have a bivariate normal distribution*

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}\right),$$

and consider $y_1 = \exp(x_1)$ *and* $y_2 = \exp(x_2)$*. From Fact 21.56 we know that the conditional distribution of* x_1 *given* y_2 *or* x_2 *is then normal*

$$x_1|(y_2 \text{ or } x_2) \sim N\left[\mu_1 + \frac{\rho\sigma_1}{\sigma_2}(x_2 - \mu_2), \sigma_1^2(1 - \rho^2)\right].$$

It follows that the conditional distribution of y_1 *given* y_2 *or* x_2 *is lognormal (with the parameters in the last equation).*

Fact 21.72 *(Moments of a bivariate lognormal distribution) In the case of Fact 21.71, we also get*

$$\begin{aligned}\mathrm{Cov}(y_1, y_2) &= [\exp(\rho\sigma_1\sigma_2) - 1]\exp\left[\mu_1 + \mu_2 + (\sigma_1^2 + \sigma_2^2)/2\right], \textit{ and} \\ \mathrm{Corr}(y_1, y_2) &= [\exp(\rho\sigma_1\sigma_2) - 1]\,/\sqrt{\left[\exp(\sigma_1^2) - 1\right]\left[\exp(\sigma_2^2) - 1\right]}.\end{aligned}$$

$\mathrm{Cov}(y_1, y_2$ *and* $\mathrm{Corr}(y_1, y_2)$ *have the same sign as* ρ *and are increasing in* ρ*. In addition,* $\mathrm{Corr}(y_1, y_2)$ *is closer to zero than* ρ*.*

21.7.3 The Chi-Square Distribution

Fact 21.73 *(The* χ_n^2 *distribution) If* $Y \sim \chi_n^2$*, then the pdf of* Y *is* $f(y) = \frac{1}{2^{n/2}\Gamma(n/2)} y^{n/2-1}e^{-y/2}$*, where* $\Gamma()$ *is the gamma function. The moment generating function is* $mgf_Y(t) = (1 - 2t)^{-n/2}$ *for* $t < 1/2$*. The first moments of* Y *are* $\mathrm{E}\,Y = n$ *and* $\mathrm{Var}(Y) = 2n$*.*

Fact 21.74 *(Quadratic forms of normally distribution random variables) If the* $n \times 1$ *vector* $X \sim N(0, \Sigma)$*, then* $Y = X'\Sigma^{-1}X \sim \chi_n^2$*. Therefore, if the* n *scalar random variables* X_i*,* $i = 1, ..., n$*, are uncorrelated and have the distributions* $N(0, \sigma_i^2)$*,* $i = 1, ..., n$*, then* $Y = \Sigma_{i=1}^n X_i^2/\sigma_i^2 \sim \chi_n^2$*.*

Fact 21.75 *(Distribution of* $X'AX$*) If the* $n \times 1$ *vector* $X \sim N(\mathbf{0}, I)$*, and* A *is a symmetric idempotent matrix (*$A = A'$ *and* $A = AA = A'A$*) of rank* r*, then* $Y = X'AX \sim \chi_r^2$*.*

Fact 21.76 *(Distribution of* $X'\Sigma^+X$*) If the* $n \times 1$ *vector* $X \sim N(0, \Sigma)$*, where* Σ *has rank* $r \leq n$ *then* $Y = X'\Sigma^+X \sim \chi_r^2$ *where* Σ^+ *is the pseudo inverse of* Σ*.*

Proof. Σ is symmetric, so it can be decomposed as $\Sigma = C\Lambda C'$ where C are the orthogonal eigenvector ($C'C = I$) and Λ is a diagonal matrix with the eigenvalues along the main diagonal. We therefore have $\Sigma = C\Lambda C' = C_1\Lambda_{11}C_1'$ where C_1 is an $n \times r$ matrix associated with the r non-zero eigenvalues (found in the $r \times r$ matrix Λ_{11}). The generalized inverse can be shown to be

$$\Sigma^+ = \begin{bmatrix} C_1 & C_2 \end{bmatrix}\begin{bmatrix} \Lambda_{11}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}\begin{bmatrix} C_1 & C_2 \end{bmatrix}' = C_1\Lambda_{11}^{-1}C_1',$$

We can write $\Sigma^+ = C_1\Lambda_{11}^{-1/2}\Lambda_{11}^{-1/2}C_1'$. Consider the $r \times 1$ vector $Z = \Lambda_{11}^{-1/2}C_1'X$, and note that it has the covariance matrix

$$\mathrm{E}\,ZZ' = \Lambda_{11}^{-1/2}C_1'\,\mathrm{E}\,XX'C_1\Lambda_{11}^{-1/2} = \Lambda_{11}^{-1/2}C_1'C_1\Lambda_{11}C_1'C_1\Lambda_{11}^{-1/2} = I_r,$$

since $C_1'C_1 = I_r$. This shows that $Z \sim N(\mathbf{0}_{r\times 1}, I_r)$, so $Z'Z = X'\Sigma^+X \sim \chi_r^2$. ∎

Fact 21.77 *(Convergence to a normal distribution) Let* $Y \sim \chi_n^2$ *and* $Z = (Y - n)/n^{1/2}$*. Then* $Z \overset{d}{\to} N(0, 2)$*.*

Example 21.78 *If* $Y = \Sigma_{i=1}^n X_i^2/\sigma_i^2$*, then this transformation means* $Z = (\Sigma_{i=1}^n X_i^2/\sigma_i^2 - 1)/n^{1/2}$*.*

