APPLIED STOCHASTIC METHODS SERIES

Mathematical Methods in Survival Analysis, Reliability and Quality of Life

Catherine Huber
Nikolaos Limnios
Mounir Mesbah and Mikhail Nikulin

iSTE

WILEY

Mathematical Methods in Survival Analysis, Reliability and Quality of Life

This page intentionally left blank

Mathematical Methods in Survival Analysis, Reliability and Quality of Life

Edited by
Catherine Huber
Nikolaos Limnios
Mounir Mesbah
Mikhail Nikulin

iSTE

WILEY

First published in Great Britain and the United States in 2008 by ISTE Ltd and John Wiley & Sons, Inc.

ISTE Ltd
6 Fitzroy Square
London W1T 5DX
UK
www.iste.co.uk

John Wiley & Sons, Inc.
111 River Street
Hoboken, NJ 07030
USA
www.wiley.com

Library of Congress Cataloging-in-Publication Data

Mathematical methods in survival analysis, reliability and quality of life / edited by Catherine Huber ... [et al.].
p. cm.
Includes bibliographical references and index.
ISBN: 978-1-84821-010-3
1. Failure time data analysis. 2. Survival analysis (Biometry) I. Huber, Catherine.
QA276.M342 2008
519.5'46--dc22

2007046232

British Library Cataloguing-in-Publication Data
A CIP record for this book is available from the British Library
ISBN: 978-1-84821-010-3

Printed and bound in Great Britain by Antony Rowe Ltd, Chippenham, Wiltshire.

FSC
Mixed Sources
Product group from well-managed
forests and other controlled sources

Cert no. SGS-COC-2953
www.fsc.org
© 1996 Forest Stewardship Council

Contents

Preface

The *European Seminar on Mathematical Methods for Survival Analysis, Reliability and Quality of Life* was created in 1997 by C. HUBER, N. LIMNIOS, M. NIKULIN and M. MESBAH, and, thanks to our respective laboratories and also to the supporting universities (see list below), it ran regularly during the last 10 years. 2007 is a special year as our European Seminar celebrates its 10$^{\text{th}}$ birthday. This seminar aims to give a review of recent research in the field of survival analysis, reliability, quality of life, and related topics, from both statistical and probabilistic points of view. Three or four sessions take place every year at the participating universities.

Besides these regular annual sessions, the European seminar supported many international conferences and workshops in France and abroad: for instance, in 2000, GOF2000 (Goodness Of Fit) in Paris and MMR2000 (Mathematical Methods in Reliability) in Bordeaux, in 2004, an international workshop on “Semi-parametric Models” organized at Mont Saint Michel and, more recently, Biostat2006 in Cyprus. More than 14 international workshops, 26 seminar sessions and about 150 talks were organized during the last ten years (see Appendix A).

Reliability and survival analysis are important applications of stochastic mathematics (probability, statistics and stochastic processes) that were usually treated separately in spite of the similarity of the involved mathematical theory. Reliability is oriented towards technical systems studies (without excluding the human factor), while survival analysis and quality of life are oriented towards biological and medical studies (without excluding the technical factor). The lifetime T of a technical system, of a human patient or of a bacteria is a non-negative random variable, and the same function of the time t, $Prob(T > t)$, is stated as the reliability function, denoted $R(t)$ in reliability theory, and as the survival function, denoted S(t), in medical applications. Nevertheless, even if the function to investigate is the same, the objectives are not always identical. In the field of reliability, most of the time, systems are ergodic and large, which is not the case in survival analysis. Thus, techniques developed in order to evaluate or to estimate the reliability/survival function are not always based on the same fundamental results. However they also include several common techniques: Cox models, degradation models, multi-state approaches (i.e., Markov and semi-Markov models), point processes, etc.

While it is recognized that quality of life is ultimately as important as quantity of life (survival time), efforts to implement quality of life measurements often fail. Statistical methods to analyze time of events are nowadays well established; opinions are largely agreed on between statisticians, clinicians and industrial professionals. Unfortunately, in the quality of life field, there is no standard instrument to measure it, no standard methodology to validate measurement instruments (questionnaires) and no standard statistical methodology to analyze obtained measurements. Specific development and application of modern psychometrical measurement models (latent variable models including the Rasch model) and connection with utility theory are important issues. A more recent issue is the joint analysis of the latent quality of life and external variables such as treatment, evolution (longitudinal analysis) and/or survival time.

The present book includes 21 chapters divided into four parts:

I. Survival analysis

II. Reliability

III. Quality of life

IV. Related topics

We would like to especially thank Julien Chiquet for his very efficient technical support on this book, and ISTE Ltd for their ongoing support.

Catherine HUBER,
Nikolaos LIMNIOS,
Mounir MESBAH,
Mikhail NIKULIN.

PART I

Survival Analysis

This page intentionally left blank

Chapter 1

Model Selection for Additive Regression in the Presence of Right-Censoring

1.1. Introduction

Statistical tools for handling regression problems when the response is censored have been developed in the last two decades. The response was often assumed to be a linear function of the covariates, but non-parametric regression models provide a very flexible method when a general relationship between covariates and response is first to be explored. In recent years, a vast literature has been devoted to non-parametric regression estimators for completely observed data. However, few methods exist under random censoring. First, Buckley and James [BUC 79] and Koul, Susarla and Van Ryzin [KOU 81] among others introduced the original idea of transforming the data to take the censoring into account, for linear regression curves. Then, Zheng [ZHE 88] proposed various classes of unbiased transformations. Dabrowska [DAB 87] and Zheng [ZHE 88] applied non-parametric methods for estimating the univariate regression curve. Later, Fan and Gijbels [FAN 94] considered a local linear approximation for the data transformed in the same way by using a variable bandwidth adaptive to the sparsity of the design points. Györfi *et al.* [GYÖ 02] also studied the consistency of generalized Stone's regression estimators in the censored case. Heuchenne and Van Keilegom [HEU 05] considered a nonlinear semi-parametric regression model with censored data. Park [PAR 04] extended a procedure suggested in Gross and Lai [GRO 96] to a general non-parametric model in the presence of left-truncation and right-censoring, by using B-spline developments.

Chapter written by Elodie BRUNEL and Fabienne COMTE.

Recently, Kohler *et al.* [KOH 03] proposed an adaptive mean-square estimator built with polynomial splines.

However, for modeling the relationship between a response and a multivariate regressor, new methodologies have to be found to solve the problem of practical implementation in higher dimension. The main objective of the article is to propose a multivariate method of model selection for an additive regression function of a low-dimensional covariate vector. In fact, the particular case of additive models seems to be more realistic in practice and may constitute a way to make the dimension of the covariate greater than 1. Suppose that $\vec{X}$ is a d-dimensional covariate in a compact set, without loss of generality we assume that $\vec{X}$ is a $[0,1]^d$-valued vector. Let $(\vec{X}_1, Y_1)$, $(\vec{X}_2, Y_2)$, ..., $(\vec{X}_n, Y_n)$ be independent identically distributed random variables. Let $T > 0$ be a fixed time for collecting the data. Therefore, the response variables *before censoring* are denoted by $Y_{i,T} = Y_i \wedge T$, where $a \wedge b$ denotes the infimum of a and b. Then, the model is defined for $i = 1, \ldots, n$, by:

$$\mathbf{E}(Y_{i,T}|\vec{X}_i) = r_T(\vec{X}_i) = r_{T,1}(X_i^{(1)}) + \cdots + r_{T,d}(X_i^{(d)}). \tag{1.1}$$

with $\vec{X}_i = (X_i^{(1)}, \ldots, X_i^{(d)})$. For identifiability we suppose that $\mathbf{E}(r_{T,j}(X_1^{(j)})) = 0$ for $j = 2, \ldots, d$. A comment on the model is required. The setting is analogous to Kohler *et al.* [KOH 03] and the fixed time T is due to the fact that functionals of the survival function under censoring cannot be estimated on the complete support, as mentioned by Gross and Lai [GRO 96]. Note that in the empirical setting, the procedure does work without any truncation. This fixed bound is introduced for technical and theoretical purposes. Of course, it is often mentioned that the function of interest would be r in the regression model $\mathbf{E}(Y|\vec{X}) = r(\vec{X})$, instead of its biased version r_T.

The method consists of building projection estimators of the d components $r_{T,1}$, ..., $r_{T,d}$ on different projection spaces. The strategy is based on a standard mean-square contrast as in Baraud [BAR 00] together with an optimized version of the data transformation proposed by Fan and Gijbels [FAN 94]. The algorithm is explained and performed through several empirical trials and real data, and in particular improved in the bivariate setting. The model and the assumptions are presented in section 1.2, the method is described in section 1.3 and the main theoretical result is given in section 1.4. Finally, practical implementation is to be found in section 1.5 together with several examples of various dimensions.

1.2. Assumptions on the model and the collection of approximation spaces

1.2.1. *Non-parametric regression model with censored data*

We consider the following censoring mechanism. Let $C_1, C_2, \ldots, C_n$ be n censoring times independent of $(\vec{X}_i, Y_i)$ and consequently independent of $(\vec{X}_i, Y_{i,T})$ as well. The $\vec{X}_i$s and the couples (Z_i, δ_i)s are observed where

$$Z_i = Y_{i,T} \wedge C_i, \quad \delta_i = \mathbb{1}_{\{Y_{i,T} \le C_i\}}$$

δ_i indicates if the observed time Z_i is a lifetime or a censoring time both occurring in the interval $[0, T]$.

Now, let $G(.)$ be the cumulative distribution function (cdf) of the C_is and F_Y be the marginal cdf of the Y_is with $\bar{F}_Y = 1 - F_Y$ and $\bar{G} = 1 - G$ being the corresponding survival functions. We suppose moreover that:

$(\mathcal{A})$ The distribution functions of the Y_is and C_is are $\mathbf{R}^+$-supported.

Under $(\mathcal{A})$, the following condition is immediately satisfied: $\mathbf{P}(Y_i \ge T) = \mathbf{P}(Y_{i,T} = T) > 0$. Moreover, we suppose that $\mathbf{P}(C_i > T) > 0$ which is satisfied for most well-known parametric survival models where the C_is are $\mathbf{R}^+$-supported. Assumption $(\mathcal{A})$ implies the existence of positive constants c_G and c_F such that: $1 - G(Y_{i,T}) \ge 1 - G(T) := c_G$, $i = 1, \ldots, n$, and $\forall t \in [0, T]$, $\quad 1 - F_Y(t) \ge 1 - F_Y(T) := c_F > 0$. Any condition ensuring these inequalities can be substituted with $(\mathcal{A})$.

1.2.2. *Description of the approximation spaces in the univariate case*

The projection spaces used in our theoretical results $(S_m)_{m \in \mathcal{M}_n}$ are described hereafter. For the sake of simplicity, we focus on the polynomial spaces. Note that it is possible to use trigonometric or wavelet spaces. Moreover, in practice the collection among which the algorithm makes its choice is much more complicated: the degrees on each bin are selected by the algorithm and the bins have not necessarily the same size. See Comte and Rozenholc [COM 04] for a description of the algorithm.

[P] *Regular piecewise polynomial spaces*: S_m is generated by $m(r+1)$ polynomials, $r+1$ polynomials of degree $0, 1, \ldots, r$ on each subinterval $[(j-1)/m, j/m]$, for $j = 1, \ldots m$, $D_m = (r+1)m$, $m \in \mathcal{M}_n = \{1, 2, \ldots, [n/(r+1)]\}$. Usual examples are the orthogonal collection in $\mathbf{L}^2([-1, 1])$ of the Legendre polynomials or the histogram basis. Dyadic collection of piecewise polynomials denoted by [DP] correspond to dyadic subdivisions with $m = 2^q$ and $D_m = (r+1)\,2^q$.

The spaces in collection [P] satisfy the following property:

$(\mathcal{H}_1)$ $(S_m)_{m \in \mathcal{M}_n}$ is a collection of finite-dimensional linear sub-spaces of $\mathbf{L}^2([0, 1])$, with dimension $\dim(S_m) = D_m$ such that $D_m \le n$, $\forall m \in \mathcal{M}_n$ and:

$$\exists \Phi_0 > 0, \forall m \in \mathcal{M}_n, \forall t \in S_m, \|t\|_\infty \le \Phi_0 \sqrt{D_m} \|t\|. \tag{1.2}$$

where $\|t\|^2 = \int_0^1 t^2(x)dx$, for t in $\mathbf{L}^2([0,1])$.

Moreover, for the results concerning the adaptive estimators, we need the following additional assumption:

$(\mathcal{H}_2)$ $(S_m)_{m\in\mathcal{M}_n}$ is a collection of nested models, and we denote by $\mathcal{S}_n$ the space belonging to the collection, such that $\forall m \in \mathcal{M}_n, S_m \subset \mathcal{S}_n$. We denote by N_n the dimension of $\mathcal{S}_n$: $\dim(\mathcal{S}_n) = N_n$ $(\forall m \in \mathcal{M}_n, D_m \leq N_n)$.

Assumption $(\mathcal{H}_1)$ is satisfied with $\Phi_0 = \sqrt{2r+1}$ for collection [P]. Moreover, [DP] satisfies $(\mathcal{H}_2)$.

1.2.3. *The particular multivariate setting of additive models*

In order to estimate the additive regression function, the approximation spaces can be described as

$$S_m = \left\{ t(x^{(1)}, \ldots, x^{(d)}) = a + \sum_{i=1}^{d} t_i(x^{(i)}),\ (a, t_1, \ldots, t_d) \in \mathbf{R} \times \Pi_{i=1}^{d} S_{m_i} \right\}$$

where S_{m_i} is chosen as a piecewise polynomial space with dimension D_{m_i}. As in the univariate case, this particular collection of multivariate spaces also satisfies $(\mathcal{H}_1)$ and $(\mathcal{H}_2)$ by taking $D_m = 1 + \sum_{i=1}^{d}(D_{m_i} - 1)$ in inequalities (1.2).

1.3. The estimation method

As usual in regression problems, a mean-square contrast can lead to an estimator of r_T. However, we need first to transform the data to take the censoring mechanism into account.

1.3.1. *Transformation of the data*

We consider the following transformation of the censored data

$$\varphi_\alpha(Z) = (1+\alpha)\int_0^Z \frac{dt}{1-G(t)} - \alpha\frac{\delta Z}{1-G(Z)}. \tag{1.3}$$

The main interest of the transformation is the following property: $\mathbf{E}(\varphi_\alpha(Z_1)|\vec{X}_1) = \mathbf{E}(Y_{1,T}|\vec{X}_1)$. Indeed

$$\begin{aligned}
\mathbf{E}\left[\frac{\delta_1 Z_1}{\bar{G}(Z_1)}|\vec{X}_1\right] &= \mathbf{E}\left[\left(\frac{\delta_1 Y_{1,T}}{\bar{G}(Y_{1,T})}|\vec{X}_1, \varepsilon_1\right) | \vec{X}_1\right] \\
&= \mathbf{E}\left[\mathbf{E}\left(\delta_1|\vec{X}_1, \varepsilon_1\right)\frac{Y_{1,T}}{\bar{G}(Y_{1,T})}|\vec{X}_1\right] = \mathbf{E}\left(Y_{1,T}|\vec{X}_1\right),
\end{aligned}$$

$$\begin{aligned}
\mathbf{E}\left[\int_0^{Z_1} \frac{dt}{1-G(t)} |\vec{X}_1\right] &= \mathbf{E}\left[\int_0^{+\infty} \frac{\mathbf{E}(\mathbb{I}_{Y_{1,T}\wedge C_1 \geq t}|\vec{X}_1, \varepsilon_1)}{1-G(t)} dt|\vec{X}_1\right] \\
&= \mathbf{E}\left[\int_0^{+\infty} \frac{\mathbb{I}_{Y_{1,T}\geq t}\mathbf{E}(\mathbb{I}_{C_1 \geq t}|\vec{X}_1, \varepsilon_1)}{1-G(t)} dt|\vec{X}_1\right] \\
&= \mathbf{E}\left[\int_0^{+\infty} \mathbb{I}_{Y_{1,T}\geq t} dt|\vec{X}_1\right] = \mathbf{E}(Y_{1,T}|\vec{X}_1).
\end{aligned}$$

The transformation φ_α was considered by Koul *et al.* [KOU 81] for $\alpha = -1$ and this case is often the only one studied in most of the theoretical results here. Leurgans [LEU 87] proposed the transformation corresponding to $\alpha = 0$ and the general form (1.3) is described in Fan and Gijbels [FAN 94]. The main problem then lies in the choice of the parameter α. We experimented with the proposition of Fan and Gijbels [FAN 94] for this choice, but we did not find it particularly satisfactory. Therefore, we performed a choice of α in order to minimize the variance $\mathrm{Var}(\varphi_\alpha(\mathrm{Z}))$ of the resulting transformed data and took the empirical version of

$$\hat{\alpha} = -\frac{\mathrm{cov}(\varphi_1(Z), \varphi_1(Z) - \varphi_2(Z))}{\mathrm{Var}(\varphi_1(Z) - \varphi_2(Z))}, \tag{1.4}$$

with

$$\varphi_1(Z) = \int_0^Z \frac{dt}{1-G(t)}, \quad \varphi_2(Z) = \varphi_1(Z) - \frac{\delta Z}{1-G(Z)}. \tag{1.5}$$

In all cases, the transformed data are unobservable since we need to define $\hat{G}$, a relevant estimator of G. We propose taking the Kaplan-Meier [KAP 58] product-limit estimator $\hat{\bar{G}}$, modified in the way suggested by Lo *et al.* [LO 89], and defined by

$$1 - \hat{G}(y) = \hat{\bar{G}}(y) = \prod_{Z_{(i)} \leq y} \left(\frac{n-i+1}{n-i+2}\right)^{1-\delta_{(i)}}. \tag{1.6}$$

Finally, by substituting G by its estimator $\hat{G}$, we obtain the empirical version of the transformed data:

$$\hat{\varphi}_{\hat{\alpha}}(Z) = (1+\hat{\alpha}) \int_0^Z \frac{dt}{1-\hat{G}(t)} - \hat{\alpha}\frac{\delta Z}{1-\hat{G}(Z)}. \tag{1.7}$$

1.3.2. *The mean-square contrast*

The mean-square strategy leads us to study the following contrast:

$$\gamma_n(t) = \frac{1}{n}\sum_{i=1}^{n}[\hat{\varphi}_{\hat{\alpha}}(Z_i) - t(\vec{X}_i)]^2. \tag{1.8}$$

In this context, it is useful to consider the empirical norm associated with the design

$$\|t\|_n^2 = \frac{1}{n}\sum_{i=1}^{n} t^2(\vec{X}_i).$$

Here we define

$$\hat{r}_m = \arg\min_{t\in S_m} \gamma_n(t). \tag{1.9}$$

The function $\hat{r}_m$ may not be easy to define but the vector $(\hat{r}_m(\vec{X}_1), \ldots, \hat{r}_m(\vec{X}_n))'$ is always well defined since it is the orthogonal projection in $\mathbf{R}^n$ of vector $(\hat{\varphi}_{\hat{\alpha}}(Z_1), \ldots,$ $\hat{\varphi}_{\hat{\alpha}}(Z_n))'$ onto the subspace of $\mathbf{R}^n$ defined by $\{(t(\vec{X}_1), \ldots, t(\vec{X}_n))',\ t \in S_m\}$. This explains why the empirical norms are particularly suitable for the mean-square contrast.

Next, model selection is performed by selecting the model $\hat{m}$ such that:

$$\hat{m} = \arg\min_{m\in\mathcal{M}_n} \left\{\gamma_n(\hat{r}_m) + \mathrm{pen}(m)\right\}, \tag{1.10}$$

where we have to determine the relevant form of pen(.) for $\hat{r}_{\hat{m}}$ to be an adaptive estimator of r.

1.4. Main result for the adaptive mean-square estimator

The automatic selection of the projection space can be performed via penalization and the following theoretical result is proved in Brunel and Comte [BRU 06], for the particular choice of $\hat{\alpha} = \alpha = -1$ i.e. for a contrast γ_n defined by (1.8) with variables $\hat{\varphi}_{-1}(Z_i) = \delta_i Z_i/(1 - \hat{G}(Z_i))$.

Theorem 1.1 *Assume that the common density f of the covariate vector $\vec{X}_i$ is such that $\forall x \in [0,1]^d, 0 < f_0 \leq f(x) < f_1 < +\infty$ and that the Y_is admit moments of order 8. Consider the collection of models [DP] with $N_n \leq n/(16 f_1 K_\varphi)$ for [DP] where K_φ is a (known) constant depending on the basis. Let $\hat{r}_m$ be the adaptive estimator defined by (1.8) with $\hat{\alpha} = -1$ and (1.10) with*

$$\mathrm{pen}(m) = \kappa \frac{\Phi_0^2}{f_0} \mathbf{E}\left[\left(\frac{\delta_1 Z_1}{\bar{G}(Z_1)}\right)^2\right] \frac{D_m}{n}, \tag{1.11}$$

where κ is a numerical constant. Then

$$\mathbf{E}(\|\hat{r}_{\hat{m}} - r_T\|_n^2) \leq C \inf_{m\in\mathcal{M}_n} \left(\|r_m - r_T\|^2 + \mathrm{pen}(m)\right) + C' \frac{\sqrt{\ln(n)}}{n}, \tag{1.12}$$

where r_m is the orthogonal projection of r_T onto S_m and C and C' are constants depending on Φ_0, $\|f\|$, c_G and $\mathbb{E}(Y_1^8)$.

The theoretical penalty pen(m) involves constants having different status. Let us recall that Φ_0 is known ($\Phi_0 = \sqrt{2r+1}$ where r is the degree of the piecewise polynomials). The unknown terms therein are f_0 and the expectation $\mathbf{E}\left[\left(\delta_1 Z_1/\bar{G}(Z_1)\right)^2\right]$ and they have to be replaced by estimators:

$$\widehat{\text{pen}}(m) = \kappa\Phi_0^2 \frac{\hat{\sigma}^2}{\hat{f}_0}\frac{D_m}{n}, \text{ where } \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\delta_i Z_i}{\hat{\bar{G}}(Z_i)}\right)^2 \tag{1.13}$$

where $\hat{f}_0$ is an estimator of the lower bound of f. Lastly, the constant κ is a numerical constant, independent of the data and for which a minimal suitable value exists. It has to be calibrated by simulation experiments: this has been done for regression problems in Comte and Rozenholc [COM 04]. It can be proved that the estimator obtained when substituting the random penalization (1.13) to the theoretical one (1.11) still satisfies inequality (1.12) under the assumptions of the theorem.

The left-hand side term of inequality (1.12) shows that an automatic and non-asymptotic trade-off is automatically performed between an unavoidable squared bias term $\|r_m - r_T\|^2$ and a term having a variance order pen(m). The non-asymptotic properties of the estimation algorithm can be appreciated when the selected model has small dimensions but allows a good adequation between the true function and the estimate.

The asymptotic rates can be deduced if a regularity assumption is set on the function to estimate r_T. It is worth emphasizing here that the rates recovered in the additive d-dimensional model correspond to the one-dimensional rates ($d = 1$), whatever the number d of covariates. Indeed, when r_T has regularity β, the resulting minimax rates should be $n^{-2\beta/(2\beta+d)}$ (see Brunel and Comte [BRU 06]), with the standard loss due to the high dimension of the problem (Stone [STO 82]). The additive model has therefore undeniable virtues from a theoretical point of view and has to be tested using some simulations.

1.5. Practical implementation

1.5.1. *The algorithm*

In practice, we consider $d = 1$, $d = 2$ or $d = 3$ covariates. We proceed by successive estimation steps. To summarize, here are the steps of our strategy for $d = 2$, for a model

$$Y_i = r_1(X_i^{(1)}) + r_2(X_i^{(2)}) + \varepsilon_i, \quad Z_i = Y_i \wedge C_i.$$

1) We obtain data: $Z_i, \delta_i, X_i^{(1)}, X_i^{(2)}$, for $i = 1, \dots n$.

2) We apply the transformation $\hat{\varphi}_{\hat{\alpha}}(.)$ defined by (1.7) for $\hat{\alpha}$ defined by (1.4), to the data (Z, δ) and build new variables for the regression $\hat{\varphi}_{\hat{\alpha}}(Z_i) := \hat{Y}_i$, for $i = 1, \ldots, n$, the δ_is being unchanged (see Figure 1.1 with both initial and transformed data where the effect of the preliminary transformation is in evidence).

3) The model must be subject to the constraint $\mathbf{E}(r_2(X_i^{(2)})) = 0$, otherwise for $b = \mathbf{E}(r_2(X_i^{(2)}))$, the estimated functions are $r_1^*(x) = r_1(x) + b$ and $r_2^*(x) = r_2(x) - b$. In practice, a good strategy may be to perform the regression algorithm of the centered data $\widehat{\hat{Y}_i} = Y_i - \bar{Y}_n$ where $\bar{Y}_n = (1/n)\sum_{i=1}^{n} \hat{Y}_i$. The output of the regression of the $\hat{Y}_i$s on the $X_i^{(1)}$ is a vector $\hat{r}_1^*(X_i^{(1)})$, $i = 1, \ldots, n$ of estimations on a space selected by contrast penalization.

4) The new variables for the regression are taken as the $\tilde{Y}_i = \hat{Y}_i - r_1^*(X_i^{(1)})$, and the output of the regression of the $\tilde{Y}_i$s on the $X_i^{(2)}$ is vector $\hat{r}_2^*(X_i^{(2)})$.

The mean-square estimation algorithm used here is the one originally implemented by Comte and Rozenholc ([COM 02], [COM 04]), which allows in addition variable degrees of the piecewise polynomials on each bin.

1.5.2. *Univariate examples*

We present in the following some univariate examples on which the procedure has first been tested.

Example 1. First we consider

$$Y_i = r(X_i) + \sigma\varepsilon_i, \quad Z_i = \inf(Y_i, C_i), \quad r(x) = 1.5 + \sin(2\pi x),$$

for iid $C_i \sim \mathcal{E}(c), \varepsilon_i \sim \mathcal{N}(0,1)$ and $X_i \sim \mathcal{U}([0,1])$. We tested different values of c, which imply different censoring rates. Figure 1.1 illustrates that whatever the proportion of censored data, the estimate is very good. The scatter plots are here to compare the original data to the transformed data by $\hat{\varphi}_{\hat{\alpha}}$ given in (1.7) with $\hat{\alpha}$ given by (1.4).

Example 2. Our second example studies another function r:

$$Y_i = r(X_i) + \sigma\varepsilon_i, \quad Z_i = \inf(Y_i, C_i), \quad r(x) = 1.5 + 5\exp(-5x),$$

for iid C_is, ε_is and U_is with $C_i \sim \mathcal{E}(c), \varepsilon_i \sim \mathcal{N}(0,1)$ and $X_i \sim \mathcal{U}([0,1])$. The results are illustrated in Figure 1.2, and the scatter plots allow here the comparison between the censored and uncensored data once the transformation $\hat{\varphi}_{\hat{\alpha}}$ has been performed.

Example 3. The third model is borrowed from Fan and Gijbels [FAN 94] and is described by

$$Y_i = r_1(X_i) + \sigma\varepsilon_i, \quad Z_i = \inf(Y_i, C_i), \quad r_1(x) = 4.5 - 64x^2(1-x)^2 - 16(x-0.5)^2$$

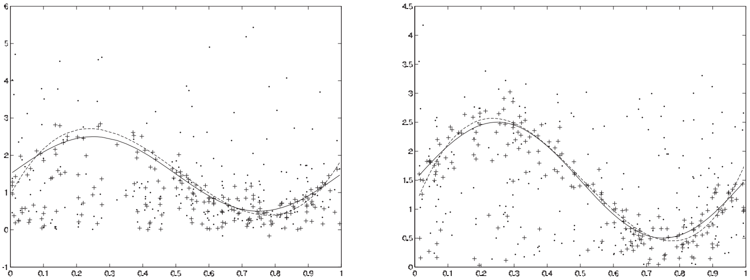


Figure 1.1. *Estimation of* r *in Example 1, true: full line, estimate: dotted line,* $. =$ *transformed data,* $+ =$ *observed data* $n = 200$*,* $\sigma = 0.25$*,* 51% *of censored data for* $c = 2$ *(left) and* 20% *of censored data for* $c = 6$ *(right)*

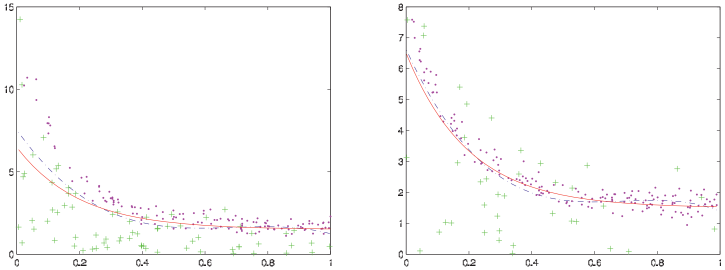


Figure 1.2. *Estimation of* r *in Example 2, true: full line, estimate: dotted line,* $. =$ *censored transformed data,* $+ =$ *uncensored tranformed data,* $n = 200$*,* $\sigma = 0.25$*,* 35% *of censored data for* $c = 4$ *(left) and* 20% *of censored data for* $c = 10$ *(right)*

and X_i iid $\mathcal{U}([0,1))$, ε_i iid $\mathcal{N}(0,1)$, $(C_i|X = c) \sim_{indep.} \mathcal{E}(c(x))$ with

$$c(x) = \begin{cases} 3(1.25 - |4x - 1|) \text{ if } 0 \leq x < 0.5 \\ 3(1.25 - |4x - 3|) \text{ if } 0.5 \leq x \leq 1. \end{cases}$$

It is worth mentioning that the method works very well for this model, even if the results proved in Brunel and Comte [BRU 06] do not encompass the framework of conditional independence between the Y_is and the C_is given the $\vec{X}_i$s. This is illustrated by Figure 1.3, for different sample sizes. It is not surprising that increasing the number of observations improves the estimation, but the results are satisfactory even for small samples ($n = 100$).

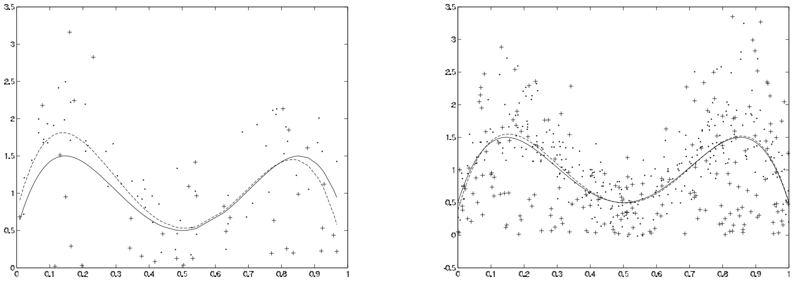


Figure 1.3. *Estimation of r in Example 3, true: full line, estimate: dotted line, $. =$ censored transformed data, $+ =$ uncensored tranformed data, $n = 100$ (left), $n = 500$ (right), $\sigma = 0.25$, about 41% of censored data*

Example 4. We also considered the classical "Stanford Heart Transplant Data", from October 1967 to February 1980, 184 patients admitted in a heart transplant program 157 with "complete tissue typing", 55 censored, originally studied by Miller and Halpern [MIL 82] and later studied by Fan and Gijbels [FAN 94] among others. Figure 1.4 illustrates our results with this data set. The estimated function seems to show that there is an optimal age for a heart transplant, which is about 35 years.

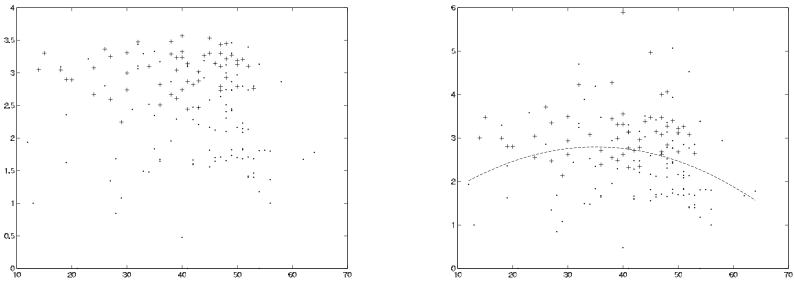


Figure 1.4. *Estimation for Stanford Heart Transplant Data, left: observed data, right: tranformed data and estimation of the regression function, $. =$ uncensored, $+ =$ censored data, $\hat{\alpha} = -0.0491$*

1.5.3. *Bivariate examples*

In this section, we consider some bivariate examples, in order to illustrate that the sequential procedure works in this setting.

Example 5. The first bivariate example is inspired by the function r_1 considered in Fan and Gijbels [FAN 94] associated with another one.

$$Y_i = r_1(X_i^{(1)}) + r_2(X_i^{(2)}) + \sigma\varepsilon_i, \varepsilon_i \sim \mathcal{N}(0,1),$$

$r_1(x) = 4.5 - 64x^2(1-x)^2 - 16(x-0.5)^2$, $r_2(x) = \exp(x/2)$, $X_i^{(1)}$ iid $\mathcal{U}([0,1))$ and $X_i^{(2)}$ iid $\mathcal{N}(0,1)$. The censoring variables are iid with exponential distribution. Note that we cannot estimate r_1 and r_2 but $r_1^* = r_1 + \mathbf{E}(r_2(X_1^{(2)}))$ and $r_2^* = r_2 - \mathbf{E}(r_2(X_1^{(2)}))$, for identifiability reasons. Note that $\mathbb{E}(r_2(X_1^{(2)})) = e^{1/8}$. We keep the model in this form for comparison with the univariate case. We correct the means to calculate the errors for the plots.

We use a "sequential" method that consists of estimating r_1^* with the regression on the $X_i^{(1)}$ and then use the residuals as new transformed data for the regression on the $X_i^{(2)}$.

We compare the results obtained for this model with the result of a univariate model $Y_i = r_1(X_i^{(1)}) + \sigma\varepsilon_i$, generated with the same function r_1 and the same σ as in Example 5. The censoring variables are taken as iid exponential distributions, with parameters adjusted to give the same proportion of censored variables for comparison. Tables 1.1 and 1.2 summarize the results of the mean squared errors (MSE) calculated for 100 simulated samples with five different sizes. It seems that estimating two functions instead of one does not greatly deteriorate the estimation of the first one and gives good results for the second one. A visualization of the orders of the MSE is given in Figure 1.5.

Example 6. Primary Biliary Cirrhosis (PBC) data. This data set is described in details in Fleming and Harrington ([FLE 91], p.2, Chapter 4) and is also studied by Fan and Gijbels [FAN 94]. The Mayo Clinic collected data on PBC, a rare but fatal chronic liver disease. From January 1974 to May 1984, 424 patients were registered, among which 312 participated in the random trial. The response variable is the logarithm of the time (in days) between registration and death, liver transplantation or time of the study analysis (July 1986). Among the 312 patients, 187 cases were censored. The covariates are first, the age, and second, the logarithm of bilirubin, which is known to be a prognostic factor. The estimated curves are given in Figure 1.6.

Censoring	**Case 1:** 60%	**Case 2:** 40%	**Case 3:** 20%
$n = 100$	0.1228 (0.0976)	0.0276 (0.0290)	0.0101 (0.0126)
$n = 200$	0.0437 (0.0593)	0.0113 (0.0141)	0.0040 (0.0039)
$n = 500$	0.0178 (0.0285)	0.0030 (0.0031)	0.0016 (0.0013)
$n = 1000$	0.0122 (0.0193)	0.0014 (0.0018)	7.3726e-004 (5.0079e-004)
$n = 2000$	0.0065 (0.0111)	7.3083e-004 (5.9819e-004)	4.5529e-004 (4.4699e-004)

Table 1.1. *MSE for the estimation of r_1 in a model of type "Example 3" with C_i following an $\mathcal{E}xp(c)$ distribution, $c = 1$ in column 1, $c = 2$ in column 2, $c = 4$ in column 3. Estimated $\hat{\alpha}$ range between 0.30 and 0.50. In parenthesis are the variances of MSE calculated over the 100 samples*

Censoring	**Case 1:** 60%		**Case 2:** 40%		**Case 3:** 20%	
function	r_1	r_2	r_1	r_2	r_1	r_2
$n = 100$	0.1777 (0.1265)	0.0893 (0.0863)	0.1221 (0.0554)	0.0821 (0.0572)	0.0582 (0.0411)	0.0754 (0.0318)
$n = 200$	0.1337 (0.0952)	0.0848 (0.0498)	0.0534 (0.0502)	0.0686 (0.0251)	0.0188 (0.0231)	0.0714 (0.0192)
$n = 500$	0.0463 (0.0476)	0.0783 (0.0351)	0.0160 (0.0229)	0.0701 (0.0185)	0.0061 (0.0087)	0.0703 (0.0130)
$n = 1000$	0.0159 (0.0216)	0.0727 (0.0204)	0.0055 (0.0064)	0.0701 (0.0117)	0.0021 (0.0017)	0.0703 (0.0089)
$n = 2000$	0.0086 (0.0141)	0.0729 (0.0153)	0.0021 (0.0021)	0.0696 (0.0084)	0.0010 (9.8779e-004)	0.0697 (0.0057)

Table 1.2. *MSE for the estimation of r_1 and r_2 in Example 5 with C_i following an $\mathcal{E}xp(c)$ distribution, $c = 2$ in column 1, $c = 4$ in column 2, $c = 8$ in column 3. Estimated $\hat{\alpha}$s range between 0.30 and 0.50. In parenthesis are the variances of MSE calculated over the 100 samples*

1.5.4. *A trivariate example*

We also tested a trivariate example. To recover all functions with a sequential estimation, the explicative variables must be ordered in the right way (decreasing variance orders), the censoring proportion must not be too great and the sample must not be too small. However, provided that all is reasonable, we still recover the regression function up to some constant. More precisely, Figure 1.7 shows the result of one estimation for $n = 500$, and respective variances of $r_i(X^i)$ are 3.1656, 1.1387 and 0.5571, for functions $r_1(x) = 5(4.5 - 64x^2(1-x)^2 - 16(x-0.5)^2)$, $r_2(x) = 4\exp(x/2)$, and $r_3(x) = 7\sin(\pi x/4)$, for $X_i^{(1)}$ iid $\mathcal{U}([0,1))$ and $X_i^{(2)}$ iid $\mathcal{N}([0,1))$ and $X_i^{(3)}$ iid

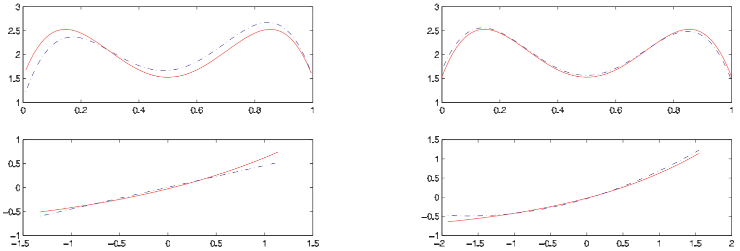


Figure 1.5. *Example of estimation (full line: true, dotted: estimate) of r_1 (top) and r_2 (bottom) of Example 5, for $n = 200$ (left) and n =1,000 (right), and about 42% of censored data, $\sigma = 0.25$. Mean squared errors are: MSE1=0.0345, MSE2=0.0025 for $n = 200$, MSE1=0.0021, MSE2*=$6.63.10^{-4}$ *for n =1,000*

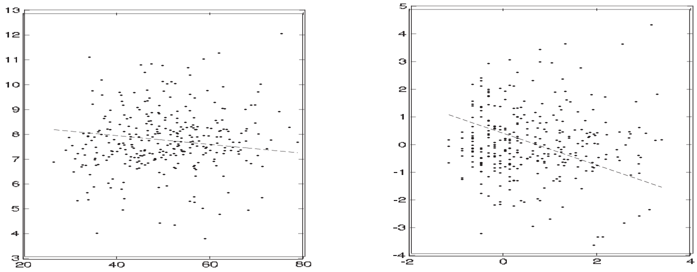


Figure 1.6. *Example of PBC data.* $\hat{\alpha} = -0.005$

$3[\mathcal{U}([0,1])]^{1/3}$. The model is thus

$$Y_i = r_1(X_i^{(1)}) + r_2(X_i^{(2)}) + r_3(X_i^{(3)}) + \sigma\varepsilon_i, \varepsilon_i \sim \mathcal{N}(0,1).$$

The functions estimated by the algorithm are $r_1^* = r_1 + a + b$, $r_2^* = r_2 - a$, $r_3^* = r_3 - b$ where $a = \mathbf{E}(r_2(X_1^{(2)}))$ and $b = \mathbf{E}(r_3(X_1^{(3)}))$.

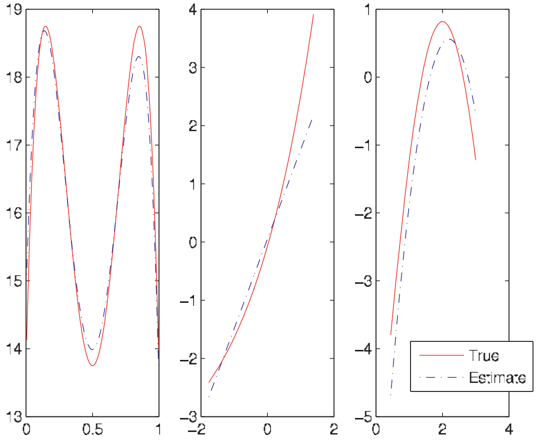


Figure 1.7. *Example of estimation of r_1, r_2 and r_3 in the trivariate example, for $n = 500$, and 29.8% of censored data, $\sigma = 0.2$*

1.6. Bibliography

[BAR 00] BARAUD Y., "Model selection for regression on a fixed design", *Probab. Theory Related Fields*, vol. 117, num. 4, p. 467–493, 2000.

[BRU 06] BRUNEL E., COMTE F., "Adaptive nonparametric regression estimation in presence of right-censoring", *Math. Methods Statist.*, vol. 15, num. 3, p. 233–255, 2006.

[BUC 79] BUCKLEY J., JAMES I., "Linear regression with censored data", *Biometrika*, vol. 66, num. 3, p. 429–464, 1979.

[COM 02] COMTE F., ROZENHOLC Y., "Adaptive estimation of mean and volatility functions in (auto-)regressive models", *Stochastic Process. Appl.*, vol. 97, num. 1, p. 111–145, 2002.

[COM 04] COMTE F., ROZENHOLC Y., "A new algorithm for fixed design regression and denoising", *Ann. Inst. Statist. Math.*, vol. 56, num. 3, p. 449–473, 2004.

[DAB 87] DABROWSKA D. M., "Nonparametric regression with censored survival time data", *Scand. J. Statist.*, vol. 14, num. 3, p. 181–197, 1987.

[FAN 94] FAN J., GIJBELS I., "Censored regression: local linear approximations and their applications", *J. Amer. Statist. Assoc.*, vol. 89, num. 426, p. 560–570, 1994.

[FLE 91] FLEMING T. R., HARRINGTON D. P., *Counting Processes and Survival Analysis*, Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics, John Wiley & Sons Inc., New York, 1991.

[GRO 96] GROSS S. T., LAI T. L., "Nonparametric estimation and regression analysis with left-truncated and right-censored data", *J. Amer. Statist. Assoc.*, vol. 91, num. 435, p. 1166–1180, 1996.

[GYÖ 02] GYÖRFI L., KOHLER M., KRZYŻAK A., WALK H., *A Distribution-free Theory of Nonparametric Regression*, Springer Series in Statistics, Springer-Verlag, New York, 2002.

[HEU 05] HEUCHENNE C., VAN KEILEGOM I., "Nonlinear regression with censored data", Discussion paper, 2005, 0512, Institut de Statistique, Université catholique de Louvain.

[KAP 58] KAPLAN E. L., MEIER P., "Nonparametric estimation from incomplete observations", *J. Amer. Statist. Assoc.*, vol. 53, p. 457–481, 1958.

[KOH 03] KOHLER M., KUL S., MÀTHÉ K., "Least squares estimates for censored regression", Preprint, http://www.mathematik.uni-stuttgart.de/mathA/lst3/kohler/hfm-pub-en.html, 2003.

[KOU 81] KOUL H., SUSARLA V., VAN RYZIN J., "Regression analysis with randomly right-censored data", *Ann. Statist.*, vol. 9, num. 6, p. 1276–1288, 1981.

[LEU 87] LEURGANS S., "Linear models, random censoring and synthetic data", *Biometrika*, vol. 74, num. 2, p. 301–309, 1987.

[LO 89] LO S. H., MACK Y. P., WANG J. L., "Density and hazard rate estimation for censored data via strong representation of the Kaplan-Meier estimator", *Probab. Theory Related Fields*, vol. 80, num. 3, p. 461–473, 1989.

[MIL 82] MILLER R., HALPERN J., "Regression with censored data", *Biometrika*, vol. 69, num. 3, p. 521–531, 1982.

[PAR 04] PARK J., "Optimal global rate of convergence in nonparametric regression with left-truncated and right-censored data", *J. Multivariate Anal.*, vol. 89, num. 1, p. 70–86, 2004.

[STO 82] STONE C. J., "Optimal global rates of convergence for nonparametric regression", *Ann. Statist.*, vol. 10, num. 4, p. 1040–1053, 1982.

[ZHE 88] ZHENG Z. K., "Strong consistency of nonparametric regression estimates with censored data", *J. Math. Res. Exposition*, vol. 8, num. 2, p. 307–313, 1988.

This page intentionally left blank

Chapter 2

Non-parametric Estimation of Conditional Probabilities, Means and Quantiles under Bias Sampling

2.1. Introduction

Models for estimating toxicity thresholds or performing diagnostic tests for the detection of diseases are still in development. Let $(X_i, Y_i)_{i \leq n}$ be a sample of the variable set (X, Y), where Y is the indicator that an event occurs and X is an explanatory variable. Conditionally on X, Y follows a Bernoulli distribution with parameter $p(x) = \mathbb{P}(Y = 1|X = x)$. The function p is supposed to be strictly monotone on the support I_X of X. A usual example is that of a response variable Y to a dose X or to an expository time X. The variable X may be observed at fixed values x_i, $i \in \{1, \ldots, m\}$ on a regular grid $\{1/m, \ldots, 1\}$ or at random values on a real interval corresponding to the values of a continuous process $(X_t)_{t \leq T}$ at fixed or random times $t_j, j \leq n$.

Exponential linear models with known link functions are often used, especially the logistic regression model defined by $p(x) = e^{\psi(x)}\{1 + e^{\psi(x)}\}^{-1}$ with a linear function $\psi(x) = \beta_0 + \beta_1 x$. For real x, the inverse function of p is easily estimated by $\widehat{q}_n(u) = \widehat{\beta}_{1n}^{-1}\{\log u - \log(1 - u) - \widehat{\beta}_{0n}\}$, with maximum likelihood estimators of β_0 and β_1. Other methods may be used for large dimensional parameter sets.

Chapter written by Odile PONS.

The observations may be biased due to X or Y and the estimators must be corrected. We present here biased designs for this model and for a continuous bivariate set (X, Y), and discuss the identifiability of the models, non-parametric estimation of the distribution functions and means, and efficiency of the estimators. Linear models with a multidimensional variable have been widely studied and the results given here may be extended to that case.

2.2. Non-parametric estimation of p

In a discrete sampling design with several independent observations for each value x_j of X, the likelihood is written

$$L_n = \prod_{i=1}^{n} p(X_i)^{Y_i}\{1 - p(X_i)\}^{1-Y_i} = \prod_{j=1}^{m}\prod_{i=1}^{n}[\{p(x_j)\}^{Y_i}\{1 - p(x_j)\}^{1-Y_i}]^{1_{\{X_i=x_j\}}}.$$

The maximum likelihood estimator of $p(x_j)$ is the proportion of individuals with $Y_i = 1$ as $X_i = x_j$,

$$\widehat{p}_{1n}(x_j) = \bar{Y}_{n;x_j} = \frac{1}{\sum_{i=1}^{n} 1_{\{X_i=x_j\}}} \sum_{i=1}^{n} Y_i 1_{\{X_i=x_j\}}, j = 1, \ldots, m.$$

Regular versions of this estimator are obtained by kernel smoothing projections on a regular basis of functions, especially if variable X is continuous. Let K denote a symmetric positive kernel with integral 1, $h = h_n$ a bandwidth and $K_h(x) = h^{-1}K(h^{-1}x)$, with $h_n \to 0$ as $n \to \infty$. A local maximum likelihood estimator of p is defined as

$$\widehat{p}_{2n}(x) = \frac{1}{\sum_{i=1}^{n} K_h(x - X_i)} \sum_{i=1}^{n} Y_i K_h(x - X_i)$$

or by higher order polynomial approximations [FAN 96], $\widehat{p}_{3n}(x) = \sum_{k=0}^{L} \widehat{\beta}_k(x)(X_i - x)^k$ where, for each value of x, the coefficients $\widehat{\beta}_k(x)$ are obtained by the minimization of $\sum_{i=1}^{n}\{Y_i - \sum_{k=0}^{L} \beta_k(x)(X_i - x)^k\}^2 K_h(X_i - x)$. The estimator $\widehat{p}_{2n}$ is P-uniformly consistent and asymptotically Gaussian if $h_n = O(n^{-1/5})$ and if p and the joint densities of X are twice continuously differentiable, under ergodicity assumptions for random sampling of processes $(X_t, Y_t)_{t \geq 0}$. Since p is assumed to be monotone, all estimators are asymptotically monotone in probability, so they may be inverted if n is sufficiently large. The inverse function q is then estimated by $\widehat{q}_n(u) = \sup\{x : \widehat{p}_n(x) \leq u\}$ if p is decreasing, or by $\widehat{q}_n(u) = \inf\{x : \widehat{p}_n(x) \geq u\}$ if p is increasing, The estimator $\widehat{q}_n$ is also P-uniformly consistent and asymptotically Gaussian [PIN 06]. For small samples, a monotone version of $\widehat{p}_n$ is required.

2.3. Bias depending on the value of Y

In case-control studies, individuals are not uniformly sampled in the population: for rare events, they are sampled so that the cases of interest (individuals with $Y_i = 1$) are sufficiently represented in the sample, but the proportion of cases in the sample differs from its proportion in the general population. Let S_i be the sampling indicator of individual i in the global population and

$$\mathbb{P}(S_i = 1|Y_i = 1) = \lambda_1,\ \mathbb{P}(S_i = 1|Y_i = 0) = \lambda_0.$$

The distribution function of (S_i, Y_i) conditionally on $X_i = x$ is given by

$$\begin{aligned}
\mathbb{P}(S_i = 1, Y_i = 1|x) &= \mathbb{P}(S_i = 1|Y_i = 1)\mathbb{P}(Y_i = 1|x) = \lambda_1 p(x),\\
\mathbb{P}(S_i = 1, Y_i = 0|x) &= \mathbb{P}(S_i = 1|Y_i = 0)\mathbb{P}(Y_i = 0|x) = \lambda_0\{1 - p(x)\},\\
\mathbb{P}(S_i = 1|x) &= \mathbb{P}(S_i = 1, Y_i = 1|x) + \mathbb{P}(S_i = 1, Y_i = 0|x)\\
&= \lambda_1 p(x) + \lambda_0\{1 - p(x)\}.
\end{aligned}$$

Let

$$\theta = \frac{\lambda_0}{\lambda_1},\ \alpha(x) = \theta\frac{1 - p(x)}{p(x)}.$$

For individual i, (X_i, Y_i) is observed conditionally on $S_i = 1$ and the conditional distribution function of Y_i is defined by

$$\begin{aligned}
\pi(x) &= \mathbb{P}(Y_i = 1|S_i = 1, X = x) = \frac{\lambda_1 p(x)}{\lambda_1 p(x) + \lambda_0\{1 - p(x)\}}\\
&= \frac{p(x)}{p(x) + \theta\{1 - p(x)\}} = \frac{1}{1 + \alpha(x)}.
\end{aligned}$$

The probability $p(x)$ is deduced from θ and $\pi(x)$ by the relation

$$p(x) = \frac{\theta\pi(x)}{1 + (\theta - 1)\pi(x)}$$

and the bias sampling is

$$\pi(x) - p(x) = \frac{(1 - \theta)\pi(x)(1 - \pi(x))}{1 + (\theta - 1)\pi(x)}.$$

The model defined by $(\lambda_0, \lambda_1, p(x))$ is over-parametrized and only the function α is identifiable. The proportion θ must therefore be known or estimated from a preliminary study before an estimation of the probability function p. In the logistic regression model, $\psi(x) = \log[p(x)\{1-p(x)\}^{-1}]$ is replaced by $\log\alpha(x) = \log[\pi(x)\{1-\pi(x)\}^{-1}] = \psi(x) - \log\theta$. Obviously, the bias sampling modifies not only the parameters of the model but also the model itself, except in the case of the logistic regression model

Let γ be the inverse of the proportion of cases in the population,

$$\gamma = \mathbb{P}(Y=0)/\mathbb{P}(Y=1) = E(1-Y)/EY = \frac{1-\int p(x)\,dF_X(x)}{\int p(x)\,dF_X(x)}. \tag{2.1}$$

Under the bias sampling,

$$\mathbb{P}(Y_i = 1|S_i = 1) = \frac{\lambda_1\int p\,dF_X}{\lambda_0(1-\int p\,dF_X) + \lambda_1\int p\,dF_X} = \frac{1}{1+\theta\gamma},$$

$$\mathbb{P}(Y_i = 0|S_i = 1) = \frac{\lambda_0(1-\int p\,dF_X)}{\lambda_0(1-\int p\,dF_X) + \lambda_1\int p\,dF_X} = \frac{\theta\gamma}{1+\theta\gamma},$$

γ is modified by the scale parameter η: it becomes $\mathbb{P}(Y=0|S=1)/\mathbb{P}(Y=1|S=1) = \theta\gamma$.

The product $\theta\gamma$ may be directly estimated from the observed Bernoulli variables Y_i by the maximization of the likelihood

$$\prod_{i=1}^{n}\{\mathbb{P}(Y_i=1|S_i=1)\}^{Y_i}\{\mathbb{P}(Y_i=0|S_i=1)\}^{1-Y_i},$$

hence,

$$\theta\widehat{\gamma}_n = 1 - \frac{\sum_i Y_i 1_{\{S_i=1\}}}{\sum_i 1_{\{S_i=1\}}}.$$

In a discrete sampling design with several independent observations for fixed values x_j of the variable X, the likelihood is

$$\prod_{i=1}^{n}\pi(X_i)^{Y_i}\{1-\pi(X_i)\}^{1-Y_i} = \prod_{j=1}^{m}\prod_{i=1}^{n}[\pi(X_i)^{Y_i}\{1-\pi(X_i)\}^{1-Y_i}]^{1_{\{X_i=x_j\}}}$$

and $\alpha_j = \alpha(x_j)$ is estimated by

$$\widehat{\alpha}_{jn} = \frac{\sum_i (1-Y_i)1_{\{S_i=1\}}1_{\{X_i=x_j\}}}{\sum_i Y_i 1_{\{X_i=x_j\}}1_{\{S_i=1\}}}.$$

For random observations of variable X, or for fixed observations without replications, $\alpha(x)$ is estimated by

$$\widehat{\alpha}_n(x) = \frac{\sum_i (1-Y_i)1_{\{S_i=1\}}K_h(x-X_i)}{\sum_i Y_i 1_{\{S_i=1\}}K_h(x-X_i)}.$$

If θ is known, non-parametric estimators of p are deduced as

$$\begin{aligned}
\widehat{p}_n(x_j) &= \frac{\theta \sum_i Y_i 1_{\{S_i=1\}} 1_{\{X_i=x_j\}}}{\sum_i (1 - Y_i + \theta Y_i) 1_{\{S_i=1\}} 1_{\{X_i=x_j\}}}, \text{ in the discret case,} \\
\widehat{p}_n(x) &= \frac{\theta \sum_i Y_i 1_{\{S_i=1\}} K_h(x - X_i)}{\sum_i (1 - Y_i + \theta Y_i) 1_{\{S_i=1\}} K_h(x - X_i)}, \text{ in the continuous case.}
\end{aligned}$$

2.4. Bias due to truncation on X

Consider that Y is observed under a fixed truncation of X: we assume that (X, Y) is observed only if $X \in [a, b]$, a sub-interval of the support I_X of the variable X, and $S = 1_{[a,b]}(X)$. Then

$$\mathbb{P}(Y_i = 1) = \int_{I_X} p(x)\, dF_X(x), \quad \mathbb{P}(Y_i = 1, S_i = 1) = \int_a^b p(x)\, dF_X(x)$$

and the conditional probabilities of sampling, given the status value, are

$$\begin{aligned}
\lambda_1 &= \mathbb{P}(S_i = 1 | Y_i = 1) = \frac{\int_a^b p(x)\, dF_X(x)}{\int_{I_X} p(x)\, dF_X(x)}, \\
\lambda_0 &= \mathbb{P}(S_i = 1 | Y_i = 0) = \frac{\int_a^b \{1 - p(x)\}\, dF_X(x)}{1 - \int_{I_X} p(x)\, dF_X(x)}.
\end{aligned}$$

If the ratio $\theta = \lambda_0/\lambda_1$ is known or otherwise estimated, the previous estimators may be used for the estimation of $p(x)$ from the truncated sample with $S_i \equiv 1$.

For a random truncation interval $[A, B]$, the sampling indicator is $S = 1_{[A,B]}(X)$ and the integrals of p are replaced by their expectation with respect to the distribution function of A and B and the estimation follows.

2.5. Truncation of a response variable in a non-parametric regression model

Consider (X, Y), a two-dimensional variable in a left-truncated transformation model: let Y denote a response to a continuous expository variable X, up to a variable of individual variations ε independent of X,

$$Y = m(X) + \varepsilon, \quad E\varepsilon = 0, \quad E\varepsilon^2 < \infty,$$

(X, ε) with distribution function (F_X, F_ε). The distribution function of Y conditionally on X is defined by

$$\begin{aligned} F_{Y|X}(y;x) &= P(Y \leq y|X = x) = F_\varepsilon(y - m(x)), \qquad (2.2) \\ m(x) &= E(Y|X = x), \end{aligned}$$

and the function m is continuous. The joint and marginal distribution functions of X and Y are denoted $F_{X,Y}$, with support $I_{Y,X}$, F_X, with bounded support I_X, and F_Y, such that $F_Y(y) = \int F_\varepsilon(y - m(s))\, dF_X(s)$ and $F_{X,Y}(x,y) = \int 1_{\{s \leq x\}} F_\varepsilon(y - m(s))\, dF_X(s)$.

The observation of Y is assumed to be left-truncated by variable T independent of (X, Y), with distribution function F_T, Y and T are observed conditionally on $Y \geq T$ and none of the variables are observed if $Y < T$. Denote $\bar{F} = 1 - F$ for any distribution function F and, under left-truncation,

$$\begin{aligned} \alpha(x) &= P(T \leq Y|X = x) = \int_{-\infty}^{\infty} \bar{F}_\varepsilon(y - m(x))\, dF_T(y), \\ A(y;x) &= P(Y \leq y|X = x, T \leq Y) \\ &= \alpha^{-1}(x) \int_{-\infty}^{y} F_T(v)\, dF_\varepsilon(v - m(x)) \qquad (2.3) \\ B(y;x) &= P(T \leq y \leq Y|X = x, T \leq Y) \\ &= \alpha^{-1}(x) F_T(y) \bar{F}_\varepsilon(y - m(x)), \qquad (2.4) \\ m^*(x) &= E(Y|X = x, T \leq Y) = \alpha^{-1}(x) \int y F_T(y)\, dF_{Y|X}(y;x). \end{aligned}$$

Obviously, the mean of Y is biased under the truncation and a direct estimation of the conditional distribution function $F_{Y|X}$ is of interest for the estimation of $m(x) = E(Y|X = x)$ instead of the apparent mean $m^*(x)$. The function $\bar{F}_\varepsilon$ is also written $\exp\{-\Lambda_\varepsilon\}$ with $\Lambda_\varepsilon(y) = \int_{-\infty}^{y} \bar{F}_\varepsilon^{-1} dF_\varepsilon$ and the expressions (2.3)-(2.4) of A and B imply that

$$\Lambda_\varepsilon(y - m(x)) = \int_{-\infty}^{y} B^{-1}(s;x)\, A(ds;x)$$

and $F_{Y|X}(y;x) = \exp\{-\Lambda_\varepsilon(y - m(x))\}$.

An estimator of $F_{Y|X}(y;x)$ is obtained as the product-limit estimator $\widehat{F}_{\varepsilon,n}(y - m(x))$ of $F_\varepsilon(y - m(x))$ based on estimators of A and B. For a sample $(X_i, Y_i)_{1 \leq i \leq n}$, let x in $I_{X,n,h} = [\min_i X_i + h, \max_i X_i - h]$ and

$$\begin{aligned}
\widehat{A}_n(y;x) &= \frac{\sum_{i=1}^n K_h(x-X_i) I_{\{T_i \le Y_i \le y\}}}{\sum_{i=1}^n K_h(x-X_i) I_{\{T_i \le Y_i\}}}, \\
\widehat{B}_n(y;x) &= \frac{\sum_{i=1}^n K_h(x-X_i) I_{\{T_i \le y \le Y_i\}}}{\sum_{i=1}^n K_h(x-X_i) I_{\{T_i \le Y_i\}}}; \\
\widehat{F}_{Y|X,n}(y;x) &= 1 - \prod_{1 \le Y_i \le y} \left\{ 1 - \frac{d\widehat{A}_n}{\widehat{B}_n}(Y_i;x) \right\} \\
&= 1 - \prod_{1 \le i \le n} \left\{ 1 - \frac{K_h(x-X_i) I_{\{T_i \le Y_i \le y\}}}{\sum_{j=1}^n K_h(x-X_j) I_{\{T_j \le Y_i \le Y_j\}}} \right\}, \qquad (2.5)
\end{aligned}$$

with $0/0 = 0$. That is a non-parametric maximum likelihood estimator of $F_{Y|X}$, as is the Kaplan-Meier estimator for the distribution function of a right-censored variable. Then an estimator of $m(x)$ may be defined as an estimator of $\int y \, F_{Y|X}(dy;x)$,

$$\begin{aligned}
\widehat{m}_n(x) &= \sum_{i=1}^n Y_i I_{\{T_i \le Y_i\}} \{ \widehat{F}_{Y|X,n}(Y_i;x) - \widehat{F}_{Y|X,n}(Y_i^-;x) \} \\
&= \frac{\sum_{i=1}^n Y_i I_{\{T_i \le Y_i\}} K_h(x-X_i) \, \widehat{F}_{Y|X,n}(Y_i^-;x)}{\sum_{j=1}^n K_h(x-X_j) I_{\{T_j \le Y_i \le Y_j\}}}. \qquad (2.6)
\end{aligned}$$

By the same arguments, from the means in (2.3)-(2.4), $\bar{F}_Y(y) = E\bar{F}_\varepsilon(y - m(X))$ is estimated by

$$\widehat{\bar{F}}_{Y,n}(y) = \prod_{1 \le i \le n} \left\{ 1 - \frac{I_{\{T_i \le Y_i \le y\}}}{\sum_{j=1}^n I_{\{T_j \le Y_i \le Y_j\}}} \right\},$$

the distribution function F_T is simply estimated by the product-limit estimator for right-truncated variables [WOO 85]

$$\widehat{F}_{T,n}(t) = \prod_{1 \le i \le n} \left\{ 1 - \frac{I_{\{t \le T_i \le Y_i\}}}{\sum_{j=1}^n I_{\{T_j \le T_i \le Y_j\}}} \right\}$$

and an estimator of F_ε is deduced from $F_{Y|X}$, F_X and m as

$$\widehat{F}_{\varepsilon,n}(s) = n^{-1} \sum_{1 \le i \le n} \widehat{F}_{Y|X,n}(s + \widehat{m}_n(X_i); X_i).$$

The means of T and C are estimated by

$$\widehat{\mu}_{T,n} = n^{-1}\sum_{i=1}^{n} \frac{T_i I_{\{T_i \le Y_i\}}\, \widehat{F}_{T,n}(T_i^-)}{\sum_{j=1}^{n} I_{\{T_j \le T_i \le Y_j\}}},\ \widehat{\mu}_{Y,n} = n^{-1}\sum_{i=1}^{n} \frac{Y_i I_{\{T_i \le Y_i\}}\, \widehat{F}_{Y,n}(Y_i^-)}{\sum_{j=1}^{n} I_{\{T_j \le Y_i \le Y_j\}}}.$$

The estimators $\widehat{F}_{Y,n}$ and $\widehat{F}_{T,n}$ are known to be P-uniformly consistent and asymptotically Gaussian. For the further convergences restricted to the interval $I_{n,h} = \{(y,x) \in I_{Y,X} : x \in I_{X,n,h}\}$, assume the following condition:

Condition 2.1 $h = h_n \to 0$ *and* $nh^3 \to \infty$ *as* $n \to \infty$,
$\alpha > 0$ *in the interior of* I_X, $\int K = 1$, $\kappa_1 = \int x^2 K(x)\,dx$ *and* $\kappa_2 = \int K^2 < \infty$.
The distribution function $F_{Y,X}$ *is twice continuously differentiable with respect to* x *with notations* $\dot{F}_{Y,X,2}(y,x) = \partial F_{Y|X}(y,x)/\partial x$, $\ddot{F}_{Y,X,2}(y,x) = \partial^2 F_{Y|X}(y,x)/\partial x^2$, *and differentiable with respect to* y *with notation* $\dot{F}_{Y|X,1}(y,x) = \partial F_{Y|X}(y,x)/\partial y$.
$E\varepsilon^{2+\delta} < \infty$ *for a* δ *in* $(1/2, 1]$.

Proposition 2.1 $\sup_{I_{n,h}} |\widehat{A}_n - A| \xrightarrow{P} 0$ *and* $\sup_{I_{n,h}} |\widehat{B}_n - B| \xrightarrow{P} 0$,

$$\begin{aligned}
b^A_{nh}(y;x) &\equiv (E\widehat{A}_n - A)(y;x) = \tfrac{h^2}{2\alpha(x)}\kappa_1 \Big\{ \textstyle\int_{-\infty}^{y} F_T(v)\ddot{F}_{Y,X,2}(dv,dx) \\
&\quad -A(y;x) \textstyle\int_{-\infty}^{\infty} F_T(v)\ddot{F}_{Y,X,2}(dv,dx) \Big\} + o(h^2), \\
b^B_{nh}(y;x) &\equiv (E\widehat{B}_n - B)(y;x) = \tfrac{h^2}{2\alpha(x)}\kappa_1 \{F_T(y) \textstyle\int \ddot{F}_{Y,X,2}(dv,dx)\,dx \\
&\quad -B(y;x) \textstyle\int F_T(v)\ddot{F}_{Y,X,2}(dv,dx)\} + o(h^2), \\
v^A_{nh}(y;x) &\equiv \mathrm{var}\widehat{A}_n(y;x) = (nh)^{-1}\kappa_2 A(1-A)(y;x)\alpha^{-1}(x) + o((nh)^{-1}), \\
v^B_{nh}(y;x) &\equiv \mathrm{var}\widehat{B}_n(y;x) = (nh)^{-1}\kappa_2 B(1-B)(y;x)\alpha^{-1}(x) + o((nh)^{-1}).
\end{aligned}$$

If $nh^5 \to 0$, $(nh)^{1/2}(\widehat{A}_n - A)$ *and* $(nh)^{1/2}(\widehat{B}_n - B)$ *converge in distribution to Gaussian processes with mean zero, variances* $\kappa_2 A(1-A)(y;x)\alpha^{-1}(x)$ *and* $\kappa_2 B(1-B)(y;x)\alpha^{-1}(x)$ *respectively, then the covariances of the limiting processes are zero.*

Proof. Let $\widehat{A}_n = \widehat{c}_n^{-1}\widehat{a}_n$ and $\widehat{B}_n = \widehat{c}_n^{-1}\widehat{b}_n$, with

$$\widehat{c}_n(x) = n^{-1}\sum_{i=1}^{n} K_h(x - X_i) I_{\{T_i \le Y_i\}},\ \widehat{a}_n(y;x) = n^{-1}\sum_{i=1}^{n} K_h(x - X_i) I_{\{T_i \le Y_i \le y\}},$$

$$\widehat{b}_n(y;x) = n^{-1}\sum_{i=1}^{n} K_h(x - X_i) I_{\{T_i \le y \le Y_i\}},$$

they satisfy

$$(nh)^{1/2}(\widehat{A}_n - A)(y;x) = (nh)^{1/2}c^{-1}(x)\{(\widehat{a}_n - a)(y;x) - A(\widehat{c}_n - c)(x)\} + o_{L^2}(1)$$

and a similar approximation for $\widehat{B}_n$. The biases and variances are deduced from those of each term and the weak convergences are proved as in [PIN 06]. ■

From proposition 2.1 and applying the results of the non-parametric regression,

Proposition 2.2 *The estimators $\widehat{F}_{Y|X,n}$, $\widehat{m}_n$, $\widehat{F}_{\varepsilon,n}$ converge P-uniformly to $F_{Y|X}$, m, F_ε, $\widehat{\mu}_{Y,n}$ and $\widehat{\mu}_{T,n}$ converge P-uniformly to EY and ET respectively.*

The weak convergence of the estimated distribution function of truncated survival data was proved in several papers ([GIL 90, LAI 91]). As in [GIL 83] and by proposition 2.1, their proof extends to their weak convergence on $(\min_i\{Y_i : T_i < Y_i\}, \max_i\{Y_i : T_i < Y_i\})$ under the conditions $\int F_T \, dF_{Y|X} < \infty$ and $\int \bar{F}^{-1}_{Y|X} \, dF_T < \infty$ on $I_{X,n,h}$, which are simply satisfied if for every x in $I_{X,n,h}$, $\inf\{t : F_T(t) > 0\} < \inf\{t : F_{Y|X}(t;x) > 0\}$ and $\sup\{t : F_{Y|X}(t;x) > 0\} < \sup\{t : F_T(t) > 0\}$.

Theorem 2.1 *$(nh)^{1/2}(\widehat{F}_{Y|X,n} - F_{Y|X})1_{I_{n,h}}$ converges weakly to a centered Gaussian process W on $I_{Y,X}$. The variables $(nh)^{1/2}(\widehat{m}_n - m)(x)$, for every x in $I_{X,n,h}$, and $(nh)^{1/2}(\widehat{\mu}_{Y,n} - EY)$ converge weakly to $EW(Y;x)$ and $E\int W(Y;x)\,dF_X(x)$.*

If m is assumed to be monotone with inverse function r, X is written $X = r(Y-\varepsilon)$ and the quantiles of X are defined by the inverse functions q_1 and q_2 of $F_{Y|X}$ at fixed y and x, respectively:

$$F_{Y|X}(y;x) = u \quad \text{and} \quad \left\{ \begin{array}{ll} x & = r(y - Q_\varepsilon(u)) \equiv q_1(u;y) \\ y & = m(x) + Q_\varepsilon(u) \equiv q_2(u;x), \end{array} \right.$$

where $Q_\varepsilon(u)$ is the inverse of F_ε at u. Finally, if m is increasing, then $F_{Y|X}(y;x)$ is decreasing in x and increasing in y, and it is the same for its estimator $\widehat{F}_{Y|X,n}$, up to a random set of small probability. The thresholds q_1 and q_2 are estimated by

$$\begin{array}{rcl} \widehat{q}_{1,n,h}(u;y) & = & \sup\{x : \widehat{F}_{Y|X,n}(y;x) \le u\}, \\ \widehat{q}_{2,n,h}(u;x) & = & \inf\{y : \widehat{F}_{Y|X,n}(u;x) \ge u\}. \end{array}$$

As a consequence of Theorem 2.1 and generalizing known results on quantiles.

Theorem 2.2 *For $k = 1, 2$, $\widehat{q}_{k,n,h}$ converges P-uniformly to q_k on $\widehat{F}_{Y,X,n}(I_{n,h})$. For every y and (respectively) x, $(nh)^{1/2}(\widehat{q}_{1,n,h} - q_1)(\cdot;y)$ and $(nh)^{1/2}(\widehat{q}_{2,n,h} - q_2)(\cdot;x)$ converge weakly to the centered Gaussian process $W \circ q_1[\dot{F}_{Y|X,1}(y;q_1(\cdot;y))]^{-1}$ and, respectively, $W \circ q_2[\dot{F}_{Y|X,2}(q_2(\cdot;x);x)]^{-1}$.*

2.6. Double censoring of a response variable in a non-parametric model

Let $Y = m(X) + \varepsilon$, where $E\varepsilon = 0$, $E\varepsilon^2 < \infty$ and ε is independent of X, be observed on an independent random interval $[T, C]$ with $T < C$. The observations are X, with a positive density f_X on a support I_X

$$W = \max\{T, U\}, \text{ with } U = \min(Y, C), \quad \eta = 1_{\{Y > T\}}, \quad \delta = 1_{\{Y \leq C\}},$$

so Y is observed only if $\eta\,\delta = 1$. We assume that the variables X, T and C have continuous distribution functions F_X, F_T and F_C such that

$$\mathbb{P}(T < Y < C | X = x) = \int (F_T - F_C)(y)\, dF_{Y|X}(y; x) > 0.$$

The distribution function of Y conditionally on X is still defined by (2.2). Let $\tau_1 = \inf\{t; F_T(t) > 0\}$ and $\tau_2 = \sup\{t; F_{Y|X}(t) F_C(t) < 1\}$, the upper bounds for Y and C. The conditional mean of the observed Y is now $m^*(x) = E(Y\eta\,\delta | X = x) = \int y (F_T - F_C)(y)\, dF_{Y|X}(y; x)$. The notations of section 2.5 become

$$\begin{aligned}
A(y; x) &= \mathbb{P}(Y \geq y, \eta = 1, \delta = 1 | X = x) \\
&= \int_y^{\tau_2} (F_T - F_C)\, dF_{Y|X}(\cdot; x), \\
B(y; x) &= \mathbb{P}(Y < T, T \geq y | X = x) = \int_y^{\tau_2} \bar{F}_{Y|X}(\cdot; x)\, dF_T, \\
C(y; x) &= \mathbb{P}(Y > C \geq y | X = x) = \int_y^{\tau_2} \bar{F}_{Y|X}(\cdot; x)\, dF_C, \\
\bar{F}_W(y; x) &= \mathbb{P}(W \geq y | X = x) = \bar{F}_T(y) + (F_T - F_C)(y) \bar{F}_{Y|X}(y; x),
\end{aligned}$$

which is the sum of A, B and C. The hazard function $\int_{\tau_1}^{\cdot} \bar{F}_W^{-1}\, dF_W$ of the observed variable W is no longer equal to the hazard function of the variable Y as it is for independent censoring variables T and C, and estimation by self-consistency equations may be used. For a sample $(W_i, \eta_i, \delta_i)_{i \leq n}$, let $\tau_{1,n} = \min_i W_i$ and $\tau_{2,n} = \max_i W_i$. Let y in $[\tau_{1,n}, \tau_{2,n}]$,

$$\begin{aligned}
Y_n^{nc}(y; x) &= \frac{\sum_{i=1}^n \eta_i \delta_i 1_{\{Y_i \geq y\}} K_h(x - X_i)}{\sum_{i=1}^n K_h(x - X_i)}, \\
Y_n^{c,1}(y; x) &= \frac{\sum_{i=1}^n (1 - \eta_i) 1_{\{T_i \geq y\}} K_h(x - X_i)}{\sum_{i=1}^n K_h(x - X_i)}, \qquad (2.7) \\
Y_n^{c,2}(y; x) &= \frac{\sum_{i=1}^n (1 - \delta_i) 1_{\{C_i \geq y\}} K_h(x - X_i)}{\sum_{i=1}^n K_h(x - X_i)},
\end{aligned}$$

and $Y_n = Y_n^{nc} + Y_n^{c,1} + Y_n^{c,2}$. The process $\sup_y |n^{-1}Y_n(y;x) - \bar{F}_W(y;x)|$ converges to zero in probability for every x in $I_{X,n,h}$.

The self-consistency equation for an estimator $\widehat{S}_n$ of $\bar{F}_{Y|X}$ is defined as a solution of

$$\begin{aligned} n\widehat{S}_n(y;x) &= Y_n(y^+;x) - \widehat{S}_n(y;x) \int_{-\infty}^{y} \frac{dY_n^{c,2}}{\widehat{S}_n} \\ &\qquad + \{1 - \widehat{S}_n(y;x)\} \int_y^{\infty} \frac{dY_n^{c,1}}{1-\widehat{S}_n}(\cdot;x) \end{aligned} \tag{2.8}$$

therefore,

$$\begin{aligned} \widehat{S}_n(y;x) &= \Big\{n + \int_{-\infty}^{y} \frac{dY_n^{c,2}}{\widehat{S}_n}(\cdot;x) + \int_y^{\infty} \frac{dY_n^{c,1}}{1-\widehat{S}_n}(\cdot;x)\Big\}^{-1} \Big\{Y_n(y^+;x) \\ &\qquad + \int_y^{\infty} \frac{dY_n^{c,1}}{1-\widehat{S}_n(\cdot;x)}\Big\}. \end{aligned}$$

This equation provides a decreasing estimator $\widehat{S}_n$, with $\widehat{S}_n(\tau_{1,n};x) = 1$ and jumping at the uncensored transitions times. Let $(W_{(1)} < W_{(2)} < \ldots)$ be the ordered sample of the variables W_i, $i = 1, \ldots, n$, and let $\delta_{(l)}$ and $\eta_{(l)}$ be the indicators related to $W_{(l)}$. The jumps $\Delta\widehat{S}_n(W_{(l)};x)$ of $\widehat{S}_n$ are equal to

$$-\frac{\eta_{(l)}\delta_{(l)}}{n - \sum_{j<l}(1-\delta_{(j)})\widehat{S}_n^{-1}(Y_{(j)};x) - \sum_{j>l}(1-\eta_{(j)})(1-\widehat{S}_n(Y_{(j)};x))^{-1}}$$

under the constraint $\Delta\widehat{S}_n(W_{(l)};x) < 0$ for every l and x. The last sum of the denominator, $I_{(l)} = \sum_{j>l}(1-\eta_{(j)})(1-\widehat{S}_n(Y_{(j)};x))^{-1}$ cannot be directly calculated from the values of $\widehat{S}_n$ at $W_{(1)}, \ldots, W_{(l-1)}$. This expression provides an iterative algorithm: $I_{(l)}$ may be omitted at the first step and the conditional Kaplan-Meier estimator for right-censored observations of Y given X is used as initial estimator of $\bar{F}_{Y|X}$, and is defined at every $W_{(l)}$. This initial estimator is used for an iterative procedure with the constraint that the estimator remains in $]0,1]$. This algorithm converges to the solution of (2.8).

Estimators of F_T and F_C are written as in [PON 06]. As a consequence of the P-uniform convergence of $n^{-1}Y_n$ to S_W and the processes $n^{-1}Y_n^{k,c}$, for $k = 1, 2$, we obtain the following theorem:

Theorem 2.3 *The estimator* $\widehat{S}_n$ *solution of* (2.8) *is* P*-uniformly consistent and* $n^{1/2}(\widehat{S}_n - S)$ *converges weakly to a centered Gaussian process.*

The weak convergence is proved by the same method as in [PON 07]. Then $m(x)$ is estimated by $\widehat{m}_n(x) = -\int y d\widehat{S}_n(y;x)$ which equals

$$\sum_{l=1}^{n} \frac{\eta_{(l)}\delta_{(l)}Y_{(l)}}{n - \sum_{j<l}(1-\delta_{(j)})\widehat{S}_n^{-1}(Y_{(j)};x) - \sum_{j>l}(1-\eta_{(j)})(1-\widehat{S}_n(Y_{(j)};x))^{-1}}$$

and $\widehat{m}_n$ converges P-uniformly to m on $I_{X,n,h}$.

2.7. Other truncation and censoring of Y in a non-parametric model

The variable Y is assumed to be left-truncated by T and right-censored by a variable C independent of (X, Y, T). The notations α and those of the joint and marginal distribution function of X, Y and T are in section 2.5, and F_C is the distribution function of C. The observations are $\delta = 1_{\{Y \le C\}}$, and $(Y \wedge C, T)$, conditionally on $Y \wedge C \ge T$. Let

$$\begin{array}{lcl}
A(y;x) & = & P(Y \le y \wedge C | X = x, T \le Y) \\
 & = & \alpha^{-1}(x) \int_{-\infty}^{y} F_T(v)\bar{F}_C(v)\, F_{Y|X}(dv;x) \\
B(y;x) & = & P(T \le y \le Y \wedge C | X = x, T \le Y) \\
 & = & \alpha^{-1}(x) F_T(y)\bar{F}_C(y)\bar{F}_{Y|X}(y;x), \\
\bar{F}_{Y|X}(y;x) & = & \exp\{-\int_{-\infty}^{y} B^{-1}(v;x)\, A(dv;x)\}.
\end{array}$$

The estimators are then written as

$$\widehat{\bar{F}}_{Y|X,n}(y;x) = \prod_{1\le i\le n} \left\{ 1 - \frac{K_h(x-X_i) I_{\{T_i \le Y_i \le y \wedge C_i\}}}{\sum_{j=1}^{n} K_h(x-X_j) I_{\{T_j \le Y_i \le Y_j \wedge C_j\}}} \right\},$$

$$\widehat{m}_n(x) = \frac{\sum_{i=1}^{n} Y_i I_{\{T_i \le Y_i \le C_i\}} K_h(x-X_i)\, \widehat{F}_{Y|X,n}(Y_i^-;x)}{\sum_{j=1}^{n} K_h(x-X_j) I_{\{T_j \le Y_i \le Y_j \wedge C_j\}}},$$

$$\widehat{\bar{F}}_{Y,n}(y) = \prod_{1\le i\le n} \left\{ 1 - \frac{I_{\{T_i \le Y_i \le y \wedge C_i\}}}{\sum_{j=1}^{n} I_{\{T_j \le Y_i \le Y_j \wedge C_j\}}} \right\}.$$

If Y is only right-truncated by C independent of (X, Y), with observations (X, Y) and C conditionally on $Y \leq C$, the expressions α, A and B are then written as

$$\begin{aligned}
\alpha(x) &= P(Y \leq C|X = x) = \int_{-\infty}^{\infty} \bar{F}_C(y)\, F_{Y|X}(dy; x), \\
A(y; x) &= P(Y \leq y|X = x, Y \leq C) = \alpha^{-1}(x) \int_{-\infty}^{y} \bar{F}_C(v)\, F_{Y|X}(dv; x), \\
B(y; x) &= P(Y \leq y \leq C|X = x, Y \leq C) = \alpha^{-1}(x)\bar{F}_C(y)F_{Y|X}(y; x), \\
A'(y; x) &= P(Y \leq C \leq y|X = x, Y \leq C) \\
&= \alpha^{-1}(x) \int_{-\infty}^{y} F_{Y|X}(v; x)\, dF_C(v).
\end{aligned}$$

The distribution function F_C and $F_{Y|X}$ are both identifiable and their expression differs from the previous ones,

$$\begin{aligned}
\bar{F}_C &= \exp\{-\int_{-\infty}^{\cdot} EB^{-1}(v; X)\, EA'(dv; X)\}, \\
F_{Y|X}(\cdot; x) &= \exp\{-\int_{\cdot}^{\infty} B^{-1}(v; x)\, A(dv; x)\}.
\end{aligned}$$

The estimators are then

$$\begin{aligned}
\widehat{F}_{Y|X,n}(y; x) &= \prod_{1 \leq i \leq n} \left\{ 1 - \frac{K_h(x - X_i)I_{\{Y_i \leq y \wedge C_i\}}}{\sum_{j=1}^{n} K_h(x - X_j)I_{\{Y_j \leq Y_i \leq C_j\}}} \right\}, \\
\widehat{\bar{F}}_{C,n}(y) &= \prod_{1 \leq i \leq n} \left\{ 1 - \frac{I_{\{Y_i \leq C_i \leq y\}}}{\sum_{j=1}^{n} I_{\{Y_j \leq Y_i \leq C_j\}}} \right\}, \\
\widehat{m}_n(x) &= \frac{\sum_{i=1}^{n} Y_i I_{\{Y_i \leq C_i\}} K_h(x - X_i)\, \widehat{F}_{Y|X,n}(Y_i^-; x)}{\sum_{j=1}^{n} K_h(x - X_j)I_{\{Y_j \leq Y_i \leq C_j\}}}.
\end{aligned}$$

If Y is left and right-truncated by mutually independent variables T and C, independent of (X, Y), the observations are (X, Y), C and T, conditionally on $T \leq Y \leq C$,

$$\begin{aligned}
\alpha(x) &= P(T \le Y \le C|X = x) = \int_{-\infty}^{\infty} F_T(y)\bar{F}_C(y)\, F_{Y|X}(dy;x),\\
A(y;x) &= P(Y \le y|X = x, T \le Y \le C)\\
&= \alpha^{-1}(x)\int_{-\infty}^{y} F_T(v)\bar{F}_C(v)\, F_{Y|X}(dv;x),\\
B(y;x) &= P(T \le y \le Y|X = x, T \le Y \le C)\\
&= \alpha^{-1}(x)F_T(y)\int_{y}^{\infty} \bar{F}_C(v)\ F_{Y|X}(dv;x),\\
A'(y) &= P(y \le T|T \le Y \le C) = \int_{y}^{\infty} dF_T(t)\int_{t}^{\infty} \bar{F}_C\, dF_Y,\\
B'(y) &= P(Y \le y \le C|T \le Y \le C) = \bar{F}_C(y)\int_{-\infty}^{y} F_T\, dF_Y,\\
B''(y) &= P(C \le y|T \le Y \le C) = \int_{-\infty}^{y}\{\int_{-\infty}^{s} F_T(v)\, dF_Y(v)\}\, dF_C(s).
\end{aligned}$$

The functions F_C, F_T and $F_{Y|X}$ are identifiable and

$$F_{Y|X}(y;x) = -\int_{-\infty}^{y} \bar{F}_C^{-1}\, dH(\cdot;x),$$

with

$$\begin{aligned}
H(y;x) &\equiv \int_{y}^{\infty} \bar{F}_C(v)\, dF_{Y|X}(dv;x) = \exp\{-\int_{-\infty}^{y} B^{-1}(v;x)\, A(dv;x)\},\\
\bar{F}_C(s) &= \exp\{-\int\int_{-\infty}^{s} B'^{-1}\, dB''\},\\
F_T(t) &= \exp[-\{\int_{t}^{\infty}(EB(\cdot;X))^{-1}\, dA'\}].
\end{aligned}$$

Their estimators are

$$\widehat{\bar{F}}_{C,n}(s) = \prod_{i=1}^{n} \left\{ 1 - \frac{I_{\{T_i \le Y_i \le C_i \le s\}}}{\sum_{j=1}^{n} I_{\{T_j \le Y_j \le C_i \le C_j\}}} \right\},$$

$$\widehat{F}_{T,n}(t) = \prod_{i=1}^{n} \left\{ 1 - \frac{I_{\{T_i \le Y_i \le C_i \le t\}}}{\sum_{j=1}^{n} I_{\{T_j \le T_i \le Y_j \le C_j\}}} \right\},$$

$$\widehat{F}_{Y|X}(y;x) = \frac{\sum_{i=1}^{n} \widehat{\bar{F}}_{C,n}^{-1}(Y_i) I_{\{T_i \le Y_i \le C_i \wedge y\}} K_h(x - X_i)\, \widehat{H}_{Y|X,n}(Y_i^-;x)}{\sum_{j=1}^{n} K_h(x - X_j) I_{\{T_j \le Y_i \le Y_j \le C_j\}}},$$

$$\widehat{H}_{Y|X}(y;x) = \prod_{i=1}^{n} \left\{ 1 - \frac{K_h(x - X_i) I_{\{T_i \le Y_i \le C_i \wedge y\}}}{\sum_{j=1}^{n} K_h(x - X_j) I_{\{T_j \le Y_i \le Y_j \le C_j\}}} \right\}.$$

The other non-parametric estimators of section 2.2 and the results of section 2.5 generalize to all the estimators of this section.

2.8. Observation by interval

Consider model (2.2) with an independent censoring variable C for Y. For observations by intervals, only C and the indicators that Y belongs to the interval $]-\infty, C]$ or $]C, \infty[$ are observed. The function $F_{Y|X}$ is not directly identifiable and efficient estimators for m and $F_{Y|X}$ are maximum likelihood estimators. Let $\delta = I_{\{Y \le C\}}$ and assume that F_ε is C^2. Conditionally on C and $X = x$, the log-likelihood of (δ, C) is

$$l(\delta, C) = \delta \log F_\varepsilon(C - m(x)) + (1 - \delta) \log \bar{F}_\varepsilon(C - m(x))$$

and its derivatives with respect to $m(x)$ and F_ε are

$$\begin{aligned} \dot{l}_{m(x)}(\delta, C) &= -\delta \frac{f_\varepsilon}{F_\varepsilon}(C - m(x)) + (1 - \delta) \frac{f_\varepsilon}{\bar{F}_\varepsilon}(C - m(x)), \\ \dot{l}_\varepsilon a(\delta, C) &= \delta \frac{\int_{-\infty}^{C - m(x)} a \, dF_\varepsilon}{F_\varepsilon(C - m(x))} + (1 - \delta) \frac{\int_{C - m(x)}^{\infty} a \, dF_\varepsilon}{\bar{F}_\varepsilon(C - m(x))} \end{aligned}$$

for every a satisfying $\int a \, dF_\varepsilon = 0$ and $\int a^2 \, dF_\varepsilon < \infty$. With $a_F = -f'_\varepsilon f_\varepsilon^{-1}$, $\dot{l}_\varepsilon a_F = \dot{l}_{m(x)}$ then $\dot{l}_{m(x)}$ belongs to the tangent space for F_ε and the estimator of $m(x) = E(Y|X = x)$ must be determined from the estimator of F_ε using the conditional probability function of the observations

$$B(t;x) = P(Y \le C \le t | X = x) = \int_{-\infty}^{t} F_\varepsilon(s - m(x)) \, dF_C(s).$$

Let $\widehat{F}_{C,n}$ be the empirical estimator of F_C and

$$\widehat{B}_n(t;x) = \frac{\sum_{i=1}^n K_h(x-X_i)I_{\{Y_i \le C_i \le t\}}}{\sum_{i=1}^n K_h(x-X_i)},$$

an estimator $\widehat{F}_{\varepsilon,n}(t-m(x))$ of $F_{\varepsilon,n}(t-m(x))$ is deduced by deconvolution and

$$\widehat{m}_n(x) = \int t\, d\widehat{F}_{\varepsilon,n}(t-m(x)).$$

2.9. Bibliography

[FAN 96] FAN J., GIJBELS I., *Local Polynomial Modelling and its Applications*, Chapman and Hall, London, 1996.

[GIL 83] GILL R., "Large sample behaviour of the product-limit estimator on the whole line", *Ann. Statist.*, vol. 11, p. 49–58, 1983.

[GIL 90] GILL R., KEIDING N., "Random truncation model and Markov processes", *Ann. Statist.*, vol. 18, p. 582–60, 1990.

[LAI 91] LAI T., YING Z., "Estimating a distribution function with truncated and censored data", *Ann. Statist.*, vol. 19, p. 417–442, 1991.

[PIN 06] PINÇON C., PONS O., "Nonparametric estimator of a quantile function for the probability of event with repeated data", *Dependence in Probability and Statistics, Lecture Notes in Statistics*, vol. 17, p. 475–489, Springer, New York, 2006.

[PON 06] PONS O., "Estimation for semi-Markov models with partial observations via self-consistency equations", *Statistics*, vol. 40, p. 377–388, 2006.

[PON 07] PONS O., "Estimation for the distribution function of one and two-dimensional censored variables or sojourn times of Markov renewal processes", *Communications in Statistics – Theory and Methods*, vol. 36, num. 14, 2007.

[WOO 85] WOODROOF M., "Estimating a distribution function with truncated data", *Ann. Statist.*, vol. 13, p. 163–177, 1985.

Chapter 3

Inference in Transformation Models for Arbitrarily Censored and Truncated Data

3.1. Introduction

In survival analysis we deal with data related to times of events (or end-points) in individual life-histories. The survival data are not amenable to standard statistical procedures used in data analysis for several reasons. One of them is that survival data is not symmetrically distributed, but the main reason is that survival times are frequently censored. This usually happens when the data from a study are to be analyzed at a point when some individuals have not yet experienced the event of interest (or not reached the end-point). Many failure time data in epidemiological studies are simultaneously truncated and interval-censored. Interval-censored data occur in grouped data or when the event of interest is assessed on repeated visits. Right and left-censored data are particular cases of interval-censored data. Right-truncated data occur in registers. For instance, an acquired immune deficiency syndrome (AIDS) register only contains AIDS cases which have been reported. This generates right-truncated samples of induction times. [TUR 76] proposed a nice method of estimating the survival function in the case of arbitrarily censored and truncated data by a non-parametric maximum likelihood estimator. [FRY 94] noted that his method needed to be corrected slightly. [ALI 96] extended previous work by fitting a proportional hazards model to arbitrarily censored and truncated data, and concentrated on hypothesis testing. [HUB 04] introduced frailty models for the analysis of arbitrarily censored and truncated data, and focused on the estimation of the parameter of interest as well as the nuisance parameter of their model.

Chapter written by Filia VONTA and Catherine HUBER.

The concept of frailty models was introduced by [VAU 79] who studied models with Gamma distributed frailties. There are many frailty distributions that could be considered, such as the Gamma which corresponds to the well-known Clayton-Cuzick model [CLA 85, CLA 86], the inverse Gaussian or the positive stable (see [HOU 84] and [HOU 86] for many examples). The choice of a Gamma distributed frailty is the most popular in other works, due to its mathematical convenience.

This work is conducting some statistical analysis of interval censored and truncated data with the use of frailty models. We intend, using this analysis, to check the performance of the model proposed by [HUB 04]. In particular, we focus on hypothesis testing about the regression parameter of the model proposed by [HUB 04], in different situations, such as the case of independent covariates and the misspecification of the truncated proportion of the population. Further research could be directed towards the case of dependent covariates and the case of misspecification of the frailty distribution producing the data.

3.2. Non-parametric estimation of the survival function S

In this section we present and follow the formulation of [TUR 76], [FRY 94] and especially [ALI 96], regarding the case of arbitrarily censored and truncated data based on independent and identically distributed positive random variables . Let $X_1, X_2, \dots, X_n$ be independent and identically distributed positive random variables with survival function $S(x)$. For every random variable X_i we have a pair of observations (A_i, B_i) where A_i is a set called the censoring set and B_i a set called the truncating set. The random variable X_i belongs to the sample only if X_i falls into the set B_i. Also, X_i is being censored by the set A_i in the sense that the only information that we know about X_i is that it belongs to the set A_i where $A_i \subseteq B_i$. The sets A_i belong to a partition $\mathcal{P}_i$ of $[0, \infty)$ and we assume that B_i and $\mathcal{P}_i$ are independent of X_i and of the parameters of interest. We assume that the censoring sets $A_i,\ i = 1, \dots, n$ can be expressed as a finite union of disjoint closed intervals, that is,

$$A_i = \cup_{j=1}^{k_i} [L_{ij}, R_{ij}]$$

where $0 \leq L_{i1} \leq R_{i1} < L_{i2} \leq R_{i2} < \dots < L_{ik_i} \leq R_{ik_i} \leq \infty$ for $i = 1, \dots, n$, $R_{i1} > 0$, $L_{ik_i} < \infty$. Moreover, we assume that the truncating sets B_i can be expressed as a finite union of open intervals

$$B_i = \cup_{j=1}^{n_i} (\mathcal{L}_{ij}, \mathcal{R}_{ij})$$

where $0 \leq \mathcal{L}_{i1} < \mathcal{R}_{i1} < \mathcal{L}_{i2} < \mathcal{R}_{i2} < \dots < \mathcal{L}_{in_i} < \mathcal{R}_{in_i} \leq \infty$ for $i = 1, \dots, n$.

The likelihood of n pairs of observations $(A_i, B_i), i = 1, 2, \dots, n$ is proportional to

$$l(S) = \prod_{i=1}^{n} l_i(S) = \prod_{i=1}^{n} \frac{P_S(A_i)}{P_S(B_i)} = \prod_{i=1}^{n} \frac{\sum_{j=1}^{k_i} \left\{ S(L_{ij}^-) - S(R_{ij}^+) \right\}}{\sum_{j=1}^{n_i} \left\{ S(\mathcal{L}_{ij}^+) - S(\mathcal{R}_{ij}^-) \right\}} \tag{3.1}$$

Let us now define the sets

$$\tilde{L} = \{L_{ij},\ 1 \le j \le k_i, 1 \le i \le n\} \cup \{\mathcal{R}_{ij},\ 1 \le j \le n_i, 1 \le i \le n\} \cup \{0\}$$

and

$$\tilde{R} = \{R_{ij},\ 1 \le j \le k_i, 1 \le i \le n\} \cup \{\mathcal{L}_{ij},\ 1 \le j \le n_i, 1 \le i \le n\} \cup \{\infty\}.$$

Notice that the above likelihood is maximized when the values of $S(x)$ are as large as possible for $x \in \tilde{L}$ and as small as possible for $x \in \tilde{R}$. A set Q is defined uniquely as the union of disjoint closed intervals whose left endpoints lie in the set $\tilde{L}$ and right endpoints in the set $\tilde{R}$ respectively, and which contain no other members of $\tilde{L}$ or $\tilde{R}$. Thus,

$$Q = \cup_{j=1}^{v} [q_j', p_j']$$

where $0 = q_1' \le p_1' < q_2' \le p_2' < \ldots < q_v' \le p_v' = \infty$. Subsequently, we denote by C the union of intervals $[q_j', p_j']$ covered by at least one censoring set, by W the union of intervals $[q_j', p_j']$ covered by at least one truncating set but not covered by any censoring set, and by $D = \overline{(\cup B_i)}$ the union of intervals $[q_j', p_j']$ not covered by any truncating set. D is actually included in the union of intervals $[q_j', p_j']$. Obviously, the set Q can be written as $Q = C \cup W \cup D$. Let us denote the set C as

$$C = \cup_{i=1}^{m} [q_i, p_i]$$

where $q_1 \le p_1 < q_2 \le p_2 < \ldots < q_m \le p_m$. Let $s_j = S_{\overline{D}}(q_j{}^-) - S_{\overline{D}}(p_j{}^+)$ where $S_{\overline{D}}(x) = P(X > x | X \in \overline{D})$. The likelihood given in (3.1) can be written as a function of $s_1, s_2, \ldots, s_m$, that is,

$$l(s_1, \ldots, s_m) = \prod_{i=1}^{n} \frac{\sum_{j=1}^{m} \mu_{ij} s_j}{\sum_{j=1}^{m} \nu_{ij} s_j} \tag{3.2}$$

where $\mu_{ij} = I_{[\,[q_j,p_j] \subset A_i]}$ and $\nu_{ij} = I_{[\,[q_j,p_j] \subset B_i]}$, $i = 1, \ldots, n$ and $j = 1, \ldots, m$. The NPMLE of $S_{\overline{D}}$ was discussed by [TUR 76], [FRY 94] and [ALI 96].

3.3. Semi-parametric estimation of the survival function S

The most widely used model in the analysis of survival data is the Cox proportional hazards model [COX 72]. The hazard rate of an individual with p-dimensional covariate vector z, for the proportional hazards model, is given as

$$h(t|z) = e^{\beta^T z} h_0(t)$$

where $\beta \in R^p$ is the parameter of interest and $h_0(t)$ is the baseline hazard rate. Assume that a positive random variable η, called frailty, is introduced to act multiplicatively on the hazard intensity function of the Cox model. Then the hazard rate of an individual with covariate vector z takes the form

$$h(t|z,\eta) = \eta e^{\beta^T z} h_0(t)$$

where $\beta \in R^p$ is the parameter of interest and $h_0(t)$ is the baseline hazard rate. Equivalently, the survival function is given by

$$S(t|z,\eta) = e^{-\eta e^{\beta^T z}\Lambda(t)}$$

where $\Lambda(t)$ is the baseline cumulative hazard function. Thus,

$$S(t|z) = \int_0^\infty e^{-xe^{\beta^T z}\Lambda(t)} dF_\eta(x) \equiv e^{-G(e^{\beta^T z}\Lambda(t))} \tag{3.3}$$

where

$$G(y) = -\ln(\int_0^\infty e^{-xy} dF_\eta(x))$$

and F_η is the distribution function of the frailty parameter assumed in what follows to be completely known. It should be mentioned here that we consider univariate data, namely, a random sample of independent frailties $\eta_1, \eta_2, \ldots, \eta_n$, each one affecting individual i, $i = 1, \ldots, n$. When $G(x) = x$, the above model reduces to the Cox model. A well known frailty model is the Clayton-Cuzick model [CLA 85, CLA 86] which corresponds to a Gamma distributed frailty. Note that in general the frailty distribution depends on a finite-dimensional parameter c, so that $G(x) = G(x, c)$, which in this chapter is assumed to be known.

The class of semi-parametric transformation models as was defined in [CHE 95] for right-censored data, namely,

$$g(S(t|z)) = h(t) + \beta^T z$$

is equivalent to our class of models (3.3) through the relations

$$g(x) \equiv log(G^{-1}(-\log(x)), \quad h(t) \equiv \log(\Lambda(t))$$

where g is known and h unknown.

Let (X_1, Z_1), ..., (X_n, Z_n) be iid random pairs of variables with marginal survival function defined in (3.3) as in [VON 96]. The function $G \in \mathcal{C}^3$ is assumed to be a known strictly increasing concave function with $G(0) = 0$ and $G(\infty) = \infty$. As in the previous section, we assume that the random variables X_i are incomplete due to arbitrary censoring and truncation. The likelihood (3.1) written for the frailty models defined in (3.3) takes the form

$$l(\Lambda, \beta | z_1, \dots, z_n)$$
$$= \prod_{i=1}^{n} \frac{\sum_{j=1}^{k_i} \left\{ e^{-G(e^{\beta^T z_i}\Lambda(L_{ij}{}^-))} - e^{-G(e^{\beta^T z_i}\Lambda(R_{ij}{}^+))} \right\}}{\sum_{j=1}^{n_i} \left\{ e^{-G(e^{\beta^T z_i}\Lambda(\mathcal{L}_{ij}{}^+))} - e^{-G(e^{\beta^T z_i}\Lambda(\mathcal{R}_{ij}{}^-))} \right\}}. \qquad (3.4)$$

Our goal is to obtain the joint NPMLE of β, the parameter of interest and Λ, the nuisance parameter. In the maximization of (3.4) with respect to Λ, we employ the following lemmas which are the analogs of Lemmas 3.1 and 3.2 given in [TUR 76] and [ALI 96].

Lemma 3.1 *Any cumulative hazard-type function Λ within model (3.3) which increases outside the set $C \cup D$ cannot be the NPMLE of Λ.*

Lemma 3.2 *For fixed values of $\Lambda(q_j{}^-)$ and $\Lambda(p_j{}^+)$, for $1 \leq j \leq m$, the likelihood is independent of how the increase actually occurs in the interval $[q_j, p_j]$, so that Λ is undefined within each interval $[q_j, p_j]$.*

We now continue to write the log-likelihood in the non-proportional hazards case in a more convenient form so that the maximization with respect to Λ and β will be possible. Since set $C = \cup_{j=1}^{m}[q_j, p_j]$, set D can be written as $D = \cup_{j=0}^{m} D_j$, where $D_j = D \cap (p_j, q_{j+1})$, $p_0 = 0$ and $q_{m+1} = \infty$. Notice that D_j is either a closed interval or a union of disjoint closed intervals. Let $\delta_j = P_\Lambda(D_j)$ denote the mass of the cumulative hazard function Λ on the set D_j. From Lemma 3.1 we find that $\Lambda(q_j^-) = \Lambda(p_{j-1}^+) + \delta_{j-1}$ for $1 \leq j \leq m+1$. The log-likelihood can then be expressed as

$$\log l(\Lambda, \beta | z_1, \dots, z_n)$$
$$= \sum_{i=1}^{n} \Bigg\{ \log \Bigg(\sum_{j=1}^{m} \mu_{ij} \big(e^{-G(e^{\beta^T z_i}(\Lambda(p_{j-1}^+)+\delta_{j-1}))} - e^{-G(e^{\beta^T z_i}\Lambda(p_j{}^+))} \big) \Bigg)$$
$$- \log \Bigg(\sum_{j=1}^{m} \nu_{ij} \big(e^{-G(e^{\beta^T z_i}(\Lambda(p_{j-1}^+)+\delta_{j-1}))} - e^{-G(e^{\beta^T z_i}\Lambda(p_j{}^+))} \big) \Bigg) \Bigg\}. \qquad (3.5)$$

In most real data problems, set D consists of the union of two intervals, namely, D_0 and D_m. If there are only right-truncated data involved, then set $D = D_m$. If there are only left-truncated data involved, then set $D = D_0$. Therefore the case $D = D_0 \cup D_m$

covers most of the problems we would encounter in practice and therefore we will deal with this case in the simulations in the next section. In the above special case, we have $\delta_1 = \delta_2 = \ldots = \delta_{m-1} = 0$ and therefore likelihood (3.5) involves the parameters $\beta, \delta_0, \Lambda(p_0), \ldots, \Lambda(p_m)$. Since $\Lambda(p_0) = 0$ we have to maximize likelihood (3.5) with respect to the $p + m + 1-$dimensional parameter $(\beta, \delta_0, \Lambda(p_1), \ldots, \Lambda(p_m))$. Notice that δ_m could be obtained directly from $\Lambda(p_m)$. Similarly to [FIN 93] and [ALI 96], we will make the reparametrization $\gamma_0 = \log(\delta_0)$ and $\gamma_j = \log(\Lambda(p_j))$ for $j = 1, \ldots, m$ for calculational convenience. Therefore, the log-likelihood becomes

$$\sum_{i=1}^{n} \Big\{ \log \Big(\sum_{j=1}^{m} \mu_{ij} \big(e^{-G(e^{\beta^T z_i + \gamma_{j-1}})} - e^{-G(e^{\beta^T z_i + \gamma_j})} \big) \Big) - \log \Big(\sum_{j=1}^{m} \nu_{ij} \big(e^{-G(e^{\beta^T z_i + \gamma_{j-1}})} - e^{-G(e^{\beta^T z_i + \gamma_j})} \big) \Big) \Big\}. \quad (3.6)$$

A second reparametrization which ensures monotonicity of the sequence γ_j was subsequently employed, that is, $\tau_1 = \gamma_1$ and $\tau_j = \log(\gamma_j - \gamma_{j-1})$ for $j = 2, \ldots, m$. The maximization in section 3.4 will be done with respect to the parameters β and $\gamma_0, \tau_1 \ldots, \tau_m$ directly with the use of existing maximization procedures found in standard software packages such as Splus and Fortran 77.

3.4. Simulations

We performed simulations in order to investigate the size and power for different values of the true parameter β of the generalized likelihood ratio test used to test hypotheses about the regression coefficient β of model (3.3) in various situations. Each simulation consists of 100 samples. The setup of the simulated data resembles the case of the right-truncated AIDS data as they appear in [KAL 89]. For each sample, 400 (induction) times x_i and 400 survival times t_i were generated. Given a time x_*, the times x_i were generated from a $Uniform(0, x_*)$ distribution while the survival times t_i were generated from a frailty model of the class defined in (3.3) with Weibull baseline hazard function. More specifically, we considered a Weibull distribution with scale parameter ρ_0 equal to 2 and shape parameter κ equal to 0.7, where the baseline cumulative hazard function is of the form $\Lambda(t) = (t/\rho_0)^\kappa$. Two binary covariates Z_1 and Z_2 for which $P(Z_i = 0) = P(Z_i = 1) = 1/2$ for $i = 1, 2$ were considered. From the data that were generated, only the data that satisfied the condition $x_i + t_i \leq x_*$ were kept in the sample giving rise to right-truncated data. The interval $[0, x_*]$ was divided into n =15 equal intervals that constitute a partition $[a_{k-1}, a_k)$ for $k = 1, \ldots, n$ where $a_0 = 0$ and $a_n = x_*^+$. The survival times t_i are not only truncated but also interval-censored as they are reported only to belong in one of those intervals of the partition. In fact, they are reported to belong to intervals of the form $[a_{k-1}, x_* - x_i)$

	$\beta_1 = 2,\ \beta_2 = 0$	$\beta_1 = 2,\ \beta_2 = 0.7$	$\beta_1 = 2,\ \beta_2 = 1$
$c = 0.5$	0.05	0.71	0.95
$c = 1.0$	0.05	0.61	0.91
$c = 2.0$	0.07	0.41	0.65

Table 3.1. *Clayton-Cuzick model*

whenever $x_* - x_i < a_k$. The true probability of belonging to the set D, $P_\Lambda(D)$ was taken to be 0.19 (as was used for example in [FIN 93]) for all simulations. Because of truncation, the sample size of the generated samples is random and so is the number of parameters to be estimated. The sample size was about 300 and the number of parameters to be estimated for each sample about 30. Also, point x_* varied according to the true values of β_1, β_2 and $P_\Lambda(D)$. In our simulations, we were faced with the situation of the likelihood having a saddle point and therefore our procedure of maximization diverged. This small proportion (about 10%) of samples was left out of the analysis.

The null hypothesis considered in all situations was $\beta_2 = 0$. Also, the true values of the regression coefficients (β_1, β_2) were taken to be either (2,0), (2,0.7) or (2,1) in all simulations. Note that we compared the value of the generalized likelihood ratio test statistic obtained by our method, with the point of the $\mathcal{X}^2$ distribution with 1 degree of freedom at the 0.05 level of significance.

We use two frailty models of the class defined in (3.3), namely the Clayton-Cuzick model and the inverse Gaussian model. For the first model, η is taken to be distributed as Gamma with mean 1 and variance c. For the inverse Gaussian model, η is taken to be distributed as inverse Gaussian with mean 1 and variance 1/2b. When we choose a value for the parameter b or c, we actually specify the variance of the frailty. The function G takes respectively the form

$$G(x, c) = \frac{1}{c} \ln(1 + cx),\ c > 0$$

and

$$G(x, b) = \sqrt{4b(b + x)} - 2b,\ b > 0.$$

The first set of simulations was generated from the Clayton-Cuzick model. In Table 3.1 we see the size and powers of the generalized likelihood ratio test when the parameter of the Clayton-Cuzick function is equal to 0.5, 1.0 or 2.0.

The results are satisfactory since the sizes are close to 0.05. The powers improve as c decreases to 0. In Tables 3.2 and 3.3 we see the mean of the estimated values of β as well as the sample variances of $\hat{\beta}$. Notice that in Table 3.1 the means are in general good estimators of the true values of β in each case. Notice that the estimator of the

	$\beta_1 = 2,\ \beta_2 = 0$		$\beta_1 = 2,\ \beta_2 = 0.7$		$\beta_1 = 2,\ \beta_2 = 1$	
	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$
$c = 0.5$	1.74913	0.00348341	1.8810800	0.6726250	2.103890	1.028380
$c = 1.0$	2.11861	-0.0503927	2.27901	0.749338	2.16962	1.0943
$c = 2.0$	2.391590	0.0352627	2.32768	0.906962	2.41705	1.34264

Table 3.2. *Mean values of estimated regression coefficients*

	$\beta_1 = 2,\ \beta_2 = 0$		$\beta_1 = 2,\ \beta_2 = 0.7$		$\beta_1 = 2,\ \beta_2 = 1$	
	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$
$c = 0.5$	0.32254	0.06889310	0.453818	0.0843178	0.653003	0.114077
$c = 1.0$	1.16533	0.175786	1.29224	0.131187	0.467791	0.146877
$c = 2.0$	0.882384	0.586621	0.653188	0.473163	0.858472	0.655379

Table 3.3. *Sample variances of estimated regression coefficients*

regression coefficient, which is bigger in value, tends in most cases (as c increases) to slightly overestimate its true value, resulting also in higher variance than the estimator of the other coefficient, which appears to be more stable as it has smaller variance than $\hat{\beta}_1$.

The second set of simulations was generated from the inverse Gaussian model. In Table 3.4 we see the size and powers of the generalized likelihood ratio test when the parameter of the inverse Gaussian function is equal to 0.05, 0.25 or 1.0.

Observe that the size in all three cases is around 0.05, which allows us to say that the size of the test is satisfactory. Notice that the powers increase as b increases. In Tables 3.5 and 3.6 we give the mean values and the sample variances of $\hat{\beta}$.

The means are in general good estimators of β. As previously noted, $\hat{\beta}_1$ appears to be more stable as it has smaller variance than $\hat{\beta}_2$. The results deteriorate (overestimation of β_1 and larger variances) as b decreases to 0.

The next two sets of simulations were generated as described in the beginning of this section, but the analysis was done under the misspecification of $P_\Lambda(D)$ being

	$\beta_1 = 2,\ \beta_2 = 0$	$\beta_1 = 2,\ \beta_2 = 0.7$	$\beta_1 = 2,\ \beta_2 = 1$
$b = 0.05$	0.06	0.40	0.82
$b = 0.25$	0.04	0.62	0.77
$b = 1.0$	0.06	0.69	0.97

Table 3.4. *Inverse Gaussian model*

	$\beta_1 = 2,\ \beta_2 = 0$		$\beta_1 = 2,\ \beta_2 = 0.7$		$\beta_1 = 2,\ \beta_2 = 1$	
	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$
$b = 0.05$	2.21307	0.00018296	2.73922	0.989606	2.2078	1.0885500
$b = 0.25$	1.95668	0.0249804	2.14617	0.747055	2.02087	1.0229300
$b = 1.0$	1.745370	0.00427109	1.896830	0.6521170	1.95841	0.9910700

Table 3.5. *Mean values of estimated regression coefficients*

	$\beta_1 = 2,\ \beta_2 = 0$		$\beta_1 = 2,\ \beta_2 = 0.7$		$\beta_1 = 2,\ \beta_2 = 1$	
	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$
$b = 0.05$	1.49436	0.577009	2.45047	0.738483	1.2071	0.339119
$b = 0.25$	0.781746	0.109733	1.60765	0.19023	0.604279	0.269023
$b = 1.0$	0.217448	0.07619790	0.295218	0.0793027	0.585862	0.0902406

Table 3.6. *Sample variances of estimated regression coefficients*

equal to 0. This imposes an additional constraint for the maximization with respect to Λ.

In Table 3.7 we used the Clayton-Cuzick model and we give the size and powers of the generalized likelihood ratio test under the misspecification of $P_\Lambda(D)$. Comparing the results with Table 3.1 where $P_\Lambda(D)$ was estimated by our procedure, we can see that the size is very satisfactory and the powers remain good (although lower in value). In Tables 3.8 and 3.9 we can see the mean values of estimated coefficients and the sample variances of the estimated coefficients.

The size and powers of the generalized likelihood ratio test remain good, a fact that justifies the use of the test even under the misspecification of $P_\Lambda(D)$. This fact was also noted in [ALI 96] for the case of the Cox model. The means of the estimators of the regression coefficients are lower than those we reported in the case with no misspecification of $P_\Lambda(D)$, while the variances of the estimators are smaller.

In Table 3.10 we see the size and powers of the tests for the case of the inverse Gaussian model under the misspecification of $P_\Lambda(D)$. The same comments apply as in the Clayton-Cuzick case, as far as the misspecification is concerned, which makes our

	$\beta_1 = 2,\ \beta_2 = 0$	$\beta_1 = 2,\ \beta_2 = 0.7$	$\beta_1 = 2,\ \beta_2 = 1$
$c = 0.5$	0.06	0.23	0.94
$c = 1.0$	0.03	0.56	0.81
$c = 2.0$	0.05	0.29	0.53

Table 3.7. *Clayton-Cuzick model* – $P_\Lambda(D) = 0$

	$\beta_1 = 2,\ \beta_2 = 0$		$\beta_1 = 2,\ \beta_2 = 0.7$		$\beta_1 = 2,\ \beta_2 = 1$	
	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$
$c = 0.5$	1.5068600	0.0514094	1.6119300	0.6080580	1.6261600	0.8301990
$c = 1.0$	1.7930100	-0.0295512	1.71777	0.633938	1.77439	0.887143
$c = 2.0$	2.1112500	0.0475518	2.064845	0.676237	2.02736	0.883918

Table 3.8. *Mean values of estimated regression coefficients*

	$\beta_1 = 2,\ \beta_2 = 0$		$\beta_1 = 2,\ \beta_2 = 0.7$		$\beta_1 = 2,\ \beta_2 = 1$	
	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$
$c = 0.5$	0.0587210	0.0739532	0.0979038	0.0556319	0.0761010	0.0576804
$c = 1.0$	0.127507	0.0738872	0.116004	0.109446	0.0890899	0.113985
$b = 2.0$	0.219691	0.204119	0.316909	0.206669	0.259659	0.189557

Table 3.9. *Sample variances of estimated regression coefficients*

results once more comparable to [ALI 96] for the Cox model. In Tables 3.11 and 3.12 we give the means and the sample variances of the estimated regression coefficients in these simulations. The means appear lower in value, although they improve as b increases, as compared with the no misspecification case while the sample variances are smaller.

The results obtained lead us to a number of conclusions. First of all, the behavior of the generalized likelihood ratio test is very good in the case of independent covariates. In the case of independent covariates along with the misspecification of $P_\Lambda(D)$, we can safely conclude that the values of the size of the test remain approximately the same as those obtained under no misspecification. Therefore, the use of the generalized likelihood ratio test is recommended even under the misspecification of $P_\Lambda(D)$. However, the estimators of the regression coefficient β are not as good as those without the misspecification, as can be seen from the relative tables. Also, the fact that the results improve in the Clayton-Cuzick case as c approaches 0 and in the inverse Gaussian case as b increases to ∞ is expected, since it is exactly then when the heterogenity in the population decreases (the variance of the frailty is tending to 0) and the frailty models approach the Cox model.

	$\beta_1 = 2,\ \beta_2 = 0$	$\beta_1 = 2,\ \beta_2 = 0.7$	$\beta_1 = 2,\ \beta_2 = 1$
$b = 0.05$	0.01	0.31	0.57
$b = 0.25$	0.04	0.56	0.79
$b = 1.0$	0.08	0.79	0.98

Table 3.10. *Inverse Gaussian model* - $P_\Lambda(D) = 0$

	$\beta_1 = 2,\ \beta_2 = 0$		$\beta_1 = 2,\ \beta_2 = 0.7$		$\beta_1 = 2,\ \beta_2 = 1$	
	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$
$b = 0.05$	1.285855	0.0297378	1.350696	0.531165	1.411863	0.689259
$b = 0.25$	1.491907	-0.0333297	1.538723	0.585153	1.524767	0.798376
$b = 1.0$	1.556253	0.00745458	1.5620740	0.5855360	1.6445200	0.8462790

Table 3.11. *Mean values of estimated regression coefficients*

	$\beta_1 = 2,\ \beta_2 = 0$		$\beta_1 = 2,\ \beta_2 = 0.7$		$\beta_1 = 2,\ \beta_2 = 1$	
	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$
$b = 0.05$	0.206824	0.107518	0.156789	0.117803	0.280216	0.120409
$b = 0.25$	0.0851329	0.747077	0.144204	0.0611628	0.0925360	0.0779742
$b = 1.0$	0.117942	0.05601970	0.0997025	0.0493194	0.0830523	0.0541475

Table 3.12. *Sample variances of estimated regression coefficients*

3.5. Bibliography

[ALI 96] Alioum, A. and Commenges, D., “A proportional hazards model for arbitrarily censored and truncated data”, *Biometrics,* 52, p. 512-524, 1996.

[CHE 95] Cheng, S. C., Wei, L. J. and Ying, Z., “Analysis of transformation models with censored data”, *Biometrika*, 82, p. 835-845, 1995.

[CLA 85] Clayton, D. and Cuzick, J., “Multivariate generalizations of the proportional hazards model” (with discussion), *J. Roy. Statist. Soc.* A 148, p. 82-117, 1985.

[CLA 86] Clayton, D. and Cuzick, J., “The semi-parametric Pareto model for regression analysis of survival times”, *Papers on Semi-Parametric Models* MS-R8614, p. 19-31. Centrum voor Wiskunde en Informatica, Amsterdam, 1986.

[COX 72] Cox, D. R., “Regression models and life tables” (with discussion), *Journal of the Royal Statistical Society, Series B* 34, p. 187-220, 1972.

[FIN 93] Finkelstein, D. M., Moore, D. F., and Schoenfeld , D. A., “A proportional hazards model for truncated AIDS data”, *Biometrics*, 49, p. 731-740, 1993.

[FRY 94] Frydman, H., “A note on non-parametric estimation of the distribution function from interval-censored and truncated observations”, *Journal of the Royal Statistical Society, Series B* 56, p. 71-74, 1994.

[HUB 04] Huber-Carol, C. and Vonta, F., “Frailty models for arbitrarily censored and truncated data”, *Lifetime Data Analysis*, 10, p. 369-388, 2004.

[HOU 84] Hougaard, P., “Life table methods for heterogeneous populations: Distributions describing the heterogenity”, *Biometrika*, 71, p. 75-83, 1984.

[HOU 86] Hougaard, P.,. “Survival models for heterogeneous populations derived from stable distributions”, *Biometrika*, 73, p. 387-396, 1986.

[KAL 89] KALBFLEISCH, J. D. AND LAWLESS, J. F., “Inference based on retrospective ascertainment: An analysis of data on transfusion-associated AIDS”, *Journal of the American Statistical Association*, 84, p. 360-372, 1989.

[TUR 76] TURNBULL, B.W., “The empirical distribution function with arbitrarily grouped, censored and truncated data”, *Journal of the Royal Statistical Society, Series B* 38, p. 290-295, 1976.

[VAU 79] VAUPEL, J. W., MANTON, K. G. AND STALLARD, E., “The impact of heterogenity in individual frailty on the dynamics of mortality”, *Demography* 16, p. 439-454, 1979.

[VON 96] VONTA, F., “Efficient estimation in a non-proportional hazards model in survival analysis”, *Scandinavian Journal of Statistics* 23, p. 49-61, 1996.

Chapter 4

Introduction of Within-area Risk Factor Distribution in Ecological Poisson Models

4.1. Introduction

We consider a class of ecological models related to environmental epidemiology in which the association between health outcomes and exposure is investigated. A set of geographical areas for which health outcome counts and exposure measurements are available. In the case of a rare disease and/or small geographical areas, the spatial variability is usually modeled using a hierarchical approach in which a Poisson distribution of the observed disease counts, Y_i, in area i is incorporated at the first level, and a log-linear relationship between the specific relative risk R_i and the exposure measurement Z_i is modeled at the second hierarchical level:

$$Y_i|R_i \sim \mathcal{P}(E_i R_i), \quad \log(R_i|\alpha) = Z_i'\alpha + \epsilon_i, \tag{4.1}$$

where E_i is the expected disease-incidence or death count in area i, possibly standardized on age and gender, and ϵ is the residual vector, with or without spatial structure. Different approaches to modeling residual spatial heterogenity have been reported in other works such: conditional dependence formulation as a Gaussian autoregressive process ([BES 91, CLA 92]) or direct parametric distribution on the residual covariance matrix ([WHI 54, BES 74, RIP 81, BES 99, WAK 99, GUI 02]). Many authors have discussed various forms of potential bias ("ecological" biases) that occur when making an inference on an individual association between disease risk and exposure based on aggregate data ([ELL 04, ELL 00, GRE 94]): firstly, a difference between the

Chapter written by Léa FORTUNATO, Chantal GUIHENNEUC-JOUYAUX, Dominique LAURIER, Margot TIRMARCHE, Jacqueline CLAVEL and Denis HÉMON.

aggregated and individual relationships may exist when the individual dose-response relationship depends non-linearly on the exposures of interest; secondly, the existence of unobserved risk factors, varying spatially or not ([RIC 92, CLA 93, PRE 95]), may introduce uncontrolled within- or between-area latent effects. We will focus on the bias arising from a non-linearity of individual relationship. Several authors ([RIC 87, RIC 96, PLU 96, LAS 00, WAK 01, JAC 06]) have investigated the relationships between the individual and aggregated forms of the risk. The case in which several exposure measurements are available per geographical area is addressed, these measurements being collected in various locations and not on the individuals. We propose a hierarchical model incorporating the within-area risk factor distribution and the integrated form of the relative risk. This approach will be compared with the classic ecological model (4.1). In order to take into account ecological latent factors independent of the studied factor, a spatial residual structure was incorporated in the models of extra-Poisson variability. A Bayesian approach was adopted because of the complexity of the model and the high number of parameters.

In section 4.2, we describe a general framework for ecological inference. Section 4.3 outlines the simulation framework. In section 4.4.1, we compare the parameter estimations for three estimation models. An analysis of the sensitivity to the Gaussian form of the within-area distribution is reported in section 4.4.2. We conclude with results obtained by applying the models to the ecological association between childhood leukemia incidences and domestic radon exposure in France ([EVR 05]).

4.2. Modeling framework

4.2.1. *Aggregated model*

Consider a domain divided into m geographical areas i. The observed number of cases in area i, Y_i, is considered to follow a Poisson distribution with mean E_iR_i where R_i is the relative risk of area i and E_i is the expected number of cases in area i, possibly standardized on age and gender. The second level consists of modeling the ecological relationship between the relative risk and risk factor. The regression model traditionally used in ecological analysis is the Clayton-Kaldor log-linear model ([CLA 87]):

$$\log(R_i) = \alpha_0 + \alpha_1 z_i + \epsilon_i \tag{4.2}$$

where α_1 is the ecological regression parameter, z_i is a measure of mean exposure to the risk factor Z within the geographical area i (empirical mean of the risk factor measurements within area i is generally used), and ϵ is the residual vector which is supposed to follow a Gaussian distribution with or without spatial structure. We consider a single risk factor Z (with mean μ_i and variance σ_i^2) uncorrelated with unmeasured risk factors. Let $H_i(.)$ be the distribution of Z in the geographical area i and $g(z)$ be the individual risk of disease D for individuals exposed at the same level

z. The expected aggregated risk for area i is:

$$E(g(Z)) = \int g(z)H_i(z)\,dz. \tag{4.3}$$

Thus, all individuals in area i share the same parametric "dose-effect" relationship given by g. We suppose a multiplicative risk model at the individual-level:

$$g(z) = \exp(\beta_0 + \beta_1 z) \tag{4.4}$$

where β_1 corresponds to the individual regression parameter. When $H_i = \mathcal{N}(\mu_i; \sigma_i^2)$, the expected aggregated risk in area i can be obtained explicitly from (4.3) and (4.4) by:

$$\exp(\beta_0 + \beta_1\mu_i + \frac{1}{2}\beta_1^2\sigma_i^2). \tag{4.5}$$

When several measurements n_i of risk factor Z are available in each geographical area i ($n_i > 1$), the proposed estimation model, hereafter referred to as the "complete" model, is as follows: if Z_{ik} is the k^{th} of the n_i measurements, then

$$\begin{cases} Y_i \sim \mathcal{P}(E_iR_i) \\ \log(R_i) = \alpha_0 + \alpha_1\mu_i + \alpha_2\sigma_i^2 + \epsilon_i \\ \alpha_2 = \frac{1}{2}\alpha_1^2 \\ Z_{ik} \sim \mathcal{N}(\mu_i; \sigma_i^2) \end{cases} \tag{4.6}$$

where μ_i and σ_i^2 will be estimated jointly with others parameters, from the n_i measurements Z_{ik}. This model (4.6) corresponds to the true aggregated model when H is Gaussian and was previously used by Best *et al.* (2001) [BES 01]. It will be compared to the classic model (4.2).

If we used the classic estimation model (4.2), the "true" mean, μ_i, is replaced by an empirical estimate without taking into account sampling fluctuations, and the individual parameter β_1 is different from the ecological parameter α_1 due to the absence of the within-area variance in the classic model (4.2). Moreover, the α_1 estimator from the classic model (4.2) will be an overestimation of β_1, as mentioned in [WAK 03], when the within-area means and variances are positively correlated. The β_1 estimator would require the first two moments of the within-area distribution of the risk factor in each area. [PLU 96] have suggested the following approximation of (4.5) by $\exp(\beta_0 + \beta_1\mu_i)$, when β_1 is "weak", the within-area variances are small, constant over the domain, or uncorrelated with the means of the risk factor. The within-area variance term will be either negligible or included in the constant term. The α_1 estimator will be an unbiased estimator of β_1.

If the underlying distribution is not Gaussian, the constraint $\alpha_2 = \frac{1}{2}\alpha_1^2$ may be inappropriate. The model is also considered without constraint on α_2. To investigate the consequences of misspecification of the within-area risk factor distribution, a Gamma distribution was used, with parameters $(\gamma_1; \gamma_2) = \left(\frac{\mu_i^2}{\sigma_i^2}; \frac{\mu_i}{\sigma_i^2}\right)$ such that $Z_i = (\mu_i; \sigma_i^2)$. The latter distribution differs from the Gaussian distribution with respect to symmetry and distribution tail, which constitute a relatively unfavorable case for the complete model (4.6). In this case, the true ecological model, as introduced in [WAK 01], differs from (4.6):

$$\begin{cases} Y_i \sim \mathcal{P}(E_i R_i) \\ \log(R_i) = \beta_0 - \frac{\mu_i^2}{\sigma_i^2} \log\left(1 - \beta_1 \frac{\sigma_i^2}{\mu_i}\right) + \epsilon_i \quad for \beta_1 \sigma_i^2 < \mu_i \\ Z_{ik} \sim G\left(\frac{\mu_i^2}{\sigma_i^2}; \frac{\mu_i}{\sigma_i^2}\right) \end{cases} \tag{4.7}$$

Using model (4.2) or (4.6) may therefore lead to a poor estimator of the individual parameter β_1, particularly when the ratio $\frac{\sigma_i^2}{\mu_i}$ is high (high relative variability of the risk factor).

Spatial structure was also incorporated in the model to take into account the global variability over the domain and/or the potential spatial structure between geographical areas. A joint distribution of residuals was used. A common parametric family models spatial dependence as a function of the distance d_{ij} between two areas i and j. Diggle *et al.* [DIG 98], for instance have proposed the following model: $A_\theta[i,j] = \exp\left(-\left(\theta d_{ij}\right)^\kappa\right)$ where θ is a scale parameter controlling the correlation decrease with distance, and κ is a parameter of smoothing magnitude set at 1 in the present study. The smoothing is then termed exponential smoothing. The three studied estimation models are restated in Table 4.1.

Classic model	$Y_i \sim \mathcal{P}(E_i R_i)$ $\log(R_i) = \alpha_0 + \alpha_1 z_i + \epsilon_i$ z_i: empirical mean $\epsilon \sim \mathcal{N}(0; \sigma_\epsilon^2 A_\theta)$
Complete model (A) (with constraint on α_2)	$Y_i \sim \mathcal{P}(E_i R_i)$ $\log(R_i) = \alpha_0 + \alpha_1 \mu_i + \frac{1}{2}\alpha_1^2 \sigma_i^2 + \epsilon_i$ $Z_{ik} \sim \mathcal{N}(\mu_i; \sigma_i^2)$ $\epsilon \sim \mathcal{N}(0; \sigma_\epsilon^2 A_\theta)$
Complete model (B) (without constraint on α_2)	$Y_i \sim \mathcal{P}(E_i R_i)$ $\log(R_i) = \alpha_0 + \alpha_1 \mu_i + \alpha_2 \sigma_i^2 + \epsilon_i$ $Z_{ik} \sim \mathcal{N}(\mu_i; \sigma_i^2)$ $\epsilon \sim \mathcal{N}(0; \sigma_\epsilon^2 A_\theta)$

Table 4.1. *Estimation Models*

4.2.2. *Prior distributions*

All the prior distributions were considered weakly informative. Parameters α followed a Gaussian distribution with mean zero and variance 100. For the prior distributions of the risk factor means and the parameter θ occurring in the residual autocorrelation matrix, a (proper) uniform distribution was chosen. Distribution supports were selected large enough in order to cover all the possible values. The choice IG(0.5;0.0005) was adopted for σ_ϵ^2 as suggested in [KEL 99]. With regard to within-area variance of the risk factor, a truncated Gaussian distribution with mean zero and variance 100 was selected, for convergence rate reasons. This distribution was consistent with the variances but sufficiently broad to remain vague. The results were compared with those obtained when prior distribution of within-area variance was IG(0.5;0.0005). The estimates of all the parameters were similar with the two priors.

4.3. Simulation framework

The performances were investigated using simulated datasets. Various situations were considered, in particular, the case of very probable ecological bias on β_1 ("strong" β_1 *and* within-area variances varying over the domain as a function of the risk factor means) and the case of a low probability of ecological bias on β_1 ("weak" β_1 *or* constant within-area variances, *or* independence of variances and means of risk factor). Six simulation frames were used taking into account the individual parameter was "strong" or "weak", and whether the within-area variances were constant, correlated with the risk-factor means or uncorrelated with them. For each frame, 100 independent datasets replications were simulated. The domain was a regular 10×10 grid ($m = 100$ areas of the same size). The distance between two adjacent areas was set to 70 km in order to obtain a range of distances consistent with those based on the French "départements". The means $\{\mu_i\}$ were randomly sampled from the observed means of the log radon (range=3.09-5.57). The number of measurements $\{n_i\}$ are also sampled from those of the radon campaigns (range=26-352). The within-area variances of Z were either constant over the domain and equal to 1, or they varied between areas (from 1 to 2.5). In the latter situation, two possibilities were considered: the variances were dependent or not on the exposure means. When dependent (referred to as the "mean-dependent case"), the variance values increased as a step function based on the risk factor means. The case in which the variances were independent of the means is referred to as the "random case". Mean empirical correlations (on the 100 datasets replications) were 0.07 and 0.79 for the "random case" and "mean-dependent case", respectively. The n_i measurements, Z_{ik}, of the risk factor in area i were simulated from the true within-area distribution (Gaussian with mean μ_i and variance σ_i^2, or Gamma with parameters $\left(\frac{\mu_i^2}{\sigma_i^2}; \frac{\mu_i}{\sigma_i^2}\right)$). The observed numbers of cases in the 100 geographical areas were simulated using a Poisson distribution with mean $E_i R_i$, in which the expected numbers of cases, $\{E_i\}$, were similar to the numbers of cases of

	$\overline{R}_2/\overline{R}_1$	$Var(R_2)/Var(R_1)$
"strong" parameter: $\beta_1 = 1$		
$\sigma^2 = 1$	1.7	2.7
$\sigma^2 =$ "random case"	2.5	8.2
$\sigma^2 =$ "mean-dependent case"	3.0	15.1
"weak" parameter: $\beta_1 = 0.3$		
$\sigma^2 = 1$	1.0	1.1
$\sigma^2 =$ "random case"	1.1	1.2
$\sigma^2 =$ "mean-dependent case"	1.1	1.5
	$R_{1i} = \exp\left(\beta_1\mu_i\right)$	$R_{2i} = \exp\left(\beta_1\mu_i + \frac{1}{2}\beta_1^2\sigma_i^2\right)$

Table 4.2. *Influence of the term $\frac{1}{2}\beta_1^2\sigma_i^2$ in the relative risk expression, when the within-area distribution is Gaussian*

childhood leukemia and ranged from 8.0 to 203.5 (average=40.6). The relative risks were simulated using the exact model defined by the expression (4.6) or (4.7) and a spatially structured residual with exponential smoothing (autocorrelation between two adjacent areas of 0.6). The values of β_1 were chosen so that the term $\frac{1}{2}\beta_1^2\sigma_i^2$ was or was not of influence in the expression (4.5), reflecting the definitions of "strong" and "weak" β_1, respectively. Two criteria were used for the two definitions and are shown in Table 4.2. The first criterion was the ratio between the mean relative risks on 100 areas with and without that term (first column), and the second criterion was the ratio between the variances of the relative risks with or without that term (second column). The values 0.3 and 1 were selected for "weak" and "strong" β_1, respectively. The residual variance σ_ϵ^2 and β_0 were determined to obtain reasonable relative risks: β_0= -1.7 and -6.4 (for "weak" and "strong" β_1) and σ_ϵ^2=0.1.

4.4. Results

Gibbs sampling ([GIL 96]) via WinBUGS software ([SPI 02b]) was used for Bayesian inferences in three models. Posterior summaries were based on 90,000 iterations for the classic model, and 300,000 and 150,000 iterations for the complete models A and B (with 10-lag), respectively, after a burn-in period of 10,000 iterations. Several criteria for convergence were checked (accessible in WinBUGS software). To compare the models, the biases and the coverages of the 95% credibility interval (95% CI) were calculated. Spielgelhalter *et al.* [SPI 02a] have suggested the Bayesian Deviance Information Criterion (DIC) as a general method for comparing complex models. In practice, if the difference between the DIC of the various models is greater than 5, the model with the smallest DIC is considered as fitting the data best, as suggested by Burnham and Anderson [BUR 98]. The results averaged on 100 datasets replications are presented in Table 4.3. Among the different simulation frames, only two of them will be presented in this section corresponding to a "strong association" with correlated within-area means and variances, where the within-area exposure distribution is Gaussian (section 4.4.1) or Gamma (section 4.4.2). Ecological bias is probable in these two cases. For the other cases where ecological bias is theoretically not possible (results not shown), biases of β_1 were very small as expected for the three models.

4.4.1. *Strong association between relative risk and risk factor, correlated within-area means and variances (mean-dependent case)*

In the 100 replications, the empirical correlation between within-area means and variances was 0.79, on average. In this situation, the term $\frac{1}{2}\beta_1^2\sigma_i^2$ is neither constant nor negligible.

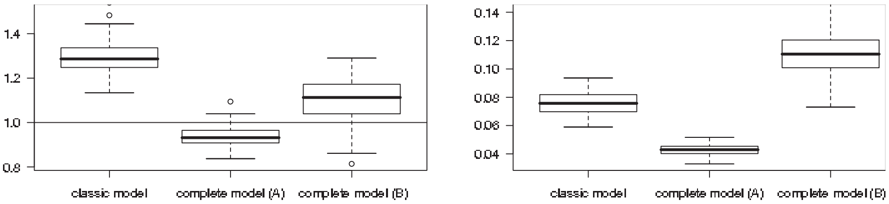


Figure 4.1. *(a) Left: Posterior means of α_1. (b) Right: Posterior sd of α_1. Mean-dependent case: the within-area distribution of the risk factor is Gaussian and the individual parameter β_1 is "strong" (true value=1)*

Table 4.3 and Figure 4.1a clearly show that the posterior mean of the ecological parameter α_1 with the classic model is a biased estimate of β_1 (mean bias about 30% with 95% confidence interval=[27.8;30.7]). However, although complete model (A), i.e. with the constraint $\alpha_2 = \frac{1}{2}\alpha_1^2$, tended to underestimate β_1, it very markedly improved the estimation both on average (mean bias about -6% with 95% confidence interval=[-7.2;-5.5]) and for each dataset. Furthermore, systematic decreases in the posterior standard deviation of α_1 and the range of its 95% CI were observed (Figure 4.1b). The true value of α_1 belonged to the 95% CI for 70 out of 100 replications with complete model (A), while with the classic model, only one 95% CI contained the true value of the parameter, showing a serious defect in the estimation. With regard to the between-area variability parameters, complete model (A) generally improved the estimations. The classic model overestimated the residual variance σ_ϵ^2 (mean bias of 36%), while complete model (A) tended to underestimate it to a lesser extent (mean bias of -6%). In addition, a considerable decrease in the posterior standard deviation, classic versus complete model (A), for the 100 replications was observed. The DIC showed a very slight systematic decrease (mean difference of about 2) between the classic model and complete model (A): for 16 datasets, complete model (A) seemed to fit better (difference of 14) while the DIC were equivalent for the other 84.

Parameters	$\mathbf{m}_{100}$[4]	$\mathbf{sd}_{100}$[4]	**%coverage**[5]	**mean bias** (%)[6]	**se bias** (%)[7]
α_1 (β_1=1)[1]					
classic model	1.30	0.08	1	29.3	0.74
complete model (A)[2]	0.94	0.04	70	-6.3	0.43
complete model (B)[3]	1.10	0.11	87	10.3	0.90
σ^2_ϵ (0.1)[1]					
classic model	0.136	0.051	87	35.9	4.64
complete model (A)[2]	0.094	0.046	87	-5.9	5.02
complete model (B)[3]	0.091	0.044	90	-8.6	4.74

[1] “True” values of parameters

[2] Complete model with the constraint $\alpha_2 = \frac{1}{2}\alpha_1^2$

[3] Complete model without the constraint $\alpha_2 = \frac{1}{2}\alpha_1^2$

[4] Average of 100 posterior means (m_{100}) and 100 posterior standard deviations (sd_{100})

[5] Percentage of 95% CI containing the true value of the parameter

[6] Average of 100 biases [100 (estimated value – “true” value)/“true” value]

[7] Standard error of mean bias

Table 4.3. *Simulations results when the true within-area distribution of the risk factor is Gaussian and the individual parameter β_1 is “strong” in the mean-dependent case*

Comparison of the estimates obtained with complete model (B), i.e. without constraint on α_2, and complete model (A) showed a slight increase in the mean bias of β_1 and its standard error: -6.3% (se=0.43) for complete model (A) versus 10.3% (se=0.90) for complete model (B). Moreover, there was a marked reduction in the precision of the estimate (Figure 4.1b): the posterior standard deviation increased threefold on average with complete model (B), with an increase in the 95% CI range. With regard to the overall fit of the model, no significant difference between the DIC of the complete models (A) and (B) was observed. In contrast, complete model (B) yielded markedly better estimates than those obtained with the classic model for parameter α_1 and for the variability parameters σ^2_ϵ and θ, without, however, there being any noteworthy difference between the DIC.

4.4.2. *Sensitivity to within-area distribution of the risk factor*

With regard to the Gamma distribution, when the individual parameter is “strong” ($\beta_1 = 1$) and the within-area variances are correlated with the means (mean-dependent case), the study of 20 dataset replications showed a mean ecological bias of 56% (se=0.48) with the classic model and 7% (se=0.25) with complete model (A). Moreover, estimation precision was markedly enhanced: the posterior standard deviation of complete model (A) was half that of the classic model. The gain in the number of 95% CI containing the true value was non-negligible with complete model (A), increasing from 0 to 11.95% CI for 20 datasets. For the other 9 datasets, the lower value of the 95% CI was relatively close to 1 (true value of β_1). Figure 4.2 clearly shows that the bias was systematically lower with complete model (A) and the 95% CI was much narrower.

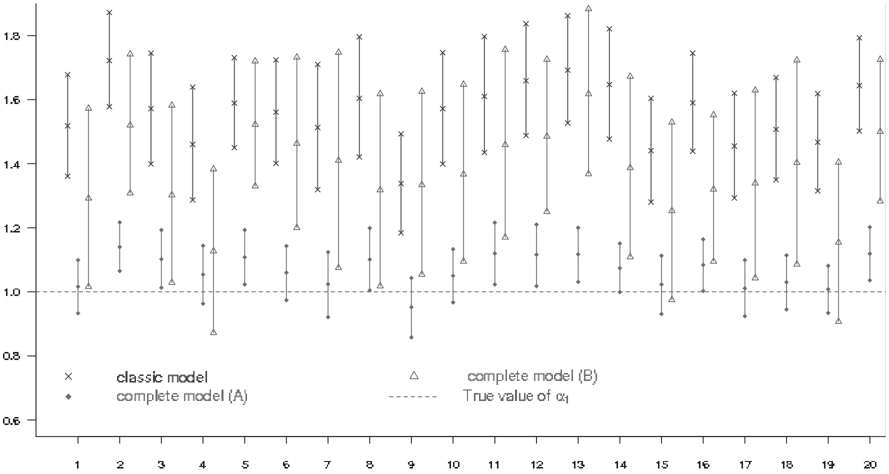


Figure 4.2. *Posterior mean and 95% CI of α_1 in the mean-dependent case (20 replications): the within-area distribution of the risk factor is Gamma, the individual parameter β_1 is "strong"*

The estimate of α_1 with complete model (B) was markedly inferior to that obtained with complete model (A). The mean bias was about 38% (se=0.62) and only three credibility intervals (95%) out of 20 contained the true value of β_1. However, the model was slightly superior to the classic model.

4.4.3. *Application: leukemia and indoor radon exposure*

This study focused on the ecological association between indoor radon exposure and childhood leukemia incidences in 94 areas ("départements") of France between 1990 and 1998. The epidemiological motivation and main results have been reported in the publication by Evrard *et al.* [EVR 05]. The observed cases of acute leukemia in children aged less than 15 years were retrieved from the French National Registry of Childhood Leukemia and Lymphoma ([CLA 04]) . Over the 9-year study period there were a total of 3,995 registered cases in our study region, including 692 cases of acute myeloid leukemia (AML) and 3,255 cases of acute lymphocytic leukemia (ALL). The expected cases are age- and gender-standardized. radon measurements were collected by the Institute for Radiation Protection and Nuclear Safety (IRSN). The measurement campaigns consisted of 12,170 measurements with an average of 130 measurements per “département”. The arithmetic mean radon concentrations ranged from 22 to 262 Bq/m^3 and thus presented with variability of interest.