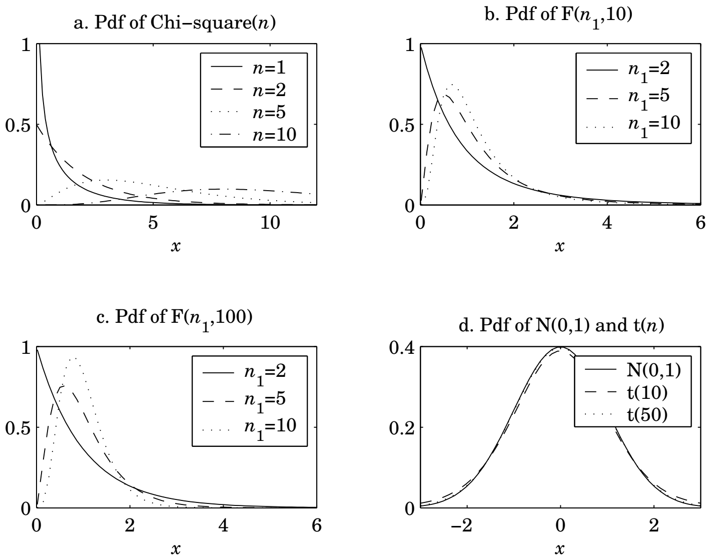


Figure 21.4: χ^2, F, and t distributions

Proof. We can directly note from the moments of a χ_n^2 variable that $\mathrm{E}\,Z = (\mathrm{E}\,Y - n)/n^{1/2} = 0$, and $\mathrm{Var}(Z) = \mathrm{Var}(Y)/n = 2$. From the general properties of moment generating functions, we note that the moment generating function of Z is

$$mgf_Z(t) = e^{-t\sqrt{n}}\left(1 - 2\frac{t}{n^{1/2}}\right)^{-n/2} \text{ with } \lim_{n\to\infty} mgf_Z(t) = \exp(t^2).$$

This is the moment generating function of a $N(0,2)$ distribution, which shows that $Z \overset{d}{\to} N(0,2)$. This result should not come as a surprise as we can think of Y as the sum of n variables; dividing by $n^{1/2}$ is then like creating a scaled sample average for which a central limit theorem applies. ■

21.7.4 The t and F Distributions

Fact 21.79 *(The $F(n_1, n_2)$ distribution) If $Y_1 \sim \chi^2_{n_1}$ and $Y_2 \sim \chi^2_{n_2}$ and Y_1 and Y_2 are independent, then $Z = (Y_1/n_1)/(Y_2/n_2)$ has an $F(n_1, n_2)$ distribution. This distribution has no moment generating function, but* $\mathrm{E}\,Z = n_2/(n_2 - 2)$ *for $n > 2$.*

Fact 21.80 *(Convergence of an $F(n_1, n_2)$ distribution) In Fact (21.79), the distribution of $n_1 Z = Y_1/(Y_2/n_2)$ converges to a $\chi^2_{n_1}$ distribution as $n_2 \to \infty$. (The idea is essentially that $n_2 \to \infty$ the denominator converges to the mean, which is* $\mathrm{E}\,Y_2/n_2 = 1$. *Only the numerator is then left, which is a $\chi^2_{n_1}$ variable.)*

Fact 21.81 *(The t_n distribution) If $X \sim N(0,1)$ and $Y \sim \chi^2_n$ and X and Y are independent, then $Z = X/(Y/n)^{1/2}$ has a t_n distribution. The moment generating function does not exist, but* $\mathrm{E}\,Z = 0$ *for $n > 1$ and Var$(Z) = n/(n-2)$ for $n > 2$.*

Fact 21.82 *(Convergence of a t_n distribution) The t distribution converges to a $N(0,1)$ distribution as $n \to \infty$.*

Fact 21.83 *(t_n versus $F(1,n)$ distribution) If $Z \sim t_n$, then $Z^2 \sim F(1,n)$.*

21.7.5 The Bernouilli and Binomial Distributions

Fact 21.84 *(Bernoulli distribution) The random variable X can only take two values: 1 or 0, with probability p and $1 - p$ respectively. The moment generating function is $mgf(t) = pe^t + 1 - p$. This gives* $\mathrm{E}(X) = p$ *and* $\mathrm{Var}(X) = p(1-p)$.

Example 21.85 *(Shifted Bernoulli distribution) Suppose the Bernoulli variable takes the values a or b (instead of 1 and 0) with probability p and $1-p$ respectively. Then* $\mathrm{E}(X) = pa + (1-p)b$ *and* $\mathrm{Var}(X) = p(1-p)(a-b)^2$.

Fact 21.86 *(Binomial distribution). Suppose $X_1, X_2, ..., X_n$ all have Bernoulli distributions with the parameter p. Then, the sum $Y = X_1 + X_2 + ... + X_n$ has a Binomial distribution with parameters p and n. The pdf is pdf$(Y) = n!/[y!(n-y)!]p^y(1-p)^{n-y}$ for $y = 0, 1, ..., n$. The moment generating function is $mgf(t) = [pe^t + 1 - p]^n$. This gives* $\mathrm{E}(Y) = np$ *and* $\mathrm{Var}(Y) = np(1-p)$.

Example 21.87 *(Shifted Binomial distribution) Suppose the Bernuolli variables $X_1, X_2, ..., X_n$ take the values a or b (instead of 1 and 0) with probability p and $1 - p$ respectively. Then, the sum $Y = X_1 + X_2 + ... + X_n$ has* $\mathrm{E}(Y) = n[pa + (1-p)b]$ *and* $\mathrm{Var}(Y) = n[p(1-p)(a-b)^2]$.

21.7.6 The Skew-Normal Distribution

Fact 21.88 *(Skew-normal distribution) Let ϕ and Φ be the standard normal pdf and cdf respectively. The pdf of a skew-normal distribution with shape parameter α is then*

$$f(z) = 2\phi(z)\Phi(\alpha z).$$

If Z has the above pdf and $Y = \mu + \omega Z$ with $\omega > 0$, then Y is said to have a $SN(\mu, \omega^2, \alpha)$ distribution (see Azzalini (2005)). Clearly, the pdf of Y is

$$f(y) = 2\phi\left[(y-\mu)/\omega\right]\Phi\left[\alpha\,(y-\mu)/\omega\right]/\omega.$$

The moment generating function is $mgf_y(t) = 2\exp\left(\mu t + \omega^2 t^2/2\right)\Phi(\delta\omega t)$ where $\delta = \alpha/\sqrt{1+\alpha^2}$. When $\alpha > 0$ then the distribution is positively skewed (and vice versa)—and when $\alpha = 0$ the distribution becomes a normal distribution. When $\alpha \to \infty$, then the density function is zero for $Y \leq \mu$, and $2\phi\left[(y-\mu)/\omega\right]/\omega$ otherwise—this is a half-normal distribution.

Example 21.89 *The first three moments are as follows. First, notice that* $\mathrm{E}\,Z = \sqrt{2/\pi}\delta$, $\mathrm{Var}(Z) = 1 - 2\delta^2/\pi$ *and* $\mathrm{E}(Z - \mathrm{E}\,Z)^3 = (4/\pi - 1)\sqrt{2/\pi}\delta^3$. *Then we have*

$$\begin{aligned}\mathrm{E}\,Y &= \mu + \omega\,\mathrm{E}\,Z\\ \mathrm{Var}(Y) &= \omega^2\,\mathrm{Var}(Z)\\ \mathrm{E}\,(Y - \mathrm{E}\,Y)^3 &= \omega^3\,\mathrm{E}(Z - \mathrm{E}\,Z)^3.\end{aligned}$$

Notice that with $\alpha = 0$ (so $\delta = 0$), then these moments of Y become μ, ω^2 and 0 respecively.

21.8 Inference

Fact 21.90 *(Comparing variance-covariance matrices) Let* $\mathrm{Var}(\hat\beta)$ *and* $\mathrm{Var}(\beta^*)$ *be the variance-covariance matrices of two estimators, $\hat\beta$ and β^*, and suppose* $\mathrm{Var}(\hat\beta) - \mathrm{Var}(\beta^*)$ *is a positive semi-definite matrix. This means that for any non-zero vector R that* $R'\,\mathrm{Var}(\hat\beta)R \geq R'\,\mathrm{Var}(\beta^*)R$*, so every linear combination of $\hat\beta$ has a variance that is as large as the variance of the same linear combination of β^*. In particular, this means that the variance of every element in $\hat\beta$ (the diagonal elements of* $\mathrm{Var}(\hat\beta)$*) is at least as large as variance of the corresponding element of β^*.*

Bibliography

Azzalini, A., 2005, “The skew-normal distribution and related Multivariate Families,” *Scandinavian Journal of Statistics*, 32, 159–188.

Davidson, J., 2000, *Econometric Theory*, Blackwell Publishers, Oxford.

DeGroot, M. H., 1986, *Probability and Statistics*, Addison-Wesley, Reading, Massachusetts.

Greene, W. H., 2000, *Econometric Analysis*, Prentice-Hall, Upper Saddle River, New Jersey, 4th edn.

Johnson, N. L., S. Kotz, and N. Balakrishnan, 1994, *Continuous Univariate Distributions*, Wiley, New York, 2nd edn.

Mittelhammer, R. C., 1996, *Mathematical Statistics for Economics and Business*, Springer-Verlag, New York.

22 Some Facts about Matrices

Some references: Greene (2000), Golub and van Loan (1989), Björk (1996), Anton (1987), Greenberg (1988).

22.1 Rank

Fact 22.1 *(Submatrix) Any matrix obtained from the $m \times n$ matrix A by deleting at most $m-1$ rows and at most $n-1$ columns is a submatrix of A.*

Fact 22.2 *(Rank) The rank of the $m \times n$ matrix A is ρ if the largest submatrix with non-zero determinant is $\rho \times \rho$. The number of linearly independent row vectors (and column vectors) of A is then ρ.*

22.2 Vector Norms

Fact 22.3 *(Vector p-norm) Let x be an $n \times 1$ matrix. The p-norm is defined as/*

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}.$$

The Euclidian norm corresponds to $p = 2$

$$\|x\|_2 = \left(\sum_{i=1}^{n} x_i^2 \right)^{1/2} = \sqrt{x'x}.$$

22.3 Systems of Linear Equations and Matrix Inverses

Fact 22.4 *(Linear systems of equations) Consider the linear system $Ax = c$ where A is $m \times n$, x is $n \times 1$, and c is $m \times 1$. A solution is a vector x such that $Ax = c$. It has a unique solution if and only if $\operatorname{rank}(A) = \operatorname{rank}([\ A \quad c\ ]) = n$; an infinite number of solutions if and only if $\operatorname{rank}(A) = \operatorname{rank}([\ A \quad c\ ]) < n$; and no solution if and only if $\operatorname{rank}(A) \neq \operatorname{rank}([\ A \quad c\ ])$.*

Example 22.5 *(Linear systems of equations, unique solution when $m = n$) Let x be 2×1, and consider the linear system*

$$Ax = c \text{ with } A = \begin{bmatrix} 1 & 5 \\ 2 & 6 \end{bmatrix} \text{ and } c = \begin{bmatrix} 3 \\ 6 \end{bmatrix}.$$

Here $\operatorname{rank}(A) = 2$ and $\operatorname{rank}([\ A \quad c\ ]) = 2$. The unique solution is $x = [\ 3 \quad 0\ ]'$.

Example 22.6 *(Linear systems of equations, no solution when $m > n$) Let x be a scalar, and consider the linear system*

$$Ax = c \text{ with } A = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \text{ and } c = \begin{bmatrix} 3 \\ 7 \end{bmatrix}.$$

Here $\operatorname{rank}(A) = 1$ and $\operatorname{rank}([\ A \quad c\ ]) = 2$. There is then no solution.

Fact 22.7 *(Least squares) Suppose that no solution exists to $Ax = c$. The best approximate solution, in the sense of minimizing (the square root of) the sum of squared errors, $\left[(c - A\hat{x})'(c - A\hat{x})\right]^{1/2} = \|c - A\hat{x}\|_2$, is $\hat{x} = (A'A)^{-1} A'c$, provided the inverse exist. This is obviously the least squares solution. In the example with $c = [\ 3 \quad 7\ ]'$, it is*

$$\begin{aligned} \hat{x} &= \left(\begin{bmatrix} 1 \\ 2 \end{bmatrix}' \begin{bmatrix} 1 \\ 2 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 \\ 2 \end{bmatrix}' \begin{bmatrix} 3 \\ 7 \end{bmatrix} \\ &= \frac{17}{5} \text{ or } 3.4. \end{aligned}$$

This is illustrated in Figure 22.1. *(Translation to OLS notation: c is the vector of dependent variables for m observations, A is the matrix with explanatory variables with the t^{th} observation in row t, and x is the vector of parameters to estimate).*

Fact 22.8 *(Pseudo inverse or generalized inverse) Suppose that no solution exists to $Ax = c$, and that $A'A$ is not invertible. There are then several approximations, $\hat{x}$, which all minimize $\|c - A\hat{x}\|_2$. The one with the smallest $\|\hat{x}\|_2$ is given by $\hat{x} = A^+c$, where A^+ is the Moore-Penrose pseudo (generalized) inverse of A. See Fact 22.54.*