	$\hat{\alpha}_1$[3]	sd[4]	95% CI	
All leukemia				
classic model	0.082	0.035	0.013	0.151
complete model (A)[1]	0.075	0.032	0.010	0.136
complete model (B)[2]	0.073	0.037	0.001	0.145
ALL				
classic model	0.057	0.039	-0.020	0.134
complete model (A)[1]	0.056	0.039	-0.019	0.132
complete model (B)[2]	0.046	0.040	-0.035	0.124
AML				
classic model	0.200	0.082	0.038	0.360
complete model (A)[1]	0.190	0.079	0.034	0.344
complete model (B)[2]	0.178	0.085	0.010	0.343

[1] complete model with the constraint $\alpha_2 = \frac{1}{2}\alpha_1^2$

[2] complete model without the constraint $\alpha_2 = \frac{1}{2}\alpha_1^2$

[3] posterior mean of α_1

[4] posterior standard deviation of α_1

Table 4.4. *Estimates of the ecological parameter α_1 for the radon-leukemia application data using the three models*

The association between indoor radon exposure and leukemia was assessed using the classic and complete (A and B) models. Radon exposure was log transformed because the within-area distribution on log scale was close to a Gaussian distribution. We assumed here a multiplicative (log-linear) model between relative risk and logarithm of radon. In this situation, complete model (A) is the appropriate model. Table 4.4 gives the posterior means of α_1 for the three models and their 95% CI for all types of leukemia, ALL and AML. The three models yielded concordant results: significant positive association for all leukemia and AML, and a non-significant association for ALL. However, complete model (A) resulted in a slightly narrower 95% CI. The estimate of α_1 was systematically smaller with complete model (B), compared to the two other models, while the 95% CI was wider. These findings were consistent with the results found in the simulation studies but to a lesser degree. The ecological parameter was much weaker, as was the correlation between the within-area means and variances of log-radon. In the simulated datasets replications, the correlation was 0.78

when the within-area variances were mean-dependent while the correlation was 0.46 for the real dataset (Figure 4.3). The incorporation of the within-area variance term in the model thus had less marked consequences than in the simulation frames. It is noteworthy that, particularly for all types of leukemia, incorporation of this variance term led to a slight decrease of the ecological association, which could be related to a bias reduction. However, the possibility that the ecological association could be due to some unknown confounding factors correlated with radon concentration cannot be excluded. There was no significant difference of the fit criterion and DIC between the three models.

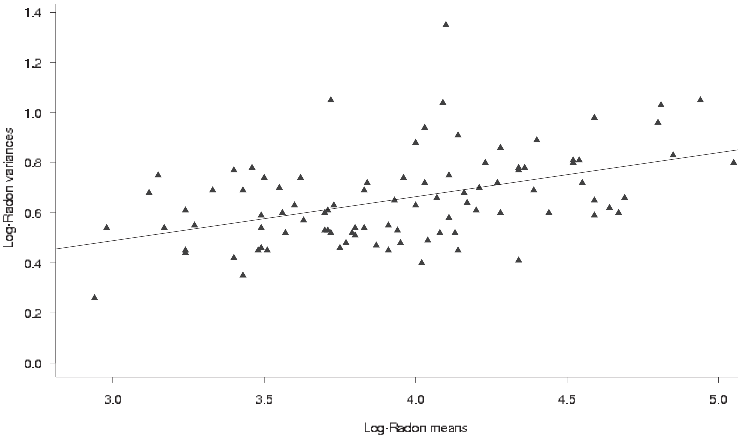


Figure 4.3. *Within-area variances and means of the log-radon*

4.5. Discussion

The consequences of incorporating the within-area risk factor variability in an ecological regression when several measurements of this factor per area were available have been studied. The performances of the three models, in various simulation frames, enabled determination of the situations in which it is important to include the within-area variability of the risk factor in the model in order to markedly reduce the bias in estimating the individual parameter β_1. If the individual parameter β_1 is "strong" and the within-area risk factor means and variances are correlated, the omission of the covariate $\left\{\sigma_i^2\right\}$ in the regression model leads to marked bias of the parameter associated with the risk factor. A very marked decrease in the ecological bias was observed with the complete model with the constraint $\alpha_2 = 1/2\alpha_1^2$, both on average (mean bias divided by 5) and with respect to bias variation. Additionally, this model allowed greater estimate precision. Overall, for complete model (A), the mean quadratic error (on 100 datasets) was 0.006 versus 0.091 for the classic model: the estimation of the individual parameter β_1 with complete model (A) seems markedly better. The complete model without constraint on α_2 also decreased the mean bias in comparison with the classic model, but markedly reduced estimate precision. With that model, the quadratic error was 0.020 and, thus, greater than that obtained with complete model (A), but a marked improvement on that obtained with the classic model. The marked reduction in the ecological bias of β_1 and the results for the 95% CI clearly demonstrate the importance of including the within-area variance term in the model. The previous results were based on 100 datasets replications. The observed differences between models were not due to random fluctuations and/or limited number of replications. For the mean-dependent case, mean differences between biases from classic and complete (A) models were 36% (95% confidence interval=[34.8;36.4]) and 17% (95% confidence interval=[14.8;18.3]) from complete (A and B) models, showing that the differences were highly statistically significant. Generally speaking, if the within-area distribution of the risk factor is Gaussian, the model including within-area variability never damages the estimations, even when inclusion is not necessary (situations in which bias theoretically does not exist).

The misspecification analysis addressed Gaussian distribution versus Gamma distribution. The classic model generated extremely-biased posterior means of the ecological parameter. In contrast, complete model (A) markedly improved the results. In particular, the mean quadratic error was 0.003 for complete model (A) versus 0.165 for the classic model. The complete model without the constraint on α_2 was less successful than the complete model with $\alpha_2 = \frac{1}{2}\alpha_1^2$, even though the latter is no longer the "correct" model in that situation. The improvement contributed by the constraint may be explained as follows. The second-order limited development of the log-relative risk expression in the ecological model (4.7) contains the term $\frac{1}{2}\beta_1^2$: $\log(R_i) \approx \beta_0 + \beta_1\mu_i + \frac{1}{2}\beta_1^2\sigma_i^2$. This development is justified only for small values of the ratio $\frac{\sigma_i^2}{\mu_i}$, which is often the case in our data. For example, in the mean-dependent case, the values of that ratio were lower than 0.53.

Acknowledgements: the authors thank Virginie Lasserre and Anne-Sophie Evrard for invaluable discussions and Solenne Billon for providing the radon data files. We are grateful to referees for valuable comments. This work was supported by grants from AFSSE (RD2004004) and INSERM-ATC (A03150LS).

4.6. Bibliography

[BES 74] BESAG J., "Spatial interaction and the statistical analysis of lattice systems", *Journal of Royal Statistical Society Series B*, vol. 36, p. 192-236, 1974.

[BES 91] BESAG J., YORK J., MOLLIÉ A., "Bayesian image restoration, with two applications in spatial statistics", *Annals of the Institute of Statistics and Mathematics*, vol. 43, p. 1-51, 1991.

[BES 99] BEST N., ARNOLD R., THOMAS A., WALLER L., CONLON E., "Bayesian models for spatially correlated disease and exposure data", *Bayesian Statistics 6*, J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith, (eds), Oxford University Press, Oxford, 1999.

[BES 01] BEST N., COCKINGS S., BENNETT J., WAKEFIELD J., ELLIOTT P., "Ecological regression analysis of environmental benzene exposure and childhood leukemia: sensitivity to data inaccuracies, geographical scale and ecological bias", *Journal of Royal Statistical Society Series A*, vol. 164, num. 1, p. 155-174, 2001.

[BUR 98] BURNHAM K. P., ANDERSON D. R., *Model Selection and Inference*, New York: Wiley, 1998.

[CLA 87] CLAYTON D., KALDOR J., "Empirical bayes estimates of age-standardized relative incidence rates for use in disease mapping", *Biometrics*, vol. 43, p. 671-681, 1987.

[CLA 92] CLAYTON D., BERNARDINELLI L., *Bayesian Methods for Mapping Disease Risk.*, English D, Elliott P, Cuzick J, Stern R (eds). Oxford University Press, Oxford, 1992.

[CLA 93] CLAYTON D., BERNARDINELLI L., MONTOMOLI C., "Spatial correlation in ecological analysis", *International Journal of Epidemiology*, vol. 22, num. 6, p. 1193-1202, 1993.

[CLA 04] CLAVEL J., GOUBIN A., AUCLERC M., AUVRIGNON A., WATERKEYN C., PATTE C., BARUCHEL A., LEVERGER G., NELKEN B., PHILIPPE N., SOMMELET D., VILMER E., BELLEC S., PERRILLAT-MENEGAUX F., HÉMON D., "Incidence of childhood leukemia and non-hodgkin's lymphoma in France – national registry of childhood leukemia and lymphoma, 1990-1999", *European Journal of Cancer Prevention*, vol. 13, p. 97-103, 2004.

[DIG 98] DIGGLE P., TAWN J., MOYEED R., "Model-based geostatistics", *Applied Statistics*, vol. 47, num. 3, p. 299-350, 1998.

[ELL 00] ELLIOT P., WAKEFIELD J., BEST N., BRIGGS D., *Spatial Epidemiology, Methods and Applications*, Oxford University Press, Oxford, 2000.

[ELL 04] ELLIOTT P., WARTENBERG D., "Spatial epidemiology: current approaches and future challenges", *Environmental Health Perspectives*, vol. 112, p. 998-1006, 2004.

[EVR 05] EVRARD A., HÉMON D., BILLON S., LAURIER D., JOUGLA E., TIRMARCHE M., CLAVEL J., "Ecological association between indoor radon concentration and childhood leukemia incidence in France, 1990-1998", *European Journal of Cancer Prevention*, vol. 14, p. 147-157, 2005.

[GIL 96] GILKS W., RICHARDSON S., SPIEGELHALTER D., *Markov Chain Monte Carlo in Practice*, Chapman & Hall, London, 1996.

[GRE 94] GREENLAND S., ROBINS J., "Invited commentary: ecologic studies–biases, misconceptions, and counterexamples", *American Journal of Epidemiology*, vol. 139, num. 8, p. 747-760, 1994.

[GUI 02] GUIHENNEUC-JOUYAUX C., RICHARDSON S., "Spatial regression: a flexible bayesian model for the autocorrelated error structure", *Proceedings of the 17th International Workshop on Statistical Modelling*, Chania, Greece, 2002.

[JAC 06] JACKSON C., BEST N., RICHARDSON S., "Improving ecological inference using individual-level data", *Statistics in Medicine*, vol. 25, num. 12, p. 2136-2159, 2006.

[KEL 99] KELSALL J., WAKEFIELD J., "Discussion of "Bayesian models for spatially correlated disease and exposure data" by Best, N.G., Arnold, R.A., Thomas, A., Waller, L.A. and Conlon, E.M.", *Bayesian Statistics 6*, J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith, (eds), Oxford University Press, Oxford, 1999.

[LAS 00] LASSERRE V., GUIHENNEUC-JOUYAUX C., RICHARDSON S., "Biases in ecological studies: utility of including within-area distribution of confounders", *Statistics in Medicine*, vol. 19, num. 1, p. 45-59, 2000.

[PLU 96] PLUMMER M., CLAYTON D., "Estimation of population exposure in ecological studies (with discussion)", *Journal of Royal Statistical Society Series B*, vol. 58, num. 1, p. 113-126, 1996.

[PRE 95] PRENTICE R., SHEPPARD L., "Aggregate data studies of disease risk factors", *Biometrika*, vol. 82, p. 113-125, 1995.

[RIC 87] RICHARDSON S., STUCKER I., HEMON D., "Comparison of relative risks obtained in ecological and individual studies: some methodological considerations", *International Journal of Epidemiology*, vol. 16, num. 1, p. 111-120, 1987.

[RIC 92] RICHARDSON S., *Statistical Methods for Geographical Correlation Studies*, English D, Elliott P, Cuzick J, Stern R (eds), Oxford University Press, Oxford, 1992.

[RIC 96] RICHARDSON S., GUIHENNEUC-JOUYAUX C., LASSERRE V., "Ecologic studies–biases, misconceptions, and counterexamples", *American Journal of Epidemiology*, vol. 143, num. 5, p. 522-523, 1996.

[RIP 81] RIPLEY B., *Spatial Stastistics*, John Wiley, New York, 1981.

[SPI 02a] SPIEGELHALTER D. J., BEST N., CARLIN B., VAN DER LINDE A., "Bayesian measures of model complexity and fit", *Journal of Royal Statistical Society Series B*, vol. 64, num. 3, p. 583-639, 2002.

[SPI 02b] SPIEGELHALTER D., THOMAS A., BEST N., CARLIN B., VAN DER LINDE A., *WinBUGS Version 1.4, User Manual*, Medical Research Council Biostatistics Unit, Cambridge, and Imperial College School of Medicine, London, UK, 2002.

[WAK 99] WAKEFIELD J., MORRIS S., "Spatial dependance and errors–in–variables in environmental epidemiology", *Bayesian Statistics 6*, J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith, (eds), Oxford University Press, Oxford, 1999.

[WAK 01] WAKEFIELD J., SALWAY R., "A statistical framework for ecological and aggregate studies", *Journal of Royal Statistical Society Series A*, vol. 164, num. 1, p. 119-137, 2001.

[WAK 03] WAKEFIELD J., "Sensitivity analyses for ecological regression", *Biometrics*, vol. 59, num. 1, p. 9-17, 2003.

[WHI 54] WHITTLE P., "On stationary process in the plan", *Biometrika*, vol. 41, p. 434-449, 1954.

Chapter 5

Semi-Markov Processes and Usefulness in Medicine

5.1. Introduction

Multi-states models play an important role in describing the evolution of a process of interest. Multi-states models naturally generalize the survival models, and deal not only with a final event (generally the death of patients in the medical field), but with different relevant events. A multi-state model is a stochastic process, namely $(X(t), t \geq 0)$, which at a given time t occupies one state $\{i\}$ from a discrete set $E = \{1, 2, \ldots, K\}$. It is characterized by its transition intensities from one state $\{i\}$ to another $\{j\}$ at time t conditional to its history F_{t^-}, namely $\lambda_{ij}(t|F_{t^-})$. Some hypotheses may be done on this transition structure, which leads us to the specific class of Markov and semi-Markov (SM) models. A non-homogenous Markov (NHM) process is defined by dependent functions of the chronological time, $\lambda_{ij}(t)$. We talk about a homogenous Markov (HM) model when the transition intensities are constant with time, λ_{ij}. Markov models have been extensively used in medicine ([KEI 89], [AAL 97], [JOL 98], [COM 02], [MAT 06]) and have delivered interesting results. However, another time scale reveals to be important in the medical field: the duration of time in the previous state. Indeed, generally, the longer patients stay in critical disease states, the more severe their evolution is. In this context, the SM model is well adapted. Let $T_{N(t^-)}$ be the time of the previous transition just before time t, and let

Chapter written by Eve MATHIEU-DUPAS, Claudine GRAS-AYGON and Jean-Pierre DAURÈS.

$x = t - T_{N(t^-)}$ be the time spent in the previous state. In non-homogenous semi-Markov (NHSM) models, the transition intensities $\lambda_{ij}(t, x)$ depend on both chronological and duration times. When the transition intensities depend only on the duration time, namely $\lambda_{ij}(x)$, we talk about homogenous semi-Markov (HSM) models. Figure 5.1 illustrates the generalization of SM models. SM models were first presented by [LEV 54], [SMI 55], [PYK 61], [PYK 64] and found applications in the counting theory [TAK 54]. Interest grew for this new class of stochastic processes [KOR 82], [ONI 83], [JAN 86], [LIM 99]. For a complete presentation of these models, see [LIM 01] and [JAN 99]. The SM models have been commonly used in the biological, medical, financial and insurance fields. For interesting references about their applications, see [GRU 89], [STE 99], [JOL 99], [DAV 00] [PAP 99], [PAP 98], [VAS 92], [JAN 01].

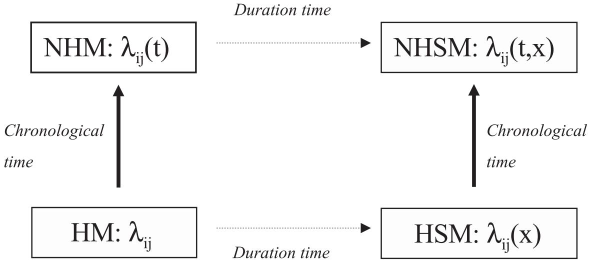


Figure 5.1. *Extension of the HM process*

The objective of this chapter is to present SM models and their usefulness in medicine. We want to show how, from these modelings, some relevant indicators may be constructed. New decision tools may then be given to the public health sector or patient care. This chapter is organized as follows: in the next section, the model and associated notation are introduced and health indicators are constructed. An application to HIV control is considered in section 5.3. Section 5.4 illustrates an application to the breast cancer. Finally, section 5.5 is a summary and discussion.

5.2. Methods

5.2.1. *Model description and notation*

Let us consider $E = \{1, \ldots, K\}$ a discrete state space and $(\Omega, \mathcal{F}, P)$ a probability space. We define the following random variables [JAN 01], $\forall n \geq 0$:

$$J_n : \Omega \to E, \qquad S_n : \Omega \to [0, +\infty),$$

where J_n represents the state at the n-th transition and S_n represents the chronological time of the n-th transition. Let $N(t)$ be the counting process $(N(t), t \geq 0)$ associated with the point process $(S_n)_{n\in \mathrm{N}}$ defined for any time $t \geq 0$ by:

$$N(t) = \sup \{n : S_n \leq t\}.$$

The random variable $N(t)$ represents the number of transitions that occurred in the interval of time $[0, t]$. Let us define the $(X_n)_{n\in \mathrm{N}}$ duration process by:

$$\begin{aligned} &X_0 = 0, \\ &X_{n+1} = S_{n+1} - S_n, \quad n \in \mathrm{N}^* \end{aligned}$$

where X_{n+1} represents the duration time spent in state J_n.

The $(J_n, S_n)_{n\in \mathrm{N}}$ process is called an NHM renewal process if

$$\begin{aligned} P(J_{n+1} = j, S_{n+1} \leq t | J_n = i, S_n = s, J_{n-1}, S_{n-1}, \ldots, J_0, S_0) \\ = P(J_{n+1} = j, S_{n+1} \leq t | J_n = i, S_n = s). \end{aligned}$$

For $j \neq i$, the following transition structure

$$Q_{ij}(t, x) = P(J_{N(t)+1} = j, X_{N(t)+1} \leq x | J_{N(t)} = i, S_{N(t)} = t).$$

is the associated NHSM kernel Q. The second component of Q, namely x, represents the duration time whereas t represents the chronological time. As is well known [WAD 92],

$$\begin{aligned} p_{ij}(t) &= \lim_{x\to\infty} Q_{ij}(t, x), \quad i, j \in E, j \neq i \\ &= P(J_{N(t)+1} = j | J_{N(t)} = i, S_{N(t)} = t), \end{aligned}$$

represents the probability of the process making its next transition to state j, given that it entered state i at time t, and $\mathbf{P}(t) = [p_{ij}(t)]_{i,j}$ is the $(K \times K)$ transition probability matrix of the embedded NHM chain $(J_n)_{n\in \mathrm{N}}$. However, before the entrance into j, the process "holds" for a time x in state i. The conditional cumulative distribution function of the waiting time in each state, given the state subsequently occupied, is defined by:

$$F_{ij}(t, x) = P(X_{N(t)+1} \leq x | J_{N(t)+1} = j, J_{N(t)} = i, S_{N(t)} = t).$$

This probability function is obtained by

$$F_{ij}(t, x) = \begin{cases} \frac{Q_{ij}(t,x)}{p_{ij}(t)} & \text{if } p_{ij}(t) \neq 0 \\ 1 & \text{if } p_{ij}(t) = 0 \end{cases}$$

Let $f_{ij}(t, x)$ be the probability density function of the waiting time, and $\mathbf{D}(t, x) = [f_{ij}(t, x)]_{i,j}$ be the $(K \times K)$ duration matrix. Let us introduce the probability that the process stays in state i for at least a duration time x, given state i has entered at chronological time t :

$$H_i(t,x) = P(X_{N(t)+1} \leq x | J_{N(t)} = i, S_{N(t)} = t) = \sum_{j \neq i}^{K} Q_{ij}(t,x).$$

Therefore, the marginal cumulative distribution function of the waiting time in each state depends on both times. Let us define $S_i\,(t,x) = 1 - H_i(t,x)$. Now it is possible to define the continuous time non-homogenous semi-Markov process $Z(t)$ [COX 80], [JAN 86], as:

$$Z(t) = J_{N(t)}, \quad t \in \mathrm{R}_+$$

with:

$$P[Z(t) = j] = P[S_{N(t)} \leq t < S_{N(t)+1}, J_{N(t)} = j].$$

The process $Z(t)$ represents the state occupied by the process at time t. This SM process is characterized both by a set of Markov transition matrices $\{\mathbf{P}(t)\}_{t \geq 0}$ and a set of duration matrices $\{\mathbf{D}(t,x)\}_{x \geq 0}$. The chronological time, namely t, is relative to an arbitrary origin. The internal time, namely x, is relative to the duration time in each state [DAV 00]. The SM model presented in this section is non-homogenous with time since the jump and duration processes depend on the chronological time.

In this non-homogenous context, transition intensities are defined by $\forall j \neq i$:

$$\lambda_{ij}(t,x) =$$

$$\lim_{h \to 0} \frac{P\left[x \leq X_{N(t)+1} < x+h; J_{N(t)+1} = j | X_{N(t)+1} \geq x; J_{N(t)} = i; S_{N(t)} = t\right]}{h},$$

$$\Lambda_{ij}(t,x) = \frac{p_{ij}(t)}{S_i(t,x)} \times f_{ij}(t,x).$$

Let us define the cumulative transition intensity from state i to state j, given the state i has entered at time t and the duration time in state i equals x:

$$\Lambda_{ij}(t,x) = \int_0^x \lambda_{ij}(t,\tau) d\tau.$$

The global cumulative transition intensity from state i to state j is given by:

$$\Lambda_i(t,x) = \sum_{j=1}^{K} \Lambda_{ij}(t,x).$$

The expressions of the SM kernel according to the previous transition intensities are the following: $\forall n \in \mathrm{N}$, $\forall i, j \in E$, such that $j \neq i$,$\forall x \in \mathrm{R}_+$:

$$\begin{aligned} Q_{ij}(t,x) &= P[J_{N(t)+1} = j, X_{N(t)+1} \leq x | J_{N(t)} = i, S_{N(t)} = t], \\ &= \int_0^x P[X_{N(t)+1} > \tau | J_{N(t)} = i, S_{N(t)} = t] \\ &\times P[J_{N(t)+1} = j, X_{N(t)+1} \in (\tau, \tau + d\tau) | J_{N(t)} = i, X_{N(t)+1} > \tau, S_{N(t)} = t], \end{aligned}$$

hence:

$$Q_{ij}(t,x) = \int_0^x \exp[-\Lambda_i(t,\tau)]\lambda_{ij}(t,\tau)d\tau. \tag{5.1}$$

$$Q^{'}_{ij}(t,x) = p_{ij}(t)f_{ij}(t,x) = \exp[-\Lambda_i(t,x)]\lambda_{ij}(t,x). \tag{5.2}$$

5.2.2. *Construction of health indicators*

5.2.2.1. *Interval transition probabilities for predictions*

For more effective patient care, physicians need tools of prediction and reference points. Let us define $\phi_{ij}(t,x)$ as the interval transition probability of the SM process [PAP 99] $\forall i, j = 1, \ldots\ldots, K$:

$$\begin{aligned} \phi_{ij}(t,x) &= P\,[\text{the NHSM process is in state } j \text{ at time } t + x \mid \text{it entered state } i \text{ at time } t] \\ &= P\left[Z(t+x) = j \mid J_{N(t)} = i\ ;\ S_{N(t)} = t\right] \end{aligned}$$

Let $Q'_{il}(t,x)$ be the product $p_{il}(t)f_{il}(t,x)$. With careful reasoning, we can prove that $\forall t, x \geq 0$:

$$\phi_{ij}(t,x) = \delta_{ij} \times S_{i.}(t,x) + \sum_{\substack{l=1 \\ l\neq i}}^{K} \int_0^x Q'_{il}(t,u)\phi_{lj}(t+u, x-u)du. \tag{5.3}$$

This equation represents the evolution equation of a continuous NHSM model. Obviously $\phi_{ij}(t,0)$=0 for $j \neq i$, 1 otherwise. If k is the index of the number of transitions in the interval of time $]t, t+x[$, then the previous equation is written as follows: $\forall i, j \in \{1, \ldots, K\}$,

$$\phi_{ij}(t,x) = \sum_{k=0}^{\infty} \phi_{ij}^k(t,x) \tag{5.4}$$

where

$$\begin{aligned}\phi_{ij}^{k}(t,x) = P[&\text{the process is in state } j \text{ at time } t+x;\\ &k \text{ transitions during the interval }]t,t+x[\\ &\mid \text{it entered state } i \text{ at time } t]\end{aligned}$$

and $\phi_{ij}^{0}(t,x) = \delta_{ij} \times S_i(t,x)$.

It can be expressed in a matrix form, with $\mathbf{\Phi}(t,x) = (\phi_{ij}(t,x))_{i,j}$ and $\mathbf{\Phi}^k(t,x) = \left(\phi_{ij}^{k}(t,x)\right)_{i,j}$:

$$\mathbf{\Phi}(t,x) = \sum_{k=0}^{\infty} \mathbf{\Phi}^k(t,x). \tag{5.5}$$

The interval transition probabilities make it possible to make predictions for patients in the medical practice. In order to make long term predictions, we use a Monte Carlo algorithm for realizing NHSM trajectories. The algorithm is based on the embedded Markov chain and gives realizations of the SM process into the interval of time $[0,C]$. It is based on five steps:

1) let $k = 0$, $S_0 = 0$, and sample j_0 according to the initial distribution;

2) sample the next state $J \sim p_{j_k.}(s_k)$ and set $j_{k+1} = J(w)$;

3) sample the sojourn time in j_k before the transition to j_{k+1}: $X \sim F_{j_k j_{k+1}}(s_k,.)$ and set $x = X(w)$;

4) let $k := k+1$ and set $s_k = s_{k-1} + x$. If $s_k \geq C$ then end;

5) set $j_k = j_{k+1}$ and continue to step 2.

The output of the algorithm will be the successive visited states and jump times, namely $(j_0, s_0, \ldots, j_k, s_k, \ldots)$.

5.2.2.2. *Distribution of probability at a chronological time t*

5.2.2.2.1. Definition

The probability to be in a subset of E at the chronological time t may lead to public health indicators as prevalence. Let us consider three subsets of E: $V = \{v_l\}_{l\in \mathrm{N}}$, $U = \{u_m\}_{m\in \mathrm{N}}$ and $W = \{w_n\}_{n\in \mathrm{N}}$ with $V \subset U$. Let us define the probability of interest:

$$\begin{aligned}\Psi_{V/U,W}(t_0,t) &= P\,[\text{NHSM process is in } V \text{ at time } t\ .\\ &\quad \mid \text{process is in } U \text{ at time } t \text{ and it entered in } W \text{ at time } t_0]\,,\\ &= P\left[Z(t) \in V \mid Z(t) \in U; J_{N(t_0)} \in W\ ;\ S_{N(t_0)} = t_0\right], \\ &= \tfrac{P\left[Z(t)\in V \mid J_{N(t_0)}\in W\ ;S_{N(t_0)}=t_0\right]}{P\left[Z(t)\in U \mid J_{N(t_0)}\in W\ ;S_{N(t_0)}=t_0\right]},\end{aligned} \tag{5.6}$$

From the definition of the interval transition probabilities, we obtain:

$$\Psi_{V/U,W}\left(t_0,t\right) = \frac{\phi_{\mathbf{W},\mathbf{V}}(t_0,t-t_0)}{\phi_{\mathbf{W},\mathbf{U}}(t_0,t-t_0)}$$

with

$$\begin{aligned}\phi_{\mathbf{W},\mathbf{V}}\left(t_0,t-t_0\right) &= \sum_l\sum_n \phi_{w_n;v_l}(t_0,t-t_0)\\ &= \sum_l\sum_n\sum_{k=0}^{\infty} \phi^k_{w_n;v_l}(t_0,t-t_0)\end{aligned}$$

and

$$\begin{aligned}\phi_{\mathbf{W},\mathbf{U}}\left(t_0,t-t_0\right) &= \sum_n\sum_m \phi_{w_n;u_m}(t_0,t-t_0)\\ &= \sum_n\sum_m\sum_{k=0}^{\infty} \phi^k_{w_n;u_m}(t_0,t-t_0)\end{aligned}$$

5.2.2.2.2. Adaptation to the notion of prevalence in public health

In a medical context, the previous probability (5.6) leads us to the specific notions of prevalence defined by [GAI 99] [KEI 91]. Let us consider the illness death model in Figure 5.2. The health state at time t is an NHSM process with the kernel structure $Q'_{ij}(t,x) = \exp[-\Lambda_i(t,x)]\lambda_{ij}(t,x)$. The chronological time represents age.

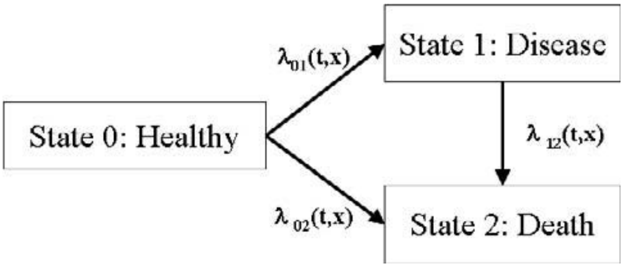


Figure 5.2. *NHSM process of the illness-death model*

The age-specific total prevalence [GAI 99] [KEI 91] refers to all persons in a given population diagnosed in the past with cancer and is alive at a determinate age. It may be expressed as follows

$$\begin{aligned}\pi\left(t\right) &= P\left(\text{a subject at age } t \text{ is diseased} \mid \text{he is alive at age t and healthy at birth}\right).\\ &= P\left[Z\left(t\right)\in V \mid Z\left(t\right)\in U; J_{N(t_0)}\in W\ ;\ S_{N(t_0)}=t_0\right],\end{aligned} \tag{5.7}$$

where V is the disease state (state 1), U represents all the alive state (state 0 and state 1) and W is the healthy state (states 0). The origin $t_0 = 0$ represents the birth of the individual. From (5.7), we obtain

$$\pi\left(t\right) = \frac{\phi_{0;1}(0,t)}{\phi_{0;0}(0,t)+\phi_{0;1}(0,t)}. \tag{5.8}$$

The age-specific partial prevalence at age t is defined by

$$\begin{aligned}\pi\left(t,x\right) \quad &= P\left(\text{a subject at age } t \text{ is diseased and diagnosed in } \left[t-x,t\right]\right.\\ &\qquad\qquad\left.\mid \text{he is alive at age } t \text{ and healthy at birth}\right) .\end{aligned}$$

Finally, the age-specific partial prevalence may be completely expressed according to the characteristics of the SM kernel:

$$\pi\left(t,x\right) = \frac{\Gamma_{0;1}(0,t,x)}{\phi_{0;0}(0,t)+\phi_{0;1}(0,t)}, \tag{5.9}$$

with

$$\Gamma_{0;1}(0,t,x) = \int_{t-x}^{t} Q'\left(0,x_1\right) S_1\left(x_1,t-x_1\right) dx_1.$$

The prevalence and the age-specific prevalence used in [GAI 99] [KEI 91] are in fact indicators derived from an NHSM kernel (see (5.8) and (5.9)). Section 5.4 will consider an application to breast cancer.

5.3. An application to HIV control

5.3.1. *Context*

SM processes are well adapted to model the evolution of HIV-1 infected patients and have already been used in this context [WIL 94], [SAT 99], [JOL 99], [FOU 05]. These models were homogenous with time and yet the follow-up time probably has an impact on the patient's evolution. With that in mind, we present a NHSM model of the HIV biological process [MAT 07] and give different indicators of the patients' health state. Infection by HIV has two fundamental markers: the viral load and the CD4 lymphocyte count. Based both on currently information and physicians' opinion, four immunological and virological states are considered (Figure 5.3). The process of interest Z_t represents the immuno-virological state of patients at time t, with $E = \{1,2,3,4\}$. During their follow-up, the patients experience immunological and virological changes and move through the four states according to the ten transitions given in Figure 5.3. The time t is considered as continuous and the origin $t = 0$ represents the first immuno-virological measure in the hospital. A total of 1,313 patients were considered. The amount of measurements is 17,888 (7% for state 1; 48% for state 2; 34.5% for state 3 and 10.5% for state 4).

5.3.2. *Estimation method*

Over a period of time $[0,C]$, M patients are observed ($p = 1,\ldots,M$). Let us assume that the p^{th} subject changes state n_p times in the instants $s_{p,1} < s_{p,2} <$

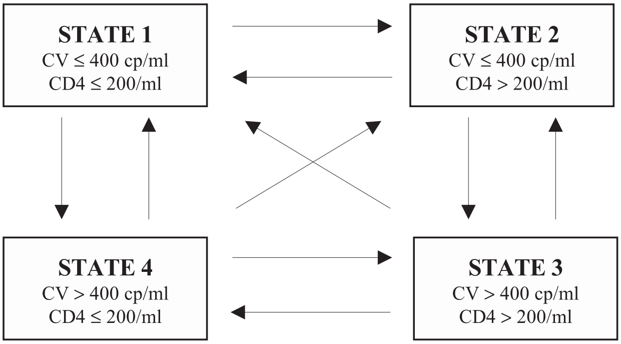


Figure 5.3. *An HIV multi-state model*

$\cdots < s_{p,n_p}$ and successively occupies states $J_{p,1}$, $J_{p,2}$,…,J_{p,n_p} with $J_{p,n} \neq J_{p,n+1}$, $\forall n \geq 1$. For more feasibility, we suppose that the NHSM process of interest is characterized by the kernel $Q_{ij}(t,x) = p_{ij}(t)F_{ij}(x)$. In this context, the contribution for an observed transition $i \longrightarrow j$, after a duration time x spent in state i, equals $p_{ij}(t)F_{ij}(x)$, namely $P[$duration time $= x; next = j|$ state i is entered at time $t]$. If the transition from state i is right-censored, after a staying time x, then the contribution is the function $S_i(t,x)$. The likelihood function for all times and transition times observed, is written as follows

$$L = \prod_{p=1}^{M}\prod_{n=1}^{n_p}[p_{J_{p,n-1},J_{p,n}}(s_{p,n-1}) \times f_{J_{p,n-1},J_{p,n}}(s_{p,n-1}, s_{p,n} - s_{p,n-1})]$$

$$\times S_{J_{p,n_p}}(s_{p,n_p},\ C - s_{p,n_p})$$

We use a parametric estimation method which consists of a linear jump process and a Weibull duration process respectively defined by:

$$\begin{array}{ll} p_{ij}(t|\theta_{ij}) = a_{ij}t + b_{ij} & \forall j \neq i \\ p_{ii}(t) = 0 & \forall i = 1, \ldots, 4 \end{array} \tag{5.10}$$

and

$$f_{ij}(x|\gamma_{ij}) = \nu_{ij}\sigma_{ij}^{\nu_{ij}} x^{\nu_{ij}-1} Exp[-(\sigma_{ij}x)^{\nu_{ij}}] \quad \forall j \neq i \tag{5.11}$$

where $\theta_{ij} = (a_{ij}, b_{ij})'$ and $\gamma_{ij} = (\nu_{ij}, \sigma_{ij})$. Regarding the jump process, the linear modeling offers the best adequacy and corresponds to the empirical estimation of jump probabilities. The maximum likelihood estimation has to be performed under constraints so that the probability functions would be bounded.

5.3.3. *Results: new indicators of health state*

The estimations of the parameters and their standard errors are given in Table 5.1. Mathematical computing was preformed on R software version 1.9.1 and simulations for the NHSM model are performed on the `Mathematica` software.

Therefore, the trend of the evolution of the health state of patients can be evaluated by estimating the 4×4 matrices of the interval transition probabilities for distinct chronological times t and duration times x. For example, the estimations of $\mathbf{\Phi}(0,1)$, $\mathbf{\Phi}(0,3)$ and $\mathbf{\Phi}(0,5)$ are respectively given in Tables 5.2, 5.3 and 5.4.

Transition	Estimators of the duration process	Estimators of the jump process
$i \to j$	$f_{ij}(x)$	$p_{ij}(t)$
$1 \to 2$	$Weibull$ (1.1069 , 1.6795)	$(0.0450 \times t)$+ 0.4748
	$std = (\pm 0.0511 , \pm 0.0996)$	$std = (\pm 0.0129 , \pm 0.0319)$
$1 \to 3$	$Weibull$ (1.4460 , 1.8283)	0.1111
	$std = (\pm 0.1373 , \pm 0.1811)$	$std = (\pm 0.0138)$
$1 \to 4$	$Weibull$ $(1.0971, 1.7254)$	$(-0.0450 \times t) - 0.4141$
	$std = (\pm 0.0642 , \pm 0.1361)$	$std = (\pm 0.0129 , \pm 0.0457)$
$2 \to 1$	$Weibull$ (0.5878 , 0.0940))	$(-0.0213 \times t)$+ 0.3148
	$std = (\pm 0.0426 , \pm 0.0187)$	$std = (\pm 0.0098 , \pm 0.0235)$
$2 \to 3$	$Weibull$ $(1.0500, 0.8844)$	$(0.0213 \times t) + 0.6852$
	$std = (\pm 0.0291 , \pm 0.0382)$	$std = (\pm 0.0098 , \pm 0.0235)$
$3 \to 2$	$Expo$ (1.0841)	0.8496
	$std = (\pm 0.0351)$	$std = (\pm 0.0099)$
$3 \to 4$	$Weibull$ $(0.7842, 0.7597$)	0.1504
	$std = (\pm 0.0456 , \pm 0.0891)$	$std = (\pm 0.0099)$
$4 \to 1$	$Weibull$ $(0.9095, 1.0556)$	$(-0.0276 \times t)$+ 0.4779
	$std = (\pm 0.0410 , \pm 0.0727)$	$std = (\pm 0.0122 , \pm 0.0236)$
$4 \to 2$	$Weibull$ $(1.1866, 1.5765)$	0.1605
	$std = (\pm 0.0847 , \pm 0.1401)$	$std = (\pm 0.0141)$
$4 \to 3$	$Expo$ (1.8410)	$(0.0276 \times t)$+ 0.3616
	$std = (\pm 0.1221)$	$std = (\pm 0.0122 , \pm 0.0377)$

Table 5.1. *Estimations of parameters and standard errors in the NHSM process defined by the linear jump process $\{p_{ij}(t)\}_{i,j}$ and the duration process $\{f_{ij}(x)\}_{i,j}$.*

Let U be the subset of the successful virological states $U = \{1, 2\}$ and D be the subset of the failure states (high VL states), namely $D = \{3, 4\}$. Let us consider specific predictions for each group of patients. Firstly, the global probability of "failure" is quite stable over five years for patients initially in U. Indeed we have $\phi_{1D}(0, 1) = 0.385$, $\phi_{1D}(0, 3) = 0.384$, $\phi_{1D}(0, 5) = 0.341$ and $\phi_{2D}(0, 1) = 0.293$,

state j state i	1	2	3	4
1	0.212	0.403	0.213	0.172
2	0.036	0.671	0.265	0.028
3	0.027	0.444	0.481	0.048
4	0.144	0.289	0.252	0.315

Table 5.2. *The* 4×4 *interval transition matrix* $(\phi_{ij}(0,1))_{i,j}$.

state j state i	1	2	3	4
1	0.058	0.558	0.303	0.081
2	0.041	0.632	0.284	0.043
3	0.039	0.577	0.331	0.053
4	0.068	0.535	0.301	0.096

Table 5.3. *The* 4×4 *interval transition matrix* $(\phi_{ij}(0,3))_{i,j}$.

$\phi_{2D}(0,3) = 0.327$, $\phi_{2D}(0,5) = 0.311$. Secondly, if we consider patients who are initially in “failure” states, we can note that their probability of staying in failure decreases ($\phi_{3D}(0,1) = 0.529$, $\phi_{3D}(0,3) = 0.384$, $\phi_{3D}(0,5) = 0.344$ and $\phi_{4D}(0,1) = 0.567$, $\phi_{4D}(0,3) = 0.397$, $\phi_{4D}(0,5) = 0.36$) to the benefit of transitions to successful states. These results suggest the improvement in the health status of patients with the follow-up time.

The notions of availability and reliability give global indicators to comprehend the entire cohort of patients. The availability function $A(t) = P[ZtU]$ suggests a good evolution of the cohort from the first year of the following time, since we have $A(1) = 0.51, A(3) = 0.62, A(5) = O.65$. The reliability function represents the time of the first virological failure occurred in the hospital for patients who are initially in successful states (Figure 5.4). The first virological failure occurs during the first two years of the following time. Indeed, we have $R(0.5) \simeq 0.6$, $R(1) \simeq 0.5$, $R(2) \simeq 0.3$.

state j state i	1	2	3	4
1	0.041	0.618	0.294	0.047
2	0.038	0.651	0.271	0.040
3	0.038	0.618	0.298	0.046
4	0.044	0.596	0.303	0.057

Table 5.4. *The* 4×4 *interval transition matrix* $(\phi_{ij}(0,5))_{i,j}$.

However, it is interesting to note that at least 17% of patients experience no failure during the 5 first years of their following ($R(5) \simeq 0.17$).

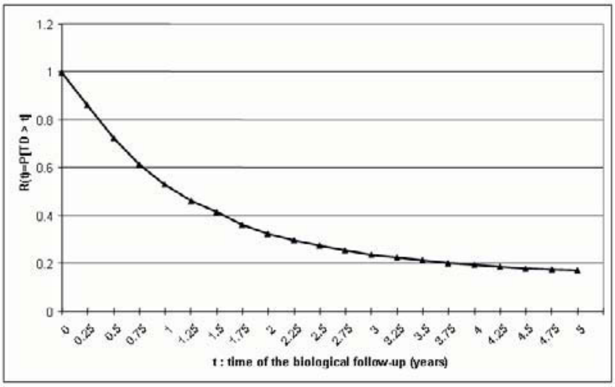


Figure 5.4. *Graph of the survival function relative to the 1st virological failure*

5.4. An application to breast cancer

5.4.1. *Context*

Breast cancer is the most common cancer for women in developed countries. Incidence and survival are not sufficient factors when the objective is the evaluation of health care needs. Indeed, the knowledge of cancer prevalence rates according to the stages of the disease provides more precise indications of the healthcare system. The duration times reveal to be crucial in the cancer modelings. Indeed, the longer the time since diagnosis, the less health care services are required by prevalent patients. In this study, data on breast cancer are provided by the Hérault (France) Cancer Registry which is a population-based registry. All newly diagnosed cases of cancer occurring among residents of Hérault and their follow-up are notified to the registry. The period of observation is five years (1994-1999). Breast cancer has two main stages that leads to different prognoses: the loco-regional stage that is linked to a slower evolution of the disease and the distant stage to a quick evolution of the disease. Let us assume that the disease is irreversible, that is, persons who have cancer remain prevalent cases during the rest of their life. Considering the natural history of breast cancer we use an NHM model, namely Z_t with the state space $E = \{0, 1, 2, 3\}$ characterized in Figure 5.5. The chronological time t is the age of the patients and the origin $t = 0$ represents the birth.

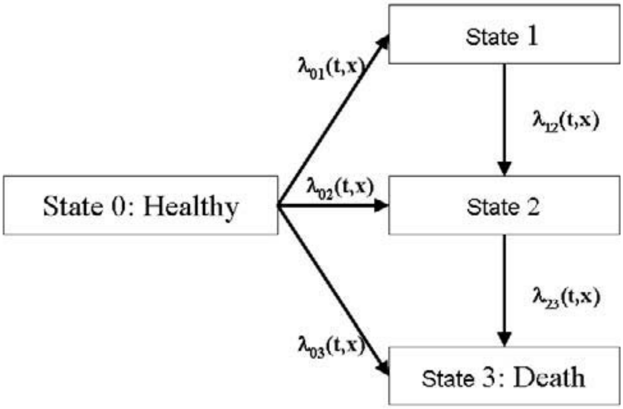


Figure 5.5. *Compartmental representation of the irreversible process of first state, second state of the disease and death*

5.4.2. *Age and stage-specific prevalence*

The age and stage-specific prevalence, namely $\pi_i\,(t,z)$, represents, at a determinate date, the proportion of people diagnosed in the past with cancer and alive at a specific stage $\{i\}$ of the disease. Then, it is formulated as follows

$$\begin{aligned}\pi_i\,(t)\; &= P(\text{a subject at age } t \text{ diseased in the state } \{i\}\\ &\quad \mid \text{he is alive at age } t \text{ and healthy at birth}),\\ &= P\left[Z\,(t) \in V \mid Z\,(t) \in U; J_{N(t_0)} \in W\;;\; S_{N(t_0)} = t_0\right],\end{aligned}$$

with $V = \{i\}, U = \{0,1,2\}, W = \{0\}$ and $t_0 = 0$. Let us consider the age-specific prevalence in the state $\{2\}$, namely $\pi_2\,(t)$ can be built as follows

$$P\left[Z\,(t) \in \{2\} \mid J_0 \in \{0\}\;;\; S_{N(0)} = 0\right] \;= \phi^1_{0;2}(t,x) + \phi^2_{0;2}(t,x)\,,$$

where

$$\phi^1_{0;2}(0,t) \;= \textstyle\int_0^t \exp\left[-\Lambda_0\,(0,x_1)\right]\lambda_{02}\,(0,x_1)\exp\left[-\Lambda_2\,(x_1,t-x_1)\right]dx_1\,,$$

and

$$\begin{aligned}\phi^2_{0;2}(0,t) \;= \textstyle\int_0^t\int_0^{t-x_1} &\exp\left[-\Lambda_0\,(0,x_1)\right]\lambda_{01}\,(0,x_1)\exp\left[-\Lambda_1\,(x_1,x_2)\right]\\ &\lambda_{12}\,(x_2,x_2-x_1)\exp\left[-\Lambda_2\,(x_2,t-x_1-x_2)\right]dx_1dx_2,\end{aligned}$$

$$\phi_{0;0}(0,t) \;= \exp\left[-\Lambda\,(t)\right].$$

Then

$$\pi_2\left(t\right) = \frac{\phi^1_{0;2}(0,t)+\phi^2_{0;2}(0,t)}{\phi_{0;0}(0,t)+\phi_{0;1}(0,t)+\phi^1_{0;2}(0,t)+\phi^2_{0;2}(0,t)}$$

The previous denominator can be approximated by the global survival of the population at age t, which leads to the same expression as [GRA 04]. Considering the section the partial prevalence is

$$\pi_2\left(t,x\right) = \frac{\Gamma^1_{0;2}(0,t,x)+\Gamma^2_{0;2}(0,t,x)}{\phi_{0;0}(0,t)+\phi_{0;1}(0,t)+\phi^1_{0;2}(0,t)+\phi^2_{0;2}(0,t)}$$

where

$$\Gamma^1_{0;2}(0,t,x) = \int_{t-x}^{t} \exp\left[-\Lambda_0\left(0,x_1\right)\right]\lambda_{02}\left(0,x_1\right)\exp\left[-\Lambda_{.2}\left(x_1,t-x_1\right)\right]dx_1$$

and

$$\begin{aligned}\Gamma^2_{0;2}(0,t,x) = \int_{t-x}^{t}\int_{t-x}^{t-x_1} &\exp\left[-\Lambda_{.0}\left(0,x_1\right)\right]\lambda_{01}\left(0,x_1\right)\exp\left[-\Lambda_{.1}\left(x_1,x_2\right)\right]\\ &\lambda_{12}\left(x_2,x_2-x_1\right)\exp\left[-\Lambda_{.2}\left(x_2,t-x_1-x_2\right)\right]dx_1dx_2\end{aligned}$$

5.4.3. *Estimation method*

Our estimation method consists of a semi-parametric approach where the transition intensities of the NHSM process are considered as piecewise constant functions [GRA 04]. We assume that $\lambda_{ij}\left(t,x\right)$ with $i \neq 0$ is a piecewise constant function according to the two time scales. We construct a finite partition of the age axis and a finite partition of the duration in the state i. The hazard $\widehat{\lambda}^{k,l}_{ij}$ is estimated as

$$\widehat{\lambda}^{k,l}_{ij} \approxeq \frac{G_{kl}}{R_{kl}}, \tag{5.12}$$

where G_{kl} represents the number of transitions from state i to state j, in the l^{th} year following the cancer diagnosis in state i, among patients who were in the k age interval at the time of diagnosis. The count R_{kl} is the corresponding number at risk of transiting to state j at the beginning of the l^{th} interval.

5.4.4. *Results: indicators of public health*

This method is applied to the Herault (France) cancer registry database. The data consists of 1,749 women diagnosed with breast cancer between 1994 and 1996, and followed up until 1999. At the time of diagnosis, there were 1,672 in state 1 and 77 in state 2. Using the previous method, the age-specific partial prevalence is estimated. Results are given in Table 5.5. In 1999, in Herault, the age-specific partial prevalence rate of loco-regional breast cancer for 50-54 year old women was 1,166 per 100,000 and of metastasis breast cancer for 50-54 year old women was 84 per 100,000.

6 years partial prevalence of breast cancer per 100,000			
Age	Total	Loco-Regional	Distant
40-44	557	518	56
45-49	1,084	1,010	91
50-54	1,249	1,166	84
60-64	1,569	1,433	98
65-69	1,549	1,465	119
70-74	1,451	1,282	157
75-79	1,218	1,015	158

Table 5.5. *Partial prevalence of breast cancer*

5.5. Discussion

The objective of this study was to present and establish the SM process in a multi-state context.The NHSM process represents a key modeling in the medical field since it takes into account two relevant elements. Indeed, chronological and duration times are two fundamental scales in medicine. The NHSM model captures the main features of different chronic diseases and provides a reasonable approximation of very complicated processes. Therefore, interesting reference points may be derived for physicians. In patient care, some predictions can be made as regards their biological evolution. In the public health context, indicators as prevalence represents tools for decision. A natural perspective to these applications consists of considering some covariates in order to make the modeling more pertinent and informative. For a dependant covariate of time, namely $Z_{ij}(t)$, the following kernel may be considered:

$$Q_{ij}(t,x|\, Z_{ij}(t)) = p_{ij}(t|\, Z_{ij}(t))F_{ij}(t,x|\, Z_{ij}(t)).$$

Therefore, the covariate may have an impact on the jump and duration processes. However, an accuracy-parsimony compromise always has to be respected in modeling steps. The structure of transition must not be too complex so that the model can be useful, pragmatic and interpretable in practice.

5.6. Bibliography

[AAL 97] AALEN O., FAREWELL V., ANGELIS D. D., N.E. DAY O. G., "A Markov model for HIV disease progression including the effect of HIV diagnosis and treatment: application to AIDS prediction in England and Wales", *Statistics in Medicine*, vol. 16, p. 2191–2210, 1997.

[COM 02] COMMENGES D., "Inference for multi-state models from interval-censored data", *Statistical Methods in Medical Research*, vol. 11, p. 167–182, 2002.

[COX 80] COX D., ISHAM V., *Point Processes*, Chapman and Hall, 1980.

[DAV 00] DAVIDOV O., ZELEN M., “Designing cancer prevention trials: a stochastic model approach”, *Statistics in Medicine*, vol. 19, p. 1983–1995, 2000.

[FOU 05] FOUCHER Y., MATHIEU E., SAINT-PIERRE P., DURAND J., DAURÈS J., “A Semi-Markov Model Based on Generalized Weibull Distribution with an Illustration for HIV Disease”, *Biometrical Journal*, vol. 47, p. 825–833, 2005.

[GAI 99] GAIL M., KESSLER D., MIDTHUNE, SCOPPA S., “Two approaches for estimating disease prevalence from population-based registries of incidence and total mortality”, *Biometrics*, vol. 55, p. 1137–1144, 1999.

[GRA 04] GRAS C., DAURÈS J., TRÉTARRE B., “Age- and stage-specific prevalence estimate of cancer from population-based cancer Registry using Inhomogeneous Poisson process”, *Statistical Methods in Medical Research*, vol. 13, p. 1–17, 2004.

[GRU 89] GRUTTOLA V. D., LAGAKOS S., “Analysis of doubly-censored survival data, with application to AIDS”, *Biometrics*, vol. 45, p. 1–11, 1989.

[JAN 86] JANSSEN J., *Semi-Markov Models. Theory and Applications*, Plenum Press, 1986.

[JAN 99] JANSSEN J., LIMNIOS N., *Semi-Markov Models and Applications*, Kluwer Academic Publishers, 1999.

[JAN 01] JANSSEN J., MANCA R., “Numerical solution of non-homogeneous semi-Markov processes in transient case”, *Methodology and Computing in Applied Probability*, vol. 3, p. 271–293, 2001.

[JOL 98] JOLY P., COMMENGES D., LETENNEUR L., “A penalized likelihood approach for arbitrarily censored and truncated data: application to age-specific incidence of dementia”, *Biometrics*, vol. 54, p. 203–212, 1998.

[JOL 99] JOLY P., COMMENGES D., “A penalized likelihood approach for a progressive three-state model with censored and truncated data: application to AIDS”, *Biometrics*, vol. 55, p. 887–890, 1999.

[KEI 89] KEIDING N., ANDERSEN P., “Nonparametric estimation of transition intensities and transition probabilities: a case study on a two-state Markov process”, *Applied Statistics*, vol. 38, p. 319–329, 1989.

[KEI 91] KEIDING N., “Age-specific incidence and prevalence: A statistical perspective”, *Journal of the Royal Statistical Society, Series A*, vol. 154, p. 371–412, 1991.

[KOR 82] KOROLYUK V., TURBIN A., *Markov Renewal Processes in Problems of Systems Reliability*, Naukova Dumka, Kiev, 1982.

[LEV 54] LEVY P., “Processus Semi Markoviens”, *Proceedings of the International Congress of Mathematics*, p. 416–426, Cong. Math. Amsterdam, 1954.

[LIM 99] LIMNIOS N., OPRISAN G., “A general framework for reliability and performability analysis of Semi-markov systems”, *Appl. Stochastic Models Business Indust.*, vol. 15, 4, p. 353–368, 1999.

[LIM 01] LIMNIOS N., OPRISAN G., *Semi-Markov Processes and Reliability*, Statistics for Industry and Technology Birkhauser, 2001.

[MAT 06] MATHIEU E., LOUP P., DELLAMONICA P., DAURES J., "Markov modelling of immunological and virological states in HIV-1 infected patients", *Biometrical Journal*, vol. 48, p. 834–846, 2006.

[MAT 07] MATHIEU E., LOUP P., DELLAMONICA P., DAURES J., "Parametric and Non-homogeneous Semi-Markov process for HIV control", *Methodology and Computing in Applied Probability*, vol. 9, num. 3, 2007.

[ONI 83] ONICESCU O., OPRIŞAN G., POPESCU G., "Renewal processes with complete connections", *Rev. Roumaine Math. Pures Appl*, vol. 28, p. 985–998, 1983.

[PAP 98] PAPADOPOULOU A., "Counting transitions-entrance probabilities in non-homogeneous semi-Markov systems", *Applied Stochastic Models and Data Analysis*, vol. 13, p. 199–206, 1998.

[PAP 99] PAPADOPOULOU A., VASSILIOU P., *Semi-Markov Models and Applications*, J. Janssen and N. Limnios, editors, 1999.

[PYK 61] PYKE R., "Markov renewal processes: definitions and preliminary properties", *Ann. of Math. Statist*, vol. 32, p. 1231–1242, 1961.

[PYK 64] PYKE R., SCHAUFELE R., "Limit theorems for Markov renewal processes", *Ann. of Math. Statist*, vol. 35, p. 1746–1764, 1964.

[SAT 99] SATTEN G., STERNBERG M., "Fitting semi-Markov models to interval-censored data with unknown initiation times", *Biometrics*, vol. 55, p. 507–513, 1999.

[SMI 55] SMITH W., "Regenerative stochastic processes", *Proc. Roy. Soc. Ser. A*, vol. 232, p. 6–31, 1955.

[STE 99] STERNBERG M., SATTEN G., "Discrete-time nonparametric estimation for Semi-Markov models of chains-of-events data subject to interval censoring and truncation", *Biometrics*, vol. 55, p. 514–522, 1999.

[TAK 54] TAKACS L., "Some investigations concerning recurrent stochastic processes of a certain type", *Magyar Tud. Akad. Mat. Kutado Int. Kzl.*, vol. 3, p. 115–128, 1954.

[VAS 92] VASSILIOU P., PAPADOPOULOU A., "Non-homogeneous Semi-Markov systems and maintainability of the state sizes", *J. Appl. Prob.*, vol. 29, p. 519–534, 1992.

[WAD 92] WADJDA W., "Uniformly strong ergodicity for non-homogeneous semi-Markov processes", *Demonstration Mathematica*, vol. 25, p. 755–764, 1992.

[WIL 94] WILSON S., SOLOMON P., "Estimates for different stages of HIV/AIDS disease", *Comput Appl. Biosci*, vol. 10, p. 681–683, 1994.

This page intentionally left blank

Chapter 6

Bivariate Cox Models

6.1. Introduction

This chapter introduces a new class of Cox models for bivariate data. We discuss the role of copulae and the effect of covariates.

It is shown that dependence indicators can easily be regressed on the covariates when the bivariate hazard baseline is positively quadrant dependent. The role of extreme-value copulae is highlighted in this respect. As an example, we regress Spearman's rho on covariate inside this model.

6.2. A dependence model for duration data

Suppose each individual in a homogeneous population is subject to two failure times X and Y which are both observed. Assume also that X and Y are both absolutely continuous (ac) non-negative random variables (rv) with joint cdf (cumulative distribution function) $H_{X,Y}(x,y) \equiv \Pr X \leq x, Y \leq y$ and pdf (probability density function) $h_{X,Y}(x,y)$. The corresponding sdf (survival distribution function) is $\bar{H}_{X,Y}(x,y) \equiv \Pr X > x, Y > y$, and the margins of $H_{X,Y}$ will be denoted by H_X and H_Y, from which $\bar{H}_{X,Y}(x,y) = 1 - H_X(x) - H_Y(y) + H_{X,Y}(x,y)$. The hazard rate of X, also known as the instantaneous or age-specific failure rate is

$$\lambda_X(x) \equiv \lim_{\Delta \to 0} \frac{\Pr x < X \leq x + \Delta | X > x}{\Delta}$$

Chapter written by Michel BRONIATOWSKI, Alexandre DEPIRE and Ya'acov RITOV.

which may conveniently be written as

$$\lambda_X(x) = -\frac{d \log \bar{H}_X(x)}{dx}. \tag{6.1}$$

Regression models aim at modeling dependence upon explanatory variables. In the proportional hazard model, the cause specific hazard functions satisfy

$$\begin{cases} \lambda_X(x) = \lambda_X^0(x)\Phi(z) \\ \lambda_Y(y) = \lambda_Y^0(y)\Psi(z). \end{cases} \tag{6.2}$$

Often the positive functions $\Phi(z)$ and $\Psi(z)$ are assumed to be parametric functions, with the standard $\Phi(z) = \exp(\alpha' z)$ and $\Psi(z) = \exp(\beta' z)$, $\alpha, \beta \in \mathbb{R}^d$. Classical references include [OAK 89, HOU 87].

Investigators are rarely only interested in marginal behaviors as described in (6.2). This model ignores the dependence between X and Y and can only be of poor interest for practical applications. In this chapter we consider models for dependence under explanatory variable. Let

$$\lambda_{Y|X=x}(y) \equiv \lim_{\Delta \to 0} \frac{\Pr y < Y \leq y + \Delta | X = x, Y > y}{\Delta} \tag{6.3}$$

and

$$\lambda_{Y|\,X>x}(y) \equiv \lim_{\Delta \to 0} \frac{\Pr y < Y \leq y + \Delta | X > x, Y > y}{\Delta}. \tag{6.4}$$

Modeling the multivariate dependency through (6.3) and (6.4) leads to different types of models, namely Cox or multiplicative models

$$\begin{cases} \lambda_X(x) = \lambda_X^0(x)\Phi(z) \\ \lambda_{Y|X=x}(y) = \lambda_{Y|X=x}^0(y)\Psi(z) \end{cases} \tag{6.5}$$

or

$$\begin{cases} \lambda_X(x) = \lambda_X^0(x)\Phi(z) \\ \lambda_{Y|X>x}(y) = \lambda_{Y|X>x}^0(y)\Psi(z). \end{cases} \tag{6.6}$$

Let H^z be the joint distribution of (X, Y) for z fixed, if it exists. For $Z = z$ fixed and (X, Y) satisfying (6.6), if (X, Y) has a joint probability distribution, then we denote by H^z this distribution. Direct approaches based on regression type models cannot deserve our purpose. Indeed, consider, for example, a model defined through

$$\begin{cases} X = f(z, U) \\ Y = g(z, V) \end{cases}$$

with $f(z,\cdot)$ and $g(z,\cdot)$ strictly increasing for all z. Then following [NEL 98] Theorem 2.4.3, (X,Y) has same copula as (U,V) for all z, which forms the basis of this chapter.

Both models (6.5) and (6.6) characterize the property that the failure of one component puts an extra load on the other component, as for example in studies involving a two organs system. Are such models mathematically valid? That is, is there a cdf H^z for the rv (X,Y) and a baseline cdf F^0 which are compatible with them? Ignoring technicalities, this is indeed always the case for model (6.5). Model (6.6) is valid under certain restrictions on the baseline hazard, as will be seen later. Model (6.5) has been widely studied, e.g. [DEM 94]. Indeed, the usual paradigm in regression analysis is conditioned upon the observed value of the variable X. Note, however, that in the present setting the covariate is in fact z. The main difficulty with the first model is that marginally it does not satisfy (6.2), the Cox paradigm, which trivially holds true for the model (6.6) with the setting $x = 0$. For statistical estimation, both models have respective interest. The likelihood function in the first one can easily be written, and hence the parameters can be estimated by partial likelihood maximization. The second model allows a straightforward estimation of the parameters, which are defined independently upon the dependence structure induced by the model, setting $x = 0$ in the second equation, and applying standard univariate estimation for the parameters of the functions Φ and Ψ. Clearly, model (6.6) leads to an easy description of the sdf of the rv (X,Y), while (6.5) is more adequate to handle the properties of the pdf. This difference prompts our interest in model (6.6). Since the sdf can easily be written as a function of the baseline and the covariate, it represents a natural model for the regression of basic dependence indices on the covariate.

6.3. Some useful facts in bivariate dependence

Throughout this section, F generically denotes a cdf on $\mathbb{R}^{+2}$. We describe the toolbox which is needed in order to study the bivariate model (6.6). For simplicity, we consider only distributions which are ac.

Let (X,Y) be two rv with a joint cdf F and marginal distributions F_X and F_Y. Their dependence properties are summarized using the copula function C, defined on $I \equiv [0,1] \times [0,1]$ by:

$$C(u,v) = F\left(F_X^{\leftarrow}(u), F_Y^{\leftarrow}(v)\right).$$

where the quasi-inverse $F_X^{\leftarrow}$ of F is defined by $F_X^{\leftarrow}(y) = \inf\{x :\ F(x) \geq y\}$.

The notion of copula is introduced in [SKL 59].

Let (U,V) be the bivariate rv $(F_X(X), F_Y(Y))$. Clearly, we have $C(u,v) = \Pr U \leq u, V \leq v$. A related notion useful for survival analysis is the survival copula $\hat{C}$ which is defined by $\hat{C}(u,v) \equiv \Pr U > 1-u, V > 1-v = C(1-u, 1-v) + u + v - 1$.

Definition 6.1 *A non-negative function G is totally positive of order 2 (TP2) if and only if for all $x_1 < x_2,\ y_1 < y_2$, $G(x_1, y_1)G(x_2, y_2) \geq G(x_1, y_2)G(x_2, y_1)$. If G is a C^2 function, then G is TP2 if and only if:*

$$\frac{\partial G}{\partial x}(x, y) \times \frac{\partial G}{\partial y}(x, y) \leq \frac{\partial^2 G}{\partial x \partial y}(x, y) \times G(x, y).$$

If F is TP2 then F is Positive Quadrant Dependent (PQD), i.e. for all (x, y), we have $F(x, y) \geq F_X(x)F_Y(y)$ (Theorem 2.3 p. 26 [JOE 97]). Thus, the following theorem holds (see [JOE 97], p. 30).

Theorem 6.1 *Let F be an ac bivariate cdf. Then: (i) F is MAXID (F^t is a cdf for all $t > 0$) if and only if F is TP2; (ii) F is MINID ($\bar{F}^t$ is an sdf for all $t > 0$) if and only if $\bar{F}$ is TP2.*

The following simple lemma expresses the intuitive fact that TP2, MAXID and MINID are properties of the copula and not of the margins, and therefore reflect properties of the dependence between X and Y.

Lemma 6.1 *Let F be a bivariate ac cdf. Then F is MAXID, if and only if C is ac and MAXID if and only if C is ac and TP2. Further F is ac and MINID if and only if $\hat{C}$ is ac and MAXID if and only if $\hat{C}$ is ac and TP2.*

Remark 6.1 *If F has a density which is TP2 (implying that F is also TP2) then its copula inherits the same property as checked by direct calculation.*

Among various classes of copulae, the family of *extreme value* copulae (EVC) is of peculiar interest. These copulae are characterized by the property that for all positive t,

$$C(u^t, v^t) = C^t(u, v),$$

for all (u, v) in I; see [DEH 79], lemma 2. Any EVC is TP2; see [JOE 97], p. 177. We will also make use of the *Pickands' representation* of EVC's; see [PIC 81]. It appears that EVC's can be described through a function of only one variable; and as such their complexity is weaker compared to the whole class of copulae. Another class enjoying this property is the *Archimedean* class of copulae. We will not consider Archimedean copulae here. The following fact holds (see [JOE 97], Theorems 6.3 and 6.4).

Proposition 6.1 *A copula C is an EVC if and only if*

$$C(u,v) \equiv \exp\left[(\log uv)\, A\left(\frac{\log v}{\log uv}\right)\right] \tag{6.7}$$

where the Pickands' function or dependence function A is convex on $[0,1]$, $A(0) = A(1) = 1$, and $\max(t, 1-t) \leq A(t) \leq 1$.

Lemma 6.2 *If F is a cdf with EVC copula and Pickands' function A, so is F^t for any $t > 0$.*

Any PQD copula can be approximated by some EVC. Let us describe the procedure. Let us parametrize the points in I in an adequate way. For any s in $]0,1]$ the arc $\{(v^{1/s-1}, v) : \ v \in [0,1]\}$ is denoted by $\mathfrak{s}$. This arc is the set of points (u,v) on which any dependence function A is constant and equals $A(s)$; for $s = 0$, define $\mathfrak{o} = \{(0,v),\ v \in [0;1]\}$. The family of the arcs $\mathfrak{s}$ covers I. For any copula C, define on $\mathfrak{s}$ the function

$$\mathbb{A}(s,v) \equiv \frac{s}{\log v} \log C(v^{(1/s)-1}, v). \tag{6.8}$$

Function $\mathbb{A}$ mimics the dependence function A; indeed, when C is an EVC, then $\mathbb{A}(s,v) \equiv A(s)$ as seen by direct substitution. Using $\mathbb{A}$ and the copula C, analogously to (6.7), write

$$C(u,v) = \exp\left[\log(uv)\mathbb{A}\left(\frac{\log v}{\log uv}, v\right)\right].$$

In order to construct a dependence function close to $\mathbb{A}$, average the values of $\mathbb{A}$ on $\mathfrak{s}$ through $\bar{\mathbb{A}}(s) \equiv \int_0^1 \mathbb{A}(s,v)dv$. Obviously other choices are possible which do not change in any way the results in the proposition hereafter. Define now

$$\widetilde{C}(v^{1/s-1}, v) = \exp\left[\frac{\log v}{s}\bar{\mathbb{A}}(s)\right] \tag{6.9}$$

which, by a change of variables, is also $\widetilde{C}(u,v) = \exp\left[\log(uv)\bar{\mathbb{A}}\left(\frac{\log v}{\log uv}\right)\right]$. We assume that

(H1) C is PQD

(H2) the function $x \to x\frac{\frac{\partial}{dx}C(x,y)}{C(x,y)}$ is non-increasing for all y in $[0,1]$.

For completeness, define $\mathbb{A}(0,v) = 1$, $\mathbb{A}(s,1) = 1$ and $\mathbb{A}(1,v) = 1$.

Remark 6.2 *Y is stochastically increasing (SI) in X (denoted $F_{Y|X}$ is SI) if and only if $\Pr Y > y|X = x$ is non-decreasing in x for all y. Furthermore, Y is left-tail decreasing (LTD) in X if and only if $\Pr Y > y|X > x$ is non-decreasing in x for all y. SI implies LTD (see [JOE 97] p. 26, Theorem 2.3). Hence, a sufficient condition for (H2) to hold is $C_{V|U}$ is SI.*

When the density of C is TP2, then necessarily both (H1) and (H2) hold.

From Remark 6.1, it follows that if F^0_{XY} has a density which is TP2, then both (H1) and (H2) hold.

The following proposition can be proved.

Proposition 6.2 *Under (H1) and (H2), the copula $\widetilde{C}$ is an EVC, and for all s in $[0, 1]$,*

$$\sup_{(u,v)\in\mathfrak{s}} \left|C(u,v) - \widetilde{C}(u,v)\right| \leq \frac{1}{2} \operatorname*{osc}_{\mathfrak{s}} \mathbb{A}$$

where $\operatorname{osc}_{\mathfrak{s}} \mathbb{A}$ denotes the oscillation of the function $\mathbb{A}$ on the arc $\mathfrak{s}$.

The upper bound in Proposition 6.2 is indeed 0 when C is an EVC.

Calculations for the C12 (with $\theta = 6$) and C13 (with $\theta = 2$) in [JOE 97] show that the upper bounds in Proposition 6.2 are, respectively, 0.015 and 0.05.

6.4. Coherence

Not all baseline survival dfs $F^0_{X,Y}$ define a coherent model, so that λ_X and $\lambda_{Y|X>x}$ are the marginal and conditional specific cause hazards for some bivariate sdf $\bar{H}^z$ under a given covariate z. We conclude from the first equation of (6.6) that $\bar{H}^z_X(x) = \bar{F}_X^{0\,\Phi(z)}(x)$. By the second equation, plugging $x = 0$, we get $\bar{H}^z_Y(y) = \bar{F}_Y^{0\,\Psi(z)}(y)$. The model is coherent if $\bar{F}_X^{0\,\Phi(z)}(x)\bar{F}_{Y|X>x}^{0\quad\Psi(z)}(y)$ defines an sdf.

Notice that

$$\begin{aligned} \bar{H}^z_{X,Y}(x,y) &= \left(\bar{F}_X^{0\,\Psi(z)}(x)\bar{F}_{Y|X>x}^{0\quad\Psi(z)}(y)\right)\bar{F}_X^{0\,\Phi(z)-\Psi(z)}(x) \\ &= \left(\bar{F}^0_{X,Y}\right)^{\Psi(z)}(x,y)\left(\bar{F}^0_X\right)^{\Phi(z)-\Psi(z)}(x) \end{aligned} \qquad (6.10)$$

which is indeed an sdf when $\Phi(z) \geq \Psi(z)$ and $\bar{F}_{X,Y}^{0\ \Psi(z)}$ is an sdf.

Note, however, that $F^0_{XY}(x,y)^t$ is indeed an sdf for all $t \geq 1$. Hence when $\Psi(z) \geq 1$, (6.6) is defined when $\Phi(z) \geq \Psi(z)$ without further restrictions pertaining to $\bar{F}^0_{XY}$.

As already stated, not all bivariate survival dfs $\bar{F}^0_{X,Y}$ are such that for all positive t, $\left(\bar{F}^0_{X,Y}\right)^t$ is an sdf. Min-infinite divisibility of the baseline hazard seems to be a natural assumption here. Assume therefore that

$$F^0_{X,Y} \text{ is min-infinitely divisible.} \tag{H3}$$

Then by (6.10), $\bar{H}^z_{X,Y}$ is an sdf as soon as $\Phi(z) \geq \Psi(z) \geq 0$.

Let us consider the case when $0 \leq \Phi(z) \leq \Psi(z)$. Analogously with (6.6) the model may then be written

$$\begin{cases} \lambda_Y(y) = \lambda^0_Y(y)\Psi(z) \\ \lambda_{X|Y>y}(x) = \lambda^0_{X|Y>y}(x)\Phi(z) \end{cases} \tag{6.11}$$

permuting the role of X and Y. In a similar way to the above, we have

$$\bar{H}^z_{X,Y}(x,y) = \bar{F}^0_{X,Y}{}^{\Phi(z)}(x,y)\bar{F}^0_Y{}^{\Psi(z)-\Phi(z)}(y).$$

is a proper sdf. To summarize the above arguments we state the following.

Let the model be defined by (6.10) if $\Phi(z) \geq \Psi(z)$ and by (6.11) if $\Phi(z) \leq \Psi(z)$. Let (M) be the model defined in this way. When (H) holds, then for all z , $\bar{H}^z_{X,Y}$ is an sdf and

$$\begin{aligned} \bar{H}^z_{X,Y}(x,y) = \mathbf{1}\left\{\Phi(z) \geq \Psi(z)\right\} \bar{F}^0_{X,Y}{}^{\Psi(z)}(x,y)\bar{F}^0_X{}^{\Phi(z)-\Psi(z)}(x) \\ + \mathbf{1}\left\{\Phi(z) \leq \Psi(z)\right\} \bar{F}^0_{X,Y}{}^{\Phi(z)}(x,y)\bar{F}^0_Y{}^{\Psi(z)-\Phi(z)}(y). \end{aligned} \tag{6.12}$$

Min-infinite divisibility of the baseline will also make any $H^z_{X,Y}$ min-infinitely divisible, showing that this class in *stable* under (M). Indeed for any positive t

$$\begin{aligned} \left(\bar{H}^z_{X,Y}\right)^t = \mathbf{1}\left\{\Phi(z) \geq \Psi(z)\right\} \left(\bar{F}^0_{X,Y}(x,y)\right)^{t\Psi(z)} \left(\bar{F}^0_X(x)\right)^{t(\Phi(z)-\Psi(z))} \\ + \mathbf{1}\left\{\Phi(z) \leq \Psi(z)\right\} \left(\bar{F}^0_{X,Y}(x,y)\right)^{t\Phi(z)} \left(\bar{F}^0_Y(y)\right)^{t(\Psi(z)-\Phi(z))}, \end{aligned}$$

which still is an sdf.

By Lemma 6.1, min-infinite divisibility is not a property of the cdf but of its copula. Formula (6.12) can be written for copulae through

$$\begin{aligned} \hat{C}_{H^z}(u,v) = \mathbf{1}\left\{\Phi(z) \geq \Psi(z)\right\} u^{\frac{\Phi(z)-\Psi(z)}{\Phi(z)}} \hat{C}_{F^0}\left(u^{\frac{1}{\Phi(z)}}, v^{\frac{1}{\Psi(z)}}\right)^{\Psi(z)} \\ + \mathbf{1}\left\{\Phi(z) \leq \Psi(z)\right\} u^{\frac{\Psi(z)-\Phi(z)}{\Psi(z)}} \hat{C}_{F^0}\left(u^{\frac{1}{\Psi(z)}}, v^{\frac{1}{\Psi(z)}}\right)^{\Phi(z)}. \end{aligned} \tag{6.13}$$

Since $\bar{H}^z_{X,Y}(x,y) = \hat{C}_{H^z}\left(\bar{H}^z_{X,Y}(x,0), \bar{H}^z_{X,Y}(0,y)\right)$ and when $\Phi(z) \geq \Psi(z)$, $\bar{H}^z_{X,Y}(x,0) = \bar{H}^z_X(x) = \bar{F}^0_X{}^{\Phi(z)}(x)$, it then holds, when $\Phi(z) \geq \Psi(z)$

$$\begin{aligned}
&\hat{C}_{H^z}(u,v) \\
&= \bar{H}^z_{X,Y}\left(\bar{H}^{z\,\leftarrow}_X(u), \bar{H}^{z\,\leftarrow}_Y(v)\right) \\
&= \left(\bar{F}^0_{X,Y}\right)^{\Psi(z)}\left(\bar{H}^{\leftarrow}_X(u), \bar{H}^{\leftarrow}_Y(v)\right)\left(\bar{F}^0_X\right)^{\Phi(z)-\Psi(z)}\left(\bar{H}^{\leftarrow}_X(u)\right) \\
&= \left(\bar{F}^0_{X,Y}\right)^{\Psi(z)}\left(\bar{F}0^{\leftarrow}_X(u^{1/\Phi(z)}), \bar{F}0^{\leftarrow}_Y(v^{1/\Psi(z)})\right)\left(\bar{F}0_X\right)^{\Phi(z)-\Psi(z)}\left(\bar{F}0^{\leftarrow}_X(u^{1/\Phi(z)})\right) \\
&= \hat{C}_{\bar{F}^0}\left(u^{1/\Phi(z)}, v^{1/\Psi(z)}\right)^{\Psi(z)} .u^{\frac{\Phi(z)-\Psi(z)}{\Psi(z)}}.
\end{aligned}$$

The expression in (6.13) is obtained by joining the latest result together with the similar result when $\Phi(z) \leq \Psi(z)$. Although (H3) is in full accordance with univariate Cox models it is only a sufficient condition for coherence.

Our interest lies in good classes of min-infinitely divisible copulae which we intend to regress on the covariate z. Among all subclasses of TP2 copulae, the EVCs enjoy nice properties as seen in Proposition 6.2, since they can nicely approximate PQD copulae and are parametrized using a smooth function of only one variable. When F^0_{XY} has an EVC $\hat{C}_{F^0_{XY}}$ with Pickands' function A ($\hat{C}_{F^0}(u,v) = \exp\left(\log(uv)A\left(\frac{\log(v)}{\log(uv)}\right)\right)$) then, denoting $\hat{C}_{H^z_{XY}}$ as the survival copula of H^z_{XY} and using (6.13), we have

$$\hat{C}_{H^z_{XY}}(u,v) = \exp\left[\log(uv)B^z\left(\frac{\log v}{\log uv}\right)\right]$$

with

$$B^z(s) = \frac{\Phi(z)-\Psi(z)}{\Phi(z)}(1-s) + \left[\frac{\Psi(z)}{\Phi(z)}(1-s)+s\right]A\left(\frac{1}{\frac{\Psi(z)}{\Phi(z)}\frac{1}{s} - \left(\frac{\Psi(z)}{\Phi(z)}-1\right)}\right)$$

when $\{\Phi(z) \geq \Psi(z)\}$, and a similar expression under $\{\Phi(z) \leq \Psi(z)\}$. To summarize, after some algebra, noting $K(z) = \min\left(\frac{\Phi(z)}{\Psi(z)}, \frac{\Psi(z)}{\Phi(z)}\right)$, we have

$$B^z(s) = \left[1-K(z)\right](1-s) + \left[K(z)(1-s)+s\right]A\left(\frac{1}{\frac{K(z)}{s} - \left[K(z)-1\right]}\right). \quad (6.14)$$

B^z is a convex function defined on $[0,1]$ which satisfies $B^z(0) = B^z(1) = 1$ and $\max(s, 1-s) \leq B^z(s) \leq 1$ because $A(t) \leq 1$ and $K(z) \leq 1$. This basic result shows that $\hat{C}_{H^z_{XY}}$ is an EVC with Pickands' function B^z. We have proved that the class of EVC's is *stable* under (M).

Notice that we do not require F^0_{XY} to have a bivariate extreme value distribution, since only its copula should be an EVC. As such, F^0_{XY} can have arbitrary margins and still fits in (M).

From (6.14), we deduce a transitive expression that links $B^{z'}$ and B^z for any z and z'. It holds that

$$\begin{aligned} B^{z'}(t) &= \left[1 - \tilde{K}(z,z')\right](1-t) \\ &\quad + \left[\tilde{K}(z,z')(1-t) + t\right] B^z\left(\frac{1}{\frac{\tilde{K}(z,z')}{t} - \left[\tilde{K}(z,z') - 1\right]}\right), \end{aligned} \tag{6.15}$$

where

$$\tilde{K}(z,z') = K(z') \cdot K(z)^{-1} \tag{6.16}$$

independently of the Pickands' function of the baseline. This formula can be seen as a kind of expression of the proportional hazard property, which links two hazard rates independently on the baseline.

When the covariate acts equally on X and Y, i.e. $\Phi(z) = \Psi(z)$ for all z, then $B^z(s) = A(s)$ for all values of s. Thus, the copula of H^z_{XY} equals that of the baseline hazard F^0_{XY} – the dependency structure of X and Y should not be altered through (M). Only the marginal distributions of X and Y in that case reflect the role of the covariate.

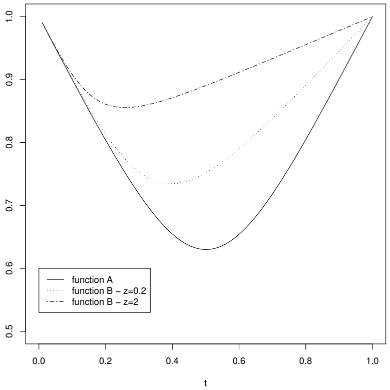


Figure 6.1. *Illustration of formula* (6.14)

Figure 6.1 illustrates the transition formula (6.14). The baseline copula is the Gumbel-Hougaard extreme-value copula with $\theta = 3$. We use $\Phi(z) = \exp(\alpha_1 z + \alpha_0)$, with $(\alpha_0, a_1) = (1, 2)$, and $\Psi(z) = \exp(\beta_1 z + \beta_0)$, with $(\beta_0, \beta_1) = (0.5, 1.5)$.

The dependence function A is the solid line, the dotted line represents the dependence function B^z for $z = 0.2$ and the dashed line is the dependence function for $z = 2$.

One consequence of the model is that the covariate has an impact on the marginal distribution, and also on the copula. If $\Phi(z) = \Psi(z)$, the copula does not change. However, if $\Phi(z) > 1 = \Psi(z)$, then a change in the distribution of Z, which made X stochastically smaller, is not followed by a corresponding change in the distribution of Y; thus, it should follow by a change in their dependency, however this is measured.

6.5. Covariates and estimation

A proper statistical study of the model (M) must include covariates. Let z be a variable which can obtain a finite number of values. Suppose we have n_i independent observations $(X_j^{z_i}, Y_j^{z_i})_{j=1}^{n_i}$ from of the bivariate rv (X^{z_i}, Y^{z_i}) with unknown cdf $H_{XY}^{z_i}$.

We suggest the following estimation procedure. First estimate consistently the function Φ using the sample of the X^z values, and Ψ using the samples of the Y^z's. This provides consistent estimates of $\Phi\,(z)$ and $\Psi\,(z)$ for any z. The standard maximum Cox partial likelihood estimator can be used at this stage, if the Cox model is assumed to hold. Define next $\tilde{K}(z, z')$ as in (6.16) where $\Phi\,(z)$ and $\Psi\,(z)$ are substituted by their estimators. The asymptotic normality of the parameters is trivial. Their joint distribution can be easily assessed using an asymptotic linearization of the estimators.

For each z_i, estimate consistently the Pickands' function B^{z_i} of $\hat{C}_{XY}^z$. Standard estimates found in other works suppose that the marginal distributions of (X^{z_i}, Y^{z_i}) are known, which we do not assume to hold. We therefore proceed as follows. First, we estimate $\hat{C}_{XY}^z$ using the non-parametric strongly consistent estimates

$$\widehat{C_n^{z_i}}(u, v) \equiv \frac{1}{n} \sum_{j=1}^{n_i} 1 \left\{ n - \mathrm{rank}(X_j^{z_i}) \leq un; n - \mathrm{rank}(Y_j^{z_i}) \leq vn \right\}$$

We then use the same procedure as in section 6.3 which was designed in order to provide an EVC approximation of a PQD copula. For s in $[0, 1]$, define a function $\mathbb{A}_n$ on the arc $\mathfrak{s}$ by

$$\mathbb{A}_n(v^{\frac{1}{s}-1}, v) \equiv \frac{s}{\log v} \log \widehat{C_n^{z_i}}(v^{\frac{1}{s}-1}, v).$$

Average $\mathbb{A}_n$ on $\mathfrak{s}$ and get

$$B_n^{z_i}(s) \equiv \bar{\mathbb{A}}_n(s) \equiv \min\left\{1, \max\left(\frac{1}{k_{n_i}} \sum_{l=1}^{k_{ni}} \mathbb{A}_n(v_l^{\frac{1}{s}-1}, v_l); s; 1-s\right)\right\} \tag{6.17}$$

for $0 < v_1 < \cdots < v_{k_{n_i}} < 1$, a uniform grid on $[0,1]$. The reason we introduce the extra max operation is to get a properly defined estimate on the segment $[0,1]$. It is easily checked that $B_n^{z_i}(0) = B_n^{z_i}(1) = 1$ and by construction $B_n^{z_i}(s) \geq \max(s, 1-s)$. When k_{n_i} tends to infinity together with n_i and when $\hat{C}_{XY}^z$ is an EVC continuity arguments show that $B_n^{z_i}(s)$ point-wisely converges almost surely to $B^{z_i}(s)$. Asymptotic properties of those estimators are postponed to a further work.

We illustrate (6.17) in Figure 6.2. The dataset consists of $n = 200$ iid points from the Gumbel-Hougaard extreme-value copula $\hat{C}(u,v) \;=\; \exp\left\{-\left[(-\log u)^\theta + (-\log v)^\theta\right]^{1/\theta}\right\}$ with $\theta = 3$. The estimate B_n^z (refered to as “empirical estimator”) is compared to the classical Pickands’ estimate of the dependence function.

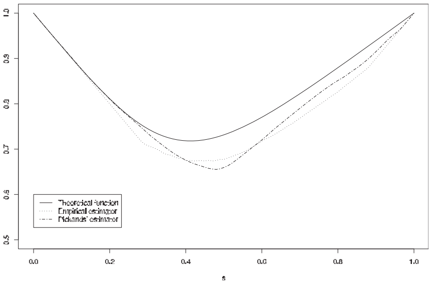


Figure 6.2. *Two estimators of B^z*

Let z be some value of the covariate for which no data are observed. Use (6.15) for each z_i so that

$$B^z(s) = \left[1 - \tilde{K}(z_i, z)\right](1-s)$$
$$+ \left[\tilde{K}(z_i, z)(1-s) + s\right] B^{z_i}\left(\frac{1}{\frac{\tilde{K}(z_i, z')}{s} - \left[\tilde{K}(z_i, z) - 1\right]}\right)$$

holds for any i.

Define the forecast of $B^z(s)$ using the interpolation

$$B_n^z(s) = \sum_{i=1}^{k} \frac{W(|z-z_i|)}{\sum_{j=1}^{k} W(|z-z_i|)} \left[1-\tilde{K}(z_i,z)\right](1-s) \tag{6.18}$$
$$+ \left[\tilde{K}(z_i,z)(1-s)+s\right] B_n^{z_i}\left(\frac{1}{\frac{\tilde{K}(z_i,z')}{s} - \left[\tilde{K}(z_i,z)-1\right]}\right)$$

where $W(|z-z_i|)$ is positive and tends to 0 as the distance between z and z_i increases. This is done in order to give more weight to the B^{z_i} functions for z_i close to z.

We consider a data set consisting of 10 values of the covariate z. The baseline hazard has a Gumbel-Hougaard survival EVC. For each one we simulated 200 couples (X_i, Y_i) in the following way: $\Phi(z) = \exp(\alpha_1 z + \alpha_0)$, $\Psi(z) = \exp(\beta_1 z + \beta_0)$, $\alpha_1 = 3$, $\alpha_0 = 2$, $\beta_1 = 2$, $\beta_0 = 1$, and $\theta = 3$. We used formula (6.14) to define the copulae for each z_i, $1 \leq i \leq 10$. The values of the z_is range between 0 and 1. The prediction is set for $z = 1.2$.

We used formula (6.18) with $W(x) = \exp(-2x)$. In Figure 6.3, the dashed line is the forecasted B function when the B^{z_i}s are estimated using Pickands' method. The dotted line is similar, but the $B_n^{z_i}$s are the estimates defined by (6.17). The solid line is the theoretical B^z as deduced from the model.

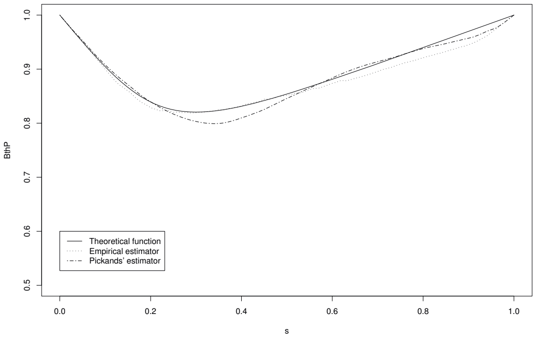


Figure 6.3. *Forecasted dependence functions*

6.6. Application: regression of Spearman's rho on covariates

We illustrate the above procedure regressing the Spearman correlation coefficient on covariate, taking F_{XY}^0 as the Gumbel-Hougaard copula, defined by $F_{XY}^0(x,y) \equiv$

$\exp\left\{-\left[(-\log u)^{\theta}+(-\log v)^{\theta}\right]^{1/\theta}\right\}$ with $\theta \geq 1$. This copula is an EVC; see [NEL 98] p. 94.

First, we recall some properties of Spearman's rho in the context of extreme-value distribution. See, for example, [HÜR 03] for details. As usual, let C be a copula, and $\hat{C}$ its survival copula. We have

$$\rho_S = 12\iint C(u,v)dudv - 3 = 12\int_0^1 \frac{dt}{(1+A(t))^2} - 3,$$

and simple calculation leads to $\rho_S = 12\iint \hat{C}(u,v)dudv - 3$.

We calculate ρ_S^z for several values of z through its empirical counterpart

$$\rho_S^z = 12\int \frac{dt}{\left[1+B_n^z(t)\right]^2} - 3 \tag{6.19}$$

where B_n^z is given in (6.18).

Ten values of the covariate z have been considered. The baseline hazard has a Gumbel-Hougaard survival EVC. For each value we simulate 500 couples (X_i, Y_i) in the following way: $\Phi(z) = \exp(\alpha_1 z + \alpha_0), \Psi(z) = \exp(\beta_1 z + \beta_0)$ $\alpha_1 = 3, \alpha_0 = 2, \beta_1 = 2, \beta_0 = 1$, $\theta = 3$. We use equation (6.14) to define the copulae for each z_i, $1 \leq i \leq 10$. The simulation step uses the above algorithm, and $W(x) = \exp(-2x)$. The computation is illustrated in Figure 6.4. The solid black line is ρ_S^z. The dashed line is the forecast (for $z's$ in the range of the $z_i's$). The dots are the empirical values on the samples for the $z_i's$.

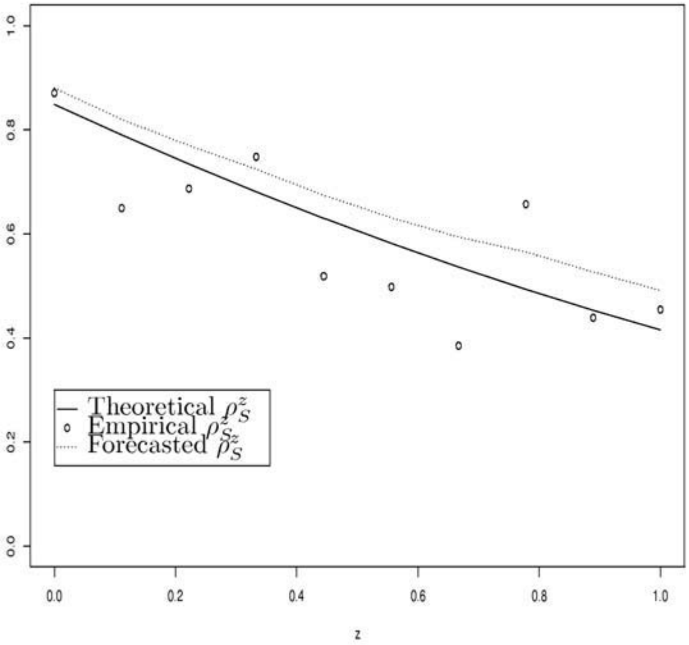


Figure 6.4. *Forecasted Spearman's rho*

6.7. Bibliography

[DEH 79] Deheuvels P., "Détermination complète du comportement asymptotique en loi des valeurs extrêmes multivariées d'un échantillon de vecteurs aléatoires indépendants", *C. R. Acad. Sci. Paris Sér. A-B*, vol. 288, num. 3, p. A217–A220, 1979.

[DEM 94] DeMasi R., Proportional Hazards Models for Multivariate Failure Time Data with Generalized Competing Risks, PhD thesis, University of North Carolina, 1994.

[HOU 87] Hougaard P., "Modelling multivariate survival", *Scandinavian Journal of Statistics*, vol. 14, num. 4, p. 291–304, 1987.

[HÜR 03] Hürlimann W., "Hutchinson-Lai's conjecture for bivariate extreme value copulas", *Statist. Probab. Lett.*, vol. 61, num. 2, p. 191–198, 2003.

[JOE 97] Joe H., *Multivariate Models and Dependence Concepts*, Chapman & Hall, 1997.

[NEL 98] Nelsen R. B., *An Introduction to Copulas*, vol. 139 of *Lectures Notes in Statistics*, Springer-Verlag, 1998.

[OAK 89] OAKES D., "Bivariate models induced by frailties", *Journal of the American Statistical Association*, vol. 84, num. 406, p. 487–493, 1989.

[PIC 81] PICKANDS III J., "Multivariate extreme value distributions", *Proceedings of the 43rd session of the International Statistical Institute, Vol. 2 (Buenos Aires, 1981)*, vol. 49, p. 859–878, 894–902, 1981, With a discussion.

[SKL 59] SKLAR M., "Fonctions de répartition à n dimensions et leurs marges", *Publ. Inst. Statist. Univ. Paris*, vol. 8, p. 229–231, 1959.

This page intentionally left blank

Chapter 7

Non-parametric Estimation of a Class of Survival Functionals

7.1. Introduction

Time to event data are encountered in many fields: medical follow-up, engineering, epidemiology, demography, economics, biology, actuarial science, etc. There is a vast number of other books dedicated to the analysis of these events, for instance [MIL 81], [FLE 91], [HOS 99]. Our aim in this chapter is to put many classical univariate survival analysis problems into a single framework and then provide a common smoothing-based estimation procedure.

Henceforth, we shall consider a non-negative continuous lifetime (or age at failure) variable T. There is a plethora of functionals which describe the distribution of lifetime data. The most important one is the survival function. It represents the chance that an individual (a unit) survives beyond time t:

$$\overline{F}(t) \equiv 1 - F(t) = \text{Pr}\{T > t\} = \int_0^t f(u)du$$

where F and f stand for the cumulative distribution function (cdf) and probability density function (pdf) respectively. The hazard function is of prime importance as well. For an individual surviving past time t, the hazard rate (or *intensity rate* or *force of mortality*) gives the instantaneous rate of failure at time t:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{\text{Pr}\{t < T \leq t + \Delta t | T > t\}}{\Delta t} = \frac{f(t)}{\overline{F}(t)}, t \geq 0,$$

Chapter written by Belkacem ABDOUS.

Note that for a small Δt, the quantities $f(t)\Delta t$ and $\lambda(t)\Delta t$ give the unconditional probability of failure (death) in $(t, t+\Delta t)$ and the conditional probability of failure (death) in $(t, t+\Delta t)$ given surviving past time t respectively. An alternative presentation of the hazard function is provided by the integrated hazard rate also known as *cumulative hazard function*

$$\Lambda(t) = -\log \overline{F}(t) = \int_0^t \lambda(u)\, du$$

Closely related to these functions is the residual lifetime distribution defined by

$$F(s|t) = \Pr\{T \leq t+s | T > t\} = \frac{F(s+t) - F(t)}{1 - F(t)} = \frac{\overline{F}(t) - \overline{F}(s+t)}{\overline{F}(t)}, s > 0$$

Clearly, $\overline{F}(s|t) = 1 - F(s|t)$ is the survival function of $T_t = T - t$, the residual lifetime of an individual who survived up to time t. The expected (or mean) residual lifetime is given by

$$e(t) = \mathbb{E}(T - t | T > t) = \begin{cases} \int_t^{+\infty} \overline{F}(u)du / \overline{F}(t) & \text{if } \overline{F}(t) > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{7.1}$$

Note that the expected lifetime is simply $e(0)$.

Each of these functions provides its own specific characterization of the lifetime T, but once one of these functions is known the others are uniquely determined. Besides, the survival function $\overline{F}(\cdot)$ has been used to construct several criteria for ageing, reliability, economics curves, social inequalities, etc. Among them, one is the *Lorenz curve*, which is defined in the Cartesian plane by $(F(t), L_F(t))$ with

$$L_F(t) = \frac{\int_0^t s dF(s)}{\int_0^\infty s dF(s)}, \quad t > 0.$$

In the context of income inequality within a given population, it depicts the simultaneous changes in income and in population see [KOT 85] for more details and applications. Besides, a closely related concept is given by the *scaled total time on test function T (or total time of test transform)*. It is defined in the Cartesian plane by the parametric form $(F(t), T_F(t))$ with

$$T_F(t) = \frac{\int_0^t \overline{F}(s)ds}{\int_0^\infty \overline{F}(s)ds}, \quad t > 0.$$

Note that $T_F(t)$ and $L_F(t)$ are linked by the following relationship

$$T_F(t) = L_F(t) + t\overline{F}(t)/\mu.$$

More details and reasons for these functionals can be found in [KOT 85], [KOT 88], and [SHO 86].

There are still some other survival functionals remaining. In the following sections, we will use the notation $\Phi(t, \overline{F})$ for $t \geq 0$ to refer to any of the above functionals or any survival functional not listed here. Next, the survival function $\overline{F}(\cdot)$ being unknown, we must rely on parametric or non-parametric estimates. Both parametric and non-parametric estimates of each of the above survival functionals have been studied extensively in other works. Our main scope in this chapter is to cast these functionals estimation problems in the same framework. This will provide us with a unified treatment of existing estimators and new ones as well. The adopted approach will be based on the weighted local polynomial smoothing technique described in section 7.2. Consistency of the proposed estimators will be investigated in section 7.3. Finally, an automatic bandwidth selection procedure is described in section 7.4.

7.2. Weighted local polynomial estimates

Suppose that the lifetime T is right-censored by a censoring time variable C having a cdf G. Denote the right-censored data by the pairs (Z_i, δ_i), $i = 1, \ldots, n$, with

$$Z_i = T_i \wedge C_i = \min(T_i, C_i) \quad \text{and} \quad \delta_i = \mathbb{I}(T_i \leq C_i).$$

The indicator function $\mathbb{I}(T_i \leq C_i)$ shows whether the observed Z_i is the failure $(\delta_i = 1)$ or the censoring time $(\delta_i = 0)$. The C_i are assumed to be independent of the T_is. Their survival functions are related to H, the common cdf of Z_is, by the following relationship:

$$\overline{H}(t) \equiv 1 - H(t) = (1 - F(t))(1 - G(t)) \equiv \overline{F}(t)\overline{G}(t)$$

Let $\overline{F}_n(t)$ stand for an estimator of the survival function. In an uncensored setting, this will be merely the standard empirical survival function, while if censoring is present then the classical Kaplan-Meier estimate (KM-estimate for short) will be adopted ([KAP 58])

$$\overline{F}_n(t) = \begin{cases} \displaystyle\prod_{i=1}^{n} \left(\frac{N(Z_i)}{1 + N(Z_i)} \right)^{\mathbb{I}(Z_i \leq t, \delta_i = 1)} & \text{if } t \leq Z_{(n)} \\ 0 & \text{otherwise} \end{cases}$$

where $N(t) = \sum_{i=1}^{n} \mathbb{I}(Z_i > t)$ and $Z_{(n)} = \max(Z_1, \ldots, Z_n)$.

Now, as mentioned earlier, let us rewrite all the previous functionals under a unique form: $\Phi(t, \overline{F})$, with $t \in \mathbb{R}^+$. First, we will merely require that for any fixed t, the functional $\Phi(t, \cdot)$ is defined for all survival functions and is continuous with respect to the sup-norm topology. This ensures that an estimator (a possibly naive one) of $\Phi(t, \overline{F})$ might be obtained by substituting the KM-estimate $\overline{F}_n$ for $\overline{F}$, and by using the uniform strong consistency of the KM-estimate $\overline{F}_n$ we should end up with $\Phi(t, \overline{F}_n) \to \Phi(t, \overline{F})$. Furthermore, if for a fixed survival $\overline{F}$, the function $\Phi(\cdot, \overline{F})$ has $r \geq 0$ derivatives which we are interested to estimate, then the assumed continuity of $\Phi(t, \cdot)$ implies that $\Phi(t, \overline{F}_n)$ converges to $\Phi(t, \overline{F})$ but it does not follow that the derivatives $\Phi(\cdot, \overline{F}_n)$ converge to the derivatives of $\Phi(\cdot, \overline{F})$. Instead, we propose estimating the derivatives of $\Phi(\cdot, \overline{F})$ by the minimizers (for each fixed t) $\hat{a}_0(t), \hat{a}_1(t), \ldots, \hat{a}_r(t)$ of the objective function

$$\int_0^\infty \frac{1}{h} K\left(\frac{s-t}{h}\right) \left\{ \Phi(s, \overline{F}_n) - \sum_{i=0}^{r} \frac{a_i(t)}{i!} (s-t)^i \right\}^2 ds \tag{7.2}$$

where the weight function K is an arbitrary probability density and the bandwidth $h = h(n)$ denotes a sequence of smoothing parameters tending to 0 as n goes to ∞. The bandwidth h controls the size of t neighborhood. This criteria might be caused by the fact that for any fixed t, Taylor expansion yields

$$\Phi(s, \overline{F}) \simeq \sum_{i=0}^{r} \frac{(s-t)^i}{i!} \Phi^{(i)}(t, \overline{F})$$

for s belonging to a neighborhood of t. Thus, criteria (7.2) simply locally fits a polynomial of order r to the empirical estimate $\Phi(\cdot, \overline{F}_n)$. Consequently, the obtained solutions $\hat{a}_j(t)$ should reasonably estimate $\Phi^{(j)}(t, \overline{F})$, for $j = 0, \ldots, r$. This estimation scheme has been applied by [ABD 03] to estimate smooth functionals of a distribution function and their derivatives. For an overview on local polynomial smoothing and various application, see [FAN 96] and [LOA 99]. Note, however, that our approach is somehow different from the classical local polynomial smoothing which is mainly based on a discrete and weighted least squares criteria.

Survival times being positive, we implicitly assumed that the support of $\Phi(\cdot, \overline{F})$ is $\mathbb{R}^+$. By restricting the above integral to the domain $\mathbb{R}^+$, we alleviate the well-known problem of boundary effects. Whenever this support is known to be confined to a subset Ω, then it suffices to restrict the integration in (7.2) to Ω.

Next, there are several ways to express $\widehat{\mathfrak{a}}_r(t) = (\hat{a}_0(t), \ldots, \hat{a}_r(t))^T$, the solution of (7.2). Indeed, if the polynomial order r is small (say $r \leq 2$) then we can simply calculate the solution of the following linear system

$$\int_{-t/h}^{\infty} s^i \Phi(t+sh, \overline{F}_n) K(s) ds = \sum_{j=0}^{r} \frac{a_j(t)}{j!} h^j \int_{-t/h}^{\infty} s^{i+j} K(s) ds, \qquad i = 0, \ldots, r, \tag{7.3}$$

and end up with

$$\widehat{\mathfrak{a}}_r(t) = (\mathbb{H}\mathbb{M})^{-1} \Phi_n(t,h)$$

where $\mathbb{H}$ denotes the $(r+1) \times (r+1)$ diagonal matrix $\mathrm{diag}(1, h/1!, \ldots, h^r/r!)$ and $\mathbb{M}$ stands for the $(r+1) \times (r+1)$ matrix with entries

$$\mu_{i,j} = \int_{-t/h}^{\infty} s^{i+j} K(s)\, ds, \qquad i, j = 0, \ldots, r,$$

and $\Phi_n(t,h) = (\Phi_n^0(t,h), \ldots, \Phi_n^r(t,h))^T$ with

$$\Phi_n^i(t,h) = \int_{-t/h}^{\infty} s^i \Phi(t+sh, \overline{F}_n) K(s) ds, \quad i = 0, \ldots, r$$

An elegant and different representation of $\widehat{\mathfrak{a}}_r(t)$ could be obtained by making use of the reproducing kernel Hilbert spaces theory (see [BER 04]). Indeed, let $L_2(K,t)$ denote the space of functions ψ such that $\int_{-t/h}^{\infty} \psi^2(s) K(s) ds < \infty$, endowed with the inner product

$$< \psi, \gamma >= \int_{-t/h}^{\infty} \psi(s)\gamma(s)K(s)ds,$$

set $\mathbb{P}_r$ for the reproducing kernel Hilbert subspace of $L_2(K,t)$ with reproducing kernel

$$\mathcal{K}_r(u,v) = \sum_{i=0}^{r} P_i(u) P_i(v)$$

where $(P_i)_{0 \le i \le r}$ is an arbitrary sequence of orthonormal polynomials in $L_2(K,t)$. Then, by rewriting (7.2) as

$$\int_{-t/h}^{\infty} \left\{ \Phi(t+sh, \overline{F}_n) - \sum_{i=0}^{r} a_i(t) \frac{(hs)^i}{i!} \right\}^2 K(s) ds, \tag{7.4}$$

it can be shown that the polynomial that minimizes this criteria is the projection of $\Phi(t + h\cdot, \overline{F}_n)$ onto the space of polynomials of degree, at most, r $\mathbb{P}_r$ (see [BER 04], Theroem 75). More precisely, we find that the minimizing coefficients $\widehat{\mathfrak{a}}_r(t)$ are merely expressed in terms of a hierarchy of high order kernels generated by K, i.e.

$$\widehat{a}_{in}^{[r]}(t) = \frac{1}{h^i} \int_0^{\infty} \Phi(s, \overline{F}_n) \frac{1}{h} \mathcal{K}_i^{[r]} \left(\frac{s-t}{h} \right) ds. \tag{7.5}$$

with

$$\mathcal{K}_i^{[r]}(u) = \left(\sum_{k=0}^{r} P_k(u) \, \frac{d^i}{dw^i} P_k(w) \bigg|_{w=0} \right) K(u).$$

The kernels $\mathcal{K}_i^{[r]}(u)$ do not depend on the chosen orthonormal basis for $\mathbb{P}_r$. We can take advantages of the numerous techniques of orthogonal polynomials to calculate these kernels. For instance, by the Christoffel-Darboux formula (for more details see, for example, [SZE 75] or [BRE 80])

$$\mathcal{K}_i^{[r]}(u) = \left(\frac{\partial^i}{\partial w^i} \, \frac{P_{r+1}(u)P_r(w) - P_{r+1}(w)P_r(u)}{u-w} \bigg|_{w=0} \right) K(u).$$

7.3. Consistency of local polynomial fitting estimators

The following consistency results rely heavily on the results presented in [ABD 03]. Hereafter, we will only stick to the estimation of $\Phi(\cdot, \overline{F})$. Consider the associated estimator

$$\widehat{a}_{0n}^{[r]}(t) = \int_{-t/h}^{\infty} \Phi(t+sh, \overline{F}_n) \mathcal{K}_0^{[r]}(s) ds.$$

and define

$$\overline{a}_{0n}^{[r]}(t) = \int_{-t/h}^{\infty} \Phi(t+sh, \overline{F}) \mathcal{K}_0^{[r]}(s) ds.$$

We have the usual decomposition of the error $\widehat{a}_{0n}^{[r]}(t) - \Phi(t, \overline{F})$: a stochastic component

$$\Delta_1(t) := \widehat{a}_{0n}^{[r]}(t) - \overline{a}_{0n}^{[r]}(t)$$

and a deterministic one

$$\Delta_2(t) := \overline{a}_{0n}^{[r]}(t) - \Phi(t, \overline{F})$$

The consistency of the deterministic part $\Delta_2(t)$ is easily obtained from the following extension of the classical Bochner's lemma (see [ABD 03] for its proof). However, first, let us recall that for a real valued function g on $\mathbb{R}$, the Lebesgue set is the collection of all $x \in \mathbb{R}$ such that

$$\lim_{s \to 0} \frac{1}{s} \int_{|y|<s} |g(x+y) - g(x)| dy = 0.$$

Note that this set includes almost all points of $\mathbb{R}$. In particular, it contains points of continuity of g (see [STE 71]).

Lemma 7.1 *Let $k \geq 0$ and set $M_k^{[r]}(t) = ess.sup._{|s| \geq |t|} |s^k \mathcal{K}_0^{[r]}(s)|$. Let g be a given real function. If $M_k^{[r]} \in L^1(\mathbb{R})$ and g is such that $g \in L^p(\mathbb{R})$, $1 \leq p \leq \infty$, then*

$$\lim_{h \to 0} \int_{-t/h}^{\infty} \mathcal{K}_0^{[r]}(y) g(t + sh) ds = g(t)$$

for all points in the Lebesgue set of g.

As for the consistency of the stochastic term $\Delta_1(t)$, we need to ascertain that the KM-estimate $\overline{F}_n(\cdot)$ and the corresponding functional $\Phi(\cdot, \overline{F}_n)$ are consistent. Indeed, strong representation of the KM-estimate has been studied by several authors, see, for example, [DEH 00], [STU 93] and [LO 89]. For instance, by Corollary 1.2 in [STU 93], if T's survival function $\overline{F}$ has no jumps in common with C's survival function $\overline{G}$, then

$$\sup_{t \leq \tau_H} |\overline{F}_n(t) - \widetilde{F}(t)| \to 0 \text{ with probability 1}$$

where $\tau_H := \inf\{x : \overline{H}(x) = 0\}$ and

$$\widetilde{F}(t) = \begin{cases} \overline{F}(t), & \text{if } t < \tau_H \\ \overline{F}(\tau_H^-) - \mathbb{I}(\tau_H \in \mathfrak{A}) F\{\tau_H\}, & \text{if } t \geq \tau_H. \end{cases}$$

with $F\{a\} = F(a) - F(a-)$ and $\mathfrak{A}$ is the set of all atoms of H. Furthermore, to show that $\Phi(\cdot, \overline{F}_n)$ is consistent as well, we might apply Theorem 1.1 of [STU 93] which states that

$$\lim_{n \to \infty} \int \phi(t) F_n(dt) = \int_{t \notin \mathfrak{A}, t < \tau_H} \phi(t) F(dt) + \sum_{a_i \in \mathfrak{A}} \phi(a_i) F\{a_i\} \text{ with probability 1}$$

for any Borel-measurable function ϕ satisfying $\int |\phi| dF < \infty$.

An alternative way to establish strong consistency or asymptotic normality of functionals $\Phi(\cdot, \overline{F}_n)$ might be based on the Von Mises calculus technique. We refer to [BOO 80] or [FER 83] for a thorough discussion of this technique. Roughly, we use Hadamard differentiability to write

$$\Phi(t, \overline{F}_n) = \Phi(t, \overline{F}) + \int \Delta_F(t, s)(\overline{F}_n - \overline{F})(ds) + r(F_n, F), \tag{7.6}$$

and prove that the remainder $r(F_n, F) = o(\|F_n - F\|_\infty)$. Then, existing results for the consistency of $\overline{F}_n - \overline{F}$ are used to deduce the consistency of

$$\int \Delta_F(t, s)(\overline{F}_n - \overline{F})(ds).$$

An adaptation of Theorem 4.1 in [ABD 03] leads to the following strong consistency result.

Theorem 7.1 *Suppose that* $\Phi(t,\overline{F})$ *belongs to* $L^p(\mathbb{R}^+)$ *for some* $1 \leq p \leq \infty$. *Assume that the point* t *is a Lebesgue point of* $\Phi(t,\overline{F})$ *and there exists an open neighborhood* $\mathcal{N}_t$ *of* t *such that*

$$\lim_{n\to\infty} \sup_{s\in\mathcal{N}_t} |\Phi(s,\overline{F}_n) - \Phi(s,\overline{F})| = 0 \tag{7.7}$$

almost surely. Then $\widehat{a}_{0n}^{[r]}(t)$ *converges almost surely to* $\Phi(t,\overline{F})$ *whenever the bandwidth converges to zero slowly enough.*

7.4. Automatic selection of the smoothing parameter

It is broadly accepted that the performance of kernel-based estimates heavily depends on the choice of the bandwidth parameter. Given this crucial importance, we will give a brief guide to selecting this parameter appropriately. Common criteria to assess the performance of a non-parametric estimate are based on L_p-errors, $p \geq 1$ or their expectations.

For standard functionals with either censored or uncensored data (e.g. density, hazard functions, regression, etc.), several bandwidth selection procedures has been proposed in other works. These procedures are either automatic such as cross-validation, bootstrap and double kernel techniques, or they minimize an asymptotic expression of a local or integrated error and estimate any unknowns by using *"rules of thumb"*, *"solve-the-equation"* or *"plug-in"* techniques. For details and references therein, see [FAN 96], [WAN 95] for uncensored data and [HOR 06], [MAR 04] and [GON 96] for censored observations. The problem is more challenging for general functionals. For instance, when estimating the MRL function (7.1), the bandwidth selection based on MISE's asymptotic expansion involves too many unknowns. Indeed, if we consider a simple kernel estimate of $e(x)$ defined on $\mathbb{R}$, i.e.

$$\hat{e}_n(x) = \int_{-\infty}^{\infty} \frac{1}{h} K\left(\frac{x-v}{h}\right) e_n(v)dv$$

where K is an arbitrary symmetric probability density with support $[-1,1]$ and

$$e_n(x) = \frac{\int_x^\infty \overline{F}_n(y)dy}{\overline{F}_n(x)} \mathbb{I}(X_{(n)} > x),$$

with $X_{(n)} = \max(X_1, \dots, X_n)$, then we can write the associated MISE as follows (see [ABD 05])

$$\begin{aligned}
&\int_{\mathbb{R}} E(\hat{e}_n(x) - e(x))^2 dx = \\
&\qquad \int \left(\int K(u)e(x+hu)(1 - F^n(x+hu))du - e(x) \right)^2 dx \\
&\quad + 2 \int K(u)K(v)S_n(x+hu, x+hv) \left(2\frac{\mu_1(x+hv)}{\overline{F}(x+hv)} - e^2(x+hv) \right) dudvdx \\
&\quad + 2 \int K(u)K(v)e(x+hu)e(x+hv)(1 - F^n(x+hu))F^n(x+hv)dudvdx \\
&\qquad + 2 \int K(u)K(v) \frac{F^n(x+hv) - F^n(x+hu)}{F(x+hv) - F(x+hu)} \mu_0(x+hv) \\
&\qquad\qquad \times \left\{ h(v-u) + \frac{\mu_0(x+hv) - \mu_0(x+hu)}{\overline{F}(x+hv)} \right\} dudvdx
\end{aligned}$$

where

$$S_n(y,z) = \sum_{i=1}^{n} \frac{F^{n-i}(y)(1 - F^i(z))}{i}, \quad \mu_i(y) = \int_y^\infty (t-y)^i \bar{F}(t)dt, \quad \text{for} \quad i = 0, 1.$$

Clearly, a bandwidth selection by means of *"rules of thumb"*, *"solve-the-equation"* or *"plug-in"* techniques is a difficult task in this context. Hereafter, we adopt the L_1-double kernel technique to provide a general and automatic tool for selecting h when a weighted local polynomial fitting technique is used to estimate a functional like those listed in section 7.1. This technique is fully automatic and has proven to be powerful in density estimation context (see [BER 94] and [DEV 89]). The L^2 version of the double kernel technique has been investigated in [ABD 99] and [JON 98]. Its adaptation to survival functionals estimation problems listed in section 7.2 follows the following lines: let $r \geq 1$ be an arbitrary integer and use (7.5) to define an estimate of the functional $\Phi(t, \overline{F})$ by

$$\widehat{a}_{0n}^{[r]}(t) = \int_0^\infty \Phi(s, \overline{F}_n) \frac{1}{h} \mathcal{K}_0^{[r]} \left(\frac{s-t}{h} \right) ds.$$

Recall that a kernel K is said to be of order m if

$$\int v^i K(v)dv = \begin{cases} 1 & \text{if } i = 0 \\ 0 & \text{if } i = 1, \dots, m-1 \\ C \neq 0 & \text{if } i = m \end{cases}$$

In light of [BER 93], the kernel function $\mathcal{K}_0^{[r]}$ is a kernel of order $(r+1)$ if r is odd and of order $r+2$ otherwise.

Next, let $p > q \geq 1$ be two arbitrary integers and note that since the order of the kernel $\mathcal{K}_0^{[p]}$ is bigger than $\mathcal{K}_0^{[q]}$, the bias of the estimate $\widehat{a}_{0n}^{[p]}(\cdot)$ should be negligible in comparison to that of $\widehat{a}_{0n}^{[q]}(\cdot)$. Thus, a rough estimate of the L_1 error, given by

$$L_1 = \int_0^\infty \left| \widehat{a}_{0n}^{[q]}(s) - \Phi(s, \overline{F}) \right| ds,$$

could be

$$\begin{aligned} DK(p,q,h) &= \int_0^\infty \left| \hat{a}_{0n}^{[q]}(t) - \hat{a}_{0n}^{[p]}(t) \right| dt \\ &= \int_0^\infty \left| \int_{-t/h}^\infty \left[\mathcal{K}_0^{[q]}(s) - \mathcal{K}_0^{[p]}(s) \right] \Phi(t+sh, \overline{F}_n) ds \right| dt. \end{aligned}$$

The bandwidth h_{DK} which minimizes $DK(p,q,h)$ should approximate h_{L_1}, the L_1 optimal bandwidth.

The practical implementation of the double kernel approach is quite straightforward. For instance, if we are willing to adopt the criteria $DK(1,2,h)$ and use a kernel K that is a symmetric pdf with support $[-1,1]$, then the involved kernels $\mathcal{K}_0^{[1]}(\cdot)$ and $\mathcal{K}_0^{[2]}(\cdot)$ have different expressions which depend on the point t where the estimates $\hat{a}_{0n}^{[\cdot]}(t)$ are to be evaluated. As a matter of fact, if we split the assumed support $\mathbb{R}^+$ into two regions: an interior region $\Omega_1 = \{t : t > h\}$ and a boundary region $\Omega_2 = \{t : t = \alpha h, \text{ with } \alpha \in [0,1]\}$, then

$$\mathcal{K}_0^{[1]}(s) = \begin{cases} K(s) \text{ if } t \in \Omega_1 \\ \dfrac{\mu_2 - \mu_1 s}{\mu_0 \mu_2 - \mu_1^2} K(s), \quad \text{if } t \in \Omega_2 \end{cases}$$

and

$$\mathcal{K}_0^{[2]}(s) = \begin{cases} \dfrac{\nu_4 - \nu_2 s^2}{\nu_4 - \nu_2^2} K(s) \text{ if } t \in \Omega_1 \\ \dfrac{(\mu_2\mu_4 - \mu_3^2) - (\mu_1\mu_4 - \mu_2\mu_3)s + (\mu_1\mu_3 - \mu_2^2)s^2}{(\mu_2\mu_4 - \mu_3^2)\mu_0 - (\mu_1\mu_4 - \mu_2\mu_3)\mu_1 + (\mu_1\mu_3 - \mu_2^2)\mu_2} K(s) \text{ if } t \in \Omega_2 \end{cases}$$

where

$$\mu_i := \mu_i(\alpha) = \int_{-\alpha}^1 s^i K(s) ds \text{ and } \nu_i = \mu_i(1) \quad i = 0, \ldots, 4.$$

Theoretical aspects of this version of the double kernel technique is beyond the scope of this chapter. Its finite and asymptotic properties will be investigated in future studies.

7.5. Bibliography

[ABD 99] ABDOUS B., "L_2 version of the double kernel method", *Statistics*, vol. 32, num. 3, p. 249–266, 1999.

[ABD 03] ABDOUS B., BERLINET A., HENGARTNER N., "A general theory for kernel estimation of smooth functionals of the distribution function and their derivatives", *Rev. Roumaine Math. Pures Appl.*, vol. 48, num. 3, p. 217–232, 2003.

[ABD 05] ABDOUS B., BERRED A., "Mean residual life estimation.", *J. Stat. Plann. Inference*, vol. 132, num. 1-2, p. 3-19, 2005.

[BER 93] BERLINET A., "Hierarchies of higher order kernels", *Probab. Theory Related Fields*, vol. 94, num. 4, p. 489–504, 1993.

[BER 94] BERLINET A., DEVROYE L., "A comparison of kernel density estimates", *Publ. Inst. Statist. Univ. Paris*, vol. 38, num. 3, p. 3–59, 1994.

[BER 04] BERLINET A., THOMAS-AGNAN C., *Reproducing kernel Hilbert spaces in probability and statistics*, Kluwer Academic Publishers, Boston, MA, 2004, with a preface by Persi Diaconis.

[BOO 80] BOOS D., SERFLING R. J., "A note on differentials and the CLT and LIL for statistical functions, with application to M-estimates", *Ann. Statist.*, vol. 8, p. 618–624, 1980.

[BRE 80] BREZINSKI C., *Padé-type Approximation and General Orthogonal Polynomials*, vol. 50 of *International Series of Numerical Mathematics*, Birkhäuser Verlag, Basel, 1980.

[DEH 00] DEHEUVELS P., EINMAHL J. H. J., "Functional limit laws for the increments of Kaplan-Meier product-limit processes and applications", *Ann. Probab.*, vol. 28, num. 3, p. 1301–1335, 2000.

[DEV 89] DEVROYE L., "The double kernel method in density estimation", *Ann. Inst. H. Poincaré Probab. Statist.*, vol. 25, num. 4, p. 533–580, 1989.

[FAN 96] FAN J., GIJBELS I., *Local Polynomial Modelling and its Applications*, vol. 66 of *Monographs on Statistics and Applied Probability*, Chapman & Hall, London, 1996.

[FER 83] FERNHOLZ L. T., *Von Mises Calculus for Statistical Functionals*, vol. 19 of *Lecture Notes in Statistics*, Springer-Verlag, New York, 1983.

[FLE 91] FLEMING T. R., HARRINGTON D. P., *Counting Processes and Survival Analysis*, Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics, John Wiley & Sons Inc., New York, 1991.

[GON 96] GONZÁLEZ-MANTEIGA W., CAO R., MARRON J. S., "Bootstrap selection of the smoothing parameter in nonparametric hazard rate estimation", *J. Amer. Statist. Assoc.*, vol. 91, num. 435, p. 1130–1140, 1996.

[HOR 06] HOROVÁ I., ZELINKA J., BUDÍKOVÁ M., "Kernel estimates of hazard functions for carcinoma data sets", *Environmetrics*, vol. 17, num. 3, p. 239–255, 2006.

[HOS 99] HOSMER JR. D. W., LEMESHOW S., *Applied Survival Analysis*, Wiley Series in Probability and Statistics: Texts and References Section, John Wiley & Sons Inc., New York, 1999, Regression modeling of time to event data, a Wiley-Interscience Publication.

[JON 98] JONES M. C., “On some kernel density estimation bandwidth selectors related to the double kernel method”, *Sankhyā Ser. A*, vol. 60, num. 2, p. 249–264, 1998.

[KAP 58] KAPLAN E. L., MEIER P., “Nonparametric estimation from incomplete observations”, *J. Amer. Statist. Assoc.*, vol. 53, p. 457–481, 1958.

[KOT 85] KOTZ S., JOHNSON N. L., READ C. B., Eds., *Encyclopedia of Statistical Sciences. Vol. 5*, a Wiley-Interscience Publication, John Wiley & Sons Inc., New York, 1985, Lindeberg condition to multitrait-multimethod matrices.

[KOT 88] KOTZ S., JOHNSON N. L., READ C. B., Eds., *Encyclopedia of statistical sciences. Vol. 9*, a Wiley-Interscience Publication, John Wiley & Sons Inc., New York, 1988, Strata chart to Zyskind-Martin models. Cumulative index, Volumes 1–9.

[LO 89] LO S. H., MACK Y. P., WANG J. L., “Density and hazard rate estimation for censored data via strong representation of the Kaplan-Meier estimator”, *Probab. Theory Related Fields*, vol. 80, num. 3, p. 461–473, 1989.

[LOA 99] LOADER C., *Local Regression and Likelihood*, Statistics and Computing, Springer-Verlag, New York, 1999.

[MAR 04] MARRON J. S., DE UÑA-ÁLVAREZ J., “SiZer for length biased, censored density and hazard estimation”, *J. Statist. Plann. Inference*, vol. 121, num. 1, p. 149–161, 2004.

[MIL 81] MILLER JR. R. G., *Survival Analysis*, John Wiley & Sons Inc., New York, 1981.

[SHO 86] SHORACK G. R., WELLNER J. A., *Empirical Processes with Applications to Statistics*, Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics, John Wiley & Sons Inc., New York, 1986.

[STE 71] STEIN E. M., WEISS G., *Introduction to Fourier Analysis on Euclidean Spaces*, Princeton University Press, Princeton, N.J., 1971, Princeton Mathematical Series, No. 32.

[STU 93] STUTE W., WANG J.-L., “The strong law under random censorship”, *Ann. Statist.*, vol. 21, num. 3, p. 1591–1607, 1993.

[SZE 75] SZEGŐ G., *Orthogonal Polynomials*, American Mathematical Society, Providence, R.I., fourth edition, 1975, American Mathematical Society, Colloquium Publications, Vol. XXIII.

[WAN 95] WAND M. P., JONES M. C., *Kernel Smoothing*, vol. 60 of *Monographs on Statistics and Applied Probability*, Chapman & Hall Ltd., London, 1995.

Chapter 8

Approximate Likelihood in Survival Models

8.1. Introduction

We consider a random lifetime Y which depends on some explanatory variable X. For describing the dependence between Y and X, there are different possibilities. One well-known model is a proportional hazard model which was introduced by Cox [COX 72], [COX 75] who considered the partial likelihood and conditional likelihood estimates. Anderson *et al.* [AND 93] discussed a lot of new ideas for the inference in survival models. Bagdonavičius and Nikulin [BAG 02] investigated accelerated life models and several models for time depending covariates. Dabrowska [DAB 97] considered models where the baseline hazard rate also depends on the covariates. Nonparametric estimates are considered by Liero [LIE 03]. We give some results and proposals for estimating the influence of covariates. The problem is formulated as the estimation of finite dimensional parameters β if nuisance parameters η are included. Here, the proportional hazard model is a good example.

If $L_n(\beta, \eta)$ is the full likelihood, then the profile likelihood for β is defined by

$$pL_n(\beta) = \sup_{\eta} L_n(\beta, \eta).$$

This profile likelihood has nice properties if it is finite. The aim of the chapter is to give a common frame for the different estimators in survival models. The starting point is the profile likelihood and with different approximations of the profile likelihood we obtain corresponding estimates. We discuss the resulting estimates in examples. One of these examples is the proportional hazard model.

Chapter written by Henning LÄUTER.

8.2. Likelihood in proportional hazard models

We study the problem of estimating the conditional distribution of Y given $X = x$. Let C be a random censoring time independent from Y. Assuming there are independent copies (Y_i, C_i, X_i), $i = 1, \ldots, n$ of (Y, C, X), we observe (T_i, Δ_i, X_i), $\quad i = 1, \ldots, n$ for $T_i = \min(Y_i, C_i)$, $\Delta_i = \mathbf{1}(Y_i \leq C_i)$. The conditional hazard rate of Y_i given $X = x$ is $\lambda(y_i \mid x)$. For continuous random lifetimes the likelihood function is proportional to

$$\prod_{i=1}^{n} \lambda(t_i \mid x_i)^{\delta_i} \, \mathrm{e}^{-\Lambda(t_i \mid x_i)}$$

for

$$\Lambda(z \mid x) = \int_0^z \lambda(\xi \mid x) \, \mathrm{d}\xi.$$

For the proportional hazard model we have

$$\lambda(y_i \mid x) = \lambda_0(y_i)\psi(x, \beta)$$

where the baseline hazard rate λ_0 and the finite dimensional parameter β are unknown. The parametric form of the function ψ is known. Then this leads to the full likelihood function (up to factors)

$$L_n(\beta, \lambda_0) = \prod_{i=1}^{n} \lambda_0(t_i)^{\delta_i} \psi(x_i, \beta)^{\delta_i} \exp\Big(-\psi(x_i, \beta) \int_0^{t_i} \lambda_0(\xi) \, \mathrm{d}\xi \Big), \tag{8.1}$$

where β is the finite dimensional parameter of interest, and λ_0 is a infinite dimensional nuisance parameter η.

8.3. Likelihood in parametric models

In parametric models, some expressions are given explicitly. Moreover, we use the representations given here in section 8.4.1. Consider parametric models of independent and identically distributed Y_i, $i = 1, \ldots, n$ where the distribution is known up to an unknown parameter $\mu \in \mathbb{R}^{k+s}$. The asymptotic inference in parametric models goes back to LeCam and Hajek. A summary is given in Bickel *et al.* [BIC 93]. The MLE $\hat{\mu}_n$ maximizes the log-likelihood $l_n(\mu)$ and under mild conditions it is an efficient estimator, thus we have

$$\sqrt{n}(\hat{\mu}_n - \mu) \longrightarrow \mathsf{N}_{k+s}(0, \mathcal{J}^{-1}(\mu)),$$

where $\mathcal{J}(\mu)$ is the Fisher information matrix. We estimate $\mathcal{J}(\mu)$ by the "observed" information matrix

$$\mathcal{J}_n(\mu) = -\frac{1}{n} \left(\frac{\partial^2}{\partial \mu_i \partial \mu_j} l_n(\mu) \right)_{i,j=1,\ldots,k+s}. \tag{8.2}$$

If the parameter μ is partitioned as $\mu = (\beta, \eta)$ and $\beta \in \mathbb{R}^k$ is a parameter of interest while $\eta \in \mathbb{R}^s$ is a nuisance parameter, then

$$[A_n(\hat{\mu}_n) - B_n(\hat{\mu}_n)C_n^{-1}(\hat{\mu}_n)B_n^t(\hat{\mu}_n)]^{-1} \tag{8.3}$$

with

$$\mathcal{J}_n(\mu) = \begin{pmatrix} A_n(\mu) & B_n(\mu) \\ B_n^t(\mu) & C_n(\mu) \end{pmatrix}$$

for a $k \times k$ matrix A_n, an $s \times s$ matrix C_n is an asymptotic unbiased estimator for the asymptotic variance of $\hat{\beta}_n$. With the similar block representation for $\mathcal{J}(\mu)$, we find

$$\sqrt{n}(\hat{\beta}_n - \beta) \longrightarrow \mathsf{N}_k(0, [A(\mu) - B(\mu)C^{-1}(\mu)B^t(\mu)]^{-1}).$$

In order to calculate the observed information matrix or the mentioned variances, we have to know the second derivatives of l_n and especially $\frac{\partial^2}{\partial\mu_i\partial\mu_j} l_n(\hat{\mu}_n)$ are good approximations of $\frac{\partial^2}{\partial\mu_i\partial\mu_j} l_n(\mu)$. This is one reason why we want to work with this full log-likelihood function. In the parametric case, we have under general mild conditions

$$\frac{1}{n}\, l_n(\hat{\mu}_n) \longrightarrow \mathsf{E}_\mu \ln f(Y_1, \mu).$$

8.4. Profile likelihood

Consider distributions $\mathsf{P}_{(\beta,\eta)}$ where $\beta \in \mathbb{R}^k$ and η is a high dimensional nuisance parameter. The profile likelihood $pL_n(\beta, \eta)$ has many properties of the original likelihood, at least if the nuisance parameter is finite dimensional (Barndorff-Nielsen and Cox [BAR 94], Murphy and van der Vaart [MUR 00]). In any case, the MLE $\hat{\beta}_n$ maximizes pL_n. In some models, pL_n is infinite, then the likelihood principle fails. For instance, in the proportional hazard model we have

$$pL_n(\beta) = \infty \quad \forall\beta$$

with $\beta = \beta$, $\eta = \lambda_0$. We encounter a similar result to that found in a standard situation of non-parametric regression.

Example: consider a non-parametric regression model with non-random regressors

$$Y_i = m(x_i) + \varepsilon_i, \quad \varepsilon_i \sim \mathsf{N}(0, \sigma^2),\ i.i.d.$$

$\beta = \sigma^2$ is the parameter of interest and $\eta = m$ is the nuisance parameter. The points $x_i, i = 1, \ldots, n$ should be different. Then we have

$$\sup_\beta pL_n(\beta) = \infty \quad \forall n,$$

where a function $\hat{m}$ with $\hat{m}(x_i) = y_i$ is an estimate for m. Thus, we cannot estimate the variance of the errors with such an unrestricted estimate $\hat{m}$. Obviously, all information from the observations is used for the estimation of the nuisance parameter, and the estimation of β is impossible.

We learn from these two cases that the estimation of the nuisance parameter without restrictions about η can lead to undesired problems. we have to work with restrictions, and this is expressed by approximations of the likelihood or profile likelihood. A possibility for finding an estimation of the parameter β is to restrict the space of the nuisance parameter.

8.4.1. *Smoothness classes*

Let $\mathcal{M}$ be a linear space which contains all possible η. We then define a sequence of d_j-dimensional sets $\mathcal{M}_j$ and

$$\mathcal{M}_j \subset \mathcal{M}_{j+1}, \quad \mathcal{M}_j \longrightarrow \mathcal{M}, \quad j = 1, 2, \ldots,$$

where $\mathcal{M}_j \longrightarrow \mathcal{M}$ is understood to be the convergence of some norm in $\mathcal{M}$. Then, L_n can be maximized over the $k + d_j$-dimensional space $\mathbb{R}^k \times \mathcal{M}_j$. Let

$$\max_{\mathbb{R}^k} \max_{\mathcal{M}_j} L_n(\beta, \eta) = L_n(\hat{\beta}_n^{(j)}, \hat{\eta}_n^{(j)}). \tag{8.4}$$

Of course, $\hat{\beta}_n^{(j)}$ is an approximation of β and at the same time $\hat{\eta}_n^{(k)}$ is an estimation for η. The set $\mathcal{M}_j$ is to be chosen in such a way that $(\beta, \eta) \in \mathbb{R}^k \times \mathcal{M}_j$ is identifiable. The rate of convergence of $(\hat{\beta}_n^{(j)}, \hat{\eta}_n^{(j)})$ to (β, η) is determined by the dimension d_j of $\mathcal{M}_j$.

Usually the Fisher information $\mathcal{J}(\beta, \eta)$ or $\mathcal{J}_n(\beta, \eta)$ are defined as linear operators. We choose a basis in $\mathcal{M}$ in such a way that $\eta \in \mathcal{M}_j$ is represented by a infinite dimensional vector where all components are 0 except the first d_j components. So the elements of $\mathcal{M}_j$ are "smoother" in comparison with the elements of $\mathcal{M}$. Then, under $(\beta, \eta) \in \mathbb{R}^k \times \mathcal{M}_j$, the observed information matrix as in (8.2) has the form

$$\mathcal{J}_n(\beta, \eta) = \begin{pmatrix} A_n(\beta, \eta) & \tilde{B}_n(\beta, \eta) & 0 \\ \tilde{B}_n^t(\beta, \eta) & \tilde{C}_n(\beta, \eta) & 0 \\ 0 & 0 & 0 \end{pmatrix}, \tag{8.5}$$

for a $k \times k$ matrix A_n, $d_j \times d_j$ matrix $\tilde{C}_n$, $k \times d_j$ matrix $\tilde{B}_n$ and is considered as an approximation for $\mathcal{J}_n(\beta, \eta)$ for $(\beta, \eta) \in \mathbb{R}^k \times \mathcal{M}$. Then, as in (8.3), we approximate the variance of $\hat{\beta}_n^{(j)}$ by

$$\mathsf{Var}\hat{\beta}_n^{(j)} \approx [A_n(\hat{\beta}_n^{(j)}, \hat{\eta}_n^{(j)}) - \tilde{B}_n(\hat{\beta}_n^{(j)}, \hat{\eta}_n^{(j)})\tilde{C}_n^{-1}(\hat{\beta}_n^{(j)}, \hat{\eta}_n^{(j)})\tilde{B}_n^t(\hat{\beta}_n^{(j)}, \hat{\eta}_n^{(j)})]^{-1}. \quad (8.6)$$

We are really only interested in the convergence of the estimate of the parameter of interest. Consider again the non-parametric regression example.

Example: (continuation)
We consider the non-parametric normal regression model. Assuming the regression function is square-integrable and with an orthonormal basis $\{g_s, s = 1, \ldots\}$ we have the representation

$$m(t) = \sum_{s=1}^{\infty} \eta_s \, g_s(t).$$

With the $n \times d_j$ matrix

$$X_{(j)} = \begin{pmatrix} g_1(t_1) & \cdots & g_{d_j}(t_1) \\ & \ddots & \\ g_1(t_n) & \cdots & g_{d_j}(t_n) \end{pmatrix}, \quad y_{(j)} = \begin{pmatrix} y_1 \\ \vdots \\ y_{d_j} \end{pmatrix}$$

we have

$$\hat{\sigma}^2 = \frac{1}{n} \mid (I - P_j) y_{(j)} \mid^2$$

where P_j is the orthogonal projection on the space spanned by the columns of $X_{(j)}$ and $| \cdot |$ is the Euclidean norm. We see that with increasing n and j under $d_j/n \longrightarrow 0$, $\hat{\sigma}^2$ is an efficient estimate for σ^2.

In other problems the rate of convergence depends on d_j in a more complicated way. This was an approach for finding estimates of β without changing the likelihood. We restricted the space of nuisance parameters in such a way that the whole space is approximated by a sequence of finite dimensional spaces. Another possibility for a similar approximation was proposed by Huang and Stone [HUA 98]. They considered classes of splines and found convergence rates for the estimates.

Up to this point we worked with an approximation of the profile likelihood. Another approach for constructing estimates of the parameter of interest is based on the replacement of the full likelihood L_n by an appropriate function $\tilde{L}_n$. Then, the desired estimator is the maximizer of $\tilde{L}_n$.

8.4.2. *Approximate likelihood function*

We approximate the likelihood function in two different ways. We formulate this for proportional hazard models.

8.4.2.1. Λ_0 *is approximated by a stepwise constant function.*

Let the steps of Λ be in observed points $t_1, \ldots, t_n$. This leads to

$$\tilde{\Lambda}_0(t_i) = \sum_{j:t_j \le t_i} \lambda_j$$

and let $\lambda_0(t_j) = \lambda_j$ (see [AND 93], [MUR 00], [MUR 97], [OWE 01], [LAW 03]).

Substituting this into (8.1), we get the approximated likelihood

$$\tilde{L}_n(\beta, \lambda_1, \ldots, \lambda_n) = \prod_{i=1}^{n} \lambda_i^{\delta_i} \psi(x_i, \beta)^{\delta_i} \exp\left(-\psi(x_i, \beta)\tilde{\Lambda}_0(t_i)\right),$$

where $\lambda_1, \ldots, \lambda_n$, and β are unknown. Maximizing this with respect to $\lambda_1, \ldots, \lambda_n$, we obtain the profile likelihood. If $\mathcal{R}(t)$ denotes the set of individuals which are alive and uncensored to time t and $V_s(t) = \mathbf{1}(s \in \mathcal{R}(t))$, then the profile likelihood is

$$p\tilde{L}_n(\beta) \propto \prod_{k=1}^{n} \Big(\frac{\psi(x_k, \beta)}{\sum_j V_j(t_k)\psi(x_j, \beta)}\Big)^{\delta_k}.$$

This coincides with the partial likelihood of Cox. Firstly, Breslow [BRE 74] remarked that the conditional likelihood estimate of Cox is also a partial likelihood estimate. By construction, it is also a non-parametric maximum likelihood estimate. Without censoring, the profile likelihood is rewritten as

$$p\tilde{L}_n(\beta) \propto \prod_{k=1}^{n} \frac{\psi(x_k, \beta)}{\sum_{j:t_j \ge t_k} \psi(x_j, \beta)}.$$

From this representation it is clear that the estimate of β is a rank statistic.

8.4.2.2. Λ_0 *is approximated by a continuous piecewise linear function.*

Let Λ_0 be a continuous piecewise linear function. Because of

$$\tilde{\Lambda}_0(t) = \int_0^t \tilde{\lambda}_0(s)\, \mathrm{d}s$$

the corresponding hazard rate $\tilde{\lambda}_0(s)$ is piecewise constant. Let $t_{(1)}, \ldots, t_{(n)}$ be the ordered observations; we then have:

$$\tilde{\Lambda}_0(t) = \sum_{i=1}^{k-1} \tilde{\lambda}_0(t_{(i)})(t_{(i)} - t_{(i-1)}) + \tilde{\lambda}_0(t_{(k)})(t - t_{(k-1)})$$

for $t_{(k-1)} \leq t \leq t_{(k)}, \quad k = 1, \ldots, n$. Here $t_{(0)} = 0$.

Consequently

$$\tilde{\Lambda}_0(t_{(i)}) = \sum_{j=1}^{i} \tilde{\lambda}_0(t_{(j)})(t_{(j)} - t_{(j-1)})$$

holds. The likelihood function is approximated by

$$\tilde{L}_n(\beta, \lambda_1, \ldots, \lambda_n) = \prod_{i=1}^{n} \tilde{\lambda}_0(t_{(i)})\psi(x_{[i]}, \beta)\, \mathrm{e}^{-\psi(x_{[i]},\beta)\sum_{j=1}^{[i]} \tilde{\lambda}_0(t_{(j)})(t_{(j)}-t_{(j-1)})}$$

where $\lambda_i := \tilde{\lambda}_0(t_{(i)})$ and $[i] = j$ if $t_{(i)} = t_j$. We consider λ_i as nuisance parameters and determine the profile likelihood. For fixed β, the function $\tilde{L}_n$ is maximized by

$$\frac{1}{\lambda_k} = (t_{(k)} - t_{(k-1)}) \sum_{i \geq k} \psi(x_{[i]}).$$

Thus, the profile likelihood is

$$p\tilde{L}(\beta) :=$$

$$\max_{\lambda_1,\ldots,\lambda_n} \tilde{L}(\beta, \lambda_1, \ldots, \lambda_n) \propto \prod_{k=1}^{n} \Big(\frac{1}{t_{(k)} - t_{(k-1)}}\ \frac{\psi(x_{[k]}, \beta)}{\sum_i\ V_{[i]}(t_{(k)})\psi(x_{[i]}, \beta)}\ \mathrm{e}^{-n}\Big)^{\delta_{[k]}}$$

and therefore

$$p\tilde{L}(\beta) \propto \prod_{k=1}^{n} \Big(\frac{\psi(x_k, \beta)}{\sum_j V_j(t_k)\psi(x_j, \beta)}\Big)^{\delta_k}.$$

Consequently, the maximal $\tilde{\beta}_n$ of the profile (approximated) likelihood is the same as in the previous approximation, i.e. in both cases the Cox estimator is the corresponding solution.

We note that $\tilde{\beta}_n$ depends on the $t_1, \ldots, t_n$ only in a restricted way: it is not important how large the differences $t_{(i+1)} - t_{(i)}$ are. Without censoring, it is obvious that the solution $\tilde{\beta}_n$ is a rank statistic in the observations, but these differences can have a lot of information about the regression part in which we are interested.

8.5. Statistical arguments

In section 8.4.2, both approximations for Λ used n parameters and the second approximation is a continuous function. However, both had the same solution. From the asymptotic point of view, the resulting Cox estimator is an efficient estimator ([EFR 77]). For relatively small sample sizes from the statistical point of view, it is very important to include the distances between different failure times because there is an information about the influence of the covariates. Thus, it is better – at least in these cases – to work with estimates from section 8.4.1. But here arises the problem of choosing an appropriate class $\mathcal{M}_j$ of smooth nuisance parameters. Often this is a step of experience or prior knowledge. In general, we cannot judge which approach is the better one. A broad simulation study can help for giving recommendations. For different types of baseline rate functions and moderate sample sizes n – sample sizes four or six times the unknown parameter number – samples are generated and the variances of the resulting estimates are calculated.

We now give a numerical example from Feigl and Zelen, given by Cox and Oakes [COX 84].

Example: two groups of leukaemia patients are considered and the failure time (time to death) in weeks is observed. For any patient the white blood counts (WBC) are given. The two groups are characterized by a positive (17 patients) or negative (16 patients) gene AG. In [COX 84] a proportional hazard model with 3 exploratory variables is taken:

$$\begin{aligned}
x_{(1)} &= \mathbf{1}_{AG=pos.},\\
x_{(2)} &= \ln(WBC) - 9.5,\\
x_{(3)} &= (x_{(1)} - 0.5152)x_{(2)}.
\end{aligned}$$

The proposed model is

$$\begin{aligned}
\lambda(t,x,\beta) = \lambda_0(t)\psi(x,\beta)&\\
=&\lambda_0(t)\exp\Big(\beta_1 x_{(1)} + \beta_2 x_{(2)} + \beta_3 x_{(3)}\Big).
\end{aligned}$$

Using the approach in section 8.4.1, we choose $\mathcal{M}_1$ as the exponential of quadratic polynomials and so we use

$$\tilde{\lambda}_0(t) \approx \exp(\eta_1 + \eta_2 t + \eta_3 t^2).$$

The resulting estimations are with $n = 33$

$$\hat{\beta}_{33}^{(1)} = [-1.398, 0.413, 0.44] \tag{8.7}$$

$$\hat{\eta}_{33}^{(1)} = [-0.832, 0.125, 0]. \tag{8.8}$$

Using the approach in section 8.4.2 then the estimate is the Cox estimator

$$\widehat{Cox} = [-1.14, 0.4, 0.5]. \tag{8.9}$$

The plots of the predictor $\beta_1 x_{(1)} + \beta_2 x_{(2)} + \beta_3 x_{(3)}$ with (8.7) and (8.9) in Figure 8.1 show some differences between both estimates but no other tendency. We point out that in the estimate $\hat{\beta}_{33}^{(1)}$ the failure times are really used not only their ranks. With the first method for a larger region, the WBC are less for the group of positive genes than for those with negative genes.