Example 22.9 *(Linear systems of equations, unique solution when $m > n$) Change c in Example 22.6 to $c = [\ 3 \quad 6\ ]'$. Then $\operatorname{rank}(A) = 1$ and $\operatorname{rank}([\ A \quad c\ ]) = 1$, and the unique solution is $x = 3$.*

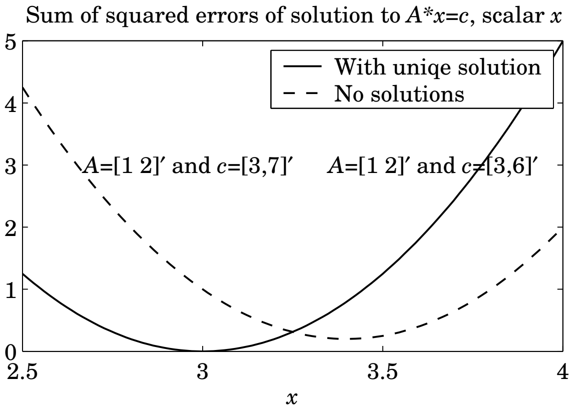


Figure 22.1: Value of qudratic loss function.

Example 22.10 *(Linear systems of equations, infinite number of solutions, $m < n$) Let x be 2×1, and consider the linear system*

$$Ax = c \text{ with } A = \begin{bmatrix} 1 & 2 \end{bmatrix} \text{ and } c = 5.$$

Here rank $(A) = 1$ *and rank*$([\ A \quad c\ ]) = 1$*. Any value of x_1 on the line $5 - 2x_2$ is a solution.*

Example 22.11 *(Pseudo inverses again) In the previous example, there is an infinite number of solutions along the line $x_1 = 5 - 2x_2$. Which one has the smallest norm $\|\hat{x}\|_2 = [(5 - 2x_2)^2 + x_2^2]^{1/2}$? The first order condition gives $x_2 = 2$, and therefore $x_1 = 1$. This is the same value as given by $\hat{x} = A^+ c$, since $A^+ = [0.2, 0.4]$ in this case.*

Fact 22.12 *(Rank and computers) Numerical calculations of the determinant are poor indicators of whether a matrix is singular or not. For instance,* $\det(0.1 \times I_{20}) = 10^{-20}$*. Use the condition number instead (see Fact 22.51).*

Fact 22.13 *(Some properties of inverses) If A, B, and C are invertible, then $(ABC)^{-1} = C^{-1}B^{-1}A^{-1}$; $(A^{-1})' = (A')^{-1}$; if A is symmetric, then A^{-1} is symmetric; $(A^n)^{-1} = (A^{-1})^n$.*

Fact 22.14 *(Changing sign of column and inverting) Suppose the square matrix A_2 is the same as A_1 except that the i^{th} and j^{th} columns have the reverse signs. Then A_2^{-1} is the same as A_1^{-1} except that the i^{th} and j^{th} rows have the reverse sign.*

22.4 Complex matrices

Fact 22.15 *(Modulus of complex number) If $\lambda = a + bi$, where $i = \sqrt{-1}$, then $|\lambda| = |a + bi| = \sqrt{a^2 + b^2}$.*

Fact 22.16 *(Complex matrices) Let A^H denote the transpose of the complex conjugate of A, so that if*

$$A = \begin{bmatrix} 1 & 2 + 3i \end{bmatrix} \text{ then } A^H = \begin{bmatrix} 1 \\ 2 - 3i \end{bmatrix}.$$

A square matrix A is unitary *(similar to orthogonal) if $A^H = A^{-1}$, for instance,*

$$A = \begin{bmatrix} \frac{1+i}{2} & \frac{1+i}{2} \\ \frac{1-i}{2} & \frac{-1+i}{2} \end{bmatrix} \text{ gives } A^H = A^{-1} = \begin{bmatrix} \frac{1-i}{2} & \frac{1+i}{2} \\ \frac{1-i}{2} & \frac{-1-i}{2} \end{bmatrix}.$$

and it Hermitian *(similar to symmetric) if $A = A^H$, for instance*

$$A = \begin{bmatrix} \frac{1}{2} & \frac{1+i}{2} \\ \frac{1-i}{2} & \frac{-1}{2} \end{bmatrix}.$$

A Hermitian matrix has real elements along the principal diagonal and A_{ji} is the complex conjugate of A_{ij}. Moreover, the quadratic form $x^H Ax$ is always a real number.

22.5 Eigenvalues and Eigenvectors

Fact 22.17 *(Homogeneous linear system). Consider the linear system in Fact 22.4 with $c = \mathbf{0}$: $A_{m\times n}x_{n\times 1} = \mathbf{0}_{m\times 1}$. Then* $\text{rank}(A) = \text{rank}([\ A \quad c\ ])$*, so it has a unique solution if and only if* $\text{rank}(A) = n$*; and an infinite number of solutions if and only if* $\text{rank}(A) < n$*. Note that $x = \mathbf{0}$ is always a solution, and it is the unique solution if* $\text{rank}(A) = n$*. We can thus only get a nontrivial solution (not all elements are zero), only if* rank $(A) < n$*.*

Fact 22.18 *(Eigenvalues) The n eigenvalues, λ_i, $i = 1, \ldots, n$, and associated eigenvec-*

tors, z_i, of the $n \times n$ matrix A satisfy

$$(A - \lambda_i I)\, z_i = \mathbf{0}_{n\times1}.$$

We require the eigenvectors to be non-trivial (not all elements are zero). From Fact 22.17, an eigenvalue must therefore satisfy

$$\det(A - \lambda_i I) = 0.$$

Fact 22.19 *(Right and left eigenvectors) A "right eigenvector" z (the most common) satisfies $Az = \lambda z$, and a "left eigenvector" v (seldom used) satisfies $v'A = \lambda v'$, that is, $A'v = \lambda v$.*

Fact 22.20 *(Rank and eigenvalues) For any $m \times n$ matrix A, $\operatorname{rank}(A) = \operatorname{rank}(A') = \operatorname{rank}(A'A) = \operatorname{rank}(AA')$ and equals the number of non-zero eigenvalues of $A'A$ or AA'.*

Example 22.21 *Let x be an $n \times 1$ vector, so $\operatorname{rank}(x) = 1$. We then have that the outer product, xx' also has rank 1.*