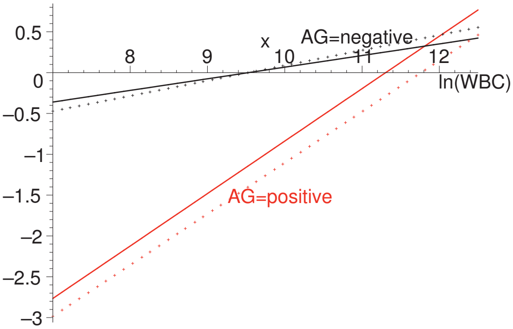


Figure 8.1. *Data of Feigl and Zelen*
with Cox estimator ——, with approx. MLE · · · · · ·

Acknowledgement: the author is very grateful to Professor H. Liero for her helpful comments and suggestions.

8.6. Bibliography

[AND 93] ANDERSON P., BORGAN O., GILL R., KEIDING N., *Statistical Models Based on Counting Processes*, Springer-Verlag, New York, 1993.

[BAG 02] BAGDONAVICIUS V., NIKULIN M., *Accelerated Life Models: Modelling and Statistical Analysis*, Chapman & Hall, London, 2002.

[BAR 94] BARNDORFF-NIELSEN O., COX D., *Inference and Asymptotics*, Chapman & Hall, London, 1994.

[BIC 93] BICKEL P., KLAASEN C., RITOV Y., WELLNER J., *Efficient and Adaptive Estimation for Semiparametric Models*, Johns Hopkins, Baltimore, 1993.

[BRE 74] BRESLOW N., “Covariance analysis of censored survival data.”, *Biometrics 30, 89-99*, 1974.

[COX 72] Cox D., “Regression models and life-tables (with discussion)”, *J. Roy. Statist. Soc. Ser. B 34,187-220*, 1972.

[COX 75] Cox D., “Partial likelihood”, *Biometrika 62 269-276.*, 1975.

[COX 84] Cox D., Oakes D., *Analysis of Survival Data*, Chapman & Hall, London, 1984.

[DAB 97] Dabrowska D., “Smoothed Cox regression”, *Ann. Statist. 25 1510-1540.*, 1997.

[EFR 77] Efron B., “The efficiency of Cox’s likelihood function for censored data”, *J. Amer. Statist. Assoc. 72, 557-565*, 1977.

[HUA 98] Huang J., Stone C., “The L^2 rate of convergence for event history regression with time-dependent covariates”, *Scand. J. Statist., 603-620*, 1998.

[LAW 03] Lawless J., *Statistical Models and Methods for Lifetime Data*, John Wiley, 2003.

[LIE 03] Liero H., “Goodness of fit tests of L_2 type”, in: *Statistical Inference for Semiparametric Models and Applications,* ed. by Nikulin, M. and Balakrishnan, N. and Limnios, N. and Mesbah, M., Birkhäuser, Boston, 2003.

[MUR 97] Murphy S., van der Vaart A., “Semiparametric likelihood ratio inference”, *Ann. Statist. 25, 1471-1509*, 1997.

[MUR 00] Murphy S., van der Vaart A., “On profile likelihoodpages”, *J. Amer. Statist. Ass.*, vol. 9, p. 449–465, 2000.

[OWE 01] Owen A., *Empirical Likelihood*, Chapman & Hall, London, 2001.

Part II

Reliability

This page intentionally left blank

Chapter 9

Cox Regression with Missing Values of a Covariate having a Non-proportional Effect on Risk of Failure

9.1. Introduction

The Cox model [COX 72] is by far the most used regression model for the analysis of survival data with explanatory variables. This model postulates that the hazard function for a failure time X associated with a p-dimensional explanatory variable (or covariate) Z takes the form

$$\lambda(t|Z) = \lambda_0(t)\exp(\beta' Z), \tag{9.1}$$

where β is an unknown p-vector of regression parameters and $\lambda_0(t)$ is an unknown and unspecified baseline hazard function. The statistical problem is that of estimating the parameter of interest β on the basis of a sample of survival times and corresponding covariates, the function $\lambda_0(t)$ being considered as a nuisance parameter.

In the following, we will consider that X may be right-censored by a non-informative random censoring time C such that X and C are independent conditionally on Z. Let (X_i, C_i) indicate the ith failure and censoring times, and let us denote

$$T_i = \min(X_i, C_i),\ \Delta_i = 1\{X_i \leq C_i\},\ N_i(t) = 1\{T_i \leq t\}\Delta_i,\ Y_i(t) \\ = 1\{T_i \geq t\},$$

Chapter written by Jean-François DUPUY and Eve LECONTE.

where $1\{\cdot\}$ is the indicator function. $N_i(t)$ is the failure counting process and $Y_i(t)$ is the at risk process. Observations are assumed to be n independent replicates denoted by (T_i, Δ_i, Z_i) for $(i = 1, \dots, n)$ of the triple (T, Δ, Z). Statistical inference in model (9.1) is based on the log-partial likelihood function [COX 72, COX 75], defined as

$$\sum_{i=1}^{n} \int_0^{\tau} \left(\beta' Z_i - \ln \sum_{j=1}^{n} Y_j(t) e^{\beta' Z_j} \right) dN_i(t), \tag{9.2}$$

where $\tau < \infty$ is a time point such that $P[T \geq \tau] > 0$. The maximum partial likelihood estimator of β is defined as the maximizer of (9.2), or equivalently as the solution to the estimating equation $U_n(\beta) = 0$, where

$$U_n(\beta) = \sum_{i=1}^{n} \int_0^{\tau} \left(Z_i - \frac{\sum_{j=1}^{n} Y_j(t) Z_j e^{\beta' Z_j}}{\sum_{j=1}^{n} Y_j(t) e^{\beta' Z_j}} \right) dN_i(t). \tag{9.3}$$

The maximum partial likelihood estimator has been shown to be consistent and asymptotically normal with a covariance matrix that can be consistently estimated [TSI 81, AND 82].

A key assumption made in model (9.1) is that the relative risks are constant with time: the proportional hazards assumption. Consider for a moment the special case where $p = 1$. The relative risk is the ratio $\lambda(t|Z+1)/\lambda(t|Z)$, which is equal to $\exp(\beta)$ and therefore does not depend on time. The covariate Z is said to have a proportional effect on the hazard of failure, or to be a proportional covariate. If $p \geq 2$, the same kind of result holds when two individuals having the same covariate vector are compared, except that they differ (by the value 1) on one component of Z. The assumption that the relative risks are constant with time is stringent and may fail in practice. Some examples can be found in [KLE 97]. In such a case, the results obtained by fitting model (9.1) can be misleading.

Consider the situation where two kinds of covariates are present: let Z denote a p-vector of proportional covariates and D be a real covariate whose effect on the hazard of failure is not proportional. Assume that D is discrete and takes its values in $\{0, \dots, K-1\}$ (if D is continuous, we may discretize its range before proceeding further). One commonly used approach for accomodating the non-proportional D in a Cox model-based analysis is to use the stratified Cox model ([KAL 02], Chapter 4). The stratified Cox model generalizes the usual Cox regression model by allowing different groups of the population on study (the strata) to have distinct baseline hazard functions. Here, the strata would divide the individuals into K disjoint groups according to the K values of D, each having a distinct baseline hazard λ_{0k} but a common value for the regression parameter. The hazard function for the failure time of an individual in stratum k (i.e. such that $D = k$) takes the form

$$\lambda_k(t|Z) = \lambda_{0k}(t) \exp(\beta' Z), \tag{9.4}$$

where $\{\lambda_{0k}(t) : t \geq 0, k = 0, \ldots, K-1\}$ are K unknown baseline hazard functions. Note that the relative risk of two individuals in distinct strata is not constant with time. As in model (9.1), β is estimated by the method of maximum partial likelihood estimation. The partial likelihood for the stratified Cox model is the product over strata of the stratum-specific partial likelihoods [FLE 91].

In many applications, measurements on certain components of the covariates Z are missing on some study subjects. Several examples are given in [LIN 93]. One immediate solution to this problem is to disregard individuals with missing covariate values. This solution, usually referred to as the *complete-case analysis*, is known to result in possible asymptotic bias of the maximum partial likelihood estimator, and in substantial reduction in efficiency, which in turn reduces the power of the test of nullity of the regression parameter. This is a serious drawback of the method, when the analysis focuses on risk factor selection.

In the last decade, several methods have therefore been proposed, which aim at recovering missing information, correcting bias and improving efficiency. It is worth noting that until now, contributions to Cox regression with missing covariate values have only considered missingness of proportional covariates, that is, covariates which enter the Cox model in its exponential component $\exp(\beta' Z)$. To the best of our knowledge, there has been no study of the missing covariate problem when the missing covariate has a non-proportional effect on the hazard of failure, and therefore is used for stratification.

Thus, in this chapter, we try to fill this gap and we consider the problem of estimating the Cox model with missing values of a non-proportional covariate. To put it another way, we consider the problem of estimating the stratified Cox model with partially known stratum affectation of the study subjects. It is anticipated that the same efficiency problem as in the Cox model with missing proportional covariates will occur if the complete-case analysis is applied to the stratified Cox model with missing stratum indicators. We will confirm this intuition in a simulation study reported at the end of the chapter. Our aim is therefore to propose a more refined method for estimation in the stratified Cox model with missing stratum indicator values.

To recover information for individuals with missing stratum indicator, we assume that an auxiliary covariate – a surrogate variable – is observed instead of the true stratum indicator, and that a complete data set (including both stratum indicator and surrogate) is available for a subsample of individuals, the validation subsample. The procedure we propose is as follows: based on the validation subsample and the surrogate variable, the missing stratum indicator values are estimated, and an approximated log-partial likelihood $\hat{l}_n(\beta)$ is derived by replacing the missing indicators by their estimators in the usual log-partial likelihood. We define an estimator $\hat{\beta}_n$ of the regression parameter as the value of β that maximizes $\hat{l}_n$.

The remainder of the chapter is organized as follows. In section 9.2, some methods for estimation in the Cox model with missing proportional covariates are reviewed, and their large-sample properties are recalled. We also briefly recall simulation results obtained by various authors who have compared the respective merits of these methods. In section 9.3, we state the notations, describe our problem, and define an estimator $\hat{\beta}_n$ of the parameter β in model (9.4), when stratum indicators are missing on some study individuals. In section 9.4, we state some asymptotic properties of $\hat{\beta}_n$ and we give brief outlines of the proofs. In section 9.5, we conduct a simulation study to complete the investigation of the properties of the proposed methodology. Section 9.6 provides some discussion.

9.2. Estimation in the Cox model with missing covariate values: a short review

For the sake of completeness, we also describe some methods of estimation in the Cox model with missing time-dependent covariates.

A commonly used method for handling missing covariate values in Cox regression, which has already been mentioned in section 9.1, is the *complete-case analysis*, which omits the subjects with unobserved covariate value. The complete-case method has two drawbacks: (i) the loss of efficiency due to ignoring data from incomplete cases can be important and (ii) this method can yield inconsistent estimators when the missingness of Z_i depends on the outcome (T_i, Δ_i).

The *approximate partial likelihood approach* proposed in [LIN 93] consists of replacing the weighted empirical mean term $\sum_{j=1}^{n} Y_j(t) Z_j e^{\beta' Z_j} / \sum_{j=1}^{n} Y_j(t) e^{\beta' Z_j}$ at time t in the partial likelihood score function (9.3) by an estimator based only on the subsample of individuals at risk at t, who have complete measurements on all covariate components at time t. It is crucial that this subsample is representative for the entire subpopulation at risk at t. Therefore, the missing data are assumed to be missing completely at random, that is, the probability that an individual has a missing covariate value does not depend on the covariate or on the survival of this individual. Note that the missing completely at random assumption is the simplest situation in the missing data hierarchy described by [LIT 87]. Under this assumption, the approximate partial likelihood estimator of β is consistent and asymptotically normal. Moreover, this estimator is generally more efficient than the complete-case estimator (it may be slightly less efficient when the censoring rate is low).

The *estimated partial likelihood method* proposed in [ZHO 95] non-parametrically estimates the conditional expectation $E[\exp(\beta' Z)|T \geq t, Z^{obs}]$ (where Z^{obs} denotes the observed component of Z). This method requires auxiliary covariate information (the surrogate variable) and a validation subsample with no missing covariate measurements. The auxiliary covariates are required to be discrete. Again, it is crucial, for the consistency of the estimated partial likelihood estimator of β to hold, that the covariates are missing completely at random. Simulations show that important efficiency gains can be made with the estimated partial likelihood method relative to the complete-case analysis, when the surrogate and missing covariate are strongly correlated.

[PON 02] considers estimation in the Cox model with a covariate missing completely at random, and assumes that a continuous surrogate W and a complete validation subsample are available. The proposed method uses kernel estimators of the conditional expectations $E[Z^k(t)\exp(\beta' Z(t))|W(t) = w]$ $(k = 0, 1, 2)$. An approximated likelihood based on these estimators is obtained, and serves as a basis for deriving estimators of both the regression parameter and the cumulative baseline hazard function $\Lambda_0(t) = \int_0^t \lambda_0(s)\,ds$. These estimators are shown to be consistent and asymptotically normal.

The method proposed in [PAI 97] resembles the estimated partial likelihood approach of [ZHO 95]. It requires that some covariates are fully observed for all individuals under study. Let Z_1 denote the completely observed covariate and Z_2 denote a possibly missing completely at random covariate. [PAI 97]'s method replaces the missing term $\exp(\beta' Z_2(t))$ at time t in the Cox partial likelihood score function, for an individual i with missing $Z_2(t)$, by an average term based on subjects at risk at t, with observed $Z_2(t)$ and with covariate Z_1 equal to the observed Z_1 for individual i. [PAI 97] also considers the case where Z_2 is missing at random [LIT 87], which occurs when the missingness probability depends on observed variables, such as the completely observed covariate Z_1 and/or the failure or censoring time. In this case, the imputation process proposed by [PAI 97] is more complicated: it relies on nonparametric estimation of the missing covariate $Z_2(t)$ in neighborhoods defined by time intervals. Simulations show that in the missing completely at random situation, the estimator of the regression parameter associated with the completely observed covariate is more efficient than the approximate partial likelihood estimator.

A *regression calibration* approach for Cox regression with missing covariate has been investigated by [WAN 97] (see also [WAN 01b] and references therein). The regression calibration method basically estimates the missing data based on a validation subsample of individuals for whom the covariate measurements are all observed. Let X be a covariate which may be missing for some individuals, Z be a covariate which is always observed, and W be a surrogate measure for X. The method requires a model for the relation between X and (Z, W), that is, the investigator needs to postulate that $E[X|Z, W] = g(Z, W, \alpha)$ for some function g specified up to an unknown parameter α. Next, the method can be carried out by the following two steps: (i) estimate the missing covariate values by $g(Z, W, \hat{\alpha})$, where $\hat{\alpha}$ estimates α from the validation subsample (ii) apply a Cox regression by using the observed covariate (X, Z) in the validation set and $(g(Z, W, \hat{\alpha}), Z)$ in the non-validation set. [WAN 97] shows that the resulting estimator of β is not consistent (it converges to some β^* rather than to the true regression parameter). However, simulations conducted by the authors indicate that the asymptotic bias stays limited, provided $g(Z, W, \alpha)$ is reasonably approximated. Also, simulations indicate that regression calibration performs well, compared to the approximate and estimated partial likelihood approaches, as long as the regression parameter associated with the missing covariable is not too large. See [WAN 97] for a more detailed discussion.

Augmented inverse probability weighted estimator for Cox regression with missing covariate is discussed in [WAN 01a]. This method basically consists of weighting each individual contribution to the complete-case score function by the inverse of the selection probability of the individual (the original idea of using inverse weighting of complete cases as an estimation method is due to [HOR 52]). The selection probability is the probability that an individual has a complete covariate. In many practical situations, this probability is unknown and has to be estimated. A detailed description of augmented inverse probability weighted complete-case estimators can be found in [TSI 06].

Non-parametric maximum likelihood (NPML) estimation has been proposed by various authors, either for missing fixed-time covariates [CHE 99, MAR 99], or for missing time-dependent covariates [DUP 02, DUP 06a]. The likelihood for the observed data is obtained by integrating the complete likelihood over the unobserved covariate data. Direct maximization of such a likelihood is not possible since the baseline hazard is not specified. However, a discretized version of this likelihood (the non-parametric likelihood) is formed by discretizing the cumulative baseline hazard function as an increasing step function with jumps at the observed uncensored event times only. The method relies on the NPML interpretation of the Nelson-Aalen estimator (see [AND 93]). EM-type algorithms are then applied to solve the maximization problem. In the case of fixed-time covariates, [CHE 99] and [MAR 99] allow covariates to be missing at random. Simulations by [MAR 99] indicate that the NPML method is more efficient than the method proposed by [PAI 97]. Simulations by [CHE 99] indicate that when covariates are missing completely at random, the NPML estimator of the regression parameter is more efficient than the complete-case and approximate partial likelihood estimators. In the missing at random case, the NPML method is less biased and more efficient than the complete-case and approximate partial likelihood approaches. By considering the situation where the value of a time-dependent covariate is unobserved when the event (or its censoring) occurs, [DUP 02] and [DUP 06a] focus on the non-ignorable non-response situation, where the probability that the covariate is missing depends on the unobserved value of this covariate. Asymptotic results for the estimators of the regression parameter, of the cumulative baseline hazard function and of the parameters of the distribution of the missing covariate are obtained by [CHE 99] and [DUP 06a].

9.3. Estimation procedure in the stratified Cox model with missing stratum indicator values

We consider the problem of estimation in the stratified Cox regression model when the stratum indicator is missing for some individuals but a surrogate variable is observed instead, and a complete data set is available for a subsample of individuals (called the validation subsample).

Recall that X denotes a random failure time that may be randomly right-censored by C, $T = \min(X, C)$, and $\Delta = 1\{X \leq C\}$. We assume that the hazard function for X takes the form of a stratified Cox model with 2 strata, labeled 0 and 1. For each study subject, define the stratum indicator D which equals 0 (respectively 1) if the subject belongs to stratum 0 (respectively stratum 1).

The hazard for the failure of an individual in stratum D can be written as:

$$\lambda(t|Z, D) = \begin{cases} \lambda_{00}(t)\exp(\beta' Z) \text{ if } D = 0 \\ \lambda_{01}(t)\exp(\beta' Z) \text{ if } D = 1, \end{cases} \tag{9.5}$$

or, equivalently, as $\lambda(t|Z, D) = \lambda_{00}(t)\bar{D}\exp(\beta' Z) + \lambda_{01}(t)D\exp(\beta' Z)$, where $\bar{D} = 1 - D$. The log-partial likelihood for β from n independent copies $(T_i, \Delta_i, Z_i, D_i)$ $(i = 1, \ldots, n)$ of (T, Δ, Z, D) is

$$\sum_{i=1}^{n} \int_0^{\tau} \left(\beta' Z_i - \bar{D}_i \ln \sum_{j=1}^{n} Y_j(s)\bar{D}_j e^{\beta' Z_j} - D_i \ln \sum_{j=1}^{n} Y_j(s) D_j e^{\beta' Z_j} \right) dN_i(s). \tag{9.6}$$

The maximum partial likelihood estimator of the regression parameter is defined as the value of β maximizing (9.6), or equivalently as the solution to the estimating equation

$$\sum_{i=1}^{n} \int_0^{\tau} \left(Z_i - \bar{D}_i \frac{\sum_{j=1}^{n} Y_j(s)\bar{D}_j Z_j e^{\beta' Z_j}}{\sum_{j=1}^{n} Y_j(s)\bar{D}_j e^{\beta' Z_j}} - D_i \frac{\sum_{j=1}^{n} Y_j(s) D_j Z_j e^{\beta' Z_j}}{\sum_{j=1}^{n} Y_j(s) D_j e^{\beta' Z_j}} \right) dN_i(s) = 0. \tag{9.7}$$

Now, consider the situation where the value of the stratum indicator D may be missing for some (but not all) individuals. We assume that a surrogate variable W for D is observed for all individuals. Therefore a validation subsample is available where all variables (T, Δ, Z, D, W) are observed. In the non-validation subsample, only data on (T, Δ, Z, W) are observed. Let V denote the validation subsample, and let n_V denote its size. The surrogate adds no information when D is observed, so that X is independent of W given (Z, D). We assume that the surrogate W is categorical and takes its values in a finite set $\mathcal{W}$.

We denote by R the indicator variable which is equal to 1 if D is observed and 0 otherwise. We assume that R is independent of X, C, D, Z and W, that is, the probability that an individual has a missing stratum indicator value does not depend on the stratum indicator, or on the covariate Z and surrogate W, or on the survival time of the individual. This corresponds to the missing completely at random situation [RUB 76]. The random vectors $(T_i, \Delta_i, Z_i, W_i, D_i, R_i)$ $(i = 1, \ldots, n)$ are n independent copies of (T, Δ, Z, W, D, R). D_i is unobserved when $R_i = 0$. Let β_0 denote the true value of β.

Our method consists of replacing the missing stratum indicator values in (9.6) by some estimators based on the validation subsample and the surrogate variable. Our suggestion is as follows. If D_i is unobserved for some individual i having $W_i = w_i$, we propose to calculate $p_{w_i} := P[D = 0|W = w_i]$ and to assign this individual to stratum 0 (respectively stratum 1) if $p_{w_i} \geq 1/2$ (respectively $p_{w_i} < 1/2$). This results in the following modified conditional log-partial likelihood given the surrogate values $w_1, \ldots, w_n$:

$$\begin{aligned}
&\sum_{i=1}^{n} \int_0^{\tau} \Bigg(\beta' Z_i - (R_i \bar{D}_i + \bar{R}_i 1\{p_{w_i} \geq 1/2\}) \\
&\times \ln \left\{ \sum_{j=1}^{n} Y_j(s)(R_j \bar{D}_j + \bar{R}_j 1\{p_{w_j} \geq 1/2\}) e^{\beta' Z_j} \right\} - (R_i D_i + \bar{R}_i 1\{p_{w_i} < 1/2\}) \\
&\qquad\qquad \times \ln \left\{ \sum_{j=1}^{n} Y_j(s)(R_j D_j + \bar{R}_j 1\{p_{w_j} < 1/2\}) e^{\beta' Z_j} \right\} \Bigg) dN_i(s), \qquad (9.8)
\end{aligned}$$

where $\bar{R}_i = 1 - R_i$. Since the probabilities p_{w_i} are unknown, we cannot directly estimate β_0 by maximizing (9.8). Instead, we define an estimator of β_0 from an approximation of (9.8), obtained by replacing p_{w_i}, for any individual such that $R_i = 0$, by an estimator $\hat{p}_{w_i}$ based on the data available for individuals with $R_i = 1$.

Consider an individual i such that $R_i = 0$ and $W_i = w_i$. We propose to estimate p_{w_i} by

$$\hat{p}_{w_i} = \frac{\sum_{j=1}^{n} R_j 1\{W_j = w_i\} 1\{D_j = 0\}}{\sum_{j=1}^{n} R_j 1\{W_j = w_i\}} = \frac{\sum_{j \in V} 1\{W_j = w_i\} 1\{D_j = 0\}}{\sum_{j \in V} 1\{W_j = w_i\}},$$

which is the proportion of individuals in the validation subsample, which lies in stratum 0 and has a surrogate value equal to w_i.

From the strong law of large numbers and the assumption of independence of R and other random variables, it follows that as $n_V \to \infty$, $\hat{p}_{w_i} \longrightarrow p_{w_i}$ almost surely. This prompts us to replace p_{w_i} by $\hat{p}_{w_i}$ in (9.8).

Let us define $\xi_i = R_i D_i + \bar{R}_i 1\{p_{w_i} < 1/2\}$ and $\bar{\xi}_i = R_i \bar{D}_i + \bar{R}_i 1\{p_{w_i} \geq 1/2\}$. Let also $\hat{\xi}_i$ (respectively $\hat{\bar{\xi}}_i$) denote the approximated value of ξ_i (respectively $\bar{\xi}_i$), obtained by replacing the unknown p_{w_i} by its estimator $\hat{p}_{w_i}$.

Then, replacing ξ_i and $\bar{\xi}_i$ by $\hat{\xi}_i$ and $\hat{\bar{\xi}}_i$ in (9.8) yields the following approximated log-partial likelihood function:

$$\hat{l}_n(\beta) =$$

$$\sum_{i=1}^{n} \int_0^{\tau} \left(\beta' Z_i - \hat{\bar{\xi}}_i \ln \left\{ \sum_{j=1}^{n} Y_j(s) \hat{\bar{\xi}}_j e^{\beta' Z_j} \right\} - \hat{\xi}_i \ln \left\{ \sum_{j=1}^{n} Y_j(s) \hat{\xi}_j e^{\beta' Z_j} \right\} \right) dN_i(s),$$

from which we derive the following score function, which is an approximated version of (9.7):

$$\hat{U}_n(\beta) = \partial \hat{l}_n(\beta) / \partial \beta$$

$$= \sum_{i=1}^{n} \int_0^{\tau} \left(Z_i - \hat{\bar{\xi}}_i \frac{\sum_{j=1}^{n} Y_j(s) \hat{\bar{\xi}}_j Z_j e^{\beta' Z_j}}{\sum_{j=1}^{n} Y_j(s) \hat{\bar{\xi}}_j e^{\beta' Z_j}} - \hat{\xi}_i \frac{\sum_{j=1}^{n} Y_j(s) \hat{\xi}_j Z_j e^{\beta' Z_j}}{\sum_{j=1}^{n} Y_j(s) \hat{\xi}_j e^{\beta' Z_j}} \right) dN_i(s).$$

We finally define the approximated partial likelihood (APL) estimator $\hat{\beta}_n$ of β_0 as the root to the estimating equation $\hat{U}_n(\beta) = 0$, which can be solved by a usual Cox regression program using observed stratum indicators in the validation subsample and estimated stratum indicators in the non-validation subsample.

9.4. Asymptotic theory

In the estimation method we propose, we can view the actual stratum indicator D as being misspecified as $\xi = RD + \bar{R} 1\{p_w < 1/2\}$. Consider an individual i with missing stratum indicator (i.e. $R_i = 0$). If $\xi_i = 0$, we assign individual i to stratum 0, but the unobserved D_i may in fact be equal to 1. Similarly, if $\xi_i = 1$, individual i is assigned to stratum 1, but the unobserved D_i may actually be equal to 0.

The proposed estimator $\hat{\beta}_n$ of the regression parameter in model (9.5) is based on the misspecified version ξ of D. Hence, by analogy with [STR 86] who have established the asymptotic bias resulting from misspecifying a proportional covariate, we may expect $\hat{\beta}_n$ to be asymptotically biased. More precisely, we may expect that $\hat{\beta}_n$ actually estimates the value β^* in the model $\lambda(t|R, D, Z, W) = \lambda_0^*(t) \bar{\xi} \exp(\beta^{*'} Z) + \lambda_1^*(t) \xi \exp(\beta^{*'} Z)$, rather than the true β_0.

In the following, we show that $\hat{\beta}_n$ converges in probability to β^*. Moreover, we show that $\hat{\beta}_n$ is asymptotically normal, with a variance that can be consistently estimated. A theoretical investigation of the asymptotic bias of $\hat{\beta}_n$ appears to be a difficult task. Thus, in section 9.5, we conduct a simulation study to investigate the magnitude of the bias of the proposed estimator under various scenarios.

In order to study the asymptotic behavior of $\hat{\beta}_n$, we introduce further notations. Let $\Lambda_0^*(t) = \int_0^t \lambda_0^*(s)\,ds$ and $\Lambda_1^*(t) = \int_0^t \lambda_1^*(s)\,ds$. If u is a vector in $\mathbb{R}^d$, let $u^{\otimes 0} = 1, u^{\otimes 1} = u$, and $u^{\otimes 2} = uu'$. For $k = 0, 1, 2$, $s \in [0, \tau]$ and $\beta \in \mathbb{R}^d$, let $s^{(k)}(s, \beta) = \mathbb{E}[Y(s) Z^{\otimes k} \xi e^{\beta' Z}]$ and $\bar{s}^{(k)}(s, \beta) = \mathbb{E}[Y(s) Z^{\otimes k} \bar{\xi} e^{\beta' Z}]$. Define also

$$\hat{S}_n^{(k)}(s,\beta) = \sum_{j=1}^{n} Y_j(s) \hat{\xi}_j Z_j^{\otimes k} e^{\beta' Z_j}, \qquad \hat{\bar{S}}_n^{(k)}(s,\beta) = \sum_{j=1}^{n} Y_j(s) \hat{\bar{\xi}}_j Z_j^{\otimes k} e^{\beta' Z_j},$$

$$S_n^{(k)}(s,\beta) = \sum_{j=1}^{n} Y_j(s) \xi_j Z_j^{\otimes k} e^{\beta' Z_j}, \qquad \bar{S}_n^{(k)}(s,\beta) = \sum_{j=1}^{n} Y_j(s) \bar{\xi}_j Z_j^{\otimes k} e^{\beta' Z_j}.$$

For $t \in [0, \tau]$ and $\beta \in \mathbb{R}^d$, define the matrices

$$\bar{I}(\beta, t) = \int_0^t \bar{s}^{(0)}(s, \beta^*) \left[\frac{\bar{s}^{(2)}(s,\beta)}{\bar{s}^{(0)}(s,\beta)} - \left\{ \frac{\bar{s}^{(1)}(s,\beta)}{\bar{s}^{(0)}(s,\beta)} \right\}^{\otimes 2} \right] d\Lambda_0^*(s),$$

$$I(\beta, t) = \int_0^t s^{(0)}(s, \beta^*) \left[\frac{s^{(2)}(s,\beta)}{s^{(0)}(s,\beta)} - \left\{ \frac{s^{(1)}(s,\beta)}{s^{(0)}(s,\beta)} \right\}^{\otimes 2} \right] d\Lambda_1^*(s),$$

and let $\bar{I}(\beta) := \bar{I}(\beta, \tau)$ and $I(\beta) := I(\beta, \tau)$.

The asymptotic properties of the estimator $\hat{\beta}_n$ will be established under the following regularity conditions:

C1 The value β^* belongs to the interior of a compact and convex subset $\mathcal{B}$ of $\mathbb{R}^d$. $\Lambda_0^*(t)$ and $\Lambda_1^*(t)$ are continuous, $\Lambda_0^*(\tau) < \infty$ and $\Lambda_1^*(\tau) < \infty$.

C2 The matrices $I(\beta)$ and $\bar{I}(\beta)$ are positive-definite for every $\beta \in \mathcal{B}$.

C3 The functions $s^{(0)}(s, \beta)$ and $\bar{s}^{(0)}(s, \beta)$ are bounded away from 0 on $[0, \tau] \times \mathcal{B}$.

C4 We assume that the covariate vector Z is bounded.

C5 We assume that for all $w \in \mathcal{W}$, $p_w := P[D = 0 | W = w] \neq \frac{1}{2}$.

Conditions C1-C3 are similar to Andersen and Gill's (1982) regularity conditions for the usual Cox model. The boundedness condition C4 holds for the case of time-independent covariates Z, but the arguments below can also accomodate time-dependent covariates, provided that appropriate regularity conditions (see [AND 82]) are made. Condition C5 ensures that we can assign an individual with a missing stratum indicator to one of the strata. We also assume that the respective sizes n_V and $n - n_V$ of the validation and non-validation subsamples both tend to infinity as n tends to infinity.

We now state our results and provide brief outlines of the proofs. Detailed proofs can be found in [DUP 06b].

Theorem 9.1 *Under the conditions stated above, $\hat{\beta}_n$ converges in probability to β^*.*

Outline of the proof. This result follows from the convergence of $\hat{A}_n(\cdot) = n^{-1}(\hat{l}_n(\cdot) - \hat{l}_n(\beta^*))$ to a function having a unique maximum at β^*. More precisely, it can be shown that for each $\beta \in \mathcal{B}$, $\hat{A}_n(\beta)$ converges in probability to

$$A(\beta) = \int_0^\tau \left((\beta - \beta^*)' \bar{s}^{(1)}(s, \beta^*) - \bar{s}^{(0)}(s, \beta^*) \ln \left\{ \frac{\bar{s}^{(0)}(s, \beta)}{\bar{s}^{(0)}(s, \beta^*)} \right\} \right) d\Lambda_0^*(s)$$
$$+ \int_0^\tau \left((\beta - \beta^*)' s^{(1)}(s, \beta^*) - s^{(0)}(s, \beta^*) \ln \left\{ \frac{s^{(0)}(s, \beta)}{s^{(0)}(s, \beta^*)} \right\} \right) d\Lambda_1^*(s).$$

Moreover, calculating the first two derivatives of A with respect to β, it can be shown that A is strictly concave with a unique maximum at β^*. Since $\hat{\beta}_n$ maximizes the concave function $\hat{A}_n(\beta)$, the convex analysis argument in Appendix 2 of [AND 82] allows us to conclude that $\hat{\beta}_n$ converges in probability to β^*.
□

Then we have the following theorem for the asymptotic distribution of the approximated score $n^{-1/2}\hat{U}_n(\beta^*)$.

Theorem 9.2 *Under the conditions stated above, the random vector $n^{-1/2}\hat{U}_n(\beta^*)$ is asymptotically normal with mean 0 and covariance matrix $\Gamma(\beta^*) = \bar{I}(\beta^*) + I(\beta^*)$.*

Outline of the proof. Weak convergence of $n^{-1/2}\hat{U}_n(\beta^*)$ is proved by adapting the arguments developed by [AND 82] for the classical Cox model.

In the course of this proof, it is useful to view $\hat{U}_n(\beta)$ as the value at $t = \tau$ of the process $\{\hat{U}_n(\beta, t) : t \in [0, \tau]\}$ given by

$$\hat{U}_n(\beta, t) = \sum_{i=1}^n \int_0^t \left(Z_i - \hat{\bar{\xi}}_i \frac{\hat{\bar{S}}_n^{(1)}(s, \beta)}{\hat{\bar{S}}_n^{(0)}(s, \beta)} - \xi_i \frac{\hat{S}_n^{(1)}(s, \beta)}{\hat{S}_n^{(0)}(s, \beta)} \right) dN_i(s).$$

Define

$$U_n(\beta, t) = \sum_{i=1}^{n} \int_0^t \left(Z_i - \bar{\xi}_i \frac{\bar{S}_n^{(1)}(s,\beta)}{\bar{S}_n^{(0)}(s,\beta)} - \xi_i \frac{S_n^{(1)}(s,\beta)}{S_n^{(0)}(s,\beta)} \right) dN_i(s).$$

It follows after some algebra that we can rewrite $U_n(\beta^*, t)$ as the martingale integral

$$U_n(\beta^*, t) = \sum_{i=1}^{n} \int_0^t H_{n,i}(s)\, dM_i(s),$$

where

$$H_{n,i}(s) = Z_i - \bar{\xi}_i \frac{\bar{S}_n^{(1)}(s,\beta^*)}{\bar{S}_n^{(0)}(s,\beta^*)} - \xi_i \frac{S_n^{(1)}(s,\beta^*)}{S_n^{(0)}(s,\beta^*)},$$

and $M(t) = N(t) - \int_0^t Y(s)[\lambda_0^*(s)\bar{\xi} + \lambda_1^*(s)\xi] e^{\beta^{*'} Z} ds$ is a martingale with respect to the filtration $(\mathcal{F}_t)_{t \geq 0}$ defined by $\mathcal{F}_t = \mathcal{H}_t^R \mathcal{G}_t^{\bar{R}}$, where $\mathcal{H}_t = \sigma\{1\{X \leq s\}, 1\{C \leq s\}, Z, W, D : s \leq t\}$ and $\mathcal{G}_t = \sigma\{1\{X \leq s\}, 1\{C \leq s\}, Z, W : s \leq t\}$.

The central limit theorem for martingales (see [FLE 91], Chapter 5) allows us to show that $\{n^{-1/2} U_n(\beta^*, t) : t \in [0, \tau]\}$ converges weakly to a zero-mean d-variate Gaussian process with covariance function at t equal to $\bar{I}(\beta^*, t) + I(\beta^*, t)$. First, consider the predictable variation process $\left\langle n^{-1/2} U_n(\beta^*, .), n^{-1/2} U_n(\beta^*, .) \right\rangle (t)$. Some algebra shows that it is equal to the following expression:

$$\int_0^t n^{-1} \left[\bar{S}_n^{(2)}(s,\beta^*) - \frac{\bar{S}_n^{(1)}(s,\beta^*)^{\otimes 2}}{\bar{S}_n^{(0)}(s,\beta^*)} \right] d\Lambda_0^*(s) \\ + \int_0^t n^{-1} \left[S_n^{(2)}(s,\beta^*) - \frac{S_n^{(1)}(s,\beta^*)^{\otimes 2}}{S_n^{(0)}(s,\beta^*)} \right] d\Lambda_1^*(s),$$

which converges in probability to $\bar{I}(\beta^*, t) + I(\beta^*, t)$. Then, the Lindeberg condition of the martingale central limit theorem may also be seen to be fulfilled (see [DUP 06b] for detailed calculations), so it follows that the martingale central limit theorem applies.

Now, some algebra shows that $n^{-1/2}(\hat{U}_n(\beta, t) - U_n(\beta, t)) = o_p(1)$, hence it holds that $n^{-1/2}\hat{U}_n(\beta^*) := n^{-1/2}\hat{U}_n(\beta^*, \tau) = n^{-1/2} U_n(\beta^*, \tau) + o_p(1)$. Applying

Slutsky's theorem gives the asymptotic distribution of $n^{-1/2}\hat{U}_n(\beta^*)$ and completes the proof. □

The next theorem establishes the asymptotic normality of $\hat{\beta}_n$. A consistent estimator for $\Gamma(\beta^*)$ is given in the proof.

Theorem 9.3 *Under the conditions stated above,* $n^{1/2}(\hat{\beta}_n - \beta^*)$ *is asymptotically normal with mean* 0 *and covariance matrix* $\Gamma(\beta^*)^{-1}$.

OUTLINE OF THE PROOF. Let $D_n(\beta) = -n^{-1}\partial\hat{U}_n(\beta)/\partial\beta$. Then $D_n(\beta)$ is equal to

$$n^{-1}\sum_{i=1}^{n}\int_0^{\tau} \hat{\hat{\xi}}_i \left[\frac{\hat{\hat{S}}_n^{(2)}(s,\beta)}{\hat{\hat{S}}_n^{(0)}(s,\beta)} - \left\{\frac{\hat{\hat{S}}_n^{(1)}(s,\beta)}{\hat{\hat{S}}_n^{(0)}(s,\beta)}\right\}^{\otimes 2}\right] \\ + \hat{\xi}_i \left[\frac{\hat{S}_n^{(2)}(s,\beta)}{\hat{S}_n^{(0)}(s,\beta)} - \left\{\frac{\hat{S}_n^{(1)}(s,\beta)}{\hat{S}_n^{(0)}(s,\beta)}\right\}^{\otimes 2}\right] dN_i(s).$$

A Taylor expansion yields $D_n(\tilde{\beta}_n)n^{1/2}(\hat{\beta}_n - \beta^*) = n^{-1/2}\hat{U}_n(\beta^*)$, where $\tilde{\beta}_n$ is on the line segment between $\hat{\beta}_n$ and β^*. Therefore, to prove asymptotic normality of $n^{1/2}(\hat{\beta}_n - \beta^*)$, it is sufficient to show that $n^{-1/2}\hat{U}_n(\beta^*)$ is asymptotically normal (Theorem 9.2) and that $D_n(\tilde{\beta}_n)$ converges in probability to an invertible matrix. This latter convergence can be verified by taking the Taylor expansion of $D_n(\tilde{\beta}_n)$ at β^*, and the limiting invertible matrix is $\Gamma(\beta^*)$. The result now follows from Slutsky's theorem and Theorem 9.2. Similarly, taking the Taylor expansion of $D_n(\hat{\beta}_n)$ at β^*, we find that $D_n(\hat{\beta}_n)$ is a consistent estimator of $\Gamma(\beta^*)$. □

9.5. A simulation study

A simulation study was carried out to investigate the performance of the proposed estimator $\hat{\beta}_n$ in various situations. The interested reader may request an R program from the second author. We considered the stratified Cox model of equation (9.5) with $\lambda_{00}(t) = 0.75$ and $\lambda_{01}(t) = 1.75$, with the regression parameter β fixed to 0.3. Z was a scalar covariate generated from a univariate normal distribution with mean 1 and variance 1. Censoring times were generated from the uniform distribution on $[0, \theta]$, where θ was chosen to yield a censoring rate of approximately 30% and 60%. The case of no censoring was also considered. Two sample sizes were considered ($n = 200, 300$), with half of the sample in each stratum. Results for $n = 300$ are not reported, since they yield the same conclusions as for $n = 200$ [DUP 06b]. For each sample size, three sizes of the validation subsample were considered, by generating the indicators R_i such that $P[R_i = 1] = 0.2, 0.5, 0.8$, corresponding respectively to $80\%, 50\%$ and 20% of missingness of the stratum indicator values. The surrogate W for D was assumed to take values in $\{0, 1\}$. Two levels of association of D and W were considered : a high level by setting $P[D = 0|W = 0] = 0.75$ and $P[D = 0|W = 1] = 0.2$, and a low level by setting $P[D = 0|W = 0] = 0.4$ and $P[D = 0|W = 1] = 0.55$.

For each combination of the simulation parameters, 10,000 data sets were generated. For each sample size and censoring percentage, under the high level of association of D and W, our method misclassified on average about 23% of the individuals with missing stratum indicators (45% under the low level of association). For comparison, the method based on complete cases only (the CC analysis) was also evaluated. The results of a full-data (FD) analysis using the actual values of the missing stratum indicators as if they were known were also obtained. These latter results indeed provide a natural benchmark for evaluating the proposed method and the CC analysis.

Table 9.1 summarizes the results of the simulations. The "Bias" means the average bias of the 10,000 estimates, the "Mean($\widehat{\text{se}}$)" denotes the average of the 10,000 standard error estimates, the "Var." denotes the sample variance of the 10,000 estimates, and the "MSE" denotes the sample mean square error. Finally, "Power" refers to the power of the Wald test at the 5% level for testing nullity of the regression parameter.

Remark 9.1 *FD method shows the results that would be obtained if the missing stratum indicators were observed. CC method deletes all subjects with unobserved stratum. APL method refers to the proposed approximated partial likelihood method. APL^l refers to the case of a low level of association of W and D, APL^h refers to the case of a high level of association of W and D. % miss. refers to the percentage of missing stratum indicator values in the whole sample.*

As expected, CC analysis generally performs better than the proposed APL method in terms of bias, but it is quite inefficient, particularly in the case of heavy stratum missingness. Compared to CC analysis, the proposed method has a smaller variance for the parameter estimator. Moreover, the mean square error from the CC method is greater than the mean square error from the APL in all cases, and this superiority of the APL is particularly evident when stratum missingness is heavy. Finally, compared to the CC analysis, the proposed method has a higher power for the Wald test in all cases. Again, the gain in power is particularly evident when the stratum missingness and censoring are heavy. This should rule out CC analysis as an estimation method in the stratified Cox regression model when stratum missingness is moderate or heavy.

n=200, Censoring percentage=0%

% miss.	0	20			50			80		
Method	FD	CC	APL^l	APL^h	CC	APL^l	APL^h	CC	APL^l	APL^h
Bias	.0026	.0027	-.0094	-.0041	.0054	-.0218	-.0128	.0169	-.0292	-.0196
Mean($\widehat{\text{s.e.}}$)	.0762	.0858	.0759	.0760	.1109	.0756	.0758	.1912	.0754	.0757
Var.	.0058	.0073	.0058	.0058	.0124	.0057	.0058	.0406	.0057	.0057
MSE	.0058	.0073	.0059	.0058	.0125	.0062	.0059	.0409	.0065	.0061
Power	.9797	.9466	.9715	.9760	.7962	.9608	.9684	.3884	.9538	.9619

n=200, Censoring percentage=30%

% miss.	0	20			50			80		
Method	FD	CC	APL^l	APL^h	CC	APL^l	APL^h	CC	APL^l	APL^h
Bias	.0034	.0040	-.0062	-.0020	.0060	-.0169	-.0089	.0175	-.0235	-.0146
Mean($\widehat{\text{s.e.}}$)	.0896	.1009	.0893	.0894	.1306	.0890	.0892	.2248	.0888	.0891
Var.	.0082	.0105	.0083	.0083	.0182	.0083	.0082	.0580	.0083	.0083
MSE	.0083	.0105	.0083	.0083	.0183	.0086	.0083	.0583	.0088	.0085
Power	.9272	.8576	.9109	.9190	.6546	.8901	.9083	.2981	.8790	.8960

n=200, Censoring percentage=60%

% miss.	0	20			50			80		
Method	FD	CC	APL^l	APL^h	CC	APL^l	APL^h	CC	APL^l	APL^h
Bias	.0030	.0038	-.0023	-.0001	.0069	-.0083	-.0038	.0227	-.0122	-.0070
Mean($\widehat{\text{s.e.}}$)	.1173	.1321	.1171	.1172	.1710	.1169	.1170	.2981	.1167	.1169
Var.	.0142	.0182	.0142	.0142	.0309	.0142	.0142	.1067	.0142	.0143
MSE	.0142	.0183	.0142	.0142	.0310	.0143	.0142	.1072	.0144	.0143
Power	.7359	.6379	.7226	.7285	.4352	.7063	.7191	.1844	.6925	.7081

Table 9.1. *Simulation results for various combinations of sample size and censoring percentage*

9.6. Discussion

We propose a method to solve the estimation problem in the stratified Cox regression model with missing stratum indicator values. This problem may occur when the Cox model is stratified according to a covariate subject to missingness. Stratification is used when a covariate has a non-proportional effect on the hazard of failure. Many recent works have considered estimation in the Cox model with missing covariate, but these works have assumed that the missing covariate has a proportional effect on the hazard of failure. The proposed method non-parametrically estimates the missing stratum indicator values, and approximates the usual log-partial likelihood for the regression parameter by replacing the missing indicators by their estimators. Due to the misclassification of some individuals, the proposed estimator is asymptotically biased. However, simulation results show that our method estimates the true parameter reasonably well under a range of conditions. Moreover, the proposed estimator is asymptotically normal, which allows us to perform a Wald test for covariate selection. An intuitively natural other method for estimation in such a situation is a CC analysis. Our simulations show that compared to CC analysis, the method we propose has smaller variance for the regression parameter estimator, and provides an important increase of the power of the Wald test.

The proposed method replaces the missing stratum indicator values D_i by $1\{\hat{p}_{w_i} < 1/2\}$. An alternative is to replace each missing D_i by the realization of a Bernoulli random variable with parameter $1 - \hat{p}_{w_i}$. Simulation studies show that the APL estimator derived from this random stratum allocation performs very similarly to the proposed estimator. In particular, although the percentages of misclassified individuals are higher with the alternative random stratum affectation method (respectively 49% and 35% under the low and high levels of association), the alternative estimator produces the same mean square errors and powers as the proposed one.

We consider here the case of a categorical surrogate variable, but our method can be modified to accomodate a continuous one. For example, if W is a continuous real surrogate, we may adapt our procedure by replacing the probability $P[D = 0|W = w]$ by $P[D = 0|W \in J_k]$, where $J_k = (a_k, a_{k+1}]$ and $-\infty = a_0 < a_1 < \ldots < a_N < a_{N+1} = \infty$ is a partition of the range of W. We may estimate $P[D = 0|W \in J_k]$ by the following histogram type estimator: $\sum_{j=1}^{n} R_j 1\{W_j \in J_k\}1\{D_j = 0\}/ \sum_{j=1}^{n} R_j 1\{W_j \in J_k\}$, which can be shown to converge to $P[D = 0|W \in J_k]$, by the same arguments as above. Choosing the number and location of the a_k requires some investigation. A related question occurs when W is categorical but has a large number of categories. In such a case, we may need to combine some categories so that we have enough observed validation individuals inside each combined category, to serve as a basis for the estimation of $P[D = 0|W \in J_k]$.

The proposed method can be generalized to more than 2 strata. This would proceed by assigning any individual with missing stratum indicator to the stratum which has the greatest estimated probability to contain this individual (in case of two strata, this boils down to assigning an individual to the stratum which contains the individual with a probability greater than $1/2$).

Finally, the missing completely at random assumption may not be satisfied in some situations. A less stringent requirement is the missing at random (MAR) assumption – [RUB 76], which here would state that the probability of missing on the stratum indicator D depends on some completely observed variables but not on D. When the probability that the stratum indicator D is missing depends on D itself, the missingness is said to be non-ignorable (NI). More work is needed to investigate estimation in the stratified Cox model with missing stratum indicators in MAR and NI situations.

9.7. Bibliography

[AND 82] ANDERSEN P. K., GILL R. D., "Cox's regression model for counting processes: a large sample study", *Ann. Statist.*, vol. 10, num. 4, p. 1100–1120, 1982.

[AND 93] ANDERSEN P. K., BORGAN Ø., GILL R. D., KEIDING N., *Statistical Models Based on Counting Processes*, Springer Series in Statistics, Springer-Verlag, New York, 1993.

[CHE 99] CHEN H. Y., LITTLE R. J. A., "Proportional hazards regression with missing covariates", *J. Amer. Statist. Assoc.*, vol. 94, num. 447, p. 896–908, 1999.

[COX 72] COX D. R., "Regression models and life-tables (with discussion)", *J. Roy. Statist. Soc. Ser. B*, vol. 34, p. 187–220, 1972.

[COX 75] COX D. R., "Partial likelihood", *Biometrika*, vol. 62, num. 2, p. 269–276, 1975.

[DUP 02] DUPUY J.-F., MESBAH M., "Joint modeling of event time and nonignorable missing longitudinal data", *Lifetime Data Anal.*, vol. 8, num. 2, p. 99–115, 2002.

[DUP 06a] DUPUY J.-F., GRAMA I., MESBAH M., "Asymptotic theory for the Cox model with missing time-dependent covariate", *Ann. Statist.*, vol. 34, num. 2, p. 903–924, 2006.

[DUP 06b] DUPUY J.-F., LECONTE E., "Estimation in a partially observed stratified Cox model", vol. Technical Report 2006-10, Laboratoire de Statistique et Probabilités, University of Toulouse 3, France, 2006.

[FLE 91] FLEMING T. R., HARRINGTON D. P., *Counting Processes and Survival Analysis*, Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics, John Wiley & Sons Inc., New York, 1991.

[HOR 52] HORVITZ D. G., THOMPSON D. J., "A generalization of sampling without replacement from a finite universe", *J. Amer. Statist. Assoc.*, vol. 47, p. 663–685, 1952.

[KAL 02] KALBFLEISCH J. D., PRENTICE R. L., *The Statistical Analysis of Failure Time Data*, Wiley Series in Probability and Statistics, John Wiley & Sons, second edition, 2002.

[KLE 97] KLEIN J. P., MOESCHBERGER M. L., *Survival Analysis: Methods for Censored and Truncated Data*, Statistics for Biology and Health, Springer, New York, 1997.

[LIN 93] LIN D. Y., YING Z., "Cox regression with incomplete covariate measurements", *J. Amer. Statist. Assoc.*, vol. 88, num. 424, p. 1341–1349, 1993.

[LIT 87] LITTLE R. J. A., RUBIN D. B., *Statistical Analysis with Missing Data*, Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics, John Wiley & Sons, New York, 1987.

[MAR 99] MARTINUSSEN T., "Cox regression with incomplete covariate measurements using the EM-algorithm", *Scand. J. Statist.*, vol. 26, num. 4, p. 479–491, 1999.

[PAI 97] PAIK M. C., TSAI W.-Y., "On using the Cox proportional hazards model with missing covariates", *Biometrika*, vol. 84, num. 3, p. 579–593, 1997.

[PON 02] PONS O., "Estimation in the Cox model with missing covariate data", *J. Nonparametr. Stat.*, vol. 14, num. 3, p. 223–247, 2002.

[RUB 76] RUBIN D. B., "Inference and missing data", *Biometrika*, vol. 63, num. 3, p. 581–592, 1976.

[STR 86] STRUTHERS C. A., KALBFLEISCH J. D., "Misspecified proportional hazard models", *Biometrika*, vol. 73, num. 2, p. 363–369, 1986.

[TSI 81] TSIATIS A. A., "A large sample study of Cox's regression model", *Ann. Statist.*, vol. 9, num. 1, p. 93–108, 1981.

[TSI 06] TSIATIS A. A., *Semiparametric Theory and Missing Data*, Springer Series in Statistics, Springer, New York, 2006.

[WAN 97] WANG C. Y., HSU L., FENG Z. D., PRENTICE R. L., "Regression calibration in failure time regression", *Biometrics*, vol. 53, num. 1, p. 131–145, 1997.

[WAN 01a] WANG C. Y., CHEN H. Y., "Augmented inverse probability weighted estimator for Cox missing covariate regression", *Biometrics*, vol. 57, num. 2, p. 414–419, 2001.

[WAN 01b] WANG C. Y., XIE S. X., PRENTICE R. L., "Recalibration based on an approximate relative risk estimator in Cox regression with missing covariates", *Statist. Sinica*, vol. 11, num. 4, p. 1081–1104, 2001.

[ZHO 95] ZHOU H., PEPE M. S., "Auxiliary covariate data in failure time regression", *Biometrika*, vol. 82, num. 1, p. 139–149, 1995.

Chapter 10

Exact Bayesian Variable Sampling Plans for Exponential Distribution under Type-I Censoring

10.1. Introduction

The use of the decision-theoretic approach to define acceptance sampling plans with censoring has received considerable attention during the past two decades. Much of the work has focused on Bayesian variable sampling plans for exponential and Weibull distributions under different forms of censoring. [LAM 90, LAM 94] used the polynomial loss functions to discuss the exponential Bayesian variable sampling plans under Type-II and Type-I censoring, respectively; see [BAL 95] for a comprehensive review of these developments. Following the work of [LAM 94], [HUA 02] proposed a new decision function and considered the cost of unit time in the loss function. Similar work for the case of random censoring has been carried out by [LAM 95], [HUA 04] and [CHE 04]. For the Weibull model, [CHE 99] used an exponential loss function to determine the Bayesian variable sampling plan under Type-I censoring.

In this chapter, we consider the Bayesian sampling plan for the exponential distribution under Type-I censoring by taking a totally different point of view from the earlier work. Suppose that the lifetime of an item in the lot is distributed as exponential Exp(λ) with density function $f(x) = \lambda e^{-\lambda x}$ for $x > 0$. Let M denote the number of failures that occur before the pre-fixed censoring time, t. It can be easily verified that the maximum likelihood estimator (MLE) of the average lifetime, $\theta = 1/\lambda$, fails to

Chapter written by Chien-Tai LIN, Yen-Lung HUANG and N. BALAKRISHNAN.

exist when $M = 0$. Thus, we propose to use the conditional probability density function of the MLE of θ, $\hat{\theta}$, given $M \geq 1$, to examine the variable sampling plan. Given $M \geq 1$, we first derive in section 10.2 the conditional moment generating function of $\hat{\theta}$ and identify the corresponding conditional probability density function. The Bayes risk is then formulated by using a quadratic loss function. Employing the discretization method discussed in [LAM 94], an algorithm is also presented for obtaining the optimal sampling plan. Finally, in section 10.3, we present some numerical examples and comparisons in order to illustrate the effectiveness of the proposed plan, and to demonstrate that its minimum Bayes risk is in general less than that of the sampling plan of [LAM 94], in addition to having the life-test shorter in general.

10.2. Proposed sampling plan and Bayes risk

Suppose that a lot of items is presented for acceptance sampling. Assume that the lifetime of an item in the lot has an Exp(λ) distribution, where the scale parameter λ has a conjugate Gamma prior distribution $G(a, b)$ with density function

$$g(\lambda; a, b) = \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-\lambda b}, \qquad \lambda > 0. \tag{10.1}$$

Let d denote the action on accepting or rejecting this inspected lot. When $d = 1$, the lot is accepted, and when $d = 0$, the lot is then rejected. For a given sample of size n from the lot, the loss function is defined as

$$l(d, \lambda, n) = nC_s + d(a_0 + a_1\lambda + \cdots + a_p\lambda^p) + (1 - d)C_r,$$

where C_s is the cost of inspecting an item, C_r is the loss of rejecting the lot, and $a_0 + a_1\lambda + \cdots + a_p\lambda^p$ is the loss of accepting the lot. Here, we also assume that $a_0 + a_1\lambda + \cdots + a_p\lambda^p$ is positive for $\lambda > 0$.

Let $X_{(1)} \leq \cdots \leq X_{(n)}$ be the order statistics of the given sample $\boldsymbol{X} = (X_1, \ldots, X_n)$. Given a pre-fixed censoring time $t > 0$, it is well-known that the MLE of θ obtained from a Type-I censored sample is

$$\hat{\theta} = \frac{1}{\hat{\lambda}} = \frac{\sum_{i=1}^{M} X_{(i)} + (n - M)t}{M},$$

where M is the random number of failures that occur before time t. Since the MLE of θ fails to exist when $M = 0$, it is natural to consider using the decision function

$$\delta(\boldsymbol{X}) = \begin{cases} 1, & \hat{\theta} \geq T \\ \\ 0, & \text{otherwise} \end{cases}$$

for acceptance when $M \geq 1$. In order to derive the Bayes risk of a sampling plan (n, t, T), we first need to know the conditional distribution of $\hat{\theta}$, given $M \geq 1$. The following two results give the conditional moment generating function and probability density function of $\hat{\theta}$, given $M \geq 1$. These results are along the same lines as those used by [CHI 03] and [CHA 04] in the case of exponential distribution under hybrid censoring.

Theorem 10.1 *The conditional moment generating function of $\hat{\theta}$, given $M \geq 1$, is given by*

$$E_\theta(e^{w\hat{\theta}}) = \frac{1}{1-q^n} \sum_{m=1}^{n} \sum_{k=0}^{m} (-1)^k \binom{m}{k} \binom{n}{m} \left(1 - \frac{w}{\lambda m}\right)^{-m} q^{n-m+k} e^{w(n-m+k)t/m},$$

where $q = e^{-\lambda t}$.

Proof. It is evident that M has a binomial $Bin(n, p)$ distribution with probability mass function $P(M = m) = \binom{n}{m} p^m (1-p)^{n-m}$ for $m = 0, 1, \cdots, n$, where $p = 1 - q = 1 - e^{-\lambda t}$. Given that $M \geq 1$, we have $E_\theta(e^{w\hat{\theta}}) = (1-q^n)^{-1} \sum_{m=1}^{n} E_\theta(e^{w\hat{\theta}}|M = m) P_\theta(M = m)$. Note that the term $E_\theta(e^{w\hat{\theta}}|M = m)P_\theta(M = m)$ can be re-expressed as $E_\theta(e^{w\hat{\theta}}|X_{(m)} \leq t \leq X_{(m+1)}) P_\theta(X_{(m)} \leq t \leq X_{(m+1)})$. For $m = 1, \ldots, n$, the joint density function of $(X_{(1)}, \ldots, X_{(m)})$, given $X_{(m)} \leq t \leq X_{(m+1)}$, is given by

$$f_{X_{(1)},\ldots,X_{(m)}}(x_1, \ldots, x_m | X_{(m)} \leq t \leq X_{(m+1)}) = \frac{n!}{(n-m)! P_\theta(M = m)} \lambda^m e^{-\lambda\left\{\sum_{i=1}^{m} x_i + (n-m)t\right\}}.$$

Hence,

$$\begin{aligned}
&E_\theta(e^{w\hat{\theta}}|X_{(m)} \leq t \leq X_{(m+1)}) P_\theta(X_{(m)} \leq t \leq X_{(m+1)}) \\
&= \frac{n!}{(n-m)!} \int_0^t \int_0^{x_m} \cdots \int_0^{x_2} e^{w\hat{\theta}} \lambda^m e^{-\lambda[\sum_{i=1}^{m} x_i + (n-m)t]} dx_1 \cdots dx_{m-1} dx_m \\
&= \frac{n! \lambda^m}{(n-m)!} \int_0^t \int_0^{x_m} \cdots \int_0^{x_2} e^{-\lambda(1-\frac{w}{\lambda m})[\sum_{i=1}^{m} x_i + (n-m)t]} dx_1 \cdots dx_m \\
&= \frac{n! \lambda^m}{(n-m)!} e^{-\lambda(1-\frac{w}{\lambda m})(n-m)t} \int_0^t \int_0^{x_m} \cdots \int_0^{x_2} e^{-\lambda(1-\frac{w}{\lambda m}) \sum_{i=1}^{m} x_i} dx_1 \cdots dx_m.
\end{aligned}$$

From the identity

$$\int_0^{x_j}\int_0^{x_{j-1}}\cdots\int_0^{x_2} e^{-a\sum_{i=1}^{j-1}x_i}dx_1\cdots dx_{j-1} = \frac{(1-e^{-ax_j})^{j-1}}{a^{j-1}(j-1)!},$$

the above expression can be re-expressed as

$$\begin{aligned}
&\frac{n!\lambda^m}{(n-m)!}e^{-\lambda(1-\frac{w}{\lambda m})(n-m)t}\int_0^t e^{-\lambda(1-\frac{w}{\lambda m})x_m}\frac{[1-e^{-\lambda(1-\frac{w}{\lambda m})x_m}]^{m-1}}{[\lambda(1-\frac{w}{\lambda m})]^{m-1}(m-1)!}dx_m\\
&\quad = \frac{n!}{(n-m)!}\frac{\lambda(1-\frac{w}{\lambda m})^{-(m-1)}}{(m-1)!}e^{-\lambda(1-\frac{w}{\lambda m})(n-m)t}\\
&\quad \times\int_0^t e^{-\lambda(1-\frac{w}{\lambda m})x_m}[1-e^{-\lambda(1-\frac{w}{\lambda m})x_m}]^{m-1}dx_m\\
&\quad = \binom{n}{m}\left(1-\frac{w}{\lambda m}\right)^{-m}e^{-\lambda(1-\frac{w}{\lambda m})(n-m)t}[1-e^{-\lambda(1-\frac{w}{\lambda m})t}]^m.
\end{aligned}$$

Using binomial expansion, we then obtain the conditional moment generating function of $\hat{\theta}$, given $M \geq 1$, as

$$\begin{aligned}
&\frac{1}{1-q^n}\sum_{m=1}^{n}\sum_{k=0}^{m}(-1)^k\binom{m}{k}\binom{n}{m}\left(1-\frac{w}{\lambda m}\right)^{-m}e^{-\lambda(1-\frac{w}{\lambda m})(n-m+k)t}\\
&\quad = \frac{1}{1-q^n}\sum_{m=1}^{n}\sum_{k=0}^{m}(-1)^k\binom{m}{k}\binom{n}{m}\left(1-\frac{w}{\lambda m}\right)^{-m}q^{n-m+k}e^{w(n-m+k)t/m}.
\end{aligned}$$

■

Since $(1-\frac{w}{\lambda m})^{-m}e^{w(n-m+k)t/m}$ is the moment generating function of $Y + (n-m+k)t/m$ at w, where Y is a gamma random variable with density function $g(y; m, \lambda m)$ defined in (10.1), the conditional probability density function of $\hat{\theta}$, given $M \geq 1$, is easily derived as follows.

Theorem 10.2 *Given that $M \geq 1$, the conditional probability density function of $\hat{\theta}$ is given by*

$$\begin{aligned}
f_{\hat{\theta}}(x) &= \frac{1}{1-q^n}\sum_{m=1}^{n}\sum_{k=0}^{m}(-1)^k\binom{m}{k}\binom{n}{m}q^{n-m+k}g\left(x-t^{\star}_{k,m}; m, \lambda m\right)\\
&= \sum_{j=0}^{\infty}\sum_{m=1}^{n}\sum_{k=0}^{m}(-1)^k\binom{m}{k}\binom{n}{m}q^{n-m+k+nj}g\left(x-t^{\star}_{k,m}; m, \lambda m\right),
\end{aligned}$$

where $t^{\star}_{k,m} = (n-m+k)t/m$ and g is as defined in (10.1).

Now we are ready to evaluate the Bayes risk. For the purpose of illustration, let assume that $p = 2$ in the loss function, i.e.,

$$l(\delta(\boldsymbol{X}), \lambda, n) = nC_s + \delta(\boldsymbol{X})(a_0 + a_1\lambda + a_2\lambda^2) + (1 - \delta(\boldsymbol{X}))C_r.$$

Then, the Bayes risk for $\delta(\boldsymbol{X})$ is

$$\begin{aligned}
&R(n,t,T) = E[l(\delta(\boldsymbol{X}), \lambda, n)] = E_\lambda\{E_{\boldsymbol{X}|\lambda}[l(\delta(\boldsymbol{X}), \lambda, n)|\lambda]\} \\
&= E_\lambda\{E_{\boldsymbol{X}|\lambda}[nC_s + \delta(\boldsymbol{X})(a_0 + a_1\lambda + a_2\lambda^2) + (1 - \delta(\boldsymbol{X}))C_r|\lambda]\} \\
&= E_\lambda\{E_{\boldsymbol{X}|\lambda}[nC_s + (a_0 + a_1\lambda + a_2\lambda^2) + (1 - \delta(\boldsymbol{X}))(C_r - a_0 - a_1\lambda - a_2\lambda^2)|\lambda]\} \\
&= nC_s + a_0 + a_1E(\lambda) + a_2E(\lambda^2) + E\{[C_r - (a_0 + a_1\lambda + a_2\lambda^2)]P(\hat{\theta} < T)\} \\
&= nC_s + a_0 + a_1\mu_1 + a_2\mu_2 \\
&+ \int_0^\infty \int_0^T \left[(C_r - a_0) - a_1\lambda - a_2\lambda^2\right] f_{\hat{\theta}}(x)\frac{b^a}{\Gamma(a)}\lambda^{a-1}e^{-\lambda b}dxd\lambda \\
&= nC_s + a_0 + a_1\mu_1 + a_2\mu_2 + \textstyle\sum_{\ell=0}^2 C_\ell \frac{b^a}{\Gamma(a)} \int_0^\infty \int_0^T \lambda^{a+\ell-1}e^{-\lambda b} f_{\hat{\theta}}(x)dxd\lambda,
\end{aligned} \tag{10.2}$$

where μ_1 and μ_2 are the first and second moments of λ about 0, and

$$C_\ell = \begin{cases} C_r - a_\ell, & \text{if } \ell = 0 \\ \\ -a_\ell, & \text{otherwise.} \end{cases}$$

Note that

$$\begin{aligned}
&\textstyle\int_0^\infty \int_0^T \lambda^{a+\ell-1}e^{-\lambda b}f_{\hat{\theta}}(x)dxd\lambda \\
&= \textstyle\sum_{j=0}^\infty \sum_{m=1}^n \sum_{k=0}^m (-1)^k \binom{m}{k}\binom{n}{m} \int_0^\infty \int_0^T \lambda^{a+\ell-1}e^{-\lambda b}q^{n-m+k+nj}g\Big(x - t^\star_{k,m}; m, \lambda m\Big)dxd\lambda \\
&= \textstyle\sum_{j=0}^\infty \sum_{m=1}^n \sum_{k=0}^m (-1)^k \binom{m}{k}\binom{n}{m} \frac{m^m}{\Gamma(m)} \int_0^\infty \int_{t^\star_{k,m}}^T \lambda^{a+\ell+m-1}e^{-\lambda(b+njt+mx)}\Big(x - t^\star_{k,m}\Big)^{m-1}dxd\lambda \\
&= \textstyle\sum_{j=0}^\infty \sum_{m=1}^n \sum_{k=0}^m (-1)^k \binom{m}{k}\binom{n}{m} \frac{m^m}{\Gamma(m)} \int_{t^\star_{k,m}}^T \frac{\Gamma(a+\ell+m)(x-t^\star_{k,m})^{m-1}}{(b+njt+mx)^{a+\ell+m}}dx.
\end{aligned} \tag{10.3}$$

Setting $C_{k,m} = b + njt + mt^{\star}_{k,m}$, Eq. (10.3) can then be written as

$$\begin{aligned}
&\textstyle\sum_{j=0}^{\infty} \sum_{m=1}^{n} \sum_{k=0}^{m} (-1)^k \binom{m}{k}\binom{n}{m} \frac{m^m}{\Gamma(m)} \int_0^{T-t^{\star}_{k,m}} \frac{\Gamma(a+\ell+m) y^{m-1}}{(C_{k,m}+my)^{a+\ell+m}} dy \\
&= \textstyle\sum_{j=0}^{\infty} \sum_{m=1}^{n} \sum_{k=0}^{m} (-1)^k \binom{m}{k}\binom{n}{m} \frac{m^m}{\Gamma(m)} \frac{\Gamma(a+\ell+m)}{(C_{k,m})^{a+\ell+m}} \int_0^{T-t^{\star}_{k,m}} \frac{y^{m-1}}{(1+\frac{m}{C_{k,m}}y)^{a+\ell+m}} dy \\
&= \textstyle\sum_{j=0}^{\infty} \sum_{m=1}^{n} \sum_{k=0}^{m} (-1)^k \binom{m}{k}\binom{n}{m} \frac{\Gamma(a+\ell)}{(C_{k,m})^{a+\ell}} \frac{\Gamma(a+\ell+m)}{\Gamma(m)\Gamma(a+\ell)} \int_0^{\frac{m(T-t^{\star}_{k,m})}{C_{k,m}}} \frac{z^{m-1}}{(1+z)^{a+\ell+m}} dz.
\end{aligned} \tag{10.4}$$

Making a transformation $z = u/(1-u)$, we have

$$\int_0^{C^\star} \frac{z^{m-1}}{(1+z)^{a+\ell+m}} dz = \int_0^{S^\star} u^{m-1}(1-u)^{a+\ell-1} du = B_{S^\star}(m, a+\ell),$$

where $C^\star = m(T - t^{\star}_{k,m})/C_{k,m}$, $S^\star = C^\star/(1 + C^\star)$, and

$$B_x(\alpha, \beta) = \int_0^x t^{\alpha-1}(1-t)^{\beta-1} dt, \qquad 0 \leq x \leq 1,$$

is the incomplete beta function. Let $I_x(\alpha, \beta) = B_x(\alpha, \beta)/B(\alpha, \beta)$ denote the beta distribution function. Substituting (10.4) into (10.2), the Bayes risk is obtained as

$$\begin{aligned}
&R(n,t,T) \\
&\quad = R_1 + \sum_{\ell=0}^{2} \sum_{j=0}^{\infty} \sum_{m=1}^{n} \sum_{k=0}^{m} C_\ell (-1)^k \binom{m}{k}\binom{n}{m} \frac{b^a}{\Gamma(a)} \frac{\Gamma(a+\ell)}{(C_{k,m})^{a+\ell}} I_{S^\star}(m, a+\ell),
\end{aligned} \tag{10.5}$$

where $R_1 = nC_s + a_0 + a_1\mu_1 + a_2\mu_2$.

10.3. Numerical examples and comparison

We have evaluated, for selected values of a, b, a_0, a_1, a_2, C_r and C_s, the Bayes risk from (10.5), determined the minimum Bayes risk $R(n_0, t_0, T_0)$, and compared them with the results of [LAM 94]. Employing the same discretization method suggested by [LAM 94], Tables 10.1–10.6 present the minimum Bayes risks for the two Bayesian sampling plans when only one of the parameters or one of the coefficients varies. Here, Plan 1 is the plan proposed in the last section and Plan 2 is the sampling plan suggested by [LAM 94]. In these tables, we also present the efficiency of the proposed plan, which is simply the ratio of the minimum Bayes risk of Lam's plan and that of the proposed plan.