Fact 22.22 *(Determinant and eigenvalues) For any $n \times n$ matrix A, $\det(A) = \Pi_{i=1}^{n}\lambda_i$.*

22.6 Special Forms of Matrices

22.6.1 Triangular Matrices

Fact 22.23 *(Triangular matrix) A lower (upper) triangular matrix has zero elements above (below) the main diagonal.*

Fact 22.24 *(Eigenvalues of triangular matrix) For a triangular matrix A, the eigenvalues equal the diagonal elements of A. This follows from that*

$$det(A - \lambda I) = (A_{11} - \lambda)(A_{22} - \lambda)\ldots(A_{nn} - \lambda).$$

Fact 22.25 *(Squares of triangular matrices) If T is lower (upper) triangular, then TT is as well.*

22.6.2 Orthogonal Vector and Matrices

Fact 22.26 *(Orthogonal vector) The $n \times 1$ vectors x and y are orthogonal if $x'y = 0$.*

Fact 22.27 *(Orthogonal matrix) The $n\times n$ matrix A is orthogonal if $A'A = I$. Properties: If A is orthogonal, then $\det(A) = \pm1$; if A and B are orthogonal, then AB is orthogonal.*

Example 22.28 *(Rotation of vectors ("Givens rotations").) Consider the matrix $G = I_n$ except that $G_{ik} = c$, $G_{ik} = s$, $G_{ki} = -s$, and $G_{kk} = c$. If we let $c = \cos\theta$ and $s = \sin\theta$ for some angle θ, then $G'G = I$. To see this, consider the simple example where $i = 2$ and $k = 3$*

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & c & s \\ 0 & -s & c \end{bmatrix}' \begin{bmatrix} 1 & 0 & 0 \\ 0 & c & s \\ 0 & -s & c \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & c^2+s^2 & 0 \\ 0 & 0 & c^2+s^2 \end{bmatrix},$$

which is an identity matrix since $\cos^2\theta + \sin^2\theta = 1$. G is thus an orthogonal matrix. It is often used to "rotate" an $n \times 1$ vector ε as in $u = G'\varepsilon$, where we get

$$\begin{aligned} u_t &= \varepsilon_t \text{ for } t \neq i, k \\ u_i &= \varepsilon_i c - \varepsilon_k s \\ u_k &= \varepsilon_i s + \varepsilon_k c. \end{aligned}$$

The effect of this transformation is to rotate the i^{th} and k^{th} vectors counterclockwise through an angle of θ.

22.6.3 Positive Definite Matrices

Fact 22.29 *(Positive definite matrix) The $n \times n$ matrix A is positive definite if for any non-zero $n \times 1$ vector x, $x'Ax > 0$. (It is positive semidefinite if $x'Ax \geq 0$.)*

Fact 22.30 *(Some properties of positive definite matrices) If A is positive definite, then all eigenvalues are positive and real. (To see why, note that an eigenvalue satisfies $Ax = \lambda x$. Premultiply by x' to get $x'Ax = \lambda x'x$. Since both $x'Ax$ and $x'x$ are positive real numbers, λ must also be.)*

Fact 22.31 *(More properties of positive definite matrices) If B is a nonsingular $n \times n$ matrix and A is positive definite, then BAB' is also positive definite.*

Fact 22.32 *(More properties of positive definite matrices)* $\det(A) > 0$*; if A is pd, then A^{-1} is too; if $A_{m\times n}$ with $m \geq n$, then $A'A$ is pd.*

Fact 22.33 *(Cholesky decomposition) See Fact 22.41.*

22.6.4 Symmetric Matrices

Fact 22.34 *(Symmetric matrix) A is symmetric if $A = A'$.*

Fact 22.35 *(Properties of symmetric matrices) If A is symmetric, then all eigenvalues are real, and eigenvectors corresponding to distinct eigenvalues are orthogonal.*

Fact 22.36 *If A is symmetric, then A^{-1} is symmetric.*

22.6.5 Idempotent Matrices

Fact 22.37 *(Idempotent matrix) A is idempotent if $A = AA$. If A is also symmetric, then $A = A'A$.*

22.7 Matrix Decompositions

Fact 22.38 *(Diagonal decomposition) An $n \times n$ matrix A is diagonalizable if there exists a matrix C such that $C^{-1}AC = \Lambda$ is diagonal. We can thus write $A = C\Lambda C^{-1}$. The $n \times n$ matrix A is diagonalizable if and only if it has n linearly independent eigenvectors. We can then take C to be the matrix of the eigenvectors (in columns), and Λ the diagonal matrix with the corresponding eigenvalues along the diagonal.*

Fact 22.39 *(Spectral decomposition.) If the eigenvectors are linearly independent, then we can decompose A as*

$$A = Z\Lambda Z^{-1}, \text{ where } \Lambda = diag(\lambda_1, ..., \lambda_1) \text{ and } Z = \begin{bmatrix} z_1 & z_2 & \cdots & z_n \end{bmatrix},$$

where Λ is a diagonal matrix with the eigenvalues along the principal diagonal, and Z is a matrix with the corresponding eigenvalues in the columns.

Fact 22.40 *(Diagonal decomposition of symmetric matrices) If A is symmetric (and possibly singular) then the eigenvectors are orthogonal, $C'C = I$, so $C^{-1} = C'$. In this case, we can diagonalize A as $C'AC = \Lambda$, or $A = C\Lambda C'$. If A is $n \times n$ but has rank $r \leq n$, then we can write*

$$A = \begin{bmatrix} C_1 & C_2 \end{bmatrix} \begin{bmatrix} \Lambda_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} C_1 & C_2 \end{bmatrix}' = C_1\Lambda_1 C_1',$$

where the $n \times r$ matrix C_1 contains the r eigenvectors associated with the r non-zero eigenvalues in the $r \times r$ matrix Λ_1.