From the efficiency values presented in Tables 10.1–10.6, it is clear that the proposed plan is considerably more efficient in general than Lam's plan (with efficiency often significantly larger than 100%); but this plan is not uniformly better in that the efficiency in some cases is below 100%, but always remains close to 100% in these instances (with the lowest value achieved being 95.9%). However, it is important to note here that the minimum sample size n_0 for the proposed plan is always at least as large as the minimum sample size of Lam's plan which, of course, is offset by the fact that the proposed plan possesses smaller censoring time t_0 in general, which means that the life-test under the proposed plan will last shorter than the life-test under Lam's plan. Figure 10.1 provides a graphical comparison of the minimum Bayes risks of these two sampling plans when $a = 2.5$, $b = 0.8$, $a_0 = 2.0$, $a_1 = 2.0$, $a_2 = 2.0$, $C_s = 0.5$ and $C_r = 30$, which once again demonstrates the overall superiority of the proposed sampling plan.

a	b	Plan	$R(n_0, t_0, T_0)$	n_0	t_0	T_0	Efficiency
2.5	0.4	1	18.4436	10	0.0444	0.3110	161.3%
		2	29.7506	1	0.7978	0.7978	
2.5	0.6	1	8.4755	12	0.0465	0.3717	327.8%
		2	27.7834	3	0.8537	0.4268	
1.5	0.8	1	16.7533	3	0.5262	0.2631	99.2%
		2	16.6233	3	0.5262	0.2631	
2.0	0.8	1	21.2875	4	0.6051	0.3026	99.7%
		2	21.2153	3	0.6051	0.3026	
2.5	0.8	1	5.0041	10	0.0620	0.3717	498.3%
		2	24.9367	3	0.7077	0.3539	
3.0	0.8	1	23.9122	10	0.0454	0.2949	115.5%
		2	27.6136	3	0.8170	0.4085	
3.5	0.8	1	20.4594	11	0.0356	0.2492	143.1%
		2	29.2789	2	1.0037	0.5019	
2.5	1.0	1	12.0104	14	0.0774	1.0842	181.2%
		2	21.7640	3	0.5483	0.2742	
2.5	1.2	1	18.7384	3	0.4158	0.2079	99.3%
		2	18.6097	3	0.4158	0.2079	
10.0	3.0	1	29.5959	2	0.8194	0.4097	99.7%
		2	29.5166	2	0.7928	0.3964	

Table 10.1. *The minimum Bayes risks and optimal sampling plans for* $a_0 = 2.0$, $a_1 = 2.0$, $a_2 = 2.0$, $C_s = 0.5$, $C_r = 30$, *and some selected values of* a *and* b

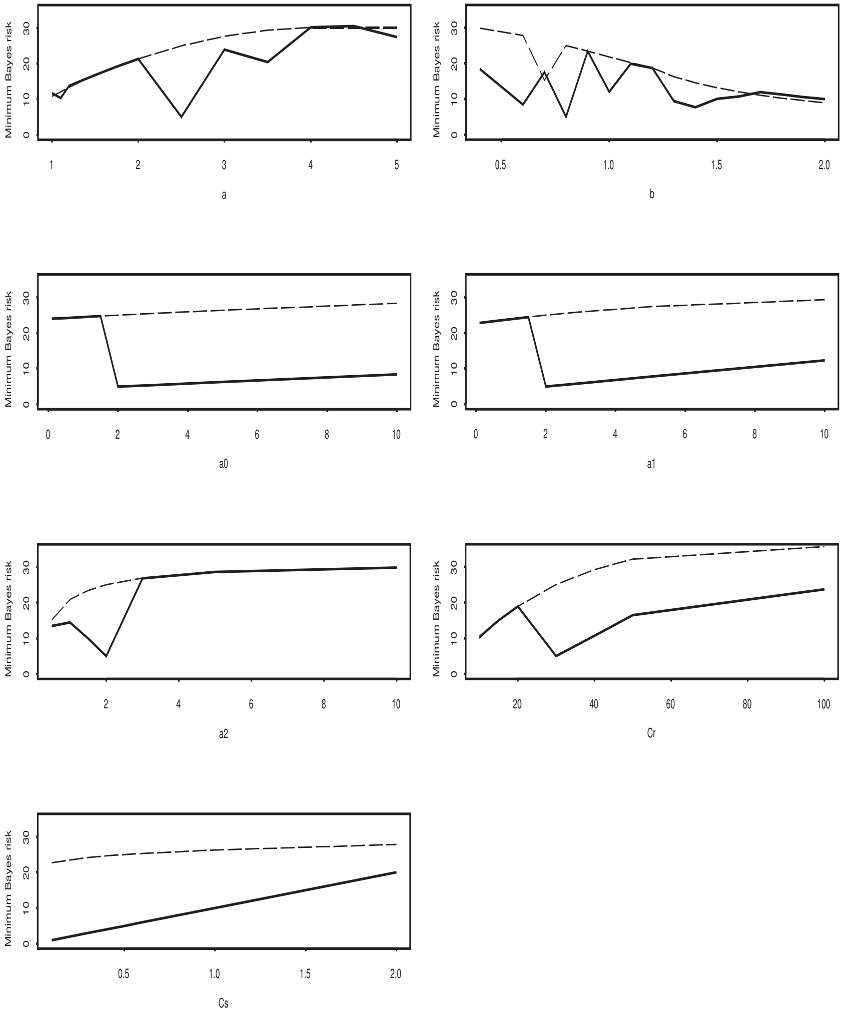


Figure 10.1. *The comparison of the minimum Bayes risks of the two sampling plans when $a = 2.5$, $b = 0.8$, $a_0 = 2.0$, $a_1 = 2.0$, $a_2 = 2.0$, $C_s = 0.5$, and $C_r = 30$. The solid and dashed lines in each plot represent the minimum Bayes risk of Plan 1 and Plan 2, respectively*

a_0	Plan	$R(n_0, t_0, T_0)$	n_0	t_0	T_0	Efficiency
0.1	1	23.9966	4	0.6539	0.3269	99.9%
	2	23.9743	4	0.6539	0.3269	
0.5	1	24.2101	4	0.6539	0.3269	99.9%
	2	24.1874	3	0.6808	0.3404	
1.5	1	24.7318	4	0.6808	0.3404	99.8%
	2	24.6925	3	0.6808	0.3404	
2.0	1	5.0041	10	0.0620	0.3717	498.3%
	2	24.9367	3	0.7077	0.3539	
3.0	1	5.4303	10	0.0620	0.3717	468.1%
	2	25.4172	3	0.7346	0.3673	
5.0	1	6.2827	10	0.0620	0.3717	419.1%
	2	26.3287	3	0.7884	0.3942	
10.0	1	8.4136	10	0.0620	0.3717	336.1%
	2	28.2745	2	1.0037	0.5018	

Table 10.2. *The minimum Bayes risks and optimal sampling plans for* $a = 2.5$, $b = 0.8$, $a_1 = 2.0$, $a_2 = 2.0$, $C_s = 0.5$, $C_r = 30$, *and some selected values of* a_0

a_1	Plan	$R(n_0, t_0, T_0)$	n_0	t_0	T_0	Efficiency
0.1	1	22.8050	4	0.5732	0.2866	99.9%
	2	22.7788	4	0.5732	0.2866	
0.5	1	23.3141	4	0.6001	0.3000	99.9%
	2	23.2897	4	0.6001	0.3000	
1.5	1	24.4684	4	0.6539	0.3269	99.9%
	2	24.4325	3	0.6808	0.3404	
2.0	1	5.0041	10	0.0620	0.3717	498.3%
	2	24.9367	3	0.7077	0.3539	
3.0	1	5.9210	10	0.0620	0.3717	436.4%
	2	25.8399	3	0.7884	0.3942	
5.0	1	7.7546	10	0.0620	0.3717	351.7%
	2	27.2715	3	0.9499	0.4749	
10.0	1	12.3387	10	0.0620	0.3717	236.8%
	2	29.2151	2	1.5687	0.7844	

Table 10.3. *The minimum Bayes risks and optimal sampling plans for* $a = 2.5$, $b = 0.8$, $a_0 = 2.0$, $a_2 = 2.0$, $C_s = 0.5$, $C_r = 30$, *and some selected values of* a_1

a_2	Plan	$R(n_0, t_0, T_0)$	n_0	t_0	T_0	Efficiency
0.5	1	13.4929	12	0.0620	0.4956	111.8%
	2	15.0859	0	0	0	
1.0	1	14.4746	12	0.0620	0.4956	143.9%
	2	20.8319	3	0.3848	0.1924	
1.5	1	9.9782	10	0.0620	0.3717	234.0%
	2	23.3494	3	0.5463	0.2731	
2.0	1	5.0041	10	0.0620	0.3717	498.3%
	2	24.9367	3	0.7077	0.3539	
3.0	1	26.8623	3	0.9768	0.4884	99.8%
	2	26.8155	3	0.9499	0.4749	
5.0	1	28.5944	3	1.3804	0.6902	99.9%
	2	28.5677	3	1.3804	0.6902	
10.0	1	29.8832	2	2.3759	1.1879	99.7%
	2	29.8049	1	1.7032	1.7032	

Table 10.4. *The minimum Bayes risks and optimal sampling plans for* $a = 2.5$, $b = 0.8$, $a_0 = 2.0$, $a_1 = 2.0$, $C_s = 0.5$, $C_r = 30$, *and some selected values of* a_2

C_r	Plan	$R(n_0, t_0, T_0)$	n_0	t_0	T_0	Efficiency
10	1	10.4252	1	2.6988	1.3494	95.9%
	2	10.0000	0	0	∞	
15	1	15.0499	2	1.5687	0.7844	98.8%
	2	14.8625	1	2.1068	1.0534	
20	1	18.9763	2	1.1382	0.5691	99.4%
	2	18.8574	2	1.1382	0.5691	
30	1	5.0041	10	0.0620	0.3717	498.3%
	2	24.9367	3	0.7077	0.3539	
40	1	10.7422	10	0.0620	0.3717	271.5%
	2	29.1674	4	0.5194	0.2597	
50	1	16.4804	10	0.0620	0.3717	194.9%
	2	32.1176	5	0.4117	0.2059	
100	1	23.7702	12	0.0620	0.2168	149.7%
	2	35.5938	0	0	0	

Table 10.5. *The minimum Bayes risks and optimal sampling plans for* $a = 2.5$, $b = 0.8$, $a_0 = 2.0$, $a_1 = 2.0$, $a_2 = 2.0$, $C_s = 0.5$, *and some selected values of* C_r

C_s	Plan	$R(n_0, t_0, T_0)$	n_0	t_0	T_0	Efficiency
0.1	1	1.0041	10	0.0620	0.3717	2257.2%
	2	22.6644	11	0.3270	0.3135	
0.3	1	3.0041	10	0.0620	0.3717	802.6%
	2	24.1116	5	0.6808	0.3404	
0.4	1	4.0041	10	0.0620	0.3717	613.6%
	2	24.5696	4	0.6808	0.3404	
0.5	1	5.0041	10	0.0620	0.3717	498.3%
	2	24.9367	3	0.7077	0.3539	
0.6	1	6.0041	10	0.0620	0.3717	420.3%
	2	25.2367	3	0.7077	0.3539	
1.0	1	10.0041	10	0.0620	0.3717	262.2%
	2	26.2303	2	0.7077	0.3539	
2.0	1	20.0041	10	0.0620	0.3717	138.8%
	2	27.7605	1	0.7884	0.3942	

Table 10.6. *The minimum Bayes risks and optimal sampling plans for* $a = 2.5$*,* $b = 0.8$*,* $a_0 = 2.0$*,* $a_1 = 2.0$*,* $a_2 = 2.0$*,* $C_r = 30$*, and some selected values of* C_s

10.4. Bibliography

[BAL 95] BALAKRISHNAN N., BASU A. P. E., Eds., *The Exponential Distribution: Theory, Methods and Applications*, The Gordon and Breach, Newark, New Jersey, 1995.

[CHA 04] CHANDRASEKAR B., CHILDS A., BALAKRISHNAN N., "Exact likelihood inference for the exponential distribution under generalized Type-I and Type-II hybrid censoring", *Naval Research Logistics*, vol. 51, p. 994–1004, 2004.

[CHE 99] CHEN J., LAM Y., "Bayesian variable sampling plan for the Weibull distribution with Type I censoring", *Acta Mathematicae Applicatae Sinica*, vol. 15, p. 269–280, 1999.

[CHE 04] CHEN J., CHOY S. T. B., LI K. H., "Optimal Bayesian sampling acceptance plan with random censoring", *Euro. Jour. of Operational Research*, vol. 155, p. 683–694, 2004.

[CHI 03] CHILDS A., CHANDRASEKAR B., BALAKRISHNAN N., KUNDU D., "Exact likelihood inference based on Type-I and Type-II hybrid censored samples from the exponential distribution", *Annals of the Institute of Statistical Mathematics*, vol. 55, p. 319–330, 2003.

[HUA 02] HUANG W., LIN Y. P., "An improved Bayesian sampling plan for exponential population with Type-I censoring", *Communications in Statistics – Theory and Methods*, vol. 31, p. 2003–2025, 2002.

[HUA 04] HUANG W. T., LIN Y. P., "Bayesian sampling plans for exponential distribution based on uniform random censored data", *Computational Statistics & Data Analysis*, vol. 44, p. 669–691, 2004.

[LAM 90] LAM Y., "An optimal single variable sampling plan with censoring", *The Statistician*, vol. 39, p. 53–67, 1990.

[LAM 94] LAM Y., “Bayesian variable sampling plans for the exponential distribution with Type I censoring”, *The Annals of Statistics*, vol. 22, p. 696–711, 1994.

[LAM 95] LAM Y., CHOY S. T. B., “Bayesian variable sampling plans for the exponential distribution with uniformly distributed random censoring”, *Journal of Statistical Planning and Inference*, vol. 47, p. 277–293, 1995.

Chapter 11

Reliability of Stochastic Dynamical Systems Applied to Fatigue Crack Growth Modeling

11.1. Introduction

The aim of this chapter is to provide some mathematical tools for the modeling of degradation processes that arise in structures operating under unstable environmental conditions. These uncertainties may come from a lack of scientific knowledge or from the intrinsic random nature of the observed physical phenomenon. The theory of stochastic processes seems to be a good way to handle the randomness of the system, thus it is widely used in many engineering fields that involve all kinds of uncertainties. This is the case for applications in the structural reliability field.

In stochastic analysis dedicated to structural mechanic, the reliability of a structure is often modeled using a classical stress-strength description (see [JOH 88, KOT 03]). That is, a stochastic process describing the "stress" applied on the structure is compared with the "strength", which represents the critical level that must not be overcome by the loads. A slightly different approach is adopted in this chapter: we describe the evolution in time of the level of degradation in a given structure using a stochastic process governed by a differential system. It is compared with a variable describing the failure domain that must not be reached, which will be a real positive constant in the following sections, for the sake of simplicity. It means that there is no renewal possible for the degradation process, that is, the system is not a reparable one. Degradation processes which may be renewed are studied in, for example, [BAG 06]. Associated

Chapter written by Julien CHIQUET and Nikolaos LIMNIOS.

statistical inference may be found in [NIK 07], where general description degradation process modeling is provided. Our approach is also close to the one adopted in [LEH 06], where threshold is considered in the degradation mechanism. Our choice is prompted by the real degradation process considered in this chapter for the validation, namely, the fatigue crack growth problem: the crack size increases randomly in time until it reaches a critical threshold upon which the structure collapses. Thus, with Z_t being the degradation process and Δ the critical level of degradation, the reliability of such a system is

$$R(t) = \mathbb{P}(Z_t < \Delta), \tag{11.1}$$

since Δ is an absorbing point. Obviously, the main issue in formulation (11.1) of the reliability consists of describing the evolution of the stochastic process Z_t. It mainly implies two problems: first, the modeling of the growth rate of Z_t using a differential system; second, the estimation of the parameters of this model, when considering real applications. A general form for a homogenous stochastic dynamical system is

$$\frac{\mathrm{d}Z_t}{\mathrm{d}t} = C(Z_t, X_t), \qquad Z_0 = z, \tag{11.2}$$

where z is an initial condition that may be random, and C a function which is generally partially known, due to some physical assumptions made from observations. The function C is positive so the paths of Z_t are monotone, since they describe the growth of the degradation process, that is, the growth of the crack in the case at hand. The process X_t is introduced to take into account the random environmental effects, and the choice of its nature is crucial for the nature of the dynamical system (11.2) itself.

The larger part of stochastic dynamical systems research is dedicated to stochastic differential equations, where the random component is modeled by the Brownian motion B_t. In this case, the dynamic of the process Z_t is governed by

$$Z_t = Z_0 + \int_0^t m(Z_s)\mathrm{d}s + \int_0^t \sigma(Z_s)\mathrm{d}B_s, \qquad Z_0 = z, \tag{11.3}$$

where, roughly speaking, m is the *drift* function, describing the mean behavior of the process, while the *diffusion* function σ describes the way the process diffuses around the drift. In that case, Z_t is known as a diffusion process.

Although stochastic differential equations remain a great tool for any kind of application fields, interesting alternatives appeared in other studies. Some authors suggested the idea of a diffusion process whose evolution should change at some discrete instants of time [GOL 51], which leads to the generalized diffusion. Another alternative are the *piecewise deterministic Markov processes* [DAV 84]. These processes are composed by the mixture of deterministic and jump motions. This is the case when X_t is a jump Markov process in the dynamical system (11.2).

While Davis [DAV 93] gave applications linked to control theory, this class of processes has also been used for the modeling of particle movements [DAU 89, LAP 98].

Dynamical systems are also widely used in engineering tools like dynamic reliability [DEV 96]. We suggest in this chapter to apply these kind of models for structural reliability analysis because it fits well with the physical description of a degradation process: indeed, a level of degradation evolves continuously in time in a real state phase, whereas its evolution changes through some shocks with random intensities, occurring at some random instants of time, modeling the jump Markov process. We will see that a formal expression of both the reliability of the system and distribution of the failure time can be obtained by means of the Markov renewal theory. In order to apply these models on real data sets, some estimation results are also required, which will be developed in the following sections.

The outline of this chapter is thus as follows: in the next section, the class of stochastic dynamical system involving jump Markov processes is presented, with the general assumptions required for the modeling of a degradation process. We give some results associated with the Markov renewal theory which are useful to calculate the reliability function. Asymptotic results based upon the Bogolyubov's averaging principle are given for the dynamical system. In the third section, a few issues are investigated concerning the estimation of the parameters of the system, particularly for the underlying jump Markov process, which is not directly observable. Finally, a numerical application is given on a real data set for fatigue crack growth modeling, which is a degradation process involved in many engineering fields.

Remark 11.1 This chapter is built from [CHI arb, CHI on, CHI 06].

11.2. Stochastic dynamical systems with jump Markov process

Let us consider a degradation process through a stochastic process $(Z_t, t \in \mathbb{R}_+)$ that increases randomly on $[z, \Delta]$, with $0 < z < \Delta$. The process Z_t starts almost surely from the point z, that is $\mathbb{P}(Z_0 = z) = 1$ and stops when reaching the absorbing point Δ. The evolution of Z_t is governed by the stochastic dynamical system

$$\dot{Z}_t = C(Z_t, X_t), \qquad Z_0 = z, \tag{11.4}$$

where $\dot{Z}_t$ stands for the derivative $\mathrm{d}Z_t/\mathrm{d}t$. Furthermore, the following additional assumptions are made.

A. 11.1 *The process $(X_t, t \in \mathbb{R}_+)$ is an irreducible Markov process with a finite state space E, initial distribution $\alpha_i = \mathbb{P}(X_0 = i)$ for all $i \in E$, stationary distribution $\pi = (\pi_i)_{i \in E}$ and a matrix generator $\mathbf{A} = (a_{ij})_{i,j \in E}$ such that*

$$\begin{array}{rlll} & a_{ij} & \geq 0, & \textit{for all } i \neq j, \\ \textit{and} & a_{ii} & = -a_i & = -\sum_{k \in E, i \neq k} a_{ik}. \end{array}$$

A. 11.2 *The function $C : \mathbb{R}_+ \times E \longrightarrow \mathbb{R}_+$ is measurable, strictly positive and Lipschitz according to the first argument, that is, there is a function $f : E \longrightarrow \mathbb{R}$ for $x, y \in \mathbb{R}_+$ and $i \in E$, such that*

$$|C(x,i) - C(y,i)| \leq f(i)\,|x-y|\,.$$

A. 11.3 *There exists a function G that makes it possible to reverse the stochastic dynamical system, thus giving an expression of X_t as a function of Z_t and its derivative, that is*

$$X_t = G(Z_t, \dot{Z}_t).$$

For instance, this is the case when X_t appears in an additive or multiplicative form in the right hand side of (11.4).

This model is known as a piecewise deterministic Markov process, for which Davis [DAV 84, DAV 93] gave the underlying theory. This designation comes from two important remarks: first it can be easily seen that the coupled process (Z_t, X_t) is Markov; second, the process Z_t evolves deterministically between the jumps of the process X_t, occurring at various instants of time $(S_n, n \in \mathbb{N})$ with random intensities. For instance, for all $t < S_1$ with S_1 being the realization of the first jump time of a path X_t and $X_0 = i$, then the Cauchy problem

$$\dot{Z}_t = C(Z_t, i), \quad Z_0 = z, \tag{11.5}$$

has a unique deterministic solution denoted by $\varphi_{z,i}(t)$, that is $Z_t = \varphi_{z,i}(t)$, for all $t \in [0, S_1)$. The whole solution Z_t for $t \in \mathbb{R}_+$ is built piecewisely on the successive jump time intervals $[S_n, S_{n+1})$ of X_t.

Rather than studying the infinitesimal generator of the couple (Z_t, X_t), we will focus on the transition probability function $P(t)$, which fully characterizes the process:

$$P_{ij}(z,B,t) = \mathbb{P}(Z_t \in B, X_t = j | Z_0 = z, X_0 = i), \qquad \forall i,j \in E, B \in \mathcal{B}, \tag{11.6}$$

where B is a subset of $\mathcal{B}$, the Borel σ-field of $\mathbb{R}_+$. We put $U = [z, \Delta)$ and $D = [\Delta, \infty)$ being respectively the set of working states and the set of failure states for Z_t. We have Δ an absorbing point for Z_t, i.e., the system is not reparable. Hence, reliability (11.1) of the system as a function of (Z_t, X_t) is:

$$R(t) = \mathbb{P}((Z_t, X_t) \in U \times E) = \sum_{i,j \in E} \alpha_i P_{ij}(z, U, t). \tag{11.7}$$

The hitting time τ to D for the process Z_t is also the failure time, which can be defined as a function of the couple by

$$\tau = \inf\{t \geq 0 : Z_t \in D\} = \inf\{t \geq 0 : (Z_t, X_t) \in D \times E\}. \tag{11.8}$$

The distribution function F_τ of τ is linked to the reliability through $F_\tau(t) = 1 - R(t)$.

Let us now study the coupled process in the framework of Markov renewal processes (see [ÇIN 69, LIM 01]). With this mean, we may build a Markov renewal equation for the transition function $P(t)$ and give a formal expression for the cumulative distribution function F_τ and for the reliability $R(t)$. For this purpose, we associate with $(Z_t, X_t, t \in \mathbb{R}_+)$ the extended renewal process $(Z_n, J_n, S_n, n \in \mathbb{N})$ such that

$$Z_n = Z_{S_n}, \qquad J_n = X_{S_n}, \tag{11.9}$$

with $(S_n, n \in \mathbb{N})$ being the jump times of X_t. The semi-Markov kernel Q is then defined by

$$Q_{ij}(z, B, t) = \mathbb{P}(Z_{n+1} \in B, J_{n+1} = j, S_{n+1} - S_n \leq t | Z_n = z, J_n = i). \tag{11.10}$$

The following proposition follows from the results of Markov renewal theory and after some calculations (see [CHI ara]).

Proposition 11.1 *The transition function $P(t)$ is given by the following Markov renewal equation:*

$$P_{ij}(z, B, t) = g_{ij}(z, B, t) + \sum_{k \in E} \int_{\mathbb{R}_+} \int_0^t Q_{ik}(z, \mathrm{d}y, \mathrm{d}s) P_{kj}(y, B, t - s), \tag{11.11}$$

where

$$Q_{ij}(z, B, \mathrm{d}t) = a_{ij} e^{-a_i t} \delta_{\phi_t(z,i)}(B) \mathrm{d}t, \tag{11.12}$$

and

$$g_{ij}(z, B, t) = \mathbb{1}_{\{i=j\}} \mathbb{1}_B(\varphi_t(z, i)) e^{-a_i t}. \tag{11.13}$$

We have used the indicator function and the Dirac distribution, that is $\mathbb{1}_B(x) = 1$ if $x \in B$, 0 otherwise, and $\delta_x(B) = 1$ if $x \in B$, 0 otherwise.

The Markov renewal equation (11.11) can be solved numerically, thus with (11.7), the reliability $R(t)$ of the system as well as the cumulative distribution function of the failure time F_τ can also be calculated.

We also want to give an interesting asymptotic result for dynamical systems, which quickly give a first approximation of the mean behavior of Z_t. Moreover, it is well adapted for any further estimation of the system, which will be investigated in the next section. Let us thus study system (11.4) in a series scheme [KOR 05], that is, the weak convergence, when $\varepsilon \to 0$, of

$$\dot{Z}_t^\varepsilon = C(Z_t^\varepsilon, X_{t/\varepsilon}), \qquad Z_0^\varepsilon = z_0. \tag{11.14}$$

The change of scale $t \to t/\varepsilon$ is performed for X_t in order to consider the behavior of the dynamical system when the random component X_t merely adds the information it would add after a very long time of observation in (11.14). This *averaging approximation* was first introduced by Bogolyubov [BOG 61] who showed that the process $\dot{Z}_t^\varepsilon$ defined by (11.14) converges weakly, when $\varepsilon \to 0$, to a deterministic process $\widetilde{z}_t$ governed by the following system

$$\dot{\widetilde{z}}_t = \overline{C}(\widetilde{z}_t), \qquad \widetilde{z}_0 = z, \tag{11.15}$$

with $\overline{C}$ being the mean function. Due to the fact that X_t is a jump Markov process, the averaging on time reduces to its stationary distribution π, and we have

$$\overline{C}(z) = \sum_{i \in E} C(z,i)\pi_i. \tag{11.16}$$

11.3. Estimation

This section is dedicated to some estimation issues concerning the dynamical system (11.4). For this purpose, we shall consider that we have K independent sample paths of Z_t from experimental feedback, which are denoted by $(Z_t^k)_{k=1,\ldots,K}$. These paths are recorded in time from $t = 0$, where $Z_0 = z$, until the process Z_t reaches the critical value Δ. The paths are thus defined on the random time interval $[0, \tau_k]$, where $(\tau_k)_{k=1,\ldots,K}$ are independent copies of the random variable τ, i.e., of the failure time.

The simplest part of the estimation consists of estimating the parameters of the function C governing the dynamical system by using the averaging principle developed above. Using a simple regression analysis (e.g., the least squares method) on the asymptotic deterministic system (11.15), we can obtain estimators of the parameters appearing in the function $\overline{C}$, whose general form is known. By this mean, we obtain the mean behavior of the dynamical system and the only unknowns that remain are the initial law α, the generator $\mathbf{A}$ and the state space E of the Markov process X_t. The stationary law π is obtained from $\mathbf{A}$ using

$$\pi\mathbf{A} = 0, \qquad \sum_{i \in E} \pi_i = 1.$$

Concerning X_t, we first need some representations of the paths $(X_t^k)_{k=1,\ldots,K}$ which cannot be directly observed during laboratory experiments. For this purpose, we refer to A.11.3, that is, we assume there exists a function G that "reverses" the dynamical system, giving explicitly the process X_t has a function of Z_t and $\dot{Z}_t$. Hence, we obtain some K estimated paths $\widetilde{X}_t^k$ by

$$\widetilde{X}_t^k = G\left(Z_t^k, \widehat{\dot{Z}}_t^k\right), \tag{11.17}$$

where the derivatives are estimated with classical finite differences

$$\widehat{\dot{Z}}_t^k = \frac{Z_{t+\Delta}^k - Z_t^k}{\Delta t}, \tag{11.18}$$

with Δt being the time discretization step of the data set. $\widetilde{X}_t^k$ may be very noisy because of the approximation of $\dot{Z}_t$. Of course, we cannot use all of the observed values of the $\widetilde{X}_t^k$ as the final state space, since it would be quite large and would not be appropriate for any further numerical estimation of the generator $\mathbf{A}$. Thus, a state space reduction is required in order to find a state space that fits well with the X_t^k with a reasonable number of states.

From the estimators $\widetilde{X}_t^k$, we have a certain number of points, corresponding to the states of the process observed at certain instants of time, that is the sampled data $\mathcal{D} = \{\widetilde{X}_{t_i}^k : k = 1, \ldots, K; i = 1, \ldots, \tau_k\}$. We suggest applying the classical K-means clustering algorithm [MAC 67] on $\mathcal{D}$ to reduce the number of clusters to a suitable amount that can be used as the state space of X_t. The *centroids*, which are the Euclidean centers of the clusters, represent in the case at hand the states of the jump Markov process X_t. They are denoted by μ_s, with $s \in \mathcal{S}$ the set of clusters. The k-means algorithm minimizes iteratively the following criteria by finding an appropriate set of centroids $(\mu_s)_{s\in\mathcal{S}}$:

$$C_{\mathcal{S}} = \sum_{x\in\mathcal{D}} \min_{s\in\mathcal{S}} ||x - \mu_s||^2. \tag{11.19}$$

Each point $x \in \mathcal{D}$ is affected by the centroid μ_s that minimizes the above criteria. It is performed using an iterative algorithm which stops when no further changes are observed in the data points assignment and in the evolution of the μ_s. The initial centroids are chosen randomly among the points $\mathcal{D}$. We also have to initially choose $\mathrm{card}(\mathcal{S})$, that is the number of centroids used for the clustering. The final number of clusters is chosen by representing the final error committed at the end of the algorithm for various numbers of clusters, and by making a compromise between the error and $\mathrm{card}(\mathcal{S})$. The estimated state space $\widehat{E}$ of X_t is composed by the final centroids, that is $\{\mu_s, s \in \mathcal{S}\}$, and the paths $\widetilde{X}_t^k$ previously obtained are then filtered on $\widehat{E}$, thus obtaining new sample paths estimations denoted by $\widehat{X}_t^k$, with values on $\widehat{E}$.

With the $\widehat{X}_t^k$s and $\widehat{E}$, we may estimate the fundamentals of the jump Markov process X_t, that is the infinitesimal generator $\mathbf{A}$ and its initial law α. The Markov process X_t will thus be completely characterized. For this purpose, let us introduce the following notation associated to the $(\widehat{X}_t^k)_{k=1,\ldots,K}$:

– $N^k(\tau_k)$, the number of jumps observed on the random time interval $[0, \tau_k)$;

– N_{ij}^k, the number of transitions from i to j observed on the kth censored path, i.e.,

$$N_{ij}^k = \sum_{n=1}^{N^k(\tau_k)} \mathbb{1}_{\left\{J_{n-1}^k = i, J_n^k = j\right\}};$$

– V_i^k, the length of time spent in state i on the k^{th} path, i.e.

$$V_i^k = \sum_{n=1}^{N^k(\tau_k)-1} (S_n^k - S_{n-1}^k)\mathbb{1}_{\{X_n^k=i\}} + \left(\tau_k - S_{N^k(\tau_k)}^k\right)\mathbb{1}_{\left\{J_{N^k(\tau_k)}^k=i\right\}}.$$

$N_{ij}(K)$ and $V_i(K)$ are the variables associated with all of the K sample paths:

$$N_{ij}(K) = \sum_{k=1}^{K} N_{ij}^k, \qquad V_i(K) = \sum_{k=1}^{K} V_i^k.$$

Concerning the infinitesimal generator of X_t, we use

$$\widehat{\mathbf{A}} = (\widehat{a}_{ij})_{i,j\in E} = \frac{N_{ij}(K)}{V_i(K)}, \qquad i \neq j, \tag{11.20}$$

and $\widehat{a}_{ii} = -\widehat{a}_i = -\sum_{k\in E, k\neq i} \widehat{a}_{ik}$.

This estimator is the one obtained when maximizing the likelihood for K paths of a jump Markov process with right random positive censoring, with the censoring variable being independent of the process. This is a straightforward generalization of the classical results obtained by Billingsley [BIL 61] and Albert [ALB 62]. In the present context, it is a first approximation for the estimation of the generator, as long as the paths of the randomizing process X_t of the dynamical system (11.4) are censored by the failure time, which is by definition (11.8) dependent of X_t. Yet, it still gives some good numerical results (see [CHI on] for further details on that issue).

For the initial law, we use the empirical estimator obtained by maximizing the likelihood when considering the initial state X_0 independent of the generator. We find

$$\widehat{\alpha}_i = \frac{1}{K}\sum_{k=1}^{K} \mathbb{1}_{\{X_0^k=i\}}. \tag{11.21}$$

11.4. Numerical application

The aim of the present section is to validate the modeling and the estimation method developed above for a real degradation process, namely, the fatigue crack growth problem. It is a typical degradation process where the use of stochastic modeling is helpful to handle with the variability of the phenomenon. Over the past 25 years, this problem has been studied due to its important implication in the aeronautic and nuclear fields. Some stochastic dynamical systems based upon a diffusion Markov process were suggested, i.e., the crack size is modeled by a process Z_t governed by a stochastic differential equation such as (11.3), interpreted in a Stratonovich sense (see [LIN 85, SOB 93]). In [TAN 99] important sampling methods are given for these models. Some similar interesting methods are developed in [IVA 04], but for applications in control. Alternative models for stochastic crack growth are available, for instance in [KIR 99], yet, to the author's knowledge, none were given based upon piecewise deterministic Markov processes. Thus, we suggest here modeling the crack growth by a stochastic dynamical system where the random part is a jump Markov process rather than the Brownian motion. To validate this approach, we employ herein the extensive set of crack growth data known as the Virkler data [VIR 79]. As seen in Figure 11.1, the 68 specimens from the Virkler data have initially the same size of $z = 9\ \mathrm{mm}$. They are exposed to the same cyclic loading until the crack size reaches the critical value of $\Delta = 49.8\ \mathrm{mm}$. Even when the laboratory experiments are well controlled, we can obviously see the intrinsically stochastic nature of this physical phenomenon.

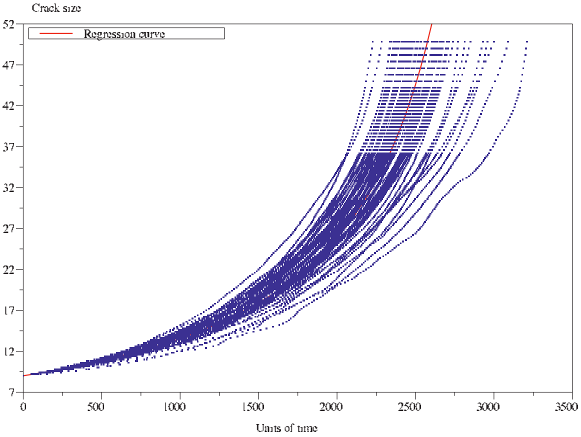


Figure 11.1. *Virkler data set with Paris-Erdogan law regression*

In fracture mechanics, deterministic models have been formulated to describe the crack growth rate as a function of the stress intensity factor range ΔK and various material parameters. The Paris-Erdogan law [PAR 63] is the most popular and versatile because of its simple formulation and the fact it suits any kind of material. It is formulated as follows:

$$\frac{\mathrm{d}z_t}{\mathrm{d}t} = p(\Delta K)^n, \tag{11.22}$$

where $\Delta K = Y(z_t)\Delta S\sqrt{\pi z_t}$, with ΔS being the applied stress range, Y a function of crack and structure geometry, and p, n some material constants. The use of lower case for process z_t means here a deterministic approach for the modeling. For the lab specimens used by Virkler, we may assume $Y \equiv 1$, hence

$$\frac{\mathrm{d}z_t}{\mathrm{d}t} = a(z_t)^b, \tag{11.23}$$

and $a = p(\Delta S)^n \pi^{n/2}, b = n/2$ are the only two remaining parameters. The idea suggested here to give a stochastic formulation of the Paris law is to add a multiplicative jump Markov process in the right hand side of (11.23), thus taking into account the variation of the crack growth rate due to many environmental factors:

$$\frac{\mathrm{d}Z_t}{\mathrm{d}t} = a(Z_t)^b \times X_t, \qquad Z_0 = z. \tag{11.24}$$

We first want to estimate the mean behavior of a crack growth path, thus we apply the Bogolyubov's averaging principle described in section 3.1. Hence, the system

$$\frac{\mathrm{d}Z_t^\varepsilon}{\mathrm{d}t} = a(Z_t^\varepsilon)^b \times X_{t/\varepsilon}, \qquad Z_0^\varepsilon = z, \tag{11.25}$$

has asymptotically the same behavior, when $\varepsilon \to 0$, as

$$\frac{\mathrm{d}\widetilde{z}_t}{\mathrm{d}t} = a_0(\widetilde{z}_t)^b, \qquad \widetilde{z}_0 = z, \tag{11.26}$$

where $a_0 = a \sum_{i \in E} \pi_i i$. The following stochastic dynamical system is equivalent to (11.24):

$$\frac{\mathrm{d}Z_t}{\mathrm{d}t} = a_0(Z_t)^b \times v(X_t), \qquad Z_0 = z, \tag{11.27}$$

where $v(X_t)$ is a jump Markov process with the same generator as X_t, but a state space E' that is linked to the state space E of X_t through the linear application v defined by $v(x) = x / \sum_{i \in E} \pi_i \times i$, for $x \in E$. By this means, the new randomizing process $v(X_t)$ is "normalized" and $\mathbb{E}v(X_t) = 1$. It is quite helpful in order to make a regression analysis to obtain the parameters a_0 and b. For instance, the simple least squares method is performed on the data set by taking the logarithm on both sides of (11.26):

$$\ln \dot{Z}_t = \ln a_0 + b \ln Z_t + \ln v(X_t). \tag{11.28}$$

The available data, that is, the Z_t^ks, can be represented as a N-sample composed by all of the points of each of the K paths, that is $\left\{(\ln Z_{t_i}, \ln \widehat{\dot{Z}}_{t_i})\right\}$, with $\widehat{\dot{Z}}_{t_i}$ being obtained using (11.18). Introducing the notation $x_i = \ln Z_{t_i}, y_i = \ln \widehat{\dot{Z}}_{t_i}$ and $\varepsilon_i = \ln v(X_{t_i})$, we have the following classical regression problem:

$$y_i = \ln a_0 + b x_i + \varepsilon_i, \quad i = 1, \ldots, N. \tag{11.29}$$

The $\ln v(X_{t_i})$s are "close to 0" in the regression equation (11.29) and play the role of the residuals ε_i. Minimizing $\sum \varepsilon_i^2$, we get the estimators associated with the well-known least-squares method:

$$\widehat{a}_0 = \exp(\overline{y} - b\overline{x}), \qquad \widehat{b} = \frac{\sum x_i y_i - N\overline{xy}}{\sum x_i^2 - N\overline{x}^2}, \qquad \text{with } \overline{x} = \frac{1}{N}\sum x_i, \qquad \overline{y} = \frac{1}{N}\sum y_i. \tag{11.30}$$

The paths of $v(X_t)$ are then estimated by successively reversing the dynamical system and applying the k-means algorithm to reduce the state space, as described in section 3. Here, the function G that gives a path of $v(X_t)$ is merely

$$(v(\widetilde{X}_t))^k = \frac{1}{\widehat{a}_0}(Z_t^k)^{-\widehat{b}} \times \widehat{\dot{Z}}_t^k. \tag{11.31}$$

We applied the K-means clustering method on the $(v(\widetilde{X}_t))^k$ for various numbers of clusters. The final error corresponding to different numbers of clusters is represented on Figure 11.2, where 13 states seems to be a good compromise.

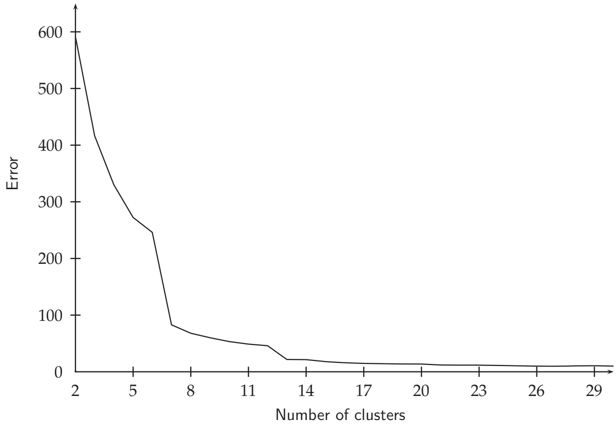


Figure 11.2. *Error committed with the k-means algorithm for various numbers of clusters*

The infinitesimal generator $\mathbf{A}$ and the initial law α are estimated with (11.20) and (11.21). We thus have a fully estimated stochastic dynamical system with which some crack growth curves can be simulated (see Figure 11.3). Coupled with some Monte-Carlo techniques, important functions can then be calculated, such as the reliability function $R(t)$ or the cumulative distribution function of Z_t, which represents the probability distribution of the crack size in time. The result is interesting for optimizing a maintenance strategy. The Monte-Carlo estimators can be compared with the empirical estimator calculated on the Virkler data set from Figure 11.1. Results are represented in Figure 11.4 for the reliability and in Figure11.5 for the distribution of Z_t for various values of t.

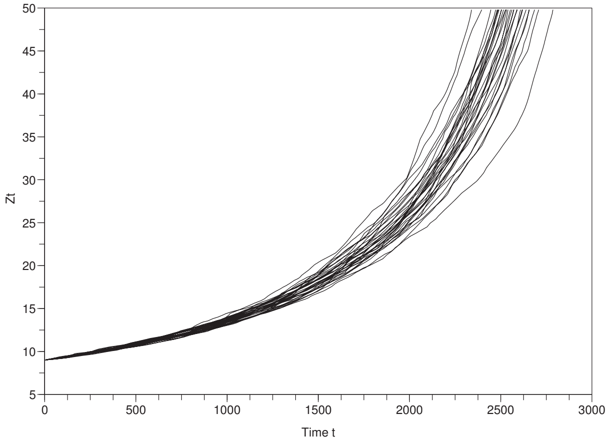


Figure 11.3. *30 crack paths simulated*

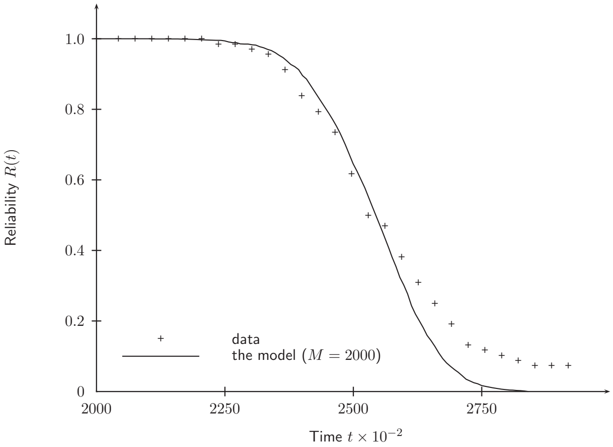


Figure 11.4. *Reliability estimation*

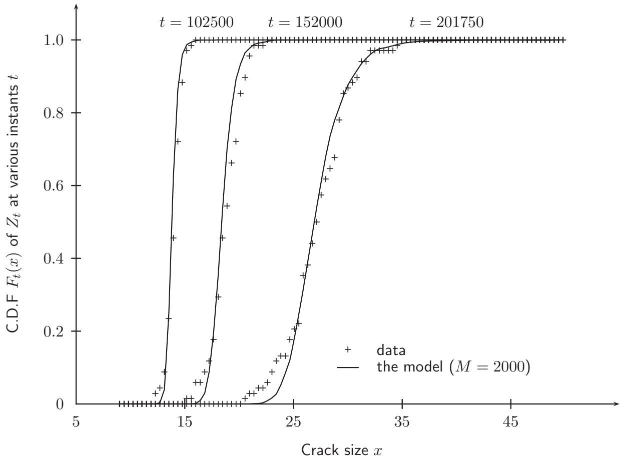


Figure 11.5. *Cumulative distribution estimation of Z_t*

11.5. Conclusion

We have seen that dynamical systems involving jump Markov processes, an alternative to diffusion processes, were efficient for applications, particularly when modeling a degradation process increasing randomly in time. Further investigations could be interesting, such as modeling the random part of the system using a semi-Markov process with a countable state space, thus making the models more precise. Moreover, taking into account the dependency between the censoring time and the jump Markov process could give an estimation of the generator that would be more accurate. We are also working on the numerical implementation associated with the calculation of the true distribution function of the failure time F_τ obtained using Proposition 11.1, developed in [CHI ara].

11.6. Bibliography

[ALB 62] ALBERT A., “Estimating the infinitesimal generator of a continuous time, finite state Markov process”, *Annals of Mathematical Statistics*, vol. 38, p. 727-753, 1962.

[BAG 06] BAGDONAVIČIUS V., BIKELIS A., KAZAKEVIČIUS V., NIKULIN M., “Non-parametric estimation in degradation-renewal-failure models”, in: *Probability, Statistics and Modelling in Public Health*, Springer, 2006.

[BIL 61] BILLINGSLEY P., *Statistical Inference for Markov Processes*, The University of Chicago, Chicago Press, 1961.

[BOG 61] BOGOLYUBOV N., MITROPOL’SKII Y., *Asymptotics Methods in the Theory of Non-linear Oscillations*, Gordon and Breach Science Publishers, 1961.

[CHI 06] Chiquet J., Limnios N., "Estimating stochastic dynamical systems driven by a continuous-time jump Markov process", *Methodology and Computing in Applied Probability*, vol. 8, 2006.

[CHI ara] Chiquet J., Limnios N., "A method to compute the reliability of a piecewise deterministic Makov process", *Statistics and Probability Letters*, to appear.

[CHI arb] Chiquet J., Limnios N., Eid M., "Modelling and estimating the reliability of stochastic dynamical systems with Markovian switching", *Reliability Engineering and System Safety:* Selected papers from ESREL 2006 - Safety and Reliability for managing Risks, to appear.

[CHI on] Chiquet J., Limnios N., Eid M., "Piecewise deterministic Markov processes applied to fatigue crack growth modelling", *Journal of Statistical Planning and Inference*, in revision.

[ÇIN 69] Çinlar E., "Markov renewal theory", *Advanced Applied Probability*, vol. 1, p. 123-187, 1969.

[DAU 89] Dautray R., Ed., *Méthodes probabilistes pour les équations de la physique*, Synthèse, Eyrolles, 1989.

[DAV 84] Davis M., "Piecewise-deterministic Markov processes: a general class of non-diffusion stochastic models", *Journal of the Royal Statistical Society*, vol. 46, num. 3, p. 353-388, 1984.

[DAV 93] Davis M., *Markov Models and Optimization*, Monographs on Statistics and Applied Probability 49, Chapman & Hall, 1993.

[DEV 96] Devooght J., Smidts C., "Probabilistic dynamics as a tool for dynamic PSA", *Reliability Engineering and System Safety*, vol. 52, p. 185-196, 1996.

[GOL 51] Goldstein S., "On diffusion by discontinuous movements, and on the telegraph equation", *Quarterly Journal of Mechanics and Applied Mathematics*, vol. 4, p. 129-156, 1951.

[IVA 04] Ivanova A., Naess A., "Importance sampling for dynamical systems by approximate calculation of the optimal control function", *Conference on Mathematical Methods in Reliability*, 2004.

[JOH 88] Johnson R., "Stress-strength models for reliability", in: *Handbook of Statistics*, vol. 7, p. 27-54, Elsevier Science Publishers, 1988.

[KIR 99] Kirkner D., Sobczyk K., Spencer B., "On the relationship of the cumulative jump model for random fatigue to empirical data", *Probabilistic Engineering Mechanics*, vol. 14, p. 257-267, 1999.

[KOR 05] Korolyuk V., Limnios N., *Stochastic Systems in Merging Phase Space*, World Scientific, 2005.

[KOT 03] Kotz S., Lumelskii Y., Pensky M., *The Stress-Strength Model and its Generalizations*, World Scientific, 2003.

[LAP 98] Lapeyre B., Pardoux E., *Méthodes de Monte-Carlo pour les équations de transport et de diffusion*, Springer, 1998.

[LEH 06] LEHMANN A., "Degradation-Threshold-Shok Models", in: *Probability, Statistics and Modelling in Public Health*, Springer, 2006.

[LIM 01] LIMNIOS N., OPRIŞAN G., *Semi-Markov Processes and Reliability*, Birkhäuser, 2001.

[LIN 85] LIN Y. K., YANG J. N., "A stochastic theory of fatigue crack propagation", *AIAA Journal*, vol. 23, p. 117-124, 1985.

[MAC 67] MACQUEEN J. B., "Some methods for classification and analysis of multivariate observations", *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, Berkeley, University of California Press, p. 281-297, 1967.

[NIK 07] NIKULIN M., GERVILLE-RÉACHE L., COUAILLIER V., *Statistique des Essais Accélérés*, Hermès, 2007.

[PAR 63] PARIS P., ERDOGAN F., "A critical analysis of crack propagation laws", *Journal of Basic Engineering*, vol. 85, p. 528-534, 1963.

[SOB 93] SOBCZYK K., Ed., *Stochastic Approach to Fatigue*, Springer-Verlag, 1993.

[TAN 99] TANAKA H., "Importance sampling simulation for a stochastic fatigue crack growth model", MELCHERS R., STEWART M., Eds., *Proceedings of ICASP8*, vol. 2, p. 907-914, 1999.

[VIR 79] VIRKLER D., HILLBERRY B., GOEL P., "The statistical nature of fatigue crack propagation", *Journal of Engineering Material Technology ASME*, vol. 101, p. 148-153, 1979.

This page intentionally left blank

Chapter 12

Statistical Analysis of a Redundant System with One Stand-by Unit

12.1. Introduction

Consider a redundant system with one operating and one stand-by unit. If the main unit fails then the stand-by unit (if it has not failed yet) is commuted and operates instead of the main one. We suppose that commuting is momentary and there are no repairs.

If the stand-by unit does not function until the failure of the main unit ("cold" reserving), it is possible that during and after commuting the failure rate increases because the stand-by unit is not "warmed" enough [VEK 87]. If the stand-by unit is functioning in the same "hot" conditions as the main unit, then usually after commuting the reliability of the stand-by unit does not change. However, "hot" redundancy has disadvantages because the stand-by unit fails earlier than the main one with the probability 0.5. So "warm" reserving is sometimes used [WAS 05]: the stand by unit functions under lower stress than the main one. In such a case, the probability of the failure of the stand-by unit is smaller than that of the main unit and it is also possible that commuting is fluent. So the main problem is to verify the hypothesis that the switch from "warm" to "hot" conditions does not do some damage to units.

Let us formulate the hypothesis strictly.

Chapter written by Vilijandas BAGDONAVIČIUS, Inga MASIULAITYTE and Mikhail NIKULIN.

12.2. The models

Suppose that the failure time T_1 of the main element has the cdf F_1 and the probability density f_1, while the failure time T_2 of the stand-by element has the cdf F_2 and the probability density f_2.

The failure time of the system is $T = max(T_1, T_2)$.

Denote by $f_2^{(y)}(x)$ the conditional density of the stand-by unit given that the main unit fails at the moment y. It is clear that $f_2^{(y)}(x) = f_2(x)$ if $0 \leq x \leq y$.

The cdf of the system failure time T is

$$F(t) = P(T_1 \leq t, T_2 \leq t) = \int_0^t \left\{ \int_0^y f_2(x)dx + \int_y^t f_2^{(y)}(x)dx \right\} f_1(y)dy. \tag{12.1}$$

When stand-by is "cold" then $f_2(x) = 0$, $f_2^{(y)}(x) = f_1(x - y)$, $x > y$, so

$$F(t) = \int_0^t \left\{ \int_y^t f_1(x-y)dx \right\} f_1(y)dy = \int_0^t F_1(t-y)dF_1(y).$$

In the case of "hot" stand-by, $f_2^{(y)}(x) = f_2(x) = f_1(x)$, so $F(t) = [F_1(t)]^2$.

In the case of "warm" reserving, the following hypothesis is assumed:

$$H_0 : f_2^{(y)}(x) = f_1(x + g(y) - y), \quad \text{for all} \quad x \geq y \geq 0, \tag{12.2}$$

where $g(y)$ is the moment which in "hot" conditions corresponds to the moment y in "warm" conditions in the sense that

$$F_1(g(y)) = P(T_1 \leq g(y)) = P(T_2 \leq y) = F_2(y),$$

so

$$g(y) = F_1^{-1}(F_2(y)).$$

Conditionally (given $T_1 = y$) the hypothesis corresponds to the well-known Sedyakin's model [SED 66]. In [BAG 97] a goodness-of-fit test of logrank-type for Sedyakin's model using experiments with fixed switch off moments is proposed (see also [BAG 02]. In the situation considered here, the switch off moments are random.

Formula (12.1) implies that under the hypothesis H_0,

$$F(t) = \int_0^t F_1(t + g(y) - y)dF_1(y). \tag{12.3}$$

In particular, if we suppose that the distribution of the units functioning in "warm" and "hot" conditions differ only in scale, i.e.

$$F_2(t) = F_1(rt), \tag{12.4}$$

for some $r > 0$, then $g(y) = ry$. In such a case, the hypothesis

$$H_0^* : \exists\, r > 0 : \quad f_2^{(y)}(x) = f_1(x + ry - y), \quad \text{for all} \quad x \geq y \geq 0, \tag{12.5}$$

is to be verified. Conditionally (given $T_1 = y$), the hypothesis corresponds to the accelerated failure time (AFT) model [BAG 78], [NEL 90]. In [BAG 90] a goodness-of-fit test for the AFT model using experiments with fixed switch off moments is proposed (see also [BAG 02]).

12.3. The tests

Suppose that the following data are available:

a) the failure times $T_{11}, \ldots, T_{1n_1}$ of n_1 units tested in "hot" conditions;

b) the failure times $T_{21}, \ldots, T_{2n_2}$ of n_2 units tested in "warm" conditions;

c) the failure times $T_1, \ldots, T_n$ of n redundant systems (with "warm" stand-by units).

The tests are based on the difference of two estimators of the cdf F. The first estimator is the empirical distribution function

$$\hat{F}^{(1)}(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{T_i \leq t\}}. \tag{12.6}$$

The second is based on Formula (12.3), i.e.

$$\hat{F}^{(2)}(t) = \int_0^t \hat{F}_1(t + \hat{g}(y) - y)\,d\hat{F}_1(y),$$

where (hypothesis H_0)

$$\hat{g}(y) = \hat{F}_1^{-1}(\hat{F}_2(y)), \quad \hat{F}_j(t) = \frac{1}{n_j}\sum_{i=1}^{n_j} \mathbf{1}_{\{T_{ji} \leq t\}}, \quad \hat{F}_1^{-1}(y) = \inf\{s : \hat{F}_1(s) \geq t\}, \tag{12.7}$$

or (hypothesis H_0^*)

$$\hat{g}(y) = \hat{r}y, \quad \hat{r} = \frac{\hat{\mu}_1}{\hat{\mu}_2}, \quad \hat{\mu}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} X_{ij}. \tag{12.8}$$

The test is based on the statistic

$$X = \sqrt{n} \int_0^\infty (\hat{F}^{(1)}(t) - \hat{F}^{(2)}(t))dt. \tag{12.9}$$

It is a natural generalization of Student's t-test for comparing the means of two populations. Indeed, the mean failure time of the system with cdf F is

$$\mu = \int_0^\infty [1 - F(s)]ds,$$

so statistic (12.9) is the normed difference of two estimators (the second being not the empirical mean) of the mean μ. Student's t-test is based on the difference of empirical means of two populations.

It will be shown that in the case of both hypothesis H_0 and H_0^* the limit distribution (as $n_i/n \to l_i \in (0,1)$, $n \to \infty$) of the statistic X is normal with mean zero and finite variance σ^2.

The test statistic is

$$T = \frac{X}{\hat{\sigma}},$$

where $\hat{\sigma}$ is a consistent estimator of σ. The distribution of the statistic T is approximated by the standard normal distribution, and the hypothesis H_0 (or H_0^*) is rejected with approximative significance value α if $\mid T \mid > z_{1-\alpha/2}$, where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of the standard normal distribution.

12.4. Limit distribution of the test statistics

Let us find the asymptotic distribution of statistic (12.9).

Theorem 12.1 *Suppose that $n_i/n \to l_i \in (0,1)$, $n \to \infty$ and the densities $f_i(x)$, $i = 1, 2$ are continuous and positive on $(0, \infty)$. Then under H_0^*, statistic (12.9) converges in distribution to the normal law $N(0, \sigma^2)$, where*

$$\sigma^2 = \mathbf{Var}(T_i) + \frac{1}{l_1}\mathbf{Var}(H(T_{1i})) + \frac{c^2 r^2}{l_2}\mathbf{Var}(T_{2i}), \tag{12.10}$$

where

$$H(x) = x[c + r - 1 - F_1(x/r) - rF_2(x)] + r\mathbf{E}(\mathbf{1}_{\{T_{1i} \le x/r\}} T_{1i}) + r\mathbf{E}(\mathbf{1}_{\{T_{2i} \le x\}} T_{2i}),$$

$$c = \frac{1}{\mu_2} \int_0^\infty y[1 - F_2(y)]dF_1(y).$$

Proof. The limit distribution of the empirical distribution functions is well known:

$$\sqrt{n}(\hat{F}_i - F_i) \overset{\mathcal{D}}{\to} U_i, \quad \sqrt{n}(\hat{F}^{(1)} - F) \overset{\mathcal{D}}{\to} U \tag{12.11}$$

on $D[0,\infty)$, where $\overset{\mathcal{D}}{\to}$ means weak convergence, U_1, U_2 and U are independent Gaussian martingales with $U_i(0) = U(0) = 0$ and covariances

$$\mathbf{cov}(U_i(s_1), U_i(s_2)) = \frac{1}{l_i}\, F_i(s_1 \wedge s_2)(1 - F_i(s_1 \vee s_2)),$$

$$\mathbf{cov}(U(s_1), U(s_2)) = F(s_1 \wedge s_2)(1 - F(s_1 \vee s_2)).$$

Under the hypothesis H_0^* the difference of the two estimators of the distribution function F can be written as follows:

$$\hat{F}^{(1)}(t) - \hat{F}^{(2)}(t) = \hat{F}^{(1)}(t) - F(t) - \int_0^t \hat{F}_1(t+\hat{g}(y)-y)d\hat{F}_1(y) + \int_0^t F_1(t+g(y)-y)\times$$

$$dF_1(y) = \hat{F}^{(1)}(t) - F(t) - \int_0^t [F_1(t+\hat{g}(y)-y) - F_1(t+g(y)-y)]dF_1(y) -$$

$$\int_0^t [(\hat{F}_1(t+\hat{g}(y)-y) - \hat{F}_1(t+g(y)-y)) - (F_1(t+\hat{g}(y)-y) - F_1(t+g(y)-y))]dF_1(y) -$$

$$\int_0^t [\hat{F}_1(t+\hat{g}(y)-y) - \hat{F}_1(t+g(y)-y)](d\hat{F}_1(y) - dF_1(y)) -$$

$$\int_0^t [\hat{F}_1(t+g(y)-y) - F_1(t+g(y)-y)]dF_1(y) -$$

$$\int_0^t [\hat{F}_1(t+g(y)-y) - F_1(t+g(y)-y)](d\hat{F}_1(y) - dF_1(y)) -$$

$$\int_0^t F_1(t+g(y)-y)(d\hat{F}_1(y) - dF_1(y)).$$

Statistic (12.9) can be written:

$$X = \int_0^\infty \sqrt{n}[\hat{F}^{(1)}(t) - F(t)]dt -$$

$$\int_0^\infty dt \int_0^t \sqrt{n}[F_1(t+\hat{g}(y)-y) - F_1(t+g(y)-y)]\, dF_1(y) -$$

$$\int_0^\infty dt \int_0^t \sqrt{n}[\hat{F}_1(t+g(y)-y) - F_1(t+g(y)-y)]\, dF_1(y) -$$

$$\int_0^\infty dt \int_0^t F_1(t+g(y)-y) d\{\sqrt{n}[\hat{F}_1(y) - F_1(y)]\} + o_P(1). \tag{12.12}$$

Set $\sigma_j^2 = \mathbf{Var}(T_{jk})$, $j = 1, 2$. The convergence

$$\sqrt{n}(\hat{\mu}_j - \mu_j) \overset{\mathcal{D}}{\to} Y_j = -\int_0^\infty U_j(y)dy \sim N(0, \sigma_j^2/l_i)$$

implies

$$\sqrt{n}(\hat{r} - r) \overset{\mathcal{D}}{\to} Y = \frac{1}{\mu_2}(Y_1 - rY_2) \sim N(0, \frac{\sigma_1^2}{\mu_2^2}(\frac{1}{l_1} + \frac{1}{l_2})). \tag{12.13}$$

Formulae (12.10)-(12.13) imply

$$\int_0^\infty \sqrt{n}[\hat{F}^{(1)}(t) - F(t)]dt \overset{\mathcal{D}}{\to} \int_0^\infty U(t)dt,$$

$$\int_0^\infty dt \int_0^t \sqrt{n}[F_1(t + \hat{g}(y) - y) - F_1(t + g(y) - y)]\, dF_1(y) \overset{\mathcal{D}}{\to}$$

$$cY_1 - rcY_2 = -c\int_0^\infty U_1(y)dy + rc\int_0^\infty U_2(y)dy,$$

$$\int_0^\infty dt \int_0^t \sqrt{n}[\hat{F}_1(t + g(y) - y) - F_1(t + g(y) - y)]\, dF_1(y) \overset{\mathcal{D}}{\to}$$

$$\int_0^\infty \int_0^t U_1(t + g(y) - y)dF_1(y)dt = \int_0^\infty dF_1(y) \int_{g(y)}^\infty U_1(u)du =$$

$$\int_0^\infty U_1(u)F_1(g^{-1}(u))du,$$

$$\int_0^\infty dt \int_0^t F_1(t + g(y) - y)d\{\sqrt{n}[\hat{F}_1(y) - F_1(y)]\} \overset{\mathcal{D}}{\to}$$

$$\int_0^\infty dt \int_0^t F_1(t + g(y) - y)dU_1(y) = \int_0^\infty F_2(t)U_1(t)dt -$$

$$\int_0^\infty U_1(y)[1 - F_2(y)]d(g(y) - y) = \int_0^\infty U_1(y)[rF_2(y) - r + 1]dy.$$

We obtained

$$X \overset{\mathcal{D}}{\to} V_1 + V_2 + V_3,$$

where

$$V_1 = \int_0^\infty U(y)dy, \quad V_2 = \int_0^\infty h(y)U_1(y)dy, \quad h(y) = c + r - 1 - F_1(g^{-1}(y)) - rF_2(y),$$

$$V_3 = -rc\int_0^\infty U_2(y)dy.$$

The variances of the random variables V_i are:

$$\mathbf{Var}(V_1) = \mathbf{Var}(T_i), \quad \mathbf{Var}(V_3) = \frac{c^2 r^2}{l_2}\mathbf{Var}(T_{2i})$$

$$\mathbf{Var}(V_2) = \frac{2}{l_1}\int_0^\infty [1 - F_1(y)]h(y)dy \int_0^y F_1(z)h(z)dz = \frac{1}{l_1}\mathbf{Var}(H(T_{1j})),$$

where

$$H(x) = \int_0^x h(y)dy = x[c + r - 1 - F_1(g^{-1}(x)) - rF_2(x)] + \int_0^x ydF_1(g^{-1}(y)) +$$
$$r\int_0^x ydF_2(y) = x[c + r - 1 - F_1(g^{-1}(x)) - rF_2(x)] +$$
$$r\mathbf{E}(\mathbf{1}_{\{T_{1i} \le x/r\}}T_{1i}) + r\mathbf{E}(\mathbf{1}_{\{T_{2i} \le x\}}T_{2i}).$$

The proof is complete.

The consistent estimator $\hat\sigma^2$ of the variance σ^2 is

$$\hat\sigma^2 = \frac{1}{n}\sum_{i=1}^n (T_i - \bar T)^2 + \frac{n}{n_1^2}\sum_{i=1}^{n_1}[\hat H(T_{1i}) - \bar H]^2 + \frac{\hat c^2 \hat r^2 n}{n_2^2}\sum_{i=1}^{n_2}(T_{2i} - \bar T_2)^2,$$

where

$$\bar T = \frac{1}{n}\sum_{i=1}^n T_i, \quad \hat c = \frac{1}{\hat\mu_2}\int_0^\infty y[1 - \hat F_2(y)]d\hat F_1(y),$$

$$\hat H(x) = x[\hat c + \hat r - 1 - \hat F_1(x/\hat r) - \hat r \hat F_2(x)] + \frac{\hat r}{n_1}\sum_{i=1}^{n_1}\mathbf{1}_{\{T_{1i} \le x/\hat r\}}T_{1i} + \frac{\hat r}{n_2}\sum_{i=1}^{n_2}\mathbf{1}_{\{T_{2i} \le x\}}T_{2i},$$

$$\hat H = \frac{1}{n_1}\sum_{i=1}^{n_1}\hat H(T_{1i}).$$

Theorem 12.2 *Suppose that $n_i/n \to l_i \in (0,1)$, $n \to \infty$ and the densities $f_i(x)$, $i = 1, 2$ are continuous and positive on $(0, \infty)$. Then under H_0, statistic (12.9) converges in distribution to the normal law $N(0, \sigma^2)$, where*

$$\sigma^2 = \mathbf{Var}(T_i) + \frac{1}{l_1}\mathbf{Var}(H(T_{1j})) +$$
$$\frac{2}{l_2}\int_0^\infty \frac{[1 - F_2(y)]^2 dF_1(y)}{f_1(g(y))}\int_0^y \frac{F_2(z)[1 - F_2(z)]dF_1(z)}{f_1(g(z))},$$

where

$$H(x) = \int_0^x \frac{[1 - F_2(y)]}{f_1(g(y))}dF_1(y) - xF_1(x/r) + g(x)[1 - F_2(x)] +$$
$$\mathbf{E}(\mathbf{1}_{\{g(T_{1i}) \le x\}}g(T_{1i}) + \mathbf{E}(\mathbf{1}_{\{T_{2i} \le x\}}g(T_{2i}) - x.$$

Proof. Similarly as in Theorem 12.1 we obtain

$$\int_0^\infty \sqrt{n}[\hat{F}^{(1)}(t) - F(t)]dt \xrightarrow{\mathcal{D}} \int_0^\infty U(t)dt,$$

$$\int_0^\infty dt \int_0^t \sqrt{n}[F_1(t+\hat{g}(y)-y) - F_1(t+g(y)-y)]\, dF_1(y) \xrightarrow{\mathcal{D}}$$

$$\int_0^\infty \frac{U_1(g(y)) - U_2(y)}{f_1(g(y))} f_1(y)[1-F_2(y)]dF_1(y),$$

$$\int_0^\infty dt \int_0^t \sqrt{n}[\hat{F}_1(t+g(y)-y) - F_1(t+g(y)-y)]\, dF_1(y) \xrightarrow{\mathcal{D}}$$

$$\int_0^\infty U_1(u)F_1(g^{-1}(u))du,$$

$$\int_0^\infty dt \int_0^t F_1(t+g(y)-y)d\{\sqrt{n}[\hat{F}_1(y) - F_1(y)]\} \xrightarrow{\mathcal{D}}$$

$$\int_0^\infty dt \int_0^t F_1(t+g(y)-y)dU_1(y) = \int_0^\infty F_2(t)U_1(t)dt -$$

$$\int_0^\infty U_1(y)[1-F_2(y)]d(g(y)-y) = \int_0^\infty U_1(y)\{F_2(y) - (g'(y)-1)[1-F_2(y)]\}dy.$$

We obtained

$$X \xrightarrow{\mathcal{D}} V_1 + V_2 + V_3,$$

where

$$V_1 = \int_0^\infty U(y)dy, \quad V_2 = \int_0^\infty h(y)U_1(y)dy,$$

$$h(y) = \frac{f_1(y)}{f_1(g(y))}[1-F_2(y)] - F_1(g^{-1}(y)) - F_2(y) + (g'(y)-1)[1-F_2(y)].$$

$$V_3 = -\int_0^\infty \frac{U_2(y)}{f_1(g(y))} f_1(y)[1-F_2(y)]dF_1(y).$$

The variances of the random variables V_i are:

$$\mathbf{Var}(V_1) = \mathbf{Var}(T_i),$$

$$\mathbf{Var}(V_3) = \frac{2}{l_2}\int_0^\infty \frac{[1-F_2(y)]^2 dF_1(y)}{f_1(g(y))} \int_0^y \frac{F_2(z)[1-F_2(z)]dF_1(z)}{f_1(g(z))},$$

$$\mathbf{Var}(V_2) = \frac{2}{l_1}\int_0^\infty [1-F_1(y)]h(y)dy \int_0^y F_1(z)h(z)dz = \frac{1}{l_1}\mathbf{Var}(H(T_{1j})),$$

where

$$H(x) = \int_0^x \frac{[1-F_2(y)]}{f_1(g(y))} dF_1(y) - \int_0^x F_1(g^{-1}(y))dy - \int_0^x F_2(y)dy +$$

$$\int_0^x [1-F_2(y)]dg(y) - \int_0^x [1-F_2(y)]dy = \int_0^x \frac{[1-F_2(y)]}{f_1(g(y))} dF_1(y) -$$

$$F_1(g^{-1}(x))x + \int_0^x y\, dF_1(g^{-1}(y)) + [1-F_2(x)]g(x) + \int_0^x g(y)dF_2(y) - x =$$

$$\int_0^x \frac{[1-F_2(y)]}{f_1(g(y))} dF_1(y) - xF_1(g^{-1}(x)) + g(x)[1-F_2(x)] +$$

$$\mathbf{E}(\mathbf{1}_{\{g(T_{1i}) \leq x\}} g(T_{1i}) + \mathbf{E}(\mathbf{1}_{\{T_{2i} \leq x\}} g(T_{2i}) - x.$$

The proof is complete.

The consistent estimator $\hat{\sigma}^2$ of the variance σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (T_i - \bar{T})^2 + \frac{n}{n_1^2} \sum_{i=1}^{n_1} [H(T_{1i} - \bar{H}]^2 +$$

$$\frac{2}{l_2} \int_0^\infty \frac{[1-\hat{F}_2(y)]^2 d\hat{F}_1(y)}{\hat{f}_1(\hat{g}(y))} \int_0^y \frac{\hat{F}_2(z)[1-\hat{F}_2(z)]d\hat{F}_1(z)}{\hat{f}_1(\hat{g}(z))},$$

and the density f_1 is estimated by the kernel estimator

$$\hat{f}_1(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - X_{1i}}{h}\right).$$

12.5. Bibliography

[BAG 78] Bagdonavičius, V. (1978) Testing the hyphothesis of the additive accumulation of damages, *Probab. Theory and its Appl.*, vol 23 (2), 403–408.

[BAG 90] Bagdonavičius,V. (1990) Accelerated life models when the stress is not constant. *Kybernetika*, vol 26, 289–295.

[BAG 97] Bagdonavičius, V. and Nikoulina, V. (1997) A goodness-of-fit test for Sedyakin's model, *Revue Roumaine de Mathématiques Pures et Appliquées*, vol 42 (1), 5–14.

[BAG 02] Bagdonavicius, V. and Nikulin, M. (2002) *Accelerated Life Models*, Chapman & Hall/CRC, Boca Raton.

[NEL 90] Nelson, W. (1990) *Accelerated Testing: Statistical Models, Test Plans, and Data Analyses,* New York: John Wiley and Sons.

[SED 66] Sedyakin, N.M. (1966) On one physical principle in reliability theory (in Russian), *Techn. Cybernetics*, vol 3, 80–87.

[VEK 87] Veklerov, E. (1987) Reliability of redundant systems with unreliable switches, *IEEE Transactions on Reliability*, vol 36, 470–472.

[WAS 05] Wasson Ch.S. (2005) *System Analysis, Design, and Development: Concepts, Principles, and Practices*, New York: Wiley.

Chapter 13

A Modified Chi-squared Goodness-of-fit Test for the Three-parameter Weibull Distribution and its Applications in Reliability

13.1. Introduction

The Weibull probability distribution plays an important role in a statistical analysis of lifetime or response data in reliability and survival studies. To extend applications of the distribution to a wider class of failure-rate situations, several modifications of the classical Weibull model were proposed. Mudholkar, Srivastava and Freimer [MUD 95] as well as Mudholkar, Srivastava and Kollia [MUD 96] introduced the exponentiated and generalized Weibull families, which include distributions with unimodal and bathtub failure rates and also possess a broad class of monotone hazard rates. Bagdonavicius and Nikulin [BAG 02] have proposed another generalization – the power generalized Weibull family, which also includes all possible failure rate functions.

One of the main problems of statistical modeling is the selection of a proper probability distribution to be used in an analysis. Different criteria, both parametric and non-parametric, may be used for such a selection. Chi-squared tests are among them and it is worth considering these tests in more detail.

Fisher [FIS 24] showed that the limit distribution of Pearson's sum essentially depends on a method of parameter estimation. He proved that using efficient estimators

Chapter written by Vassilly VOINOV, Roza ALLOYAROVA and Natalie PYA.

based on grouped data would lead to the limit chi-squared distribution with $r - s - 1$ degrees of freedom, where r stands for the number of grouping intervals and s the number of estimated parameters.

Now it is known [VOI 04] that for equiprobable cells Fisher's tests possess very low power. Up to 1954, statisticians thought that estimating parameters by the well-known method of maximum likelihood based on ungrouped data would give the same limit distribution as while using grouped data. However, Chernoff and Lehmann [CHE 54] proved that in the case of ungrouped data the limit distribution of Pearson's sum would not follow the chi-squared distribution with $r - s - 1$ degrees of freedom in the limit and, moreover, will depend on unknown parameters, thus making Pearson's test inapplicable. In 1973, Nikulin [NIK 73c], [NIK 73b], [NIK 73a] proposed to modify the standard Pearson's test in such a manner that the limit distribution of a resulting quadratic form will be chi-squared with $r - 1$ degrees of freedom in the limit and will not depend on unknown parameters. In 1974, Rao and Robson [RAO 74] by other means obtained the same result for the exponential family of distributions. Since 1988, the test is known as the Rao-Robson-Nikulin (RRN) test [DRO 88], [VAA 98]. In 1975, Moore and Spruill [MOO 75] developed the general theory of constructing modified chi-squared tests based on MLE.