Fact 22.41 *(Cholesky decomposition) Let Ω be an $n \times n$ symmetric positive definite matrix. The Cholesky decomposition gives the unique lower triangular P such that $\Omega = PP'$ (some software returns an upper triangular matrix, that is, Q in $\Omega = Q'Q$ instead). Note that each column of P is only identified up to a sign transformation; they can be reversed at will.*

Fact 22.42 *(Triangular Decomposition) Let Ω be an $n \times n$ symmetric positive definite matrix. There is a unique decomposition $\Omega = ADA'$, where A is lower triangular with ones along the principal diagonal, and D is diagonal with positive diagonal elements. This decomposition is usually not included in econometric software, but it can easily be calculated from the commonly available Cholesky decomposition since P in the Cholesky decomposition is of the form*

$$P = \begin{bmatrix} \sqrt{D_{11}} & 0 & \cdots & 0 \\ \sqrt{D_{11}}A_{21} & \sqrt{D_{22}} & & 0 \\ \vdots & & \ddots & \vdots \\ \sqrt{D_{11}}A_{n1} & \sqrt{D_{22}}A_{n2} & \cdots & \sqrt{D_{nn}} \end{bmatrix}.$$

Fact 22.43 *(Schur decomposition) The decomposition of the $n \times n$ matrix A gives the $n \times n$ matrices T and Z such that*

$$A = ZTZ^H$$

where Z is a unitary $n \times n$ matrix and T is an $n \times n$ upper triangular Schur form with the eigenvalues along the diagonal. Note that premultiplying by $Z^{-1} = Z^H$ and postmulti-

plying by Z gives

$$T = Z^H AZ,$$

which is upper triangular. The ordering of the eigenvalues in T can be reshuffled, although this requires that Z is reshuffled conformably to keep $A = ZTZ^H$, which involves a bit of tricky "book keeping."

Fact 22.44 *(Generalized Schur Decomposition) The decomposition of the $n \times n$ matrices G and D gives the $n \times n$ matrices Q, S, T, and Z such that Q and Z are unitary and S and T upper triangular. They satisfy*

$$G = QSZ^H \text{ and } D = QTZ^H.$$

The generalized Schur decomposition solves the generalized eigenvalue problem $Dx = \lambda Gx$, where λ are the generalized eigenvalues (which will equal the diagonal elements in T divided by the corresponding diagonal element in S). Note that we can write

$$Q^H GZ = S \text{ and } Q^H DZ = T.$$

Example 22.45 *If $G = I$ in the generalized eigenvalue problem $Dx = \lambda Gx$, then we are back to the standard eigenvalue problem. Clearly, we can pick $S = I$ and $Q = Z$ in this case, so $G = I$ and $D = ZTZ^H$, as in the standard Schur decomposition.*

Fact 22.46 *(QR decomposition) Let A be $m \times n$ with $m \geq n$. The QR decomposition is*

$$\begin{aligned} A_{m\times n} &= Q_{m\times m} R_{m\times n} \\ &= \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \begin{bmatrix} R_1 \\ \mathbf{0} \end{bmatrix} \\ &= Q_1 R_1. \end{aligned}$$

where Q is orthogonal ($Q'Q = I$) and R upper triangular. The last line is the "thin QR decomposition," where Q_1 is an $m \times n$ orthogonal matrix and R_1 an $n \times n$ upper triangular matrix.

Fact 22.47 *(Inverting by using the QR decomposition) Solving $Ax = c$ by inversion of A can be very numerically inaccurate (no kidding, this is a real problem). Instead, the problem can be solved with QR decomposition. First, calculate Q_1 and R_1 such that $A = Q_1 R_1$. Note that we can write the system of equations as*

$$Q_1 Rx = c.$$

Premultply by Q_1' to get (since $Q_1' Q_1 = I$)

$$Rx = Q_1' c.$$

This is an upper triangular system which can be solved very easily (first solve the first equation, then use the solution is the second, and so forth.)

Fact 22.48 *(Singular value decomposition) Let A be an $m \times n$ matrix of rank ρ. The singular value decomposition is*

$$A = U_{m\times m} S_{m\times n} V'_{n\times n}$$

where U and V are orthogonal and S is diagonal with the first ρ elements being non-zero, that is,

$$S = \begin{bmatrix} S_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \text{ where } S_1 = \begin{bmatrix} s_{11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & s_{\rho\rho} \end{bmatrix}.$$

Fact 22.49 *(Singular values and eigenvalues) The singular values of A are the nonnegative square roots of AA^H if $m \leq n$ and of $A^H A$ if $m \geq n$.*

Remark 22.50 *If the square matrix A is symmetric and idempotent ($A = A'A$), then the singular values are the same as the eigevalues. From Fact (22.40) we know that a symmetric A can be decomposed as $A = C\Lambda C'$. It follows that this is the same as the singular value decomposition.*

Fact 22.51 *(Condition number) The condition number of a matrix is the ratio of the largest (in magnitude) of the singular values to the smallest*

$$c = |s_{ii}|_{\max} / |s_{ii}|_{\min}.$$

For a square matrix, we can calculate the condition value from the eigenvalues of AA^H

or $A^H A$ (see Fact 22.49). In particular, for a square matrix we have

$$c = \left|\sqrt{\lambda_i}\right|_{\max} / \left|\sqrt{\lambda_i}\right|_{\min},$$

where λ_i are the eigenvalues of AA^H and A is square.

Fact 22.52 *(Condition number and computers) The determinant is not a good indicator of the realibility of numerical inversion algorithms. Instead, let c be the condition number of a square matrix. If $1/c$ is close to the a computer's floating-point precision (10^{-13} or so), then numerical routines for a matrix inverse become unreliable. For instance, while* $\det(0.1 \times I_{20}) = 10^{-20}$, *the condition number of* $0.1 \times I_{20}$ *is unity and the matrix is indeed easy to invert to get* $10 \times I_{20}$.

Fact 22.53 *(Inverting by using the SVD decomposition) The inverse of the square matrix A is found by noting that if A is square, then from Fact 22.48 we have*

$$\begin{aligned} AA^{-1} &= I \text{ or} \\ USV'A^{-1} &= I, \text{ so} \\ A^{-1} &= VS^{-1}U', \end{aligned}$$

provided S is invertible (otherwise A will not be). Since S is diagonal, S^{-1} is also diagonal with the inverses of the diagonal elements in S, so it is very easy to compute.