The main idea of Nikulin's modification was to recover Fisher's sample information lost while grouping data. Nowadays we know [LEM 01], [VOI 06] that for equiprobable cells Nikulin's test recovers the largest part of the information lost and, possibly, cannot be improved. The RRN test can be implemented in two main settings: equiprobable fixed or random intervals, and Neyman-Pearson classes [GRE 96]. It is well known that it does not matter which classes, fixed or random, are used because limit distributions of modified tests will be the same (see [GRE 96]).

In 1976, Hsuan and Robson [HSU 76] showed that to modify Pearson's test we may use not only efficient maximum likelihood estimates (MLEs) but non-efficient $\sqrt{n}$- consistent moment type estimates (MMEs) as well. They proved the existence of such a modification but have written it implicitly. In 2001, Mirvaliev [MIR 01] derived the corresponding quadratic form explicitly. We shall refer to this test as the HRM statistic.

In section 13.2, we present three such modifications of chi-squared type tests for the three-parameter Weibull distribution. The proposed tests are based on non-efficient moment type estimators of all unknown parameters. In section 13.3 Monte Carlo simulation is used to study the power of the tests for equiprobable random cells versus the three modifications of the Weibull family of distributions. Section 13.4 is devoted to a modified chi-squared test based on two Neyman-Pearson classes. Section 13.5 is devoted to a discussion of results that were obtained.

Throughout the chapter vectors and matrices are boldfaced.

13.2. Parameter estimation and modified chi-squared tests

Let us consider the three-parameter Weibull family with the probability density function

$$f(x;\theta,\mu,p) = \frac{p}{\theta}\left(\frac{x-\mu}{\theta}\right)^{p-1} \exp\left\{-\left(\frac{x-\mu}{\theta}\right)^{p}\right\},$$

$$x > \mu, \theta > 0, p > 0, \mu \in R^1. \qquad (13.1)$$

It is well known that there are serious problems with obtaining maximum likelihood estimates (MLEs) for this probability distribution if all three parameters are unknown. Sometimes there is no local maximum for the likelihood function, and sometimes the likelihood can be infinite [LOC 94]. If a person intends to apply the RRN test, s/he would use the following Fisher's information matrix $\mathbf{J}$:

$$\begin{pmatrix} \frac{1}{p^2}\left[(c-1)^2+\frac{\pi^2}{6}\right] & \frac{c-1}{\theta} & -\frac{1}{\theta}\Gamma\left(2-\frac{1}{p}\right)\left[\Psi\left(1-\frac{1}{p}\right)+1\right] \\ \frac{c-1}{\theta} & \frac{p^2}{\theta^2} & \frac{p^2}{\theta^2}\Gamma\left(2-\frac{1}{p}\right) \\ -\frac{1}{\theta}\Gamma\left(2-\frac{1}{p}\right)\left[\Psi\left(1-\frac{1}{p}\right)+1\right] & \frac{p^2}{\theta^2}\Gamma\left(2-\frac{1}{p}\right) & \frac{(p-1)^2}{\theta^2}\Gamma\left(1-\frac{2}{p}\right) \end{pmatrix}$$

which does not exist for infinitely many values of unknown shape parameter p ($p = \frac{1}{2+k}$ and $p = \frac{2}{1+k}, k = 0, 1, 2, \ldots$). Because of the mentioned problems the RRN test based on MLEs is not easy to apply, and instead we can use the HRM test based on moment type estimates (MMEs). In case of the three-parameter Weibull distribution, MMEs exist for any $p > 0$ (see section 13.7) and are easily found, for example, by using Microsoft Excel Solver.

From Figure 13.1 it can be seen that MMEs are $\sqrt{n}$ - consistent. Let $\bar{\boldsymbol{\theta}}_n$ be the MME of $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)^T$, where $\theta_1 = \mu$, $\theta_2 = \theta$, $\theta_3 = p$. The regularity conditions of Hsuan and Robson [HSU 76] needed for implementing the HRM test are as follows:

(1) the MMEs are $\sqrt{n}$- consistent;

(2) matrix $\mathbf{K}$(see below) is non-singular;

(3) $\int\limits_{x>\mu} g_i(x) f(x;\boldsymbol{\theta})dx \int\limits_{x>\mu} g_i(x)\frac{\partial f(x;\boldsymbol{\theta})}{\partial\theta_j}dx, \int\limits_{x>\mu} g_i(x)\frac{\partial^2 f(x;\boldsymbol{\theta})}{\partial\theta_j\partial\theta_k}dx,$
where $g_i(x) = x^i$ exist and are finite and continuous in $\boldsymbol{\theta}$ for all $i, j, k = 1, 2, 3$ in a neighborhood of the true value of the parameter $\boldsymbol{\theta}$. It can be easily verified that the condition (3) is satisfied for the three-parameter Weibull family (13.1) if $p > 2$.

Consider briefly the theory of the HRM test.

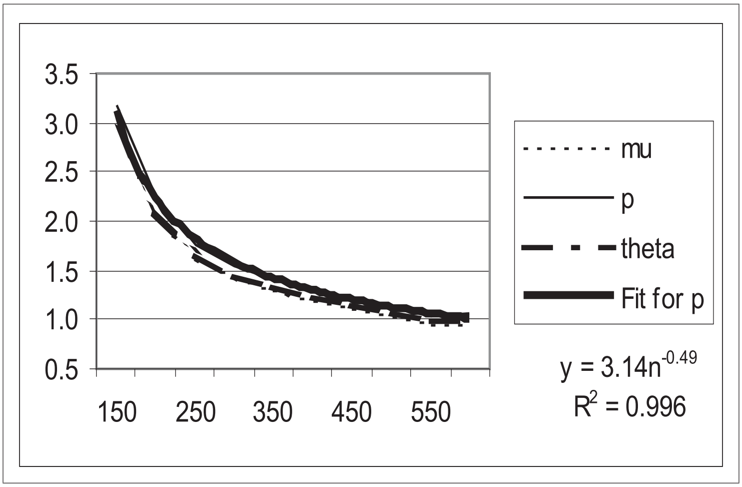


Figure 13.1. *Simulated absolute values of errors for MLEs $\bar{\mu}, \bar{\theta}, \bar{p}$ versus their true values as a function of the sample size n*

Let $\mathbf{B}$ be an $r \times 3$ matrix with elements

$$\frac{1}{\sqrt{p_i(\boldsymbol{\theta})}} \int\limits_{\triangle_i(\boldsymbol{\theta})} \frac{\partial f(x;\boldsymbol{\theta})}{\partial \theta_j} dx, i = 1, \ldots, r, j = 1, 2, 3,$$

where $\triangle_i(\boldsymbol{\theta})$ are non-intersecting grouping intervals, random if $\boldsymbol{\theta}$ is replaced by its estimate $\bar{\boldsymbol{\theta}}_n$, and $\mathbf{K}$ is a 3×3 matrix with elements

$$\int\limits_{x>\mu} g_i(x) \frac{\partial f(x;\boldsymbol{\theta})}{\partial \theta_j} dx, i, j = 1, 2, 3.$$

Direct calculations show that this matrix is non-singular. Thus, all regularity conditions of Hsuan and Robson [HSU 76] are satisfied if $p > 2$. Let the matrix

$$\mathbf{V} = (V_{ij}), V_{ij} = m_{ij}(\boldsymbol{\theta}) - m_i(\boldsymbol{\theta}) m_j(\boldsymbol{\theta}),$$

where $m_i(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}[X^i]$, $m_{ij}(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}[X^{i+j}]$, $i, j = 1, 2, 3$, and $\mathbf{C}$ is an $r \times 3$ matrix with elements

$$p_i^{-1/2}(\boldsymbol{\theta}) \Bigg(\int\limits_{\triangle_j(\boldsymbol{\theta})} x^j f(x;\boldsymbol{\theta}) dx - p_i(\boldsymbol{\theta}) m_j(\boldsymbol{\theta}) \Bigg), i = 1, \ldots, r, j = 1, 2, 3,$$

where $p_j(\boldsymbol{\theta}) = \int_{\triangle_j(\boldsymbol{\theta})} f(x;\boldsymbol{\theta}) dx, j = 1, \ldots, r$. Hsuan and Robson showed that under the above-listed regularity conditions the limit distribution of the Pearson's sum would be

$$\lim_{n\to\infty} P\{X_n^2(\bar{\boldsymbol{\theta}}_n) \geq x \mid H_0\} = P\Big\{\sum_{j=1}^{r-1} \lambda_j(\boldsymbol{\theta})\chi_j^2 \geq x\Big\},$$

where χ_j^2 are independent central chi-squared random variables with one degree of freedom, and $\lambda_j(\boldsymbol{\theta})$ are non-zero characteristic roots of the limit covariance matrix $\boldsymbol{\Sigma}$ of the r - vector of standardized frequencies $\mathbf{V}^{(n)}(\bar{\boldsymbol{\theta}}_n)$ with components $\nu_i^{(n)}(\bar{\boldsymbol{\theta}}_n) = [np_i(\bar{\boldsymbol{\theta}}_n)]^{-1/2}(N_i^{(n)} - np_i(\bar{\boldsymbol{\theta}}_n)), i = 1, \ldots, r$, where $N_i^{(n)}$ is the number of observations that fall into $\triangle_i(\boldsymbol{\theta})$. The matrix $\boldsymbol{\Sigma}$ is written as

$$\boldsymbol{\Sigma} = \mathbf{I} - \mathbf{q}\mathbf{q}^T + \mathbf{B}\mathbf{K}^{-1}\mathbf{V}(\mathbf{K}^{-1})^T\mathbf{B}^T - \mathbf{C}(\mathbf{K}^{-1})^T\mathbf{B}^T - \mathbf{B}\mathbf{K}^{-1}\mathbf{C}^T,$$

where $\mathbf{q} = \big((p_1(\boldsymbol{\theta}))^{1/2}, \ldots, (p_r(\boldsymbol{\theta}))^{1/2}\big)^T$.

Since the limit distribution of the Pearson's statistic $X_n^2(\bar{\boldsymbol{\theta}}_n)$ depends on unknown parameter $\boldsymbol{\theta}$, it cannot be implemented for hypothesis testing.

Using Wald's idea [WAL 43] (see also [MOO 77]), Hsuan and Robson proved that the statistic $\mathbf{V}^{(n)T}(\bar{\boldsymbol{\theta}}_n)\boldsymbol{\Sigma}^{-}\mathbf{V}^{(n)}(\bar{\boldsymbol{\theta}}_n)$, where $\boldsymbol{\Sigma}^{-}$ is any generalized inverse matrix of $\boldsymbol{\Sigma}$, will in the limit follow the chi-squared distribution with $r-1$ degrees of freedom and will not depend on unknown parameters. In 2001, Mirvaliev revived the idea of Hsuan and Robson and using the Moore-Penrose inversion of the matrix $\boldsymbol{\Sigma}$ obtained the following explicit formula for the modified chi-squared test based on MMEs:

$$Y2_n^2(\bar{\boldsymbol{\theta}}_n) = U_n^2(\bar{\boldsymbol{\theta}}_n) + W_n^2(\bar{\boldsymbol{\theta}}_n) + R_n^2(\bar{\boldsymbol{\theta}}_n) - Q_n^2(\bar{\boldsymbol{\theta}}_n), \tag{13.2}$$

where $U_n^2(\bar{\boldsymbol{\theta}}_n) = \mathbf{V}^{(n)T}(\bar{\boldsymbol{\theta}}_n)\big(\mathbf{I} - \mathbf{B}(\bar{\boldsymbol{\theta}}_n)\mathbf{J}^{-1}(\bar{\boldsymbol{\theta}}_n)\mathbf{B}^T(\bar{\boldsymbol{\theta}}_n)\big)\mathbf{V}^{(n)}(\bar{\boldsymbol{\theta}}_n)$ is the well-known Dzhaparidze and Nikulin (DN) statistic [DZH 92], $\mathbf{J}(\bar{\boldsymbol{\theta}}_n) = \mathbf{B}^T(\bar{\boldsymbol{\theta}}_n)\mathbf{B}(\bar{\boldsymbol{\theta}}_n)$ is the Fisher's information matrix for grouped data,

$$W_n^2(\bar{\boldsymbol{\theta}}_n) = \mathbf{V}^{(n)T}(\bar{\boldsymbol{\theta}}_n)\mathbf{B}(\bar{\boldsymbol{\theta}}_n)\mathbf{J}^{-1}(\bar{\boldsymbol{\theta}}_n)\mathbf{B}^T(\bar{\boldsymbol{\theta}}_n)\mathbf{V}^{(n)}(\bar{\boldsymbol{\theta}}_n),$$

$$U_n^2(\bar{\boldsymbol{\theta}}_n) + W_n^2(\bar{\boldsymbol{\theta}}_n) = X_n^2(\bar{\boldsymbol{\theta}}_n)$$

is the Pearson's statistic [MIR 01],

$$R_n^2(\bar{\boldsymbol{\theta}}_n) = \mathbf{V}^{(n)T}(\bar{\boldsymbol{\theta}}_n)\mathbf{C}_n(\mathbf{V}_n - \mathbf{C}_n^T\mathbf{C}_n)^{-1}\mathbf{C}_n^T\mathbf{V}^{(n)}(\bar{\boldsymbol{\theta}}_n),$$

and

$$Q_n^2(\bar{\boldsymbol{\theta}}_n) = \mathbf{V}^{(n)T}(\bar{\boldsymbol{\theta}}_n)\mathbf{A}_n(\mathbf{C}_n - \mathbf{B}_n\mathbf{K}_n^{-1}\mathbf{V}_n)\mathbf{L}_n^{-1}(\mathbf{C}_n - \mathbf{B}_n\mathbf{K}_n^{-1}\mathbf{V}_n)^T\mathbf{A}_n^T\mathbf{V}^{(n)}(\bar{\boldsymbol{\theta}}_n).$$

By $\mathbf{K}_n, \mathbf{C}_n, \mathbf{B}_n, \mathbf{L}_n, \mathbf{A}_n$ and $\mathbf{V}_n$ we mean the estimates of corresponding matrices, e.g., $\mathbf{V}_n = \mathbf{V}(\bar{\boldsymbol{\theta}}_n)$.

Mirvaliev proved that under regularity conditions of Hsuan and Robson $Y2_n^2(\bar{\boldsymbol{\theta}}_n) \sim \chi^2_{r-1}$, $U_n^2(\bar{\boldsymbol{\theta}}_n) \sim \chi^2_{r-4}$, and $W_n^2(\bar{\boldsymbol{\theta}}_n) + R_n^2(\bar{\boldsymbol{\theta}}_n) - Q_n^2(\bar{\boldsymbol{\theta}}_n) \sim \chi^2_3$. Since the limit distributions of the DN test $U_n^2(\bar{\boldsymbol{\theta}}_n)$ based on initial data and Pearson-Fisher's test $X_n^2(\bar{\boldsymbol{\theta}}_n)$ based on grouped data are asymptotically equivalent, it follows that the term $W_n^2(\bar{\boldsymbol{\theta}}_n) + R_n^2(\bar{\boldsymbol{\theta}}_n) - Q_n^2(\bar{\boldsymbol{\theta}}_n)$ recovers information lost while data grouping. Asymptotically the test $Y2_n^2(\bar{\boldsymbol{\theta}}_n) - U_n^2(\bar{\boldsymbol{\theta}}_n) = W_n^2(\bar{\boldsymbol{\theta}}_n) + R_n^2(\bar{\boldsymbol{\theta}}_n) - Q_n^2(\bar{\boldsymbol{\theta}}_n)$ is independent of the DN test and can be used in its own right. We also see that $Y2_n^2(\bar{\boldsymbol{\theta}}_n)$ equals

$$Y2_n^2(\bar{\boldsymbol{\theta}}_n) = X_n^2(\bar{\boldsymbol{\theta}}_n) + R_n^2(\bar{\boldsymbol{\theta}}_n) - Q_n^2(\bar{\boldsymbol{\theta}}_n),$$

the sum of Pearson's test $X_n^2(\bar{\boldsymbol{\theta}}_n) = U_n^2(\bar{\boldsymbol{\theta}}_n) + W_n^2(\bar{\boldsymbol{\theta}}_n)$ and a quadratic form $R_n^2(\bar{\boldsymbol{\theta}}_n) - Q_n^2(\bar{\boldsymbol{\theta}}_n)$, which is the term correcting $X_n^2(\bar{\boldsymbol{\theta}}_n)$ in such a manner that the limit distribution of the resulting quadratic form $Y2_n^2(\bar{\boldsymbol{\theta}}_n)$ will not depend on $\boldsymbol{\theta}$ and be chi-squared distributed with the maximal possible number $r-1$ degrees of freedom.

From the above theory it seems reasonable to investigate the power of three modified chi-squared type tests $Y2_n^2(\bar{\boldsymbol{\theta}}_n)$, $U_n^2(\bar{\boldsymbol{\theta}}_n)$, and $Y2_n^2(\bar{\boldsymbol{\theta}}_n) - U_n^2(\bar{\boldsymbol{\theta}}_n)$.

13.3. Power estimation

To investigate the power of the DN $U_n^2(\bar{\boldsymbol{\theta}}_n)$, the HRM $Y2_n^2(\bar{\boldsymbol{\theta}}_n)$ and $Y2_n^2(\bar{\boldsymbol{\theta}}_n) - U_n^2(\bar{\boldsymbol{\theta}}_n)$ tests for the three-parameter Weibull null hypothesis against several useful alternatives in reliability theory, we conducted Monte Carlo simulation for the different number of equiprobable random cells (parameters of the null hypothesis were $p = 3, \mu = 0, \theta = 3.5$). Figures 13.2–13.4 illustrate results that were obtained using the study.

From these figures we see that the DN $U_n^2(\bar{\boldsymbol{\theta}}_n)$ test for equiprobable random cells possesses no power for the exponentiated and power generalized Weibull alternatives and is almost insensitive with respect to the generalized Weibull alternative. On the contrary, the statistic $Y2_n^2(\bar{\boldsymbol{\theta}}_n) - U_n^2(\bar{\boldsymbol{\theta}}_n)$ is the most powerful test for all alternatives considered and all reasonable numbers r of cells. The same relation between the powers of $U_n^2(\bar{\boldsymbol{\theta}}_n)$ and $Y2_n^2(\bar{\boldsymbol{\theta}}_n) - U_n^2(\bar{\boldsymbol{\theta}}_n)$ is observed, for instance, when testing logistic null hypothesis against normal probability distribution [VOI 08]. Note that the case $r > 40$ needs further investigation because the expected cell probabilities become small and limit distributions of tests considered may not follow the chi-squared distribution.

13.4. Neyman-Pearson classes

To maximize a measure of the distance between the null and alternative probability density functions, we may use the Neyman-Pearson classes [GRE 96] for partitioning a sample space. Intuitively this should increase the power of tests. Let $f(x;\boldsymbol{\theta})$ and

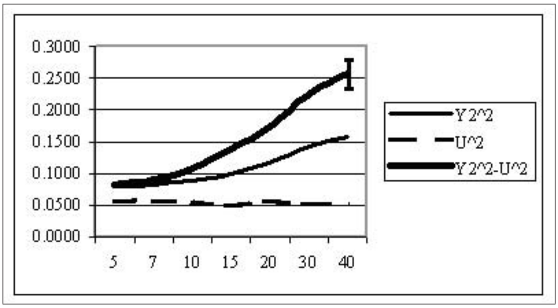


Figure 13.2. *Estimated powers as functions of the number of equiprobable cells r of* $Y2_n^2(\bar{\boldsymbol{\theta}}_n)$*,* $U_n^2(\bar{\boldsymbol{\theta}}_n)$*, and* $Y2_n^2(\bar{\boldsymbol{\theta}}_n) - U_n^2(\bar{\boldsymbol{\theta}}_n)$ *tests versus exponentiated Weibull (ExpW) alternative* $F(x) = \left[1 - \exp(1 - (x/\alpha)^\beta)\right]^\gamma$*,* $x, \alpha, \beta, \gamma > 0$*, of Mudholkar, Srivastava and Freimer [MUD 95]. Sample size* $n = 200$*, Type-1 error* $\alpha = 0.05$*. The statistical error shown is of one standard deviation*

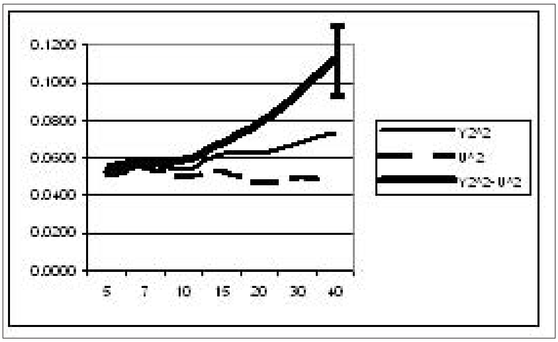


Figure 13.3. *Estimated powers of* $Y2_n^2(\bar{\boldsymbol{\theta}}_n)$*,* $U_n^2(\bar{\boldsymbol{\theta}}_n)$*, and* $Y2_n^2(\bar{\boldsymbol{\theta}}_n) - U_n^2(\bar{\boldsymbol{\theta}}_n)$ *tests versus power generalized Weibull (PGW) alternative* $F(x) = 1 - \exp\left\{1 - \left[1 + (x/\sigma)^\nu\right]^{1/\gamma}\right\}$*,* $x, \sigma, \nu, \gamma > 0$*, of Bagdonavicius and Nikulin [BAG 02]. Sample size* $n = 200$*, Type-1 error* $\alpha = 0.05$*. The statistical error shown is of one standard deviation*

$g(x; \boldsymbol{\varphi})$ be densities of the null and the alternative hypotheses correspondingly. Given parameters $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$, define two fixed Neyman-Pearson classes as follows: $I_1 = \{x : f(x; \boldsymbol{\theta}) < g(x; \boldsymbol{\varphi})\}$ and $I_2 = \{x : f(x; \boldsymbol{\theta}) > g(x; \boldsymbol{\varphi})\}$.

If density functions intersect at three points, say, x_1, x_2, and x_3 as in Figure 13.5, then $I_1 = (0, x_1] \cup [x_2, x_3]$ and $I_2 = (x_1, x_2) \cup (x_3, \infty)$.

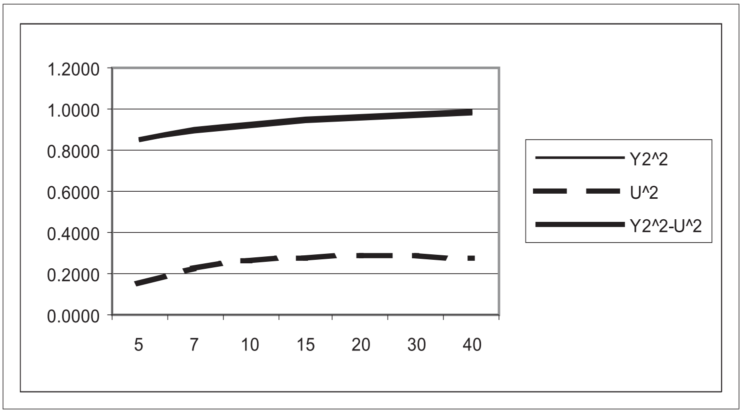


Figure 13.4. *Estimated powers of* $Y2_n^2(\bar{\boldsymbol{\theta}}_n)$, $U_n^2(\bar{\boldsymbol{\theta}}_n)$, *and* $Y2_n^2(\bar{\boldsymbol{\theta}}_n) - U_n^2(\bar{\boldsymbol{\theta}}_n)$ *tests versus generalized Weibull (GWeib) alternative* $F(x) = \left[1 - \left(1 - \lambda(x/\sigma)^{1/\alpha}\right)^{1/\lambda}\right]$, $\alpha, \sigma > 0, \lambda \in R^1$, *of Mudholkar, Srivastava, and Kollia [MUD 96]. Sample size* $n = 200$, *Type-1 error* $\alpha = 0.05$

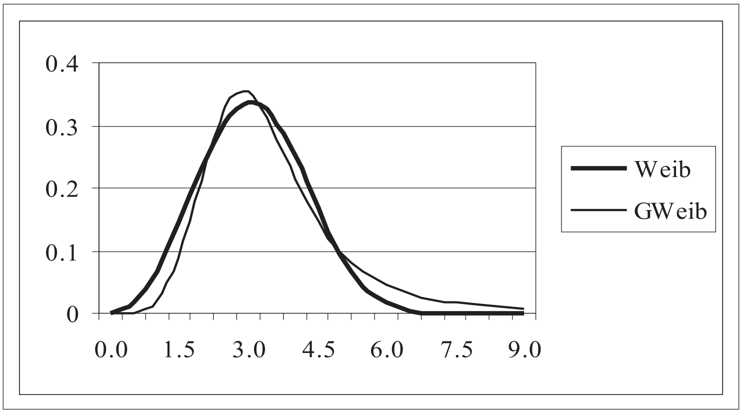


Figure 13.5. *Probability density functions of the three-parameter Weibull distribution (13.1) and the generalized Weibull distribution [MUD 96]*

The power of the HRM statistic $Y2_n^2(\bar{\boldsymbol{\theta}}_n)$ was simulated for the three-parameter Weibull family (13.1) as a null hypothesis and for all three alternatives that were considered above. Sample size and Type-1 error were analogous to the test when equiprobable cells were investigated. The results are presented in Table 13.1.

	W-PGW	W-ExpW	W-GWeib
$\alpha = 0.05$	0.141 ± 0.025	0.294 ± 0.015	1.000

Table 13.1. *Power of the HRM test for two Neyman-Pearson classes*

The statistical errors shown in Table 13.1 are of one standard deviation.

13.5. Discussion

From Figures 13.2–13.4 we may conclude that if an alternative hypothesis is not specified and we use equiprobable cells, then the DN $U_n^2(\bar{\boldsymbol{\theta}}_n)$ test is useless and the more powerful $Y2_n^2(\bar{\boldsymbol{\theta}}_n) - U_n^2(\bar{\boldsymbol{\theta}}_n)$ test is recommended for applications. If, on the contrary, the alternative hypothesis is specified, then from Table 13.1 it follows that the HRM test $Y2_n^2(\bar{\boldsymbol{\theta}}_n)$ for two Neyman-Pearson classes possesses higher power and is recommended for implementing. At the same time, the $Y2_n^2(\bar{\boldsymbol{\theta}}_n) - U_n^2(\bar{\boldsymbol{\theta}}_n)$ test for equiprobable cells and $Y2_n^2(\bar{\boldsymbol{\theta}}_n)$ for Neyman-Pearson classes do not discriminate between the three-parameter Weibull and the GPW distributions and are not very powerful against the ExpW family.

Consider the numerical data of Smith and Naylor [SMI 87] (Sample 1), which describe the strength of glass fibers of length 1.5 cm. The global maximum of the likelihood function for the three-parameter continuous Weibull probability distribution is $+\infty$ and is achieved at $\mu = X_{(1)}$. This fact creates problems in obtaining MLE. To overcome this problem Smith and Naylor proposed considering observations as being integers. Using this approach they estimated parameters of the hypothetical three-parameter Weibull distribution ($\hat{\mu} = -1.6, \hat{\theta} = 3.216, \hat{p} = 11.9$) by the local MLEs. To test the hypothesis they compared the empirical distribution function (EDF) of the original data with the approximately 95% confidence limits formed from 19 EDFs based on pseudo-random samples. Smith and Naylor did not observe a systematic discrepancy, but a correct statistical testing is still desirable. Our MME are quite different ($\bar{\mu} = -4.4, \bar{\theta} = 6.063, \bar{p} = 22.92$), but visually the density functions (13.1) calculated for these sets of estimates do not differ essentially (Figure 13.6).

The fit for MME ($\chi_9^2 = 49.5$) is slightly better than that for MLE ($\chi_9^2 = 58.0$). Having calculated for the same data set the statistics $Y2_n^2(\bar{\boldsymbol{\theta}}_n)$ and $Y2_n^2(\bar{\boldsymbol{\theta}}_n) - U_n^2(\bar{\boldsymbol{\theta}}_n)$ for the reasonable number of equiprobable cells $r = 12$, we obtained p - values 0.101

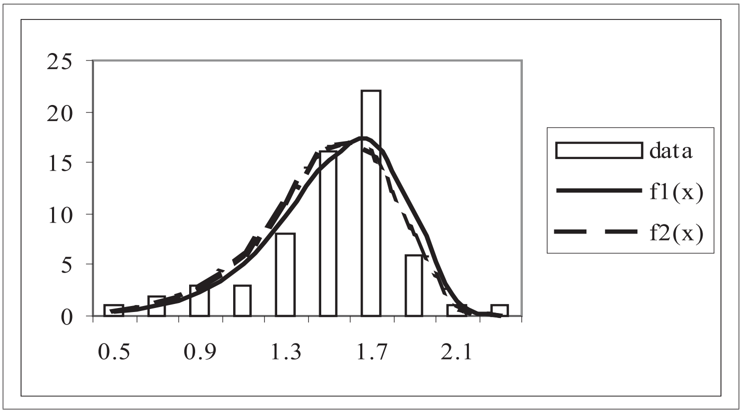


Figure 13.6. *The histogram of the data observed and densities (13.1) calculated for our MME (f1(x)) and for MLE of [SMI 87] (f2(x))*

and 0.043 respectively. From this it follows that the null hypothesis about the three-parameter Weibull distribution is not rejected at a level of significance $\alpha = 0.05$ by $Y2_n^2(\bar{\boldsymbol{\theta}}_n)$ and is rejected by the more powerful test $Y2_n^2(\bar{\boldsymbol{\theta}}_n) - U_n^2(\bar{\boldsymbol{\theta}}_n)$.

13.6. Conclusion

We tested a null hypothesis against an alternative assuming parameters to be unknown. This means that parameters of the null hypothesis were adjusted to a sample generated by the alternative model, thus making null and alternative hypotheses as close as possible. In other words, tests are sensitive only to the difference in shape of those hypothetical null probability distributions. From the results obtained it follows that shapes of the three-parameter Weibull, GPW, and ExpW distributions are very close to each other, though their hazard rate functions can be different. This suggests developing a test which will directly compare observed and hypothetical failure rate functions. Note also that implementing MMEs for testing hypotheses disproves the popular opinion that only effective MLEs may be used for modifying the Pearson's chi-squared test.

13.7. Appendix

For r equiprobable random cells, borders of intervals were defined as:

$$x_i = \mu + \theta\left[-\ln\left(1-\frac{i}{r}\right)\right]^{1/p}, i = 0, 1, \ldots, r, x_0 = \mu, x_r = \infty, p_i = \frac{1}{r}, i = 1, \ldots, r.$$

Elements of matrix **K** are:

$$K_{11}=1,\ \ K_{12}=\frac{1}{p}\Gamma\Big(\frac{1}{p}\Big),\ \ K_{13}=-\frac{\theta}{p^2}\Gamma'\Big(1+\frac{1}{p}\Big),\ \ K_{21}=2\mu+2\theta\Gamma\Big(1+\frac{1}{p}\Big),$$

$$K_{22}=2\mu\Gamma\Big(1+\frac{1}{p}\Big)+2\theta\Gamma\Big(1+\frac{1}{p}\Big),\ \ K_{23}=-\frac{2\mu\theta}{p^2}\Gamma'\Big(1+\frac{1}{p}\Big)-\frac{2\theta^2}{p^2}\Gamma'\Big(1+\frac{2}{p}\Big),$$

$$K_{31}=3\mu^2+6\mu\theta\Gamma\Big(1+\frac{1}{p}\Big)+3\theta^2\Gamma\Big(1+\frac{2}{p}\Big),$$

$$K_{32}=3\mu^2\Gamma\Big(1+\frac{1}{p}\Big)+6\mu\theta\Gamma\Big(1+\frac{2}{p}\Big)+3\theta^2\Gamma\Big(1+\frac{3}{p}\Big),$$

$$K_{33}=-\frac{3\mu^2\theta}{p^2}\Gamma'\Big(1+\frac{1}{p}\Big)-\frac{6\mu\theta^2}{p^2}\Gamma'\Big(1+\frac{2}{p}\Big)-\frac{3\theta^3}{p^2}\Gamma'\Big(1+\frac{3}{p}\Big),$$

where $\Gamma'(x)=\Gamma(x)\Psi(x)$ and $\Psi(x)$ is the psi-function. In this study, to calculate $\Psi(x)$ we used the series expansion

$$\Psi(a)=-\mathbf{C}+(a-1)\sum_{k=0}^{\infty}\frac{1}{(k+1)(k+a)},$$

where $\mathbf{C}=0.57721566\ldots$

Elements of matrix **B** are:

$$B_{i1}=\frac{p}{\theta\sqrt{p_i}}\Big\{(\frac{x_{i-1}-\mu}{\theta})^{p-1}\exp\Big\{-(\frac{x_{i-1}-\mu}{\theta})^p\Big\}-$$

$$-(\frac{x_i-\mu}{\theta})^{p-1}\exp\Big\{-(\frac{x_i-\mu}{\theta})^p\Big\}\Big\},$$

$$B_{i2}=\frac{p}{\theta\sqrt{p_i}}\Big\{(\frac{x_{i-1}-\mu}{\theta})^{p}\exp\Big\{-(\frac{x_{i-1}-\mu}{\theta})^p\Big\}-$$

$$-(\frac{x_i-\mu}{\theta})^{p}\exp\Big\{-(\frac{x_i-\mu}{\theta})^p\Big\}\Big\},$$

$$B_{i3}=\frac{1}{\sqrt{p_i}}\Big\{-(\frac{x_{i-1}-\mu}{\theta})^{p}\ln(\frac{x_{i-1}-\mu}{\theta})\exp\Big\{-(\frac{x_{i-1}-\mu}{\theta})^p\Big\}+$$

$$+(\frac{x_i-\mu}{\theta})^{p}\ln(\frac{x_i-\mu}{\theta})\exp\Big\{-(\frac{x_i-\mu}{\theta})^p\Big\}\Big\},\ \ i=1,\ldots,r.$$

Elements of matrix **C** are:

$$C_{i1}=\frac{1}{\sqrt{p_i}}\Big\{\mu\exp\Big\{-(\frac{x_{i-1}-\mu}{\theta})^p\Big\}-\mu\exp\Big\{-(\frac{x_i-\mu}{\theta})^p\Big\}-$$

$$-p_i m_1 + \theta\gamma\Big[(1+\frac{1}{p}),(\frac{x_i-\mu}{\theta})^p\Big] - \theta\gamma\Big[(1+\frac{1}{p}),(\frac{x_{i-1}-\mu}{\theta})^p\Big]\Big\},$$

$$C_{i2} = \frac{1}{\sqrt{p_i}}\Big\{\mu^2\exp\Big\{-(\frac{x_{i-1}-\mu}{\theta})^p\Big\} - \mu^2\exp\Big\{-(\frac{x_i-\mu}{\theta})^p\Big\} - p_i m_2 +$$

$$+2\mu\theta\gamma\Big[(1+\frac{1}{p}),(\frac{x_i-\mu}{\theta})^p\Big] - 2\mu\theta\gamma\Big[(1+\frac{1}{p}),(\frac{x_{i-1}-\mu}{\theta})^p\Big]+$$

$$+\theta^2\gamma\Big[(1+\frac{2}{p}),(\frac{x_i-\mu}{\theta})^p\Big] - \theta^2\gamma\Big[(1+\frac{2}{p}),(\frac{x_{i-1}-\mu}{\theta})^p\Big]\Big\},$$

$$C_{i3} = \frac{1}{\sqrt{p_i}}\Big\{\mu^3\exp\Big\{-(\frac{x_{i-1}-\mu}{\theta})^p\Big\} - \mu^3\exp\Big\{-(\frac{x_i-\mu}{\theta})^p\Big\} - p_i m_3 +$$

$$+3\mu^2\theta\gamma\Big[(1+\frac{1}{p}),(\frac{x_i-\mu}{\theta})^p\Big] - 3\mu^2\theta\gamma\Big[(1+\frac{1}{p}),(\frac{x_{i-1}-\mu}{\theta})^p\Big]+$$

$$+3\mu\theta^2\gamma\Big[(1+\frac{2}{p}),(\frac{x_i-\mu}{\theta})^p\Big] - 3\mu\theta^2\gamma\Big[(1+\frac{2}{p}),(\frac{x_{i-1}-\mu}{\theta})^p\Big]+$$

$$+\theta^3\gamma\Big[(1+\frac{3}{p}),(\frac{x_i-\mu}{\theta})^p\Big] - \theta^3\gamma\Big[(1+\frac{3}{p}),(\frac{x_{i-1}-\mu}{\theta})^p\Big]\Big\}, \quad i=1,\ldots,r,$$

where

$$m_i = \sum_{l=0}^{i}\begin{pmatrix} i \\ l \end{pmatrix}\theta^{i-l}\mu^l\Gamma\Big(1+\frac{i-l}{p}\Big), i=1,2,3,$$

and $\gamma(a,x) = \int\limits_0^x t^{a-1}e^{-t}dt$ is the incomplete gamma-function. To calculate the $\gamma(a,x)$ we used the following series expansion:

$$\gamma(a,x) = x^a\sum_{n=0}^{\infty}\frac{(-1)^n x^n}{n!(a+n)}.$$

Elements of matrix $\mathbf{V}$ are:

$$V_{ij} = m_{i+j}(\boldsymbol{\theta}) - m_i(\boldsymbol{\theta})m_j(\boldsymbol{\theta}) = \sum_{n=0}^{i+j}\begin{pmatrix} i+j \\ n \end{pmatrix}\theta^{i+j-n}\mu^n\Gamma\Big(1+\frac{i+j-n}{p}\Big)-$$

$$-\sum_{n=0}^{i}\begin{pmatrix} i \\ n \end{pmatrix}\theta^{i-n}\mu^n\Gamma\Big(1+\frac{i-n}{p}\Big)\times\sum_{n=0}^{j}\begin{pmatrix} j \\ n \end{pmatrix}\theta^{j-n}\mu^n\Gamma\Big(1+\frac{j-n}{p}\Big), i,j=1,2,3.$$

13.8. Bibliography

[BAG 02] Bagdonavicius V., Nikulin M., *Accelerated Life Models*, Boca Raton: Chapman and Hall, 2002.

[CHE 54] Chernoff H., Lehmann E., “The use of maximum likelihood estimates in χ^2 tests for goodness-of-fit”, *Annals of Mathematical Statistics*, vol. 25, p. 579–586, 1954.

[DRO 88] Drost F., *Asymptotics for Generalized Chi-square Goodness-of-fit Tests*, Amsterdam, Center for Mathematics and Computer Science, CWI Tracts, V. 48, 1988.

[DZH 92] Dzhaparidze K., Nikulin M., “On evaluation of statistics of chi-square type tests”, *Problems of the Theory of Probability Distributions, St. Petersburg: Nauka*, vol. 12, p. 59–90, 1992.

[FIS 24] Fisher R., “The condition under which χ^2 measures the discrepancy between observation and hypothesis”, *Journal of the Royal Statistical Society*, vol. 87, p. 442–450, 1924.

[GRE 96] Greenwood P., Nikulin M., *A Guide to Chi-squared Testing*, New York: John Wiley and Sons, 1996.

[HSU 76] Hsuan T., Robson D., “The χ^2 goodness-of-fit tests with moment type estimators”, *Commun. Statist. Theory and Methods*, vol. A5, p. 1509-1519, 1976.

[LEM 01] Lemeshko B., Postovalov S., Chimitova E., “On the distribution and power of Nikulin’s chi-squared test”, *Industrial Laboratory (in Russian)*, vol. 67, p. 52-58, 2001.

[LOC 94] Lockhart R., Stephens M., “Estimation and tests of fit for three-parameter Weibull distribution”, *Journal of the Royal Statistical Society*, vol. B56, p. 491-500, 1994.

[MIR 01] Mirvaliev M., “An investigation of generalized chi-squared type statistics”, *Doctoral thesis, Academy of Science of the Republic of Uzbekistan, Tashkent*, 2001.

[MOO 75] Moore D., Spruill M., “Unified large-sample theory of general chi-squared statistics for tests of fit”, *Annals of Statistics*, vol. 3, p. 599-616, 1975.

[MOO 77] Moore D., “Generalized inverses, Wald’s method and the construction of chi-squared tests of fit”, *Journal of the American Statistical Association*, vol. 72, p. 131-137, 1977.

[MUD 95] Mudholkar G., Srivastava D., Freimer M., “The exponentiated Weibull family: a reanalysis of the bus-motor-failure data”, *Technometrics*, vol. 37, p. 436-445, 1995.

[MUD 96] Mudholkar G., Srivastava D., Kollia G., “A generalization of the Weibull distribution with application to the analysis of survival data”, *Journal of the American Statistical Association*, vol. 91, p. 1575-1583, 1996.

[NIK 73a] Nikulin M., “Chi-square test for continuous distributions”, *Theory of Probability and its Applications*, vol. 18, p. 638-639, 1973.

[NIK 73b] Nikulin M., “Chi-square test for continuous distributions with shift and scale parameters”, *Theory of Probability and its Applications*, vol. 18, p. 559-568, 1973.

[NIK 73c] NIKULIN M., "Chi-square test for normality", in: *Proc. of the International Vilnius Conference on Prob. Theory and Math. Statist.*, vol. 2, p. 119-122, 1973.

[RAO 74] RAO K., ROBSON D., "A chi-squared statistic for goodness-of-fit tests within the exponential family", *Communications in Statistics*, vol. 3, p. 1139-1153, 1974.

[SMI 87] SMITH R., NAYLOR J., "A comparison of maximum likelihood and Bayesian estimators for the three-parameter Weibull distribution", *Applied Statistics*, vol. 36, p. 358-369, 1987.

[VAA 98] VAN DER VAART A., *Asymptotic Statistics. Cambridge Series in Statistics and Probabilistic Mathematics*, Cambridge: Cambridge University Press, 1998.

[VOI 04] VOINOV V., PYA N., "On the power of modified chi-squared goodness-of-fit tests for the family of logistic and normal distributions", in: *Proc. of the 7th Iranian Statist. Conf., Allameh-Tabatabaie University*, p. 385-403, 2004.

[VOI 06] VOINOV V., "On optimality of the Rao-Robson-Nikulin test", *Industrial Laboratory (in Russian)*, vol. 72, p. 65-70, 2006.

[VOI 08] VOINOV V., ALLOYAROVA R., PYA N., "Recent achievements in modified chi-squared goodness-of-fit testing", *Statistical Models and Methods for Biomedical and Technical Systems,* Eds.: Vonta F., Nikulin M., Limnios N., Huber C., Brikhäuser, forthcoming, 2008.

[WAL 43] WALD A., "Tests of statistical hypothesis when the number of observations is large", *Transactions of the A.M.S.*, vol. 54, p. 426-482, 1943.

Chapter 14

Accelerated Life Testing when the Hazard Rate Function has Cup Shape

14.1. Introduction

Most papers on Accelerated Life Testing (ALT) consider parametric models with exponential (constant hazard rates), Weibull (monotone hazard rates), lognormal and loglogistic ($\cap$-shaped hazard rates) lifetime distributions under constant-in-time stresses ([BNI 02], [LAW 03], [ME 98],[NEL 04], [VIE 88], [SIN 71]). These models can not be used if the hazard rates have a $\cup$-shaped form which is typical.

We consider the application of the generalized Weibull (GW) distribution [BNI 02] as the baseline distribution model in the accelerated failure time (AFT) model ([BAG 78], [NEL 04]) which is the most applied model in ALT. The GW distribution can be used in all of the most common situations when the hazard rates under constant in time stresses are monotone, $\cap$-shaped and even $\cup$-shaped (see Figure 14.1).

Estimation procedure for the AFT-GW model is given, and comparative analysis with AFT-Weibull and AFT-lognormal models is performed.

Chapter written by Vilijandas BAGDONAVIČIUS, Luc CLERJAUD and Mikhail NIKULIN.

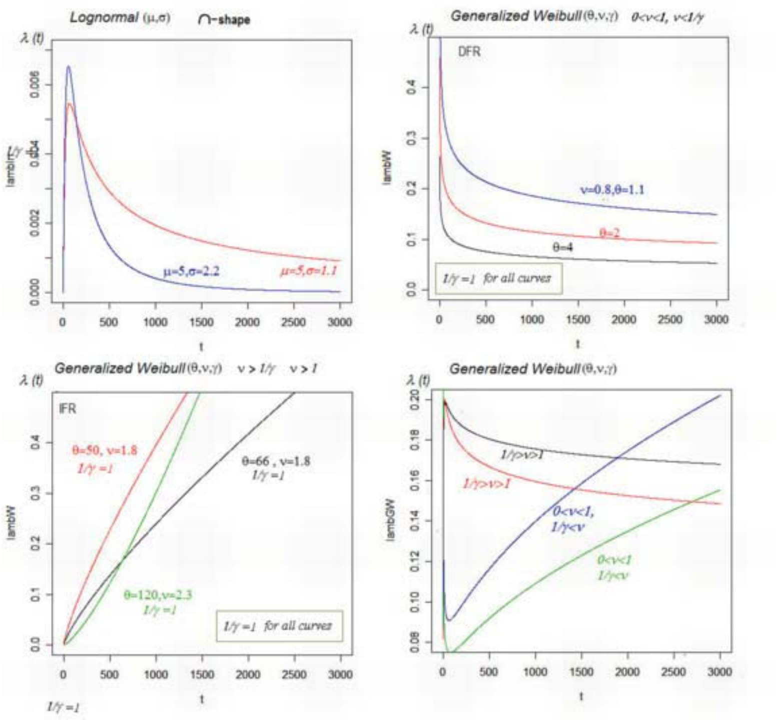


Figure 14.1. *Different shapes of hazard rate functions*

14.2. Estimation in the AFT-GW model

14.2.1. *AFT model*

Let $x(\cdot) = (x_0(\cdot), \ldots, x_m(\cdot))^T$, $x^{(0)}(t) \equiv 1$, be possibly time-varying m-dimensional stress, where x_i are one-dimensional stresses, $i = 1, \ldots, m$, and $x^{(0)}$ be the usual (standard, normal) constant-over-time m-dimensional stress. Let us suppose that the lifetime $T_{x(\cdot)}$ under stress $x(\cdot)$ is a positive random variable, and let $S_{x(\cdot)}$ be the survival function of $T_{x(\cdot)}$:

$$S_{x(\cdot)}(t) = P\left\{T_{x(\cdot)} > t\right\}. \tag{14.1}$$

The AFT model is true on a set of stresses E if there exists a function $r : E \longrightarrow R_+$ such that

$$S_{x(\cdot)}(t) = S_0\left(\int_0^t r[x(\tau)]d\tau\right), \ \forall\, x(\cdot) \in E, \tag{14.2}$$

where S_0 is the baseline survival function. It could be possible that $S_0 = S_{x^{(0)}}$.

If $x(\tau) = x \in E_1$ is constant in time, then

$$S_x(t) = S_0(r(x)t), \ \forall \, x \in E_1. \tag{14.3}$$

In parametric models S_0 belongs to a *parametric family* and $r(x)$ is parameterized. If the stress x is one-dimensional (scalar) and constant, r is often parameterized as

$$r(x) = e^{-\beta_0 - \beta_1 \varphi(x)} \tag{14.4}$$

where φ is a given function of x. The most applied are three models:

– *log-linear model* , where $r(x) = e^{-\beta_0 - \beta_1 x}, \ \varphi(x) = x,$

– *power-rule model*, where $r(x) = e^{-\beta_0 - \beta_1 \ln x}, \ \varphi(x) = \ln x, x > 0,$

– *Arrhenius model*, where $r(x) = e^{-\beta_0 - \beta_1 / x}, \ \varphi(x) = 1/x.$

14.2.2. *AFT-Weibull, AFT-lognormal and AFT-GW models*

Consider three families of survival distributions as models for the baseline survival function S_0:

$$S_0(t) = e^{-t^{\nu}}, \ \nu > 0, \ (\textit{Weibull}) \tag{14.5}$$

$$S_0(t) = 1 - \Phi\left(\nu \ln t\right), \ \nu > 0, \ \ (\textit{lognormal}) \tag{14.6}$$

where $\Phi(u)$ is the distribution function of the standard normal distribution,

$$S_0(t) = exp\left\{1 - (1 + t^{\nu})^{\gamma}\right\}, \ \nu, \gamma > 0, \ \ (\textit{GW}). \tag{14.7}$$

If $0 < \nu < 1$ and $1/\gamma < \nu$, the curve of the hazard function has a $\cup$-shape. If $1/\gamma > \nu > 1$, the curve of the hazard function has a $\cap$-shape. If $0 < \nu < 1$ and $\nu < 1/\gamma$ then the hazard function decreases from $+\infty$ to 0 (DFR). If $\nu > 1$ and $\nu > 1/\gamma$ then then hazard function increases to $+\infty$ (IFR) (see Figure 14.1).

Taking the Weibull, lognormal and GW distribution as the baseline distribution in the AFT model (14.1.2) (or (14.1.3)), we obtain the AFT-Weibull, AFT-lognormal and AFT-GW model, respectively.

14.2.3. *Plans of ALT experiments*

First plan of experiments: units are tested under the accelerated with respect to the *usual stress* $x^{(0)}$ constant over time stresses $x^{(1)}, \dots x^{(k)}$. n_i units are tested under stress $x^{(i)} > x^{(0)}$, $(i = 1, \dots, k)$. Stress x is accelerated with respect to stress y if $S_x(t) < S_y(t)$ for all $t > 0$.

Second plan of experiments: n units are tested under the step-stress:

$$x(\tau) = \begin{cases} x^{(1)}, \ 0 \le \tau < t_1, \\ x^{(2)}, \ t_1 \le \tau < t_2, \\ \dots, \dots \\ x^{(k)}, \ t_{k-1} \le \tau < t_k, \end{cases} \tag{14.8}$$

where $x^{(j)}$ are constant stresses, $t_0 = 0$, $t_k = +\infty$. If the AFT model holds on a set E_k of step-stresses, then the survival function under any stress $x(\cdot) \in E_k$ of form (14.8) can be written as follows: for any $t \in [t_{i-1}, t_i)$ $(i = 1, 2, \dots, k)$,

$$S_{x(\cdot)}(t) = S_{x^{(i)}}\{t - t_{i-1} + \frac{1}{r(x^{(i)})} \sum_{j=1}^{i-1} r(x^{(j)})(t_j - t_{j-1})\}. \tag{14.9}$$

If $r(x)$ is parametrized as $r(x) = e^{-\beta^T x}$, then for any $t \in [t_{i-1}, t_i)$,

$$S_{x(\cdot)}(t) = S_0\{\mathbf{1}\{i > 1\} \sum_{j=1}^{i-1} e^{-\beta^T x^{(j)}} (t_j - t_{j-1}) + e^{-\beta^T x^{(i)}} (t - t_{i-1})\}, \tag{14.10}$$

where $x^{(j)}$ may be $\varphi(x^{(j)})$ for one-dimensional stress.

14.2.4. *Parameter estimation: AFT-GW model*

Let us consider the first plan of experiments when the units are tested under accelerated constant stresses: n_i units are tested under accelerated stress $x^{(i)}$ $(i = 1, \dots, k)$. Let t_i be the maximal experiment duration for the ith group. Let $\beta = (\beta_0, \dots, \beta_m)^T$ be the regression parameter. Let the lifetime of the jth unit from the ith group be denoted by T_{ij}. Set $X_{ij} = T_{ij} \wedge t_i$ and $\delta_{ij} = \mathbf{1}\{T_{ij} < t_i\}$. The likelihood function is:

$$L(\beta, \nu, \gamma) = \prod_{i=1}^{k} \prod_{j=1}^{n_i} \left\{ \nu\gamma e^{-\nu\beta^T x^{(i)}} X_{ij}^{\nu-1} \left(1 + \left(e^{-\beta^T x^{(i)}} X_{ij}\right)^{\nu}\right)^{\gamma-1} \right\}^{\delta_{ij}} \times \exp\left\{1 - \left(1 + \left(e^{-\beta^T x^{(i)}} X_{ij}\right)^{\nu}\right)^{\gamma}\right\} \tag{14.11}$$

where $x^{(i)} = (x_{i0}, \dots, x_{im})$, $x_{i0} = 1$.

The expressions of the score functions $U_l(\beta, \nu, \gamma)$ and the matrix $I(\beta, \nu, \gamma) = (I_{ls}(\beta, \nu, \gamma)))_{(m+3)\times(m+3)}$ of minus second partial derivatives of $\ln L$ are given in section 14.6.

Then, the estimator of the survival function under usual stress $x^{(0)}$ is

$$\hat{S}_{x^{(0)}}(t) = exp\left\{1 - \left(1 + \left(e^{-\hat{\beta}^T x^{(0)}} t\right)^{\hat{\nu}}\right)^{\hat{\gamma}}\right\}. \qquad (14.12)$$

The $(1-\alpha)$ approximate confidence limit for $S_{x^{(0)}}(t)$ is

$$\left(1 + \frac{1 - \hat{S}_{x^{(0)}}(t)}{\hat{S}_{x^{(0)}}(t)} \exp\left\{\pm \hat{\sigma}_{Q_{x^{(0)}}} w_{1-\alpha/2}\right\}\right), \qquad (14.13)$$

where

$$\hat{\sigma}_{Q_{x^{(0)}}} = \frac{1}{\left(1 - \hat{S}_{x^{(0)}}(t)\right)^2} \sum_{l=0}^{m+2} \sum_{s=0}^{m+2} a_l\left(t, \hat{\beta}, \hat{\nu}, \hat{\gamma}\right) I^{ls}\left(\hat{\beta}, \hat{\nu}, \hat{\gamma}\right) a_s^T\left(t, \hat{\beta}, \hat{\nu}, \hat{\gamma}\right)$$

$$a_l\left(t, \hat{\beta}, \hat{\nu}, \hat{\gamma}\right) = -\hat{\nu} x^{(l)} a_{m+1}\left(\hat{\beta}, \hat{\nu}, \hat{\gamma}\right) / \left(\ln t - \hat{\beta}^T x^{(0)}\right), l = 0, \dots, m$$

$$a_{m+1}\left(t, \hat{\beta}, \hat{\nu}, \hat{\gamma}\right) = -\hat{\gamma}\left(e^{-\hat{\beta}^T x^{(0)}} t\right)^{\hat{\nu}} \left(1 + \left(e^{-\hat{\beta}^T x^{(0)}} t\right)^{\hat{\nu}}\right)^{\hat{\gamma}-1} \left(\ln t - \hat{\beta}^T x^{(0)}\right),$$

$$a_{m+2}\left(t, \hat{\beta}, \hat{\nu}, \hat{\gamma}\right) = -\left(1 - \ln \hat{S}_{x^{(0)}}(t)\right) \ln\left(1 + \left(e^{-\hat{\beta}^T x^{(0)}} t\right)^{\hat{\nu}}\right),$$

$w_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of the standard normal distribution and I^{ls} are the elements of the matrix $I^{-1}(\hat{\beta}, \hat{\nu}, \hat{\gamma})$.

14.3. Properties of estimators: simulation results for the AFT-GW model

Suppose that the AFT-GW model with power-rule parametrization is considered. The data were simulated using the following values of the parameters: $\gamma = 2.5$, $\nu = 0.8$, $\beta_0 = 6$ and $\beta_1 = -0.78$. In such a case the hazard rate function has a $\cup$-shape.

Let us consider the first plan of experiments without censoring and suppose that the units are tested under accelerated constant stresses $x^{(1)} = 2 < x^{(2)} = 4 < x^{(3)} = 7 < x^{(4)} = 13$. The usual stress level is $x^{(0)} = 1.5 < x^{(1)}$.

n_i units are tested under accelerated stress x_i. Three sample sizes were considered: $n_i = 150, 230$ and 350 $(i = 1, \dots, 4)$.

Set $\theta = (\beta_0, \beta_1, \nu, \gamma)^T$. In the case of the Weibull distribution, the component γ is absent.

The AFT-GW data were simulated $M = 1000$ times. Denote by

$$\hat{\theta}^{(i)} = (\hat{\beta_0}^{(i)}, \hat{\beta_1}^{(i)}, \hat{\nu}^{(i)}, \hat{\gamma}^{(i)})^T \quad \text{and} \quad \hat{S}^{(i)}_{x^{(0)}}(t)$$

the estimators of the parameter θ and the survival function $S_{x^{(0)}}(t)$ from the ith simulated sample, and set

$$\hat{\theta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\nu}, \hat{\gamma})^T = \frac{1}{M}\sum_{i=1}^{M} \hat{\theta}^{(i)}, \quad \hat{S}_{x^{(0)}}(t) = \frac{1}{M}\sum_{i=1}^{M} \hat{S}^{(i)}_{x^{(0)}}(t).$$

For the AFT-GW model, the values of the mean simulated maximum likelihood estimators values $\hat{\beta}_0, \hat{\beta}_1, \hat{\nu}, \hat{\gamma}$ are given in Tables 14.1–14.3 (sample sizes $n_i = 150, 230$ and 350). The simulation results confirm the well-known fact that in regular models the maximum likelihood estimators are consistent. The slowest convergence rate is in the case of the estimator of the parameter γ, which determines the difference between the Weibull and GW distributions. Using the same AFT-GW data, the values of the mean simulated maximum likelihood estimators values for the AFT-Weibull and the AFT-lognormal models are also given in Tables 14.1–14.3.

Distribution	**Weibull**	**Generalized Weibull**	**Log-normal**
$\hat{\beta}_0$	4.52	6.21	3.89
$\hat{\beta}_1$	-0.781	-0.781	-0.781
$\hat{\nu}$	0.956	0.803	0.695
$\hat{\gamma}$	None	3.41	None

Table 14.1. *Mean simulated values of parameter estimators,* $n_i = 150$

Distribution	**Weibull**	**Generalized Weibull**	**Log-normal**
$\hat{\beta}_0$	4.51	6.11	3.88
$\hat{\beta}_1$	-0.780	-0.780	-0.778
$\hat{\nu}$	0.956	0.803	0.693
$\hat{\gamma}$	None	3.01	None

Table 14.2. *Mean simulated values of parameter estimators,* $n_i = 230$

Distribution	**Weibull**	**Generalized Weibull**	**Log-normal**
$\hat{\beta}_0$	4.52	6.07	3.88
$\hat{\beta}_1$	-0.781	-0.780	-0.776
$\hat{\nu}$	0.956	0.803	0.693
$\hat{\gamma}$	None	2.78	None

Table 14.3. *Mean simulated values of parameter estimators,* $n_i = 350$

For samples of sizes $n_i = 350$ we calculated the mean relative bias $RB\left(t_j\right)$ and the relative mean square error $RMSE(t_j)$ of the estimators of $S_{x^{(0)}}(t_j)$ at several points t_j when AFT-GW, AFT-Weibull and AFT-lognormal models were used (Tables 14.4 and 14.5).

The relative bias of the mean simulated value $\hat{S}_{x^{(0)}}(t)$ from the real value of survival function $S_{x^{(0)}}(t)$ is:

$$RB\left(t\right) = (\hat{S}_{x^{(0)}}(t) - S_{x^{(0)}}(t))/S_x\left(t\right).$$

The relative mean square error of simulated values $\hat{S}^{(i)}_{x^{(0)}}(t)$:

$$RMSE(t) = \frac{1}{M}\sum_{i=1}^{M}\left[\left(\hat{S}^{(i)}_{x^{(0)}}(t) - S_{x^{(0)}}(t)\right)/S_{x^{(0)}}(t_j, \theta)\right]^2.$$

t_j	$S_{x^{(0)}}(t_j)$	$RB\left(t_j\right)$ **(Weibull)**	$RB\left(t_j\right)$ **(GW)**	$RB\left(t_j\right)$ **(Lognormal)**
15	0.781	0.0073	0.00093	-0.073
40	0.556	-0.0261	0.00125	-0.162
120	0.182	-0.0460	0.00233	0.090
200	0.052	0.125	0.00647	1.230
280	0.012	0.601	0.0210	5.129

Table 14.4. *Relative bias* $RB\left(t_j\right)$

t_j	$RMSE(t_j)$ **(Weibull)**	$RMSE(t_j)$ **(Generalized Weibull)**	$RMSE(t_j)$ **(Lognormal)**
15	0.00033	0.00023	0.006
40	0.00184	0.00102	0.02779
120	0.00935	0.00900	0.015099
200	0.04264	0.03183	1.5601
280	0.4706	0.0913	26.8

Table 14.5. *Relative mean square errors* $RMSE(t_j)$

In case of the AFT-GW model, the relative bias and the relative mean square errors are smallest for all moments t_j.

We took one sample with $n_i = 230$ to calculate confidence intervals for the values of the survival function $S_{x^{(0)}}(t)$. Let θ_W and θ_{LN} be the vectors of parameters for

Weibull and lognormal distribution. The estimators of the parameters are:

$$\hat{\theta} = (6.37,\ -0.766,\ 0.809,\ 3.29)^T \quad \text{(AFT-GW model)},$$

$$\hat{\theta}_W = (4.50,;\ -0.760,\ 0.990)^T \quad \text{(AFT-Weibull model)},$$

$$\hat{\theta}_{LN} = (3.85,\ -0.735,\ 0.732)^T \quad \text{(AFT-lognormal model)}.$$

The 99% confidence interval is narrowest in the case of GW distribution. For some values of t the theoretic curve of $S_{x^{(0)}}$ is next to the two 99% confidence limits or outside of these limits when Weibull and lognormal models are used, whereas in the case of GW model the estimated curve is very near to the theoretic curve, which is between the confidence limits.

The estimators of the survival function under usual stress with two 99% confidence limits are given in Figure 14.2. The estimators of the hazard rate function are also given. In the case of GW distribution, the hazard rate function is closest to the theoretic curve.

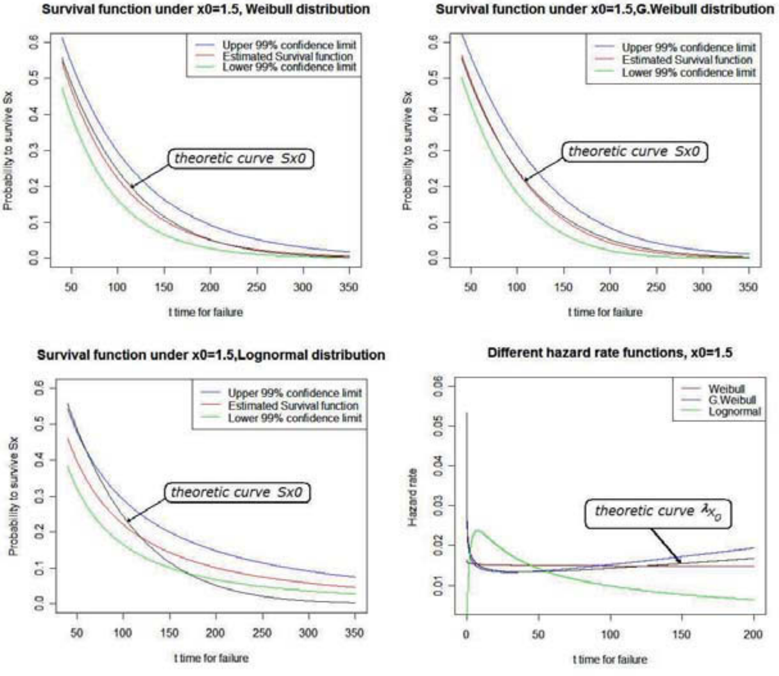


Figure 14.2. *Point and interval estimators of survival and hazard rate functions under usual stress $x^{(0)}$*

If we use the likelihood ratio test

$$Z = -2\left(\ln LL_W\left(\hat{\theta}_W\right) - \ln LL_{GW}\left(\hat{\theta}\right)\right)$$

for this simulated sample, we obtain $Z = 13.228$. P-value is very small, so the AFT-Weibull model is rejected against the AFT-GW model. For 1,200 simulations, there are 1,064 rejections of the AFT-Weibull model. The power of the test is 0.8867.

Note that in the case of IFR distributions the AFT-Weibull is usually sufficient and the AFT-GW model is not necessary: the data were simulated $M = 600$ times using the following values of the parameters: $\gamma = 1.5$, $\nu = 4.5$, $\beta_0 = 5$ and $\beta_1 = -0.5$. In this case, the baseline hazard function is increasing. The values of $\hat{\beta}_0, \hat{\beta}_1, \hat{\nu}$ and $\hat{\gamma}$ are given in Table 14.6.

Distributions	**Weibull**	**Generalized Weibull**
$\hat{\beta}_0$	4.875	5.031
$\hat{\beta}_1$	-0.499	-0.499
$\hat{\nu}$	4.991	4.517
$\hat{\gamma}$	None	1.839

Table 14.6. *AFT model, mean simulated values of parameter estimators,* $n_i = 150$

We calculated the mean relative bias $RB\left(t_j\right)$ and the relative mean square error $RMSE(t_j)$ of the estimators of $S_{x^{(0)}}\left(t_j\right)$ when AFT-GW and AFT-Weibull models were used (Tables 14.7 and 14.8). They are similar for both models.

t_j	**Weibull**	**Generalized Weibull**
50	0.00518	-0.000254
80	0.004373	0.000431
110	-0.03339	0.000747
140	0.235354	0.001957

Table 14.7. *Relative bias* $RB\left(t_j\right)$

14.4. Some remarks on the second plan of experiments

In the case of step-stresses and the AFT-GW (and other) models, the survival distribution does not belong to the same class as the baseline distribution S_0 because of transformation $\int_0^t r(x(s))ds$.

t_j	**Weibull**	**Generalized Weibull**
50	0.000048	0.00004
80	0.000967	0.00085
110	0.01341	0.014474
140	0.29829	0.24301

Table 14.8. *Relative mean square errors* $RMSE(t_j)$

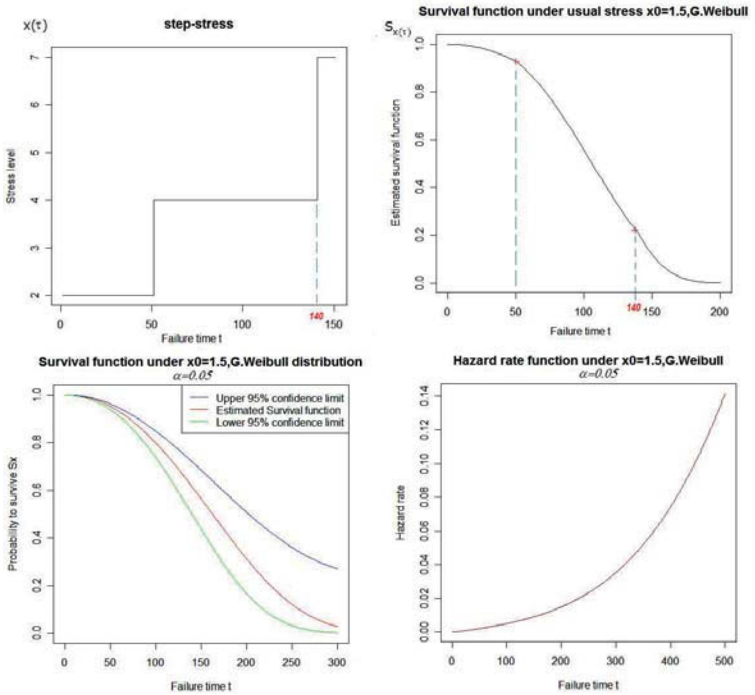


Figure 14.3. *Survival functions with step-stress, IFR hazard rate function*

Let us consider the AFT-GW model with three value step-stress (14.8) (see Figure 14.3, first graph), defined by $x^{(1)} = 2$, $x^{(2)} = 4$ and $x^{(3)} = 7$. The power-rule parametrization was used.

We simulated one sample of size n =2,000 using the AFT-GW model with parameters $\gamma = 1.5$, $\nu = 4.5$, $\beta_0 = 5$ and $\beta_1 = -0.5$. In this case the baseline hazard function is increasing.

In the case of the second plan of experiments, the estimation procedure is similar to the case of constant over time stresses. The likelihood function is:

$$L(\beta,\nu,\gamma) = \prod_{i=1}^{n} \left\{ \nu\gamma e^{-\nu\beta^T x^{(i)}(X_i)} (f_i(X_i,\beta,\gamma))^{\nu-1} \left(1 + (f_i(X_i,\beta,\gamma))^{\nu}\right)^{\gamma-1} \right\}^{\delta_i}$$
$$\times \exp\left\{1 - \left(1 + (f_i(X_i,\beta,\gamma))^{\nu}\right)^{\gamma}\right\}, \quad (14.14)$$

where X_i is the failure time of the ith unit and

$$f_i(t,\beta,\gamma) = \int_0^t e^{-\beta^T x^{(i)}(u)} du.$$

The maximum of the log-likelihood is -10022.24. In the second graph (Figure 14.3) we can see the estimated survival function under step-stress. The two 95% confidence limits of the survival function are wider than the case of constant stress, for the same size of sample (Figure 14.3, third graph). The estimator of the hazard rate is given (Figure 14.3, fourth graph).

14.5. Conclusion

By comparing AFT-Weibull, AFT-normal and AF-GW models we see that in the case of $\cup$-shaped hazard rate function, the AFT-GW model is significantly more suitable for estimation of the survival function under usual stress from ALT experiments.

14.6. Appendix

Consider the likelihood function (14.11). By derivations on the log-likelihood $\ln L(\beta,\nu,\gamma)$, we get:

$$U_l(\beta,\nu,\gamma) = \frac{\partial \ln L(\beta,\nu,\gamma)}{\partial \beta_l} = \quad (14.15)$$

$$\nu \sum_{i=1}^{k} x_{il} \sum_{j=1}^{n_i} \left(\gamma\omega_{ij}(\beta,\nu,\gamma) - \delta_{ij} u_{ij}(\beta,\nu,\gamma)\right), \; l \; = \; 0 \; , \; \ldots, m;$$

$$U_{m+1}(\beta,\nu,\gamma) = \frac{\partial \ln L(\beta,\nu,\gamma)}{\partial \nu} = \quad (14.16)$$

$$\frac{D}{\nu} - \frac{1}{\nu} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(\gamma\omega_{ij}(\beta,\nu,\gamma) - \delta_{ij} u_{ij}(\beta,\nu,\gamma)\right) \ln h_{ij}(\beta,\nu,\gamma)$$

$$U_{m+2}(\beta,\nu,\gamma)=\frac{\partial \ln L(\beta,\nu,\gamma)}{\partial\gamma}= \tag{14.17}$$

$$\frac{D}{\nu}-\frac{1}{\nu}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\left(1+h_{ij}\left(\beta,\nu\right)\right)^{\gamma}-\delta_{ij}\right)\ln\left(1+h_{ij}(\beta,\nu)\right),$$

where $D=\sum_{i=1}^{k}\sum_{j=1}^{n_i}\delta_{ij}$ and

$$h_{ij}(\beta,\nu)=\left(e^{-\beta^T x^{(i)}}X_{ij}\right)^{\nu}\ ,\omega_{ij}\left(\beta,\nu,\gamma\right)=\left(1+h_{ij}(\beta,\nu)\right)^{\gamma-1},$$

$$u_{ij}(\beta,\nu,\gamma)=1+(\gamma-1)\tfrac{h_{ij}(\beta,\nu)}{1+h_{ij}(\beta,\nu)}.$$

Let us consider the matrix $I(\beta,\nu,\gamma)=\left(I_{ls}\left(\beta,\nu,\gamma\right)\right)\right)_{(m+3)\times(m+3)}$ with the following elements:

$$I_{ls}\left(\beta,\nu,\gamma\right)=-\frac{\partial^2\ln L\left(\beta,\nu,\gamma\right)}{\partial\beta_l\partial\beta_s}=\nu\sum_{i=1}^{k}x_{il}x_{is}\times$$

$$\sum_{j=1}^{n_i}\left\{\frac{\nu\omega_{ij}\left(\beta,\nu,\gamma\right)\left(1+h_{ij}\left(\beta,\nu\right)\right)-\delta_{ij}\left(u_{ij}\left(\beta,\nu,\gamma\right)-1\right)}{1+h_{ij}\left(\beta,\nu\right)}\right\},(l,s=0,\ldots,m)$$

$$I_{l,m+1}\left(\beta,\nu,\gamma\right)=-\frac{\partial^2\ln L\left(\beta,\nu,\gamma\right)}{\partial\beta_l\partial\nu}=-\frac{1}{\nu}U_l\left(\beta,\nu,\gamma\right)-$$

$$\sum_{i=1}^{k}x_{il}\sum_{j=1}^{n_i}\left\{\frac{\ln h_{ij}\left(\beta,\nu\right)}{1+h_{ij}\left(\beta,\nu\right)}\right\}\times$$

$$\left\{\gamma\omega_{ij}\left(\beta,\nu,\gamma\right)\left(1+\gamma h_{ij}\left(\beta,\nu\right)\right)-\delta_{ij}\left(u_{ij}\left(\beta,\nu,\gamma\right)-1\right)\right\},\ (l=0,\ldots,m)$$

$$I_{l,m+2}\left(\beta,\nu,\gamma\right)=-\frac{\partial^2\ln L\left(\beta,\nu,\gamma\right)}{\partial\beta_l\partial\gamma}=-\frac{1}{\nu}\sum_{i=1}^{k}x_{il}\times$$

$$\sum_{j=1}^{n_i}\left\{\omega_{ij}\left(\beta,\nu,\gamma\right)\left(1+\gamma\ln\left(1+h_{ij}\left(\beta,\nu\right)\right)\right)-\delta_{ij}\frac{h_{ij}\left(\beta,\nu\right)}{1+h_{ij}\left(\beta,\nu\right)}\right\},(l=0,\ldots,m),$$

$$I_{m+1,m+1}\left(\beta,\nu,\gamma\right)=-\frac{\partial^2\ln L\left(\beta,\nu,\gamma\right)}{\partial\nu^2}=\frac{1}{\nu}U_{m+1}\left(\beta,\nu,\gamma\right)+$$

$$\frac{1}{\nu^2}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left\{\frac{\ln^2 h_{ij}\left(\beta,\nu\right)}{1+h_{ij}\left(\beta,\nu\right)}\right\}\left\{\gamma\omega_{ij}\left(\beta,\nu,\gamma\right)\left(1+\gamma h_{ij}\left(\beta,\nu\right)\right)-\delta_{ij}\left(u_{ij}\left(\beta,\nu,\gamma\right)-1\right)\right\}$$

$$+\frac{1}{\nu^2}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\ln h_{ij}\left(\beta,\nu,\gamma\right)\left\{\gamma\omega_{ij}\left(\beta,\nu,\gamma\right)-\delta_{ij}u_{ij}\left(\beta,\nu,\gamma\right)\right\},$$

$$I_{m+1,m+2}\left(\beta,\nu,\gamma\right)=-\frac{\partial^2\ln L\left(\beta,\nu,\gamma\right)}{\partial\nu\partial\gamma}=\frac{1}{\nu}\times\sum_{i=1}^{k}\sum_{j=1}^{n_i}\frac{h_{ij}\left(\beta,\nu\right)\ln h_{ij}\left(\beta,\nu\right)}{1+h_{ij}\left(\beta,\nu\right)}$$

$$\times \left\{ \left(1 + h_{ij}\left(\beta, \nu\right)\right)^{\gamma} + \gamma \left(1 + h_{ij}\left(\beta, \nu\right)\right)^{\gamma} + \ln\left(1 + h_{ij}\left(\beta, \nu\right)\right) - \delta_{ij} \right\}$$

$$I_{m+2,m+2}\left(\beta, \nu, \gamma\right) = -\frac{\partial^2 \ln L\left(\beta, \nu, \gamma\right)}{\partial \gamma^2} =$$

$$= \frac{D}{\gamma^2} + \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(1 + h_{ij}\left(\beta, \nu, \gamma\right)\right)^{\gamma} \ln^2\left(1 + h_{ij}\left(\beta, \nu, \gamma\right)\right).$$

14.7. Bibliography

[BAG 78] Bagdonavičius, V. (1978) Testing the hyphothesis of the additive accumulation of damages. *Probab. Theory and its Appl.*, vol. 23 (2), 403–408.

[NEL 04] Nelson, W. (2004). *Accelerated Testing, Statistical Models, Test Plans, and Data Analysis*. John Wiley & Sons, New Jersey.

[BNI 02] Bagdonavicius, V. and Nikulin, M. (2002) *Accelerated Life Models*. Chapman & Hall/CRC, Boca Raton.

[LAW 03] Lawless, J.F. (2003) *Statistical Models and Methods for Lifetime Data*. New York: John Wiley and Sons.

[ME 98] Meeker W.Q, Escobar.L. (1998). *Statistical Method for Reliability Data*. John Wiley, New York.

[SIN 71] Singpurwalla, N.D. (1971) “Inference from accelerated life tests when observations are obtained from censored samples”, *Technometrics*, vol. 13, 161–170.

[VIE 88] Viertl, R. (1988). *Statistical Methods in Accelerated Life Testing*, Vandenhoeck & Ruprecht, Gottingen.

This page intentionally left blank

Chapter 15

Point Processes in Software Reliability

15.1. Introduction

Most systems are now driven by software applications. It is well-recognized that assessing reliability of software products is a major issue in reliability engineering, particularly in terms of cost. An impressive list of system crashes due to software and their cost is reported in [PHA 00]. Thus, a huge amount of human and financial resources are devoted to the production of dependable systems which is, at the present time, the prerogative of the software engineering community. However, a central question is: does the production meet the objectives? To answer this, mathematical methods for reliability assessment of software systems must be proposed. The failure process of a software differs from that of hardware in the following specific aspects:

1) We are primarily concerned with design faults. A *fault* (or bug) refers to the manifestation in the code of a mistake made by the programmer or designer with respect to the specification of the software. Activation of a fault by an input value leads to an incorrect output. Detection of such an event corresponds to an occurrence of a *software failure*.

2) The software does not wear out. Its reliability is intended to be improved by corrective maintenance.

3) The failure process is highly dependent on the operational profile of the software. Activation of a fault depends on the input data and this fault is activated each time the corresponding input data is selected (as long as the fault is not removed).

Chapter written by James LEDOUX.

The prevalent approach in software reliability modeling is the *black-box* approach, in which only the interactions of the software with the environment are considered. More than 50 such models are listed in [PHA 00]. No real effort has been made to include all models in a single mathematical framework, which should make a comparative analysis of models easier. Decision makers in industry face the lack of clarity in the research activities for failure prediction, and have focused their resources on methods for producing dependable softwares. These methods are essentially based on fault prevention and removing. However, this does not respond to the initial question, that is, assessing if the reliability objectives are met. To the best of my knowledge, Gaudoin and Soler made the first significant attempt to clarify the relationships between various models [GAU 90]. Their idea was to use the self-excited point processes as their mathematical framework. With the help of results in Snyder's book [SNY 91], they classified a large part of the standard models and provided their main statistical properties. This was rediscovered by Singpurwalla and his co-authors in [CHE 97, SIN 99]. In fact, this framework can be included in the martingale approach of point processes which allows us to consider a much wider class of models. This was sketched in [LED 03a] and is formed in the forthcoming book [GAU 07].

In a second approach, called the *white-box* approach, information on the structure of the software is incorporated in the models. Software applications increase in size and complexity, so that a satisfactory control of their behavior cannot be expected from a single black-box view of the system. Extensive numerical experiments have shown that black-box models are suitable for capturing a trend in the dependability metrics, but not for obtaining precise estimates. This is a standard need during the preliminary phases of the life of a software. However, specifically when the software is in operation, it is intended that an architecture-based approach allows analyzing the sensitivity of the dependability of the system with respect to those of its components. To save money, such analysis must be done prior to their effective implementation in the system. Moreover, the usage profile of a component of a system during the operating and testing phases differs significantly. Therefore, the failure process of the system depends greatly on the structure of the execution process. However, it is a matter of fact that the contributions to the architecture-based reliability modeling are rare. This can be explained from certain open issues: what is the architecture of a software? What kind of data and how much data can be collected?

Input values may be considered as arriving to the software randomly. So although software failure is not generated stochastically, it is detected in such a manner. Therefore, this justifies the use of stochastic models of the underlying random process that governs the software failures. Specifically, the failure process is modeled by a point process. In section 15.2, the basic concepts of reliability of repairable systems are recalled. Then, the basic set-up for black-box models is introduced in section 15.3. Section 15.4 is devoted to white-box modeling. After introducing a benchmark model given by Littlewood, a general Markovian model is presented. Calibration of its parameters and Poisson approximation in the case of reliability growth are discussed.

15.2. Basic concepts for repairable systems

The *time-to-failure* for a non-repairable system is a random variable X taking values in $[0, +\infty]$. The probability distribution of X is assumed to have a density f over $\mathbb{R}_+ = [0, +\infty[$. The reliability function is defined by

$$\forall x \geq 0, \quad R(x) = 1 - \int_0^x f(u)\, du.$$

Note that there exist reliability models in which the system may be considered as failure-free. In this case,

$$\mathbb{P}\{X = +\infty\} = 1 - \int_0^{+\infty} f(x)\, dx > 0.$$

The *failure rate* or *hazard rate* is the main reliability metric of a non-repairable system:

$$\forall x \geq 0, \quad h(x) = \frac{f(x)}{R(x)} = \frac{f(x)}{1 - \int_0^x f(u)\, du}. \tag{15.1}$$

It characterizes the probability distribution of X via the exponentiation formula. Function $R(\cdot)$ is not very useful in the context of repairable systems. Thus, we adhere to Ascher-Feingold's definition of the reliability function for repairable systems as a generalization of the residual survival function of non-repairable systems [ASC 84]. Indeed, at time t, the probability that the system is operational up to a specified time, should depend on all observed events prior to t (failures, corrections or external factors). All these past events are gathered in the *history* (or filtration) at time t, $\mathcal{H}_t$. Thus, the reliability of a repairable system is a probability conditional to $\mathcal{H}_t$.

Definition 15.1 *At time t, the reliability of the system at horizon $\tau \geq 0$, is the probability that the system is operational on the interval $]t, t + \tau]$ given the history at time t. That is, the* ***reliability function*** *at t is the function R_t defined by*

$$\forall \tau \geq 0, \quad R_t(\tau) = \mathbb{P}\{N_{t+\tau} - N_t = 0 \mid \mathcal{H}_t\} = \mathbb{P}\{T_{N_t+1} - t > \tau \mid \mathcal{H}_t\}.$$

Let us comment on the concept of *reliability growth*, i.e. the fact that dependability is improving with time. This is interpreted as follows: when the system is observed at time t_2, posterior to t_1, the reliability at t_2 is greater than that evaluated at t_1 whatever the horizon τ: for $t_1 < t_2$,

$$\forall \tau \geq 0, \quad R_{t_1}(\tau) \leq R_{t_2}(\tau).$$

Nevertheless, function R_t is non-increasing as a function of τ.

A software is observed from instant $t_0 = 0$. Failures happen at instants $\{t_n\}_{n\geq 1}$. After each of them, the software is either corrected or not, and then rebooted. The standard models assume the following: 1) the delays to recover a safe state are not taken into account. 2) The failure instants can be identified at the moments of request (i.e. the time to execution is neglected). 3) Correction is immediate. Now, each t_n is the observed value of a random variable T_n and $\{T_n\}_{n\geq 1}$ is a *point process*, that is, a sequence of non-negative random variables, all defined on the probability space $(\Omega, \mathcal{A}, \mathbb{P})$, such that:

1) $T_0 = 0$ and $T_1 > 0$,

2) $T_n < T_{n+1}$ if $T_n < +\infty$, and $T_{n+1} = T_n$ when $T_n = +\infty$.

Note that

$$\lim_{n\to+\infty} T_n = +\infty. \tag{15.2}$$

This definition uncovers the important case in software reliability of models with a finite number N of failure times $T_1, \dots, T_N$ ($T_n = +\infty$ for $n \geq N+1$). A point process is specified by any of the two following stochastic processes:

1) the sequence of inter-failure times, $\{X_n\}_{n\geq 1}$, where $X_{n+1} = T_{n+1} - T_n$ if $T_n < +\infty$, and $+\infty$ when $T_n = +\infty$;

2) the *counting process* of failures, $\{N_t\}_{t\geq 0}$, with $N_0 = 0$ and $N_t = \sum_{n=1}^{+\infty} \mathbb{1}_{\{T_n \leq t\}}$. Note that condition (15.2) is equivalent to $\mathbb{P}\{N_t < +\infty\} = 1$ for every t.

Let $\{N_t\}_{t\geq 0}$ be adapted to some history $\boldsymbol{\mathcal{H}} = \{\mathcal{H}_t\}_{t\geq 0}$. It is assumed to be *integrable* to simplify the exposition:

$$\forall t \geq 0, \quad \mathbb{E}[N_t] < +\infty. \tag{15.3}$$

We introduce the concept of stochastic intensity which generalizes the concept of the failure rate of non-repairable systems.

Definition 15.2 *Let $\{\lambda_t\}_{t\geq 0}$ be a non-negative and $\boldsymbol{\mathcal{H}}$-predictable process. Set*

$$\forall t \geq 0, \quad M_t = N_t - \int_0^t \lambda_s \, ds.$$

*$\{\lambda_t\}_{t\geq 0}$ is called the **$\boldsymbol{\mathcal{H}}$-intensity** of $\{N_t\}_{t\geq 0}$ if $\{M_t\}_{t\geq 0}$ is an $\boldsymbol{\mathcal{H}}$-martingale.*

Recall that any $\boldsymbol{\mathcal{H}}$-adapted process with left-continuous paths is $\boldsymbol{\mathcal{H}}$-predictable. The interest in an $\boldsymbol{\mathcal{H}}$-predictable intensity is that it is unique up to a set of $\mathbb{P} \otimes l$-measure zero [BRE 81], where l is the Lebesgue measure on $\mathbb{R}_+$. The process $\{\Lambda_t\}_{t\geq 0}$ defined by

$$\forall t \geq 0, \quad \Lambda_t = \int_0^t \lambda_s \, ds$$

is $\boldsymbol{\mathcal{H}}$-adapted with non-decreasing and continuous paths. This is called the $\boldsymbol{\mathcal{H}}$ *compensator* of the counting process $\{N_t\}_{t\geq 0}$ and provides the decomposition

$$\forall t \geq 0, \quad N_t = \Lambda_t + M_t. \tag{15.4}$$

For any integrable counting process, the $\boldsymbol{\mathcal{H}}$-compensator always exists, from Doob-Meyer's theorem. Moreover, decomposition (15.4) is unique (up to a set of $\mathbb{P}$ probability zero). All counting processes in software reliability have a compensator that is absolutely continuous with respect to the Lebesgue measure and therefore have an intensity [BRE 81, Chapter 2, Theorem 13].

Remark 15.1 When the integrability condition (15.3) is not satisfied, the property of $\boldsymbol{\mathcal{H}}$-martingale only holds for the family of processes $\{N_{t\wedge T_n} - \Lambda_{t\wedge T_n}\}_{t\geq 0}$, $n \geq 1$. The concept of martingale is replaced by that of local martingale.

15.3. Self-exciting point processes and black-box models

The *internal history* of $\{N_t\}_{t\geq 0}$, $\boldsymbol{\mathcal{H}}^N = \{\mathcal{H}_t^N\}_{t\geq 0}$ with $\mathcal{H}_t^N = \sigma(N_s, s \leq t)$, is central in the point process theory. It can be verified that

$$\mathcal{H}_t^N = \sigma\left(N_t, T_{N_t}, \ldots, T_1\right) \quad \text{and} \quad \mathcal{H}_{T_n}^N = \sigma(T_n, \ldots, T_1).$$

The following simplified version of a result due to Jacod [JAC 75] gives an explicit form to the $\boldsymbol{\mathcal{H}}^N$-intensity.

Theorem 15.1 *Let $\{N_t\}_{t\geq 0}$ be an integrable counting process. Assume the conditional distribution of the inter-failure time X_{n+1} given $\mathcal{H}_{T_n}^N$ has a density $f_{X_{n+1}|\mathcal{H}_{T_n}^N}$ with $(s,\omega) \mapsto f_{X_{n+1}|\mathcal{H}_{T_n}^N}(s,\omega)$ be $\mathcal{B}(\mathbb{R}_+) \otimes \mathcal{H}_{T_n}^N$-measurable. Its hazard rate, denoted by $h_{X_{n+1}|\mathcal{H}_{T_n}^N}$, has form (15.1). Then the process $\{\widehat{\lambda}_t\}_{t\geq 0}$ defined by*

$$\widehat{\lambda}_t = \sum_{n\geq 0} h_{X_{n+1}|\mathcal{H}_{T_n}^N}(t - T_n)\mathbb{1}_{\{T_n < t \leq T_{n+1}\}}$$

is the $\boldsymbol{\mathcal{H}}^N$-intensity of $\{N_t\}_{t\geq 0}$.

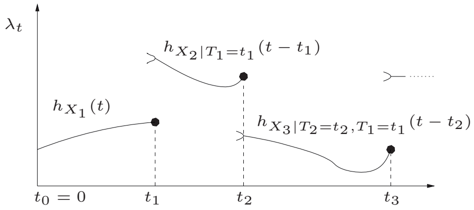


Figure 15.1. *A path of* $\widehat{\lambda}_t$

The process $\{\widehat{\lambda}_t\}_{t\geq 0}$ is a concatenation of the hazard rate functions of conditional distributions of inter-failure times. A typical path of $\{\widehat{\lambda}_t\}_{t\geq 0}$ is illustrated in Figure 15.1.

It is clear that any failure model for which the future of the failure process only depends on its past is specified by the sequence of hazard rates $\{h_{X_{n+1}|\mathcal{H}^N_{T_n}}(\cdot)\}_{n\geq 0}$. For instance, the intensity of the celebrated Jelinski-Moranda's model [JEL 72] has the form

$$\widehat{\lambda}_t = \Phi(N - N_{t-}).$$

Its paths are non-increasing and piecewise constant with jumps of size $\Phi > 0$. Here, N must be thought of as the initial number of faults in the software and Φ as a (uniform) factor of the quality of the debugging action after each failure instant. Thus, $\widehat{\lambda}_t$ is proportional to the number of residual faults at time t and Φ can also be interpreted as the manifestation rate of any of the N faults.

The concept of "concatenated failure rate" was used in [CHE 97, SIN 99] in connection with "self-exciting point processes" whose definition requires the existence of the limit [SNY 91]

$$\lim_{\Delta t\downarrow 0}\frac{1}{\Delta t}\,\mathbb{P}\big\{N_{(t+\Delta t)-} - N_{t-} = 1 \mid \mathcal{H}^N_{t-}\big\}, \tag{15.5}$$

as well as the "conditional orderliness" condition to hold :

$$\lim_{\Delta t\downarrow 0}\frac{1}{\Delta t}\,\mathbb{P}\big\{N_{(t+\Delta t)-} - N_{t-} \geq 1 \mid \mathcal{H}^N_{t-}\big\} = \lim_{\Delta t\downarrow 0}\frac{1}{\Delta t}\,\mathbb{P}\big\{N_{(t+\Delta t)-} - N_{t-} = 1 \mid \mathcal{H}^N_{t-}\big\}.$$

It can be checked that limit (15.5) is equal to the intensity in definition 15.2:

$$\widehat{\lambda}_t = \lim_{\Delta t\downarrow 0}\frac{1}{\Delta t}\,\mathbb{P}\big\{N_{(t+\Delta t)-} - N_{t-} = 1 \mid \mathcal{H}^N_{t-}\big\}.$$

Thus, the intensity expresses the "likelihood" of a system to experience a failure in $[t, t+\Delta t[$, given the number of failures as well as the failure instants just before t.

A natural way to introduce growth in reliability is to choose an intensity with non-increasing paths.

The reliability metrics have an explicit form for self-exciting point processes. ROCOF, if it exists, is defined by $\mathbb{E}[N_t] = \int_0^t \mathrm{ROCOF}_s\, ds$. Then, we have:

$$\mathbb{E}[N_t] = \mathbb{E}[\Lambda_t], \quad \mathrm{ROCOF}_t = \mathbb{E}\big[\widehat{\lambda}_t\big],$$

$$R_t(\tau) = \exp\left(- \int_t^{t+\tau} \widehat{\lambda}_s\, ds \right) = \exp\big(- (\Lambda_{t+\tau} - \Lambda_t)\big),$$

$$\mathrm{MTTF}_t = \mathbb{E}[T_{N_t+1} - t \mid N_t, T_{N_t}, \ldots, T_1] = \int_0^{+\infty} R_t(\tau)\, d\tau.$$

Potentially, any "standard" point process with an $\boldsymbol{\mathcal{H}}$-stochastic intensity has an intensity with respect to $\boldsymbol{\mathcal{H}}^N$ given by the formula [BRE 81]

$$\widehat{\lambda}_t = \mathbb{E}[\lambda_t \mid \mathcal{H}_{t-}^N].$$

However, a very large collection of software reliability models reported in [MUS 87, XIE 91, LYU 96] have been directly developed using specific forms of the sequence of failure rates involved in Theorem 15.1. The choice of failure rates is motivated by considering the software system as a single unity. Some of these models are reported in Table 15.1. Note that in general, λ_t only depends on the past of the point process through t, N_t, T_{N_t}. See [LYU 96, ALM 98, PHA 00] for numerical illustrations of black-box models.

Moranda (M)	Generalized order statistics (GOS)	Al Mutairi-Chen-Singpurwalla (ACS)
$\widehat{\lambda}_t = \lambda \phi^{N_{t-}}$	$\widehat{\lambda}_t = (N - N_{t-})\phi(t)$	$\widehat{\lambda}_t = \frac{1}{\frac{t-T_{N_{t-}}}{k} + \frac{T_{N_{t-}}}{bN_{t-}}}$
$\phi \in]0,1[, \lambda \in \mathbb{R}_+$	$N \in \mathbb{N}^*$, $\phi(\cdot)$ decreasing positive function	$k \in \mathbb{N}^*, b \in \mathbb{R}_+$

Duane (D)	Goel-Okumoto (GO)	Yamada-Obha-Osaki (YOO)
$\widehat{\lambda}_t = \alpha\beta t^{\beta-1}$	$\widehat{\lambda}_t = m\phi \exp(-\phi\, t)$	$\widehat{\lambda}_t = m\phi^2 t \exp(-\phi t)$
$\Lambda_t = \alpha t^{\beta}$	$\Lambda_t = m\big(1 - \exp(-\phi t)\big)$	$\Lambda_t = m\big(1 - (1+\phi t)\exp(-\phi t)\big)$
$\alpha \in \mathbb{R}_+, \beta \in]0,1[$	$m, \phi \in \mathbb{R}_+$	

Table 15.1. *Some black-box models*

Let us briefly comment on the models above (see [GAU 07] for details):

– All the models have a non-increasing intensity which converges to 0, so that the software will become perfect. GO and D illustrate the fact that different convergence rates to 0 of the intensity may be considered in order to take into account the different levels of growth in reliability. A few models have been proposed with a non-zero asymptotic intensity.

– D is a standard model in reliability. Note that a complete statistical analysis of this model is available.

– Model M has been introduced to relax two basic assumptions of model JM, a finite number of initial faults and an uniform factor of quality of any correction. Here, the factor decreases with the number of experienced failures. This reflects the fact that the faults detected and removed in the early phases of testing contribute more to reliability growth than those detected later. λ is the initial failure intensity.

– GOSs cover a wide class of proposed models. N is the initial number of faults in the software and $\widehat{\lambda}_t$ is assumed to be proportional to the number of residual faults at time t, so that $\phi(t)$ can be interpreted as the manifestation rate of each residual fault at time t. Note that if we consider a uniform ϕ, we retrieve the model JM.

– GO and YOO belong to the class of bounded NHPP, that is, their compensator converges to the finite value m. Thus, m is interpreted to be the expected number of initial faults. Note that such a property means that the total number of observed failures will be finite with probability 1. Specifically, it has a Poisson distribution with parameter m. Finally, let us mention that if the parameter N in a model GOS is assumed to have a Poisson distribution with parameter m, then the combination of the two random sources gives an NHPP model. For instance, GO is the NHPP version of JM.

– The intensity of YOO has an S-shaped form: it is first increasing and then decreasing. This is supported by the following experimental evidence: the tests are efficient only after a burn-in period. In others words, the ability of engineers to detect bugs in a software increases with time. Note that if $\phi(t) = \phi^2 t/(1 + \phi t)$, then $\widehat{\lambda}_t = \phi(t)(m - \Lambda_t)$ and $\phi(t)$ may be interpreted as a factor of the quality of the correction. In this last equality, the case $\phi(t) = \phi$ corresponds to the GO model.

– The ACS model is analyzed from a Bayesian point of view in [ALM 98]. It has very interesting properties. We only list some of them here:

1) The first property is that the mean time to the next failure, MTTF_t, is an increasing function of the time elapsed from the last failure instant $t - T_{N_t}$:

$$k > 1, \quad \mathrm{MTTF}_t = \frac{k}{(k-1)b} \frac{T_{N_t}}{N_t} + \frac{1}{k-1}(t - T_{N_t}).$$

This is supported by the subjective argument that the longer the elapsed time since the last failure, the longer the expected time to the next failure.

2) It can be verified that $h_{X_{n+1}|\mathcal{H}_{T_n}}(0) = b(n/T_n)$, which is proportional to the inverse of the mean of the n first inter-failures durations $T_n/n = (\sum_{i=1}^n X_i)/n$. The parameter b appears as a scaling parameter.

3) Let us consider that we have enhanced reliability going from the nst inter-failure time to the $(n+1)$th inter-failure time if and only if $h_{X_{n+1}|\mathcal{H}^N_{T_n}} \leq h_{X_n|\mathcal{H}^N_{T_{n-1}}}$ on $[0, \min(X_n, X_{n+1})[$. Then this last condition is equivalent to

$$X_n \geq \frac{1}{n-1} \sum_{i=1}^{n-1} X_i.$$

In other words, a growth in reliability implies that the next inter-failure time is greater than the average of all past inter-failure times. Moreover, it can be shown that

$$\frac{\mathbb{E}[X_{n+1}]}{\mathbb{E}[X_n]} \leq 1 \Longleftrightarrow b \geq \frac{k}{k-1},$$

that is, on average, we observe a growth in reliability in mean when $b \geq k/(k-1)$.

4) The jump of the intensity at T_n is upward provided that T_n/T_{n-1} satisfies certain conditions.

15.4. White-box models and Markovian arrival processes

The *white-box* (or architecture-based) point of view is an alternative approach in which the architecture of the system is explicitly taken into account. This is advocated for instance in [CHE 80, LIT 79]. The references in the recent reviews on the architecture-based approach [GOS 01, GOK 06] provide a representative sample of the models. We present the main features of Littlewood's model which are common to most works. Note that Cheung's model can be thought of as a discrete time counterpart of Littlewood's one. The following additional papers [SIE 88, KAÂ 92, OKA 04, LO 05, WAN 06] are concerned with discrete time Markov chain modeling. See [CHE 80, LED 99, GOK 06] for numerical illustrations of white-box models.

In the first step, Littlewood defines an execution model of the software. The basic entity is the standard software engineering concept of *module* as for instance in [CHE 80]. The software architecture is then represented by the *call graph* of the set $\mathscr{Y}$ of the modules. These modules interact by execution control transfers and, at each instant, control lies in one and only one of the modules, which is called the active one. Let us consider the continuous time stochastic process $\{Y_t\}_{t\geq 0}$ where Y_t indicates the active module at time t. $\{Y_t\}_{t\geq 0}$ is assumed to be a homogenous Markov process on the set $\mathscr{Y} = \{e_1, \ldots, e_m\}$ with generator Q. A sequential execution of modules is assumed, so that each state of the execution model corresponds to a module. However, this limitation can be relaxed using an appropriate redesign of the states in $\mathscr{Y}$ (e.g. [TRI 01, WAN 06]).

In the second step, a failure processes is defined. Failure may happen during a control transfer between two modules or during an execution period of any module.

When module e_i is active, failures are part of a Poisson process with parameter μ_i. When control is transferred from module e_i to module e_j, a failure may occur with probability $\mu_{i,j}$. Given a sequence of executed modules, all failure processes are assumed to be independent. Generation of failures is illustrated in Figure 15.2.

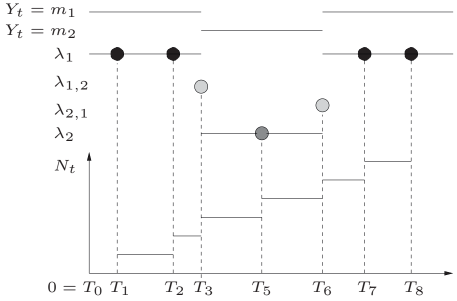


Figure 15.2. *Failure arrivals in Littlewood's model with two modules*

The execution and failure models are merged into a single model which can then be analyzed. Basically, we are interested in the counting process of failures $\{N_t\}_{t\geq 0}$. Let us make the following comments on Littlewood's model. Each failure does not affect the execution dynamics. The underlying rational for such an assumption is that the software experiences minor failures from the user point of view, so that the delay to recover a safe state is neglected. The reliability growth is not modeled. When the software is in operation, the reliability growth phenomenon is not very significant and Littlewood's model is thought of as a pessimistic model.

15.4.1. *A Markovian arrival model*

In this section, we describe a failure model that generalizes Littlewood's [LED 99]. The following assumptions are assumed to hold:

1) Two classes of failures in regards to their severity are introduced:

a) The first class is composed of failures that provoke an interruption of the delivered service. Such an event is called a *primary failure*. When the software experiences such a failure, the execution returns to a checkpoint or is assumed to be re-initialized from an input component. When component e_i is executed, a primary failure may occur according an exponential distribution with rate λ_i. A transfer of control from state e_i to e_j may also experience a primary failure with probability $\lambda_{i,j}$. After a failure occurs during the execution of component e_i or a transfer of control from e_i, component $e_j \in \mathscr{Y}$ is the first component entered with probability $p(i,j)$. Thus, the control flow may be redirected due to the use of error recovery procedures.

b) The second class gathers together failures that are not perceived by the user as a non-deliverance of service. In other words, their effects are assumed to be negligible. Such an event is called *secondary failure*. In this case, the model is exactly like Littlewood's model. A secondary and a primary failure may be simultaneously experienced, but only the primary is taken into account.

2) All failure processes are assumed to be independent given the sequence of executed components.

3) The delays to recover a safe state are assumed to be negligible.

The third assumption is supported by the fact that the breakdown periods are often much shorter than the inter-failure times in the operating phase of the software lifecycle. This assumption is relaxed in [LED 99]. The reliability growth is still not taken into account. In contrast to Littlewood's model, the failure may affect the execution dynamics of the system. Let us denote the entered component just after the nth failure instant T_n by $Y^*_{T_n}$. Then $\{Y^*_{T_n}\}_{n\geq 0}$ is a Markov chain from the assumptions and the occupation times of each component e_i is exponentially distributed with parameter $-Q(i,i)+\lambda_i+\mu_i$. We define the following jump Markov process $\{Y^*_t\}_{t\geq 0}$ as

$$Y^*_t = Y^*_{T_n} \qquad \text{when } T_n \leq t < T_{n+1}.$$

Now, it is easily seen that the bivariate process $\{(N_t, Y^*_t)\}_{t\geq 0}$, where $\{N_t\}_{t\geq 0}$ is the counting process of failures, is a jump Markov process with state space $\mathbb{N}\times\mathscr{Y}$. Its generator has the following form using the lexicographic order on $\mathbb{N}\times\mathscr{Y}$:

$$\begin{pmatrix} D_0 & D_1 & \mathbf{0} & \cdots \\ \mathbf{0} & D_0 & D_1 & \ddots \\ \vdots & \ddots & \ddots & \ddots \end{pmatrix}$$

with $m\times m$-matrices D_0, D_1 defined by : for $i,j=1,\ldots,m$,

$$D_0(i,j) = Q(i,j)(1-\lambda_{i,j})(1-\mu_{i,j})\ i\neq j, \quad D_0(i,i) = Q(i,i)-\lambda_i-\mu_i,$$

$$D_1(i,j) = Q(i,j)\big[1-\lambda_{i,j}\big]\mu_{i,j} + \big[\,\lambda_i + \sum_{k\neq i} Q(i,k)\lambda_{i,k}\,\big]\,p(i,j).$$

The generator of the Markov process $\{Y^*_t\}_{t\geq 0}$ is

$$Q^* = D_0 + D_1$$

Q^* being assumed to be irreducible so that $\mathbb{P}\{N_\infty = +\infty\} = 1$. Y^*_0, Y_0 have the same probability distribution that is denoted by the vector $\boldsymbol{\alpha} = (\alpha(i))_{i=1}^m$. It is interpreted as the usage profile of the various components. Littlewood's model corresponds to matrices D_0, D_1 with $\lambda_i = \lambda_{i,j} = 0$, so that $Q^* = Q$. The simplest model for which a failure affects the execution process is the PH-renewal process. Let us only consider primary failures in the modules, i.e. $\mu_i = 0$, $\lambda_{i,j} = \mu_{i,j} = 0$, $i, j = 1, \ldots, m$. After a failure has been experienced, the control is redirected to the component e_j with probability α_j, independently of where the failure occurred: $i, j = 1, \ldots, m$, $p(i, j) = \alpha_j$. We obtain with $\underline{\lambda} = (\lambda_i)_{i=1}^m$,

$$D_0 = Q - \mathrm{diag}(\lambda_i), \quad D_1 = \underline{\lambda}^\top \boldsymbol{\alpha} = -D_0 \mathbf{1}^\top \boldsymbol{\alpha}, \qquad Q^* = D_0 - D_0 \mathbf{1}^\top \boldsymbol{\alpha}.$$

As in the self-exciting context, explicit formulae for the various reliability metrics can be given. Let $\{X_n\}_{n \geq 0}$ be the inter-failure times. Let us denote the exponential matrix function $\exp(D_0\, t)$ by $f_0(t)$. Then the $\boldsymbol{\mathcal{H}}^N$-intensity $\{\widehat{\lambda}_t\}_{t \geq 0}$ of $\{N_t\}_{t \geq 0}$ is

$$\widehat{\lambda}_t = \frac{\boldsymbol{\alpha} f_0(X_1) D_1 \cdots f_0(X_{N_t}) D_1 f_0(t - T_{N_t}) D_1 \mathbf{1}^\top}{\boldsymbol{\alpha} f_0(X_1) D_1 \cdots f_0(X_{N_t}) D_1 f_0(t - T_{N_t}) \mathbf{1}^\top}$$

and the main reliability metrics have the following form

$$R_t(\tau) = \frac{\boldsymbol{\alpha} f_0(X_1) D_1 \cdots f_0(X_{N_t}) D_1 f_0(t - T_{N_t}) f_0(\tau) \mathbf{1}^\top}{\boldsymbol{\alpha} f_0(X_1) D_1 \cdots f_0(X_{N_t}) D_1 f_0(t - T_{N_t}) \mathbf{1}^\top},$$

$$\mathrm{MTTF}_t = \frac{\boldsymbol{\alpha} f_0(X_1) D_1 \cdots f_0(X_{N_t}) D_1 f_0(t - T_{N_t})\big) \big(- D_0\big)^{-1} \mathbf{1}^\top}{\boldsymbol{\alpha} f_0(X_1) D_1 \cdots f_0(X_{N_t}) D_1 f_0(t - T_{N_t}) \mathbf{1}^\top},$$

$$\mathbb{E}[N_t] = \int_0^t \boldsymbol{\alpha} \exp(Q^* u) du\, D_1 \mathbf{1}^\top, \qquad \mathrm{ROCOF}_t = \boldsymbol{\alpha} \exp(Q^* t)\, D_1 \mathbf{1}^\top.$$

The distribution function of random variable N_t also has an explicit form [LED 99]. The uniformization method can be used to calculate the exponential matrices [STE 94].

15.4.2. *Parameter estimation*

The multidimensional parameter $\theta = \{D_0(i, j), D_1(i, j),\ i, j = 1, \ldots, m\}$ must be estimated. The likelihood function for observations $t_1 < \cdots < t_k$ is

$$\theta \mapsto \mathcal{L}(\theta; t_1, \ldots, t_k) = \boldsymbol{\alpha} f_0(t_1) D_1 f_0(t_2 - t_1) D_1 \cdots f_0(t_k - t_{k-1}) D_1 \mathbf{1}^\top. \quad (15.6)$$

This function is highly non-linear, so that a standard numerical procedure fails to optimize it when m is large. Estimating θ can be thought of as an estimation problem in a missing-data context. Indeed, θ is associated with the Markov process $\{(N_t, Y_t^*)\}_{t \geq 0}$ and must be estimated from observations of its first component

$\{N_t\}_{t\geq 0}$. This suggests the use of the expectation-maximization (EM) methodology. Vector $\boldsymbol{\alpha}$ is assumed to be known. In the following, t denotes the previous time t_k.

Let us introduce the *complete data likelihood*, that is, the likelihood associated with the observation of the bivariate Markov process $\{(N_t, Y_t^*)\}_{t\geq 0}$ over $[0,t]$ (see [KLE 03]):

$$L_t(\theta; N, Y^*)$$
$$= \prod_{i=1}^{m} \alpha(i)^{\mathbf{1}_{\{Y_0^*=e_i\}}} \prod_{i=1}^{m} e^{D_0(i,i)\mathcal{O}_t^{(i)}} \prod_{i,j=1,j\neq i}^{m} D_0(i,j)^{\mathcal{L}_t^{0,ij}} \prod_{i,j=1}^{m} D_1(i,j)^{\mathcal{L}_t^{1,ij}}$$

where

$$\mathcal{L}_t^{1,ij} = \sum_{0<s\leq t} \Delta N_s \mathbf{1}_{\{Y_{s-}^*=e_i, Y_s^*=e_j\}} = \int_0^t \mathbf{1}_{\{Y_{s-}^*=e_i, Y_s^*=e_j\}}\, dN_s,$$

$$j \neq i,\ \mathcal{L}_t^{0,ij} = \sum_{0<s\leq t} (1-\Delta N_s)\mathbf{1}_{\{Y_{s-}^*=e_i, Y_s^*=e_j\}},\ \mathcal{O}_t^{(i)} = \int_0^t \mathbf{1}_{\{Y_{s-}^*=e_i\}}\, ds$$

and ΔN_s is the jump size $N_s - N_{s-}$ of the counting process at time s.

For $i \neq j$, $\mathcal{L}_t^{1,ij}$ is the number of failures with a transfer of control from e_i to e_j up to time t. $\mathcal{L}_t^{1,ii}$ is the number of observed failures at time t with no transition of $\{Y_t^*\}_{t\geq 0}$ from state e_i. For $i \neq j$, $\mathcal{L}_t^{0,ij}$ is the number of control transfers from e_i to e_j with no failure. The last statistics is the occupation time of state e_i by $\{Y_t^*\}_{t\geq 0}$ on $[0,t]$. Note that the complete data log-likelihood has the form

$$\log L_t(\theta; N, Y^*) = \sum_{i,j=1,j\neq i}^{m} \log D_0(i,j)\ \mathcal{L}_t^{0,ij} + \sum_{i,j=1}^{m} \log D_1(i,j)\mathcal{L}_t^{1,ij}$$
$$+ \sum_{i=1}^{m} D_0(i,i)\mathcal{O}_t^{(i)} + K$$

where K does not depend on θ.

Let $\mathbb{P}_\theta$ denote the probability model under θ. A sequence $\{\theta_l\}_{l\in\mathbb{N}}$ is obtained from the iteration of the following maximization procedure:

$$\theta_{l+1} = \underset{\theta^*}{\operatorname{argmax}}\, \mathbb{E}_{\theta_l}\big[\log L_t(\theta^*; N, Y^*) \mid \mathcal{H}_t^N\big].$$

Using the Lagrange multipliers method, $\theta_{l+1} = \{D_0^{(l+1)}(i,j), D_1^{(l+1)}(i,j),\ i,j = 1,\ldots,m\}$ is shown to be

$$D_1^{(l+1)}(i,j) = \frac{\mathbb{E}_{\theta_l}[\mathcal{L}_t^{1,ij} \mid \mathcal{H}_t^N]}{\mathbb{E}_{\theta_l}[\mathcal{O}_t^{(i)} \mid \mathcal{H}_t^N]}, \quad D_0^{(l+1)}(i,j) = \frac{\mathbb{E}_{\theta_l}[\mathcal{L}_t^{0,ij} \mid \mathcal{H}_t^N]}{\mathbb{E}_{\theta_l}[\mathcal{O}_t^{(i)} \mid \mathcal{H}_t^N]} \quad i \neq j. \quad (15.7)$$

The crucial fact is that the sequence $\{\theta_l\}_{l\in\mathbb{N}}$ makes the likelihood function (15.6) non-decreasing, i.e.

$$L(\theta_{l+1}; t_1, \ldots, t_k) \geq L(\theta_l; t_1, \ldots, t_k).$$

It can be shown that it is actually an equality if and only if $\theta_{l+1} = \theta_l$ under natural conditions [CAP 05]. Note that the zero coefficients of D_0, D_1 are preserved by the procedure.

The EM-algorithm consists of selecting an initial value θ_0 and iterates the calculation of (15.7) as long as some stopping criterion is not met. θ_0 may be evaluated from empirical methods based on data gathered in the early phases of the software lifecycle (validation phase, integration tests, etc.) [GOS 01]. Thus, the estimate θ_0 is not representative of the operating phase and must be considered as an *a priori* estimate of θ. In this context, the EM procedure must be thought of as a recalibration method of θ_0 in view of failure data observed in the operating phase of the software lifecycle. The conditional expectations in (15.7) may be calculated using the well-known forward-backward or Baum-Welch technique. The starting point is to exchange the order of integration in (15.7) to calculate the average in time, for $s \leq t$, of:

$$\mathbb{P}\{Y_s = e_i \mid \mathcal{H}_t^N\}, \qquad \mathbb{P}\{\Delta N_s = p, Y_{s-} = e_i, Y_s = e_j \mid \mathcal{H}_t^N\}\; p = 0, 1$$

recalling that $t = t_k$. These conditional probabilities are called *smoothers* and are calculated from a forward and a backward recursive formula [KLE 03, and the references therein].

For hidden Markov chains, an alternative approach based on a direct recursive calculation of the conditional expectations is proposed by Elliott [ELL 95]. The materials for implementing this approach to the Markovian model introduced here are given in [LED 07a]. In contrast to the Baum-Welch technique, only a single pass through the data is needed. Thus, the memory requirement is independent of the number of observations. Moreover, an online estimation of θ is allowed. The disadvantage of the filter-based method is that each conditional expectation requires one recursive formula to be implemented. A detailed discussion on the respective properties of the two approaches may be found in [CAP 05] for the discrete time case and in [LED 06] for the present context. Now, we turn back to the recursive equations derived in [LED 07a] using a change of probability measure. Indeed, there exists a probability measure $\mathbb{P}_0$ under which $\{N_t\}_{t\geq 0}$ is the counting process of a Poisson process with intensity 1 and $\{Y_t^*\}_{t\geq 0}$ is a Markov chain with generator $D_0 + D_1$. Let $\boldsymbol{e_i}$ be the ith vector of the canonical basis of $\mathbb{R}^m$ so that

$$\mathbb{1}_{\{Y_t^* = e_i\}} = Y_t^*(i).$$

Theorem 15.2 [LED 07a, Th 2] *Let L_t be the likelihood ratio over the interval $[0,t]$ associated with the counting process $\{N_t\}_{t\geq 0}$ of intensity $\lambda_t = Y^*_{s-} D_1 \mathbf{1}^\top$:*

$$L_t = \prod_{0<s\leq t} {\lambda_s}^{\Delta N_s} \exp\left(\int_0^t (1-\lambda_s)\, ds\right)$$

Set $\sigma(Z_t) = \mathbb{E}_0[Z_t L_t \mid \mathcal{H}^N_t]$ for any $\boldsymbol{\mathcal{H}}^{N,Y^*}$-adapted integrable process $\{Z_t\}_{t\geq 0}$. Let $\{n_t\}_{t\geq 0}$ be the process defined by $n_t = N_t - t$. We have for any $t \geq 0$,

$$\mathbb{E}[Y^*_0] + \int_0^t \sigma(Y^*_{s-}) Q\, ds + \int_0^t \sigma(Y^*_{s-})(D_1 - I)\, dn_s,$$

$$\begin{aligned}\sigma(\mathcal{O}^{(i)}_t Y^*_t) &= \int_0^t \left[\sigma(\mathcal{O}^{(i)}_{s-} Y^*_{s-}) Q + \sigma(Y^*_{s-})(i)\, \boldsymbol{e_i}\right] \mathrm{d}s \\ &\quad + \int_0^t \sigma(\mathcal{O}^{(i)}_{s-} Y^*_{s-})(D_1 - I)\, dn_s,\end{aligned}$$

$$\begin{aligned}\sigma(\mathcal{L}^{0,ij}_t Y^*_t) &= \int_0^t \left[\sigma(\mathcal{L}^{0,ij}_{s-} Y^*_{s-}) Q + D_0(i,j)\sigma(Y^*_{s-})(i)\, \boldsymbol{e_j}\right] ds \\ &\quad + \int_0^t \sigma(\mathcal{L}^{0,ij}_{s-} Y^*_{s-})(D_1 - I)\, dn_s,\end{aligned}$$

$$\sigma(\mathcal{L}^{1,ij}_t Y^*_t) = \int_0^t \left[\sigma(\mathcal{L}^{1,ij}_{s-} Y^*_{s-}) Q + D_1(i,j)\sigma(Y^*_{s-})(i)\, \boldsymbol{e_j}\right] ds$$

$$+ \int_0^t \left[\sigma(\mathcal{L}^{1,ij}_{s-} Y^*_{s-})(D_1 - I) + D_1(i,j)\sigma(Y^*_{s-})(i)\, \boldsymbol{e_j}\right] dn_s.$$

The conditional expectations of $\mathcal{O}^{(i)}_t, \mathcal{L}^{1,ij}_t, \mathcal{L}^{0,ij}_t$ are $\sigma(\mathcal{O}^{(i)}_t) = \sigma(\mathcal{O}^{(i)}_t Y^*_t)\mathbf{1}^\top$, $\sigma(\mathcal{L}^{1,ij}_t) = \sigma(\mathcal{L}^{1,ij}_t Y^*_t)\mathbf{1}^\top, \sigma(\mathcal{L}^{0,ij}_t) = \sigma(\mathcal{L}^{0,ij}_t Y^*_t)\mathbf{1}^\top$. Finally, the conditional expectations under the original probability $\mathbb{P}$, are obtained as follows:

$$\mathbb{E}[\mathcal{O}^{(i)}_t \mid \mathcal{H}^N_t] = \frac{\sigma(\mathcal{O}^{(i)}_t)}{\sigma(1)},\ \mathbb{E}[\mathcal{L}^{0,ij}_t \mid \mathcal{H}^N_t] = \frac{\sigma(\mathcal{L}^{0,ij}_t)}{\sigma(1)},\ \mathbb{E}[\mathcal{L}^{1,ij}_t \mid \mathcal{H}^N_t] = \frac{\sigma(\mathcal{L}^{1,ij}_t)}{\sigma(1)}.$$

This set of stochastic differential equations provides the following algorithm (see [LED 07a]).

RECURSIVE ALGORITHM.–

$f_0(x) = \exp(D_0 x), f_1(x) = D_1 f_0(x)$; $\Delta t_l = t_l - t_{l-1}$, $l = 1, \ldots, k$ with $t_0 = 0$.

for $l = 1, \ldots, k$
$\sigma(X_{t_l}) = \sigma(Y^*_{t_{l-1}})\, f_1(\Delta t_l)$
$\sigma(\mathcal{L}^{0,ij}_{t_l} Y^*_{t_l}) = \sigma(\mathcal{L}^{0,ij}_{t_{l-1}} Y^*_{t_{l-1}}) f_1(\Delta t_l) + \sigma(Y^*_{t_{l-1}}) \int_{t_{l-1}}^{t_l} f_0(s - t_{l-1}) e_i^\top D_0(i,j) e_j f_1(t_l - s)\, ds$
$\sigma(\mathcal{L}^{1,ij}_{t_l} Y^*_{t_l}) = f_1(\Delta t_l) \sigma(\mathcal{L}^{1,ij}_{t_{l-1}} Y^*_{t_{l-1}}) + e_j\, D_1(j,i)\, e_i^\top f_0(\Delta t_l) \sigma(Y^*_{t_{l-1}})$
$\sigma(\mathcal{O}^{(i)}_{t_l} Y^*_{t_l}) = f_1(\Delta t_l)\, \sigma(\mathcal{O}^{(i)}_{t_{l-1}} Y^*_{t_{l-1}}) + \int_{t_{l-1}}^{t_l} f_1(t_l - s) e_i e_i^\top f_0(s - t_{l-1})\, ds\, \sigma(Y^*_{t_{l-1}})$
Comment. The factor $\exp(\Delta t_l)$ is omitted in the equations above, because the estimates (15.7) of D_0, D_1 only require the knowledge of the filters up to a constant.

In general, the EM-algorithm converges to a local maxima of the likelihood function (15.6). Its main drawback is its slow convergence. See [WU 83, CAP 05] for details. Experiments show that the procedure is robust. In the specific context of the architecture-based software reliability modeling, the use of EM-algorithm was suggested in [LED 07a]. See [LED 06] for a detailed discussion in the context of discrete/continuous time software reliability modeling.

15.4.3. *Reliability growth*

The previous models do not take into account the expected reliability growth of a software. As seen in [LIT 75, LIT 79], we investigate the limit model when the failure parameters become much smaller than the parameters that drive exchanges of control. In this context, Littlewood claimed that $\{N_t\}_{t \geq 0}$ converges in distribution to a homogenous Poisson process with intensity

$$\lambda_{eq} = \sum_{i=1}^{m} \pi(i) \big[\mu_i + \sum_{j \neq i} Q(i,j) \mu_{i,j}\big], \tag{15.8}$$

where $(\pi(i))_{i=1,\ldots,m}$ is the invariant distribution of the irreducible generator Q. The next theorem states that a similar limit result holds for a version of the model of section 15.4.1 in which $\{Y_t\}_{t \geq 0}$ is assumed to be a non-homogenous Markov process with a family of measurable generators $\{Q(t)\}_{t \geq 0}$ satisfying

$$\sup_t \|Q(t)\|_1 = \sup_t \max_i \big(\sum_{j=1}^{m} |Q(t)(i,j)|\big) < \infty.$$

A direct way to introduce the problem is to multiply the failure parameters by a small parameter $\varepsilon > 0$. As a result, we obtain a Markovian model with matrices $\{D_0^{(\varepsilon)}(t), D_1^{(\varepsilon)}(t)\}_{t \geq 0}$ of the form

$$D_0^{(\varepsilon)}(t) = Q(t) + \varepsilon B_0(t) + \varepsilon^2 L_0(t) \quad \text{and} \quad D_1^{(\varepsilon)}(t) = \varepsilon B_1(t) + \varepsilon^2 L_1(t),$$

where $B_0(t), L_0(t), B_1(t)$ and $L_1(t)$ are matrices of uniformly bounded measurable functions such that

$$(B_0(t) + B_1(t))\mathbf{1}^\top = (L_0(t) + L_1(t))\mathbf{1}^\top = \mathbf{0}.$$

The Markov process $\{Y_t^*\}_{t\geq 0}$ has generator $\{Q^{*,\varepsilon}(t)\}_{t\geq 0}$ with

$$Q^{*,\varepsilon}(t) = D_0^{(\varepsilon)}(t) + D_1^{(\varepsilon)}(t) = Q(t) + \varepsilon R_0(t) + \varepsilon^2 R_1(t).$$

The main assumption is on the rate of convergence of $Q(t)$ to some irreducible generator Q with stationary distribution π:

$$\exists \beta > 1, \quad \lim_{t\to+\infty} (2t)^\beta \|Q(t) - Q\|_1 = 0. \tag{15.9}$$

Such a condition implies that $B_1(t)$ converges to a non-negative matrix B_1. Then, we investigate the asymptotic distribution of the counting process $\{N_t^{(\varepsilon)}\}_{t\geq 0}$ defined by

$$\forall t \geq 0, \quad N_t^{(\varepsilon)} = N_{t/\varepsilon}.$$

Let us introduce the distance in variation between the probability distributions $\mathcal{L}(N_{\mathbb{T}}^{(\varepsilon)})$ and $\mathcal{L}(\Pi_{\mathbb{T}})$ of random vectors $N_{\mathbb{T}}^{(\varepsilon)} = (N_{t_1}^{(\varepsilon)}, \ldots, N_{t_n}^{(\varepsilon)})$ and $P_{\mathbb{T}} = (P_{t_1}, \ldots, P_{t_n})$:

$$\mathrm{d_{TV}}\big(\mathcal{L}(N_{\mathbb{T}}^{(\varepsilon)}), \mathcal{L}(P_{\mathbb{T}})\big) = \sup_{B\subset\mathbb{N}^n} \big|\mathbb{P}\{N_{\mathbb{T}}^{(\varepsilon)} \in B\} - \mathbb{P}\{P_{\mathbb{T}} \in B\}\big|.$$

Theorem 15.3 [LED 07b, Th 4] *Let $\{P_t\}_{t\geq 0}$ be a Poisson process with intensity*

$$\lambda_{eq} = \pi B_1 \mathbf{1}^\top = \sum_{i=1}^m \pi(i)\big[\mu_i + \lambda_i + \sum_{j\neq i} Q(i,j)(\lambda_{i,j} + \mu_{i,j})\big]. \tag{15.10}$$

Under condition (15.9), for any $T > 0$ there exists a constant C_T such that

$$\mathrm{d_{TV}}\big(\mathcal{L}(N_{\mathbb{T}}^{(\varepsilon)}), \mathcal{L}(P_{\mathbb{T}})\big) \leq C_T\, \varepsilon.$$

For Littlewood's model, (15.10) reduces to (15.8). The result provides a rate at which the convergence takes place. An important fact is that the order of the convergence rate is optimal (see [LED 07b, Remark 5]).

Remark 15.2 In fact, the convergence in variation takes place in the Skorokhod space [LED 07b, Theorem 4].

Remark 15.3 In Littlewood's model, when $\{Y_t\}_{t\geq 0}$ is an irreducible semi-Markovian process, Littlewood claimed that $\{N_t\}_{t\geq 0}$ is still asymptotically a Poisson process [LIT 79]. This fact is justified in [LED 03b, Remark 2].

15.5. Bibliography

[ALM 98] AL-MUTAIRI D., CHEN Y., SINGPURWALLA N. D., "An adaptative concatenated failure rate model for software reliability", *J. Amer. Statist. Ass.*, vol. 93, num. 443, p. 1150–1163, 1998.

[ASC 84] ASCHER H., FEINGOLD H., *Repairable Systems Reliability: Modeling, Inference, Misconceptions and their Causes*, Marcel Dekker, 1984.

[BRE 81] BREMAUD P., *Point Processes and Queues*, Springer, 1981.

[CAP 05] CAPPÉ O., MOULINES E., RYDÉN T., *Inference in Hidden Markov Models*, Springer, 2005.

[CHE 80] CHEUNG R. C., "A user-oriented software reliability model", *IEEE Trans. Software Eng.*, vol. 6, p. 118–125, 1980.

[CHE 97] CHEN Y., SINGPURWALLA N. D., "Unification of software reliability models by self-exciting point processes", *Adv. in Appl. Probab.*, vol. 29, p. 337–352, 1997.

[ELL 95] ELLIOTT R. J., AGGOUN L., MOORE J. B., *Hidden Markov Models*, Springer, 1995.

[GAU 90] GAUDOIN O., "Outils statistiques pour l'évaluation de la fiabilité du logiciel", PhD thesis, Université Joseph Fourier – Grenoble I, 1990.

[GAU 07] GAUDOIN O., LEDOUX J., *Modélisation Aléatoire en Fiabilité des Logiciels*, Hermès, 2007.

[GOK 06] GOKHALE S. S., TRIVEDI K. S., "Analytical models for architecture-based software reliability prediction: a unification framework, ", *IEEE Trans. Reliab.*, vol. 55, p. 578–590, 2006.

[GOS 01] GOSEVA-POPSTOJANOVA K., TRIVEDI K. S., "Architecture-based approach to reliability assessment of software systems", *Perfor. Evaluation*, vol. 45, p. 179–204, 2001.

[JAC 75] JACOD J., "Multivariate point processes: predictable projection, Radon-Nikodym derivatives, representation of martingales", *Z. Wahrsch. Verw. Gebiete*, vol. 31, p. 235–253, 1975.

[JEL 72] JELINSKI Z., MORANDA P. B., "Software reliability research", FREIBERGER W., Ed., *Statistical computer performance evaluation*, p. 465–484, Academic Press, 1972.

[KAÂ 92] KAÂNICHE M., KANOUN K., "The discrete time hyperexponential model for software reliability growth evaluation", *Int. Symp. on Software Reliability*, p. 64–75, 1992.

[KLE 03] KLEMM A., LINDEMANN C., LOHMANN M., "Modeling IP traffic using the batch Markovian arrival process", *Performance Evaluation*, vol. 54, p. 149–173, 2003.

[LED 99] LEDOUX J., "Availability modeling of modular software", *IEEE Trans. Reliab.*, vol. 48, p. 159–168, 1999.

[LED 03a] LEDOUX J., "Software reliability modeling", in PHAM H., Eds., *Handbook of Reliability Engineering*, p. 213–234, Springer, 2003.

[LED 03b] LEDOUX J., "On the asymptotic analysis of Littlewood's reliability model for modular software", in LINDQVIST B. H., DOKSUM K. A., Eds., *Mathematical and Statistical Methods in Reliability*, p. 521–536, World Scientific, 2003.

[LED 06] LEDOUX J., "EM-algorithm in for software modeling", *Int. Conf. on Degrad., Damage, Fatigue and Accel. Life Models in Reliab. Testing (ALT'06)*, p. 278–286, 2006.

[LED 07a] LEDOUX J., "Filtering and the EM-algorithm for the Markovian arrival process", *Communications in Statistics – Theory Methods*, vol. 36 (14), 2007.

[LED 07b] LEDOUX J., "Strong convergence of a class of non-homogeneous Markovian arrival process to a Poisson process", Accepted for publication in *Statistics and Probability Letters*, 2007.

[LIT 75] LITTLEWOOD B., "A Reliability model for systems with Markov structure", *J. Roy. Statist. Soc. Ser. C*, vol. 24, p. 172–177, 1975.

[LIT 79] LITTLEWOOD B., "Software reliability model for modular program structure", *IEEE Trans. Reliab.*, vol. 28, p. 241–246, 1979.

[LO 05] LO J.-H., HUANG C.-Y., CHEN I.-Y., KUO S.-Y., LYU M. R., "Reliability assessment and sensitivity analysis of software reliability growth modeling based on software module structure", *The Journal of Systems and Software*, vol. 76, p. 3–13, 2005.

[LYU 96] LYU M. R., Ed., *Handbook of Software Reliability Engineering*, IEEE Computer Society Press and McGraw-Hill Book Company, 1996.

[MUS 87] MUSA J. D., IANNINO A., OKUMOTO K., *Software Reliability: Measurement, Prediction, Application*, McGraw-Hill International Editions, 1987.

[OKA 04] OKAMURA H., KUROKI S., DOHI T., OSAKI S., "A reliability growth model for modular software", *Electronics and Comm. in Japan*, vol. 87, p. 905–914, 2004.

[PHA 00] PHAM H., *Software Reliability*, Springer, 2000.

[SIE 88] SIEGRIST K., "Reliability of systems with Markov transfer of control, II.", *IEEE Trans. Software Eng.*, vol. 14, p. 1478–1480, 1988.

[SIN 99] SINGPURWALLA N. D., WILSON S., *Statistical Methods in Software Engineering*, Springer, 1999.

[SNY 91] SNYDER D. L., MILLER M. I., *Random Point Processes in Time and Space*, Springer, 1991.

[STE 94] STEWART W. J., *Introduction to the Numerical Solution of Markov chains*, Princeton University, 1994.

[TRI 01] TRIVEDI K. S., *Probability and Statistics with Reliability, Queuing, and Computer Science Applications*, John Wiley, 2nd Edition, 2001.

[WAN 06] WANG W.L., PAN, D., CHEN M.H., "Architecture-based software reliability modeling", *J. Systems and Software*, vol. 79, p. 132–146, 2006.

[WU 83] WU C. F. J., "On the convergence properties of the EM algorithm", *Ann. Statist.*, vol. 11, p. 95–103, 1983.

[XIE 91] XIE M., *Software Reliability Modeling*, World Scientific Publishing, 1991.

This page intentionally left blank

Part III

Quality of Life

This page intentionally left blank

Chapter 16

Likelihood Inference for the Latent Markov Rasch Model

16.1. Introduction

The main aim of this chapter is to illustrate likelihood inference for a discrete latent variable model for binary longitudinal data, which will be referred to as the latent Markov Rasch (LMR) model. This model is based on the following assumptions: (i) the response variables are conditionally independent given a sequence of latent variables which follows a first-order homogenous Markov chain; (ii) the distribution of each response variable depends on the corresponding latent variable as in the Rasch model [RAS 61]. The LMR model may then be seen as a constrained version of the latent Markov (LM) model of Wiggins [WIG 73] in which the Rasch parametrization is assumed for the conditional distribution of the response variables given the latent process and this process is time-homogenous. On the other hand, it may be seen as an extended version of the latent class Rasch (LCR) model [LEE 86, LIN 91, FOR 95] in which a subject is allowed to move between latent classes during the survey. Therefore, the LMR model may be used as an alternative model against testing with the LCR model. The null hypothesis that we are testing in this way is that the transition matrix of the latent process is diagonal and this may be of interest even outside a longitudinal context. The typical example is when a sequence of dichotomously-scored items is administered in the same order to a group of subjects, and we want to detect the presence of learning-through-training phenomena [STE 63].

Chapter written by Francesco BARTOLUCCI, Fulvia PENNONI and Monia LUPPARELLI.

For the maximum likelihood (ML) estimation of the LMR model, we illustrate an EM-algorithm [DEM 77] in which the E-step is based on certain recursions developed for hidden Markov models [MAC 97]. We also deal with the asymptotic distribution of the likelihood ratio (LR) statistic for testing hypotheses on the transition probabilities. Note that, under hypotheses of this type, the parameters may be on the boundary of the parameter space and, therefore, an inferential problem under non-standard conditions arises [SEL 87, SIL 04]. However, as shown by [BAR 06], this LR test statistic has an asymptotic chi-bar-squared distribution under the null hypothesis.

In this chapter, we also illustrate two extended versions of the LMR model. The first may be used to deal with discrete response variables having more than two levels. In this case, the distribution of each response variable given the corresponding latent variable is parameterized by using generalized logits [SAM 96] in a way that closely recalls the Rasch parametrization. The second extension is to the multivariate case, in which more response variables are observed at each time occasion. Using a terminology well known in LM studies [COL 92, LAN 94, VER 02], this model may be referred to as the multiple-indicator LMR model.

The remainder of this chapter is organized as follows. In section 16.2 we introduce the basic notation and briefly recall the LCR model. In section 16.3 we illustrate, as an extension of the LCR model, the LMR model and its basic properties. ML estimation of the latter is illustrated in section 16.4 together with LR testing of hypotheses on the transition probabilities. Section 16.5 contains an example based on a dataset coming from a study on the level of dementia of a group of elderly people. In section 16.6 we deal with the extension of the LMR model to the case of discrete response variables with more than two levels and to that of multivariate longitudinal data. The latter is illustrated by means of an application to a dataset concerning the conviction histories of a cohort of offenders. Finally, in section 16.7 we draw some conclusions.

16.2. Latent class Rasch model

Let $X_j, j = 1, \ldots, J$ denote the jth response variable. It corresponds, for instance, to a dichotomously-scored item administered to a group of subjects. The LCR model assumes that these response variables are conditionally independent given a latent variable Θ with support $\{\xi_1, \ldots, \xi_k\}$, and that

$$\text{logit}\,(\lambda_{jc}) = \xi_c - \beta_j, \qquad j = 1, \ldots, J, \quad c = 1, \ldots, k,$$

where $\lambda_{jc} = p(X_j = 1|\Theta = \xi_c)$ denotes the probability that the response to the jth item of a subject with latent trait level ξ_c equals 1. Note that this probability increases with ξ_c and decreases with β_j and, therefore, the parameters ξ_c are usually interpreted as ability levels and the parameters β_j as difficulty levels. The assumption of conditional independence between the response variables X_j given the latent variable Θ is usually referred to as local independence (LI) and characterizes the latent class model [LAZ 68, GOO 74]. This implies that Θ is the only explanatory variable of the response variables.

Because of LI, the conditional distribution of $\boldsymbol{X} = \begin{pmatrix} X_1 & \cdots & X_J \end{pmatrix}'$ given Θ may be expressed as

$$p(\boldsymbol{x}|c) = p(\boldsymbol{X} = \boldsymbol{x}|\Theta = \xi_c) = \prod_j \lambda_{jc}^{x_j}(1-\lambda_{jc})^{1-x_j}.$$

Moreover, the manifest distribution of $\boldsymbol{X}$, i.e. the marginal distribution of $\boldsymbol{X}$ with respect to Θ, may be expressed as

$$p(\boldsymbol{x}) = p(\boldsymbol{X} = \boldsymbol{x}) = \sum_c p(\boldsymbol{x}|c)\pi_c,$$

where $\pi_c = P(\Theta = \xi_c)$ is the probability that a subject has ability level ξ_c or, using an equivalent terminology, he/she belongs to the cth latent class.

16.3. Latent Markov Rasch model

The LMR model may be seen as an extension of the LCR model in which the response variables in $\boldsymbol{X}$ are assumed to be conditionally independent given a sequence of time-specific latent variables Θ_j, $j = 1, \ldots, J$, which follows a homogenous first-order Markov chain. This Markov chain has state space $\{\xi_1, \ldots, \xi_k\}$, initial probabilities $\pi_c = P(\Theta_1 = \xi_c)$, $c = 1, \ldots, k$, and transition probabilities $\pi_{cd} = P(\Theta_j = \xi_d|\Theta_{j-1} = \xi_c)$, $c, d = 1, \ldots, k$, for $j = 2, \ldots, J$. As in the LCR model, it is also assumed that

$$\text{logit}\,(\lambda_{jc}) = \xi_c - \beta_j, \qquad j = 1, \ldots, J, \quad c = 1, \ldots, k, \tag{16.1}$$

where, in this case, $\lambda_{jc} = p(X_j = 1|\Theta_j = \xi_c)$. This obviously means that the response variable X_j depends only on the corresponding latent variable Θ_j.

Note that, by assuming the existence of a latent Markov chain, we allow a subject to move between latent classes in a way that depends on the transition probabilities π_{cd}. This difference between the LCR and the LMR models is made clear by the path diagram in Figure 16.1.

Let $\boldsymbol{\Theta} = \begin{pmatrix} \Theta_1 & \cdots & \Theta_J \end{pmatrix}'$ denote the vector of latent variables and note that this vector may assume k^J configurations of type $\boldsymbol{\xi_c} = \begin{pmatrix} \xi_{c_1} & \cdots & \xi_{c_J} \end{pmatrix}'$, where $\boldsymbol{c} = \begin{pmatrix} c_1 & \cdots & c_J \end{pmatrix}'$. The assumption of conditional independence between the elements of $\boldsymbol{X}$ given $\boldsymbol{\Theta}$ implies that

$$p(\boldsymbol{x}|\boldsymbol{c}) = p(\boldsymbol{X} = \boldsymbol{x}|\boldsymbol{\Theta} = \boldsymbol{\xi_c}) = \prod_j \lambda_{jc_j}^{x_j}(1-\lambda_{jc_j})^{1-x_j}.$$

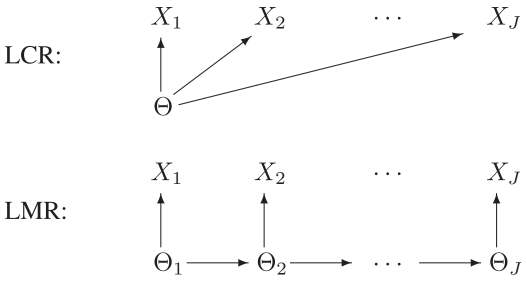


Figure 16.1. *Path diagram for the LCR and the LMR models*

Moreover, the distribution of $\boldsymbol{\Theta}$ may be expressed as

$$p(\boldsymbol{c}) = p(\boldsymbol{\Theta} = \boldsymbol{\xi}_{\boldsymbol{c}}) = \pi_{c_1} \prod_{j>1} \pi_{c_{j-1}c_j}$$

and, finally, the manifest distribution of $\boldsymbol{X}$ may be expressed as

$$p(\boldsymbol{x}) = p(\boldsymbol{X} = \boldsymbol{x}) = \sum_{\boldsymbol{c}} p(\boldsymbol{x}|\boldsymbol{c})p(\boldsymbol{c}).$$

Calculating $p(\boldsymbol{x})$ by using the sum above may be cumbersome in many contexts since the number of possible configurations of $\boldsymbol{\Theta}$ may be huge. A more efficient method of calculating this distribution, already used for hidden Markov models, is the following (for a more detailed description see [MAC 97, BAR 06]). Let $\boldsymbol{\pi}$ denote the initial probability vector and let $\boldsymbol{\Pi}$ denote the transition probability matrix, i.e.

$$\boldsymbol{\pi} = \begin{pmatrix} \pi_1 \\ \vdots \\ \pi_k \end{pmatrix}, \qquad \boldsymbol{\Pi} = \begin{pmatrix} \pi_{11} & \cdots & \pi_{1k} \\ \vdots & \ddots & \vdots \\ \pi_{k1} & \cdots & \pi_{kk} \end{pmatrix}.$$

Also, let $\boldsymbol{X}_{\leq j} = \begin{pmatrix} X_1 & \cdots & X_j \end{pmatrix}'$ denote the subvector of $\boldsymbol{X}$ made of its first j elements and let $\boldsymbol{q}(\boldsymbol{x}_{\leq j})$ denote the column vector with elements $p(\Theta_j = \xi_c, \boldsymbol{X}_{\leq j} = \boldsymbol{x}_{\leq j})$ for $c = 1, \ldots, k$. Note that this vector may be calculated for $j = 1, \ldots, J$ by using the recursion

$$\boldsymbol{q}(\boldsymbol{x}_{\leq j}) = \begin{cases} \operatorname{diag}(\boldsymbol{\phi}_{1x_1})\boldsymbol{\pi} & \text{if } j = 1, \\ \operatorname{diag}(\boldsymbol{\phi}_{jx_j})\boldsymbol{\Pi}'\boldsymbol{q}(\boldsymbol{x}_{\leq j-1}) & \text{otherwise,} \end{cases} \tag{16.2}$$

where $\boldsymbol{\phi}_{jx}$ denotes the column vector with elements $p(X_j = x|\Theta_j = \xi_c)$ for $c = 1, \ldots, k$. Then, we suggest calculating $p(\boldsymbol{x})$ as $\boldsymbol{q}(\boldsymbol{x}_{\leq J})'\boldsymbol{1}$, with $\boldsymbol{1}$ denoting a vector of suitable dimension with all elements equal to 1.

In the LMR model illustrated above, the transition matrix $\mathbf{\Pi}$ is completely unconstrained. Imposing a structure on $\mathbf{\Pi}$ allows us to formulate hypotheses which are of interest in many contexts [COL 92, BAR 06]. When $\xi_1 \leq \cdots \leq \xi_k$, so that the latent classes are ordered from the least capable subjects to the most capable subjects, it may be interesting to test the hypothesis that the transition matrix is upper triangular; with $k = 3$, for instance, we have

$$\mathbf{\Pi} = \begin{pmatrix} 1-(\alpha_1+\alpha_2) & \alpha_1 & \alpha_2 \\ 0 & 1-\alpha_3 & \alpha_3 \\ 0 & 0 & 1 \end{pmatrix}. \qquad (16.3)$$

This means that a subject in latent class c may remain in the same latent class or move only to class $d = c+1, \ldots, k$ and therefore improve in his/her ability level. Another hypothesis of interest is that the transition matrix is diagonal; with $k = 3$, for instance, we have

$$\mathbf{\Pi} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \qquad (16.4)$$

In this case, transition between latent classes is not allowed and so the LMR model specializes into the LCR model. Note that by using different constraints on the transition matrix, we generate a wide class of LM models which goes from the LCR model to the full LMR model initially illustrated.

As we will show in the following, LR testing of a hypothesis on the transition matrix may not be performed by using the standard asymptotic theory since one or more parameters of the model may be on the boundary of the parameter space under this hypothesis.

16.4. Likelihood inference for the latent Markov Rasch model

Let $n(\boldsymbol{x})$ denote the frequency of the response configuration $\boldsymbol{x}$ in the observed sample, the size of which is denoted by n. Also, let $\boldsymbol{\psi}$ denote the vector of all model parameters, i.e. $\beta_1, \ldots, \beta_J$ (difficulty levels), $\xi_1, \ldots, \xi_k$ (ability levels), $\pi_1, \ldots, \pi_k$ (initial probabilities of the latent process), $\pi_{11}, \pi_{12}, \ldots, \pi_{kk}$ (transition probabilities of the latent process).

As usual, we assume that the subjects in the sample have response patterns independent from one another, so that the log-likelihood of the LMR model may be expressed as

$$\ell(\boldsymbol{\psi}) = \sum_{\boldsymbol{x}} n(\boldsymbol{x}) \log[p(\boldsymbol{x})],$$

where $p(\boldsymbol{x})$ is calculated as a function of $\boldsymbol{\psi}$ by using recursion [16.2]. We can maximize $\ell(\boldsymbol