Fact 22.54 *(Pseudo inverse or generalized inverse) The Moore-Penrose pseudo (generalized) inverse of an $m \times n$ matrix A is defined as*

$$A^+ = VS^+U', \text{ where } S^+_{nxm} = \begin{bmatrix} S_{11}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix},$$

where V and U are from Fact 22.48. The submatrix S_{11}^{-1} contains the reciprocals of the non-zero singular values along the principal diagonal. A^+ satisfies the A^+ satisfies the Moore-Penrose conditions

$$AA^+A = A,\ A^+AA^+ = A^+,\ \left(AA^+\right)' = AA^+, \text{ and } \left(A^+A\right)' = A^+A.$$

See Fact 22.8 for the idea behind the generalized inverse.

Fact 22.55 *(Some properties of generalized inverses) If A has full rank, then $A^+ = A^{-1}$; $(BC)^+ = C^+B^+$; if B, and C are invertible, then $(BAC)^{-1} = C^{-1}A^+B^{-1}$; $(A^+)' = (A')^+$; if A is symmetric, then A^+ is symmetric.*

Fact 22.56 *(Pseudo inverse of symmetric matrix) If A is symmetric, then the SVD is identical to the spectral decomposition $A = Z\Lambda Z'$ where Z are the orthogonal eigenvector ($Z'Z = I$) and Λ is a diagonal matrix with the eigenvalues along the main diagonal. By Fact 22.54) we then have $A^+ = Z\Lambda^+Z'$, where*

$$\Lambda^+ = \begin{bmatrix} \Lambda_{11}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix},$$

with the reciprocals of the non-zero eigen values along the principal diagonal of Λ_{11}^{-1}.

22.8 Matrix Calculus

Fact 22.57 *(Matrix differentiation of non-linear functions, $\partial y/\partial x'$) Let the vector $y_{n\times 1}$ be a function of the vector $x_{m\times 1}$*

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = f(x) = \begin{bmatrix} f_1(x) \\ \vdots \\ f_n(x) \end{bmatrix}.$$

Then, let $\partial y/\partial x'$ be the $n \times m$ matrix

$$\frac{\partial y}{\partial x'} = \begin{bmatrix} \frac{\partial f_1(x)}{\partial x'} \\ \vdots \\ \frac{\partial f_n(x)}{\partial x'} \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1(x)}{\partial x_1} & \cdots & \frac{\partial f_1(x)}{\partial x_m} \\ \vdots & & \vdots \\ \frac{\partial f_n(x)}{\partial x_1} & \cdots & \frac{\partial f_n(x)}{\partial x_m} \end{bmatrix}.$$

This matrix is often called the Jacobian of the f functions. (Note that the notation implies that the derivatives of the first element in y, denoted y_1, with respect to each of the elements in x' are found in the first row of $\partial y/\partial x'$. A rule to help memorizing the format of $\partial y/\partial x'$: y is a column vector and x' is a row vector.)

Fact 22.58 *($\partial y'/\partial x$ instead of $\partial y/\partial x'$) With the notation in the previous Fact, we get*

$$\frac{\partial y'}{\partial x} = \begin{bmatrix} \frac{\partial f_1(x)}{\partial x} & \cdots & \frac{\partial f_n(x)}{\partial x} \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1(x)}{\partial x_1} & \cdots & \frac{\partial f_n(x)}{\partial x_1} \\ \vdots & & \vdots \\ \frac{\partial f_1(x)}{\partial x_m} & \cdots & \frac{\partial f_n(x)}{\partial x_m} \end{bmatrix} = \left(\frac{\partial y}{\partial x'}\right)'.$$

Fact 22.59 *(Matrix differentiation of linear systems) When $y_{n\times1} = A_{n\times m}x_{m\times1}$, then $f(x)$ is a linear function*

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nm} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix}.$$

In this case $\partial y/\partial x' = A$ and $\partial y'/\partial x = A'$.

Fact 22.60 *(Matrix differentiation of inner product) The inner product of two column vectors, $y = z'x$, is a special case of a linear system with $A = z'$. In this case we get $\partial(z'x)/\partial x' = z'$ and $\partial(z'x)/\partial x = z$. Clearly, the derivatives of $x'z$ are the same (a transpose of a scalar).*

Example 22.61 *($\partial(z'x)/\partial x = z$ when x and z are 2×1 vectors)*

$$\frac{\partial}{\partial x}\left(\begin{bmatrix} z_1 & z_2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}.$$

Fact 22.62 *(First order Taylor series) For each element $f_i(x)$ in the $n\times$ vector $f(x)$, we can apply the mean-value theorem*

$$f_i(x) = f_i(c) + \frac{\partial f_i(b_i)}{\partial x'}(x-c),$$

for some vector b_i between c and x. Stacking these expressions gives

$$\begin{bmatrix} f_1(x) \\ \vdots \\ f_n(x) \end{bmatrix} = \begin{bmatrix} f_1(c) \\ \vdots \\ f_n(c) \end{bmatrix} + \begin{bmatrix} \frac{\partial f_1(b_1)}{\partial x_1} & \cdots & \frac{\partial f_1(b_1)}{\partial x_m} \\ \vdots & & \vdots \\ \frac{\partial f_n(b_n)}{\partial x_1} & \cdots & \frac{\partial f_n(b_n)}{\partial x_m} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix} \text{ or}$$

$$f(x) = f(c) + \frac{\partial f(b)}{\partial x'}(x-c),$$

where the notation $f(b)$ is a bit sloppy. It should be interpreted as that we have to evaluate the derivatives at different points for the different elements in $f(x)$.

Fact 22.63 *(Matrix differentiation of quadratic forms) Let $x_{m\times1}$ be a vector, $A_{m\times m}$ a matrix, and $f(x)_{n\times1}$ a vector of functions. Then,*

$$\begin{aligned} \frac{\partial f(x)' A f(x)}{\partial x} &= \left(\frac{\partial f(x)}{\partial x'}\right)' \left(A + A'\right) f(x) \\ &= 2\left(\frac{\partial f(x)}{\partial x'}\right)' A f(x) \text{ if } A \text{ is symmetric.} \end{aligned}$$

If $f(x) = x$, then $\partial f(x)/\partial x' = I$, so $\partial(x'Ax)/\partial x = 2Ax$ if A is symmetric.

Example 22.64 *($\partial(x'Ax)/\partial x = 2Ax$ when x is 2×1 and A is 2×2)*

$$\begin{aligned} \frac{\partial}{\partial x}\left(\begin{bmatrix} x_1 & x_2 \end{bmatrix}\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) &= \left(\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} + \begin{bmatrix} A_{11} & A_{21} \\ A_{12} & A_{22} \end{bmatrix}\right)\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \\ &= 2\begin{bmatrix} A_{11} & A_{12} \\ A_{12} & A_{22} \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \text{ if } A_{21} = A_{12}. \end{aligned}$$

Example 22.65 *(Least squares) Consider the linear model $Y_{m\times1} = X_{m\times n}\beta_{n\times1} + u_{m\times1}$. We want to minimize the sum of squared fitted errors by choosing the $n\times1$ vector β. The fitted errors depend on the chosen β: $u(\beta) = Y - X\beta$, so quadratic loss function is*

$$\begin{aligned} L &= u(\beta)'u(\beta) \\ &= (Y - X\beta)'(Y - X\beta). \end{aligned}$$

In thus case, $f(\beta) = u(\beta) = Y - X\beta$, so $\partial f(\beta)/\partial\beta' = -X$. The first order condition for $u'u$ is thus

$$-2X'\left(Y - X\hat{\beta}\right) = \mathbf{0}_{n\times1} \text{ or } X'Y = X'X\hat{\beta},$$

which can be solved as

$$\hat{\beta} = \left(X'X\right)^{-1}X'Y.$$

22.9 Miscellaneous

Fact 22.66 *(Some properties of transposes) $(A+B)' = A' + B'$; $(ABC)' = C'B'A'$ (if conformable).*

Fact 22.67 *(Kronecker product) If A and B are matrices, then*

$$A \otimes B = \begin{bmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{bmatrix}.$$

Properties: $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$ (if conformable); $(A \otimes B)(C \otimes D) = AC \otimes BD$ (if conformable); $(A \otimes B)' = A' \otimes B'$; if a is $m \times 1$ and b is $n \times 1$, then $a \otimes b = (a \otimes I_n)b$.

Fact 22.68 *(Cyclical permutation of trace) Trace(ABC) =Trace(BCA) =Trace(CAB), if the dimensions allow the products.*

Fact 22.69 *(The* vec *operator).* vec *A where A is $m \times n$ gives an $mn \times 1$ vector with the columns in A stacked on top of each other. For instance,* $\text{vec}\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \begin{bmatrix} a_{11} \\ a_{21} \\ a_{12} \\ a_{22} \end{bmatrix}$.

Properties: $\text{vec}\,(A + B) = \text{vec}\,A+ \text{vec}\,B$*;* $\text{vec}\,(ABC) = (C' \otimes A)\,\text{vec}\,B$*; if a and b are column vectors, then* $\text{vec}\,(ab') = b \otimes a$.

Fact 22.70 *(The vech operator) vech$A$ where A is $m \times m$ gives an $m(m+1)/2 \times 1$ vector with the elements on and below the principal diagonal A stacked on top of each other (columnwise). For instance, vech*$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \begin{bmatrix} a_{11} \\ a_{21} \\ a_{22} \end{bmatrix}$*, that is, like vec, but uses only the elements on and below the principal diagonal.*

Fact 22.71 *(Duplication matrix) The duplication matrix D_m is defined such that for any symmetric $m \times m$ matrix A we have* $\text{vec}\,A = D_m vech A$*. The duplication matrix is therefore useful for "inverting" the vech operator (the step from* vec *A to A is trivial). For instance, to continue the example of the vech operator*

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} a_{11} \\ a_{21} \\ a_{22} \end{bmatrix} = \begin{bmatrix} a_{11} \\ a_{21} \\ a_{21} \\ a_{22} \end{bmatrix} \text{ or } D_2 vech A = \text{vec}\,A.$$

Fact 22.72 *(Sums of outer products) Let y_t be $m \times 1$, and x_t be $k \times 1$. Suppose we have T such vectors. The sum of the outer product (an $m \times k$ matrix) is*

$$S = \sum_{t=1}^{T} y_t x_t'.$$

Create matrices $Y_{T\times m}$ and $X_{T\times k}$ by letting $y_t^{'}$ and $x_t^{'}$ be the t^{th} rows

$$Y_{T\times m} = \begin{bmatrix} y_1^{'} \\ \vdots \\ y_T^{'} \end{bmatrix} \text{ and } X_{T\times k} = \begin{bmatrix} x_1^{'} \\ \vdots \\ x_T^{'} \end{bmatrix}.$$

We can then calculate the same sum of outer product, S, as

$$S = Y'X.$$

(To see this, let $Y(i,:)$ be the ith row of Y, and similarly for X, so

$$Y'X = \sum_{t=1}^{T} Y(t,:)'X(t,:),$$

which is precisely $\Sigma_{t=1}^{T} y_t x_t'$.)

Fact 22.73 *(Matrix geometric series) Suppose the eigenvalues to the square matrix A are all less than one in modulus. Then,*

$$I + A + A^2 + \cdots = (1 - A)^{-1}.$$

To see why this makes sense, consider $(1 - A)\ \Sigma_{t=1}^{T} A^t$ (with the convention that $A^0 = I$). It can be written as

$$(1 - A)\ \Sigma_{t=1}^{T} A^t = \left(I + A + A^2 + \cdots\right) - A\left(I + A + A^2 + \cdots\right) = I - A^{T+1}.$$

If all the eigenvalues are stable, then $\lim_{T\to\infty} A^{T+1} = \mathbf{0}$, so taking the limit of the previous equation gives

$$(1 - A) \lim_{T\to\infty} \Sigma_{t=1}^{T} A = I.$$

Fact 22.74 *(Matrix exponential) The matrix exponential of an $n \times n$ matrix A is defined* *as*

$$\exp{(At)} = \sum_{s=0}^{\infty} \frac{(At)^s}{s!}.$$

Bibliography

Anton, H., 1987, *Elementary Linear Algebra*, John Wiley and Sons, New York, 5th edn.

Björk, Å., 1996, *Numerical Methods for Least Squares Problems*, SIAM, Philadelphia.

Golub, G. H., and C. F. van Loan, 1989, *Matrix Computations*, The John Hopkins University Press, Baltimore, 2nd edn.

Greenberg, M. D., 1988, *Advanced Engineering Mathematics*, Prentice Hall, Englewood Cliffs, New Jersey.

Greene, W. H., 2000, *Econometric Analysis*, Prentice-Hall, Upper Saddle River, New Jersey, 4th edn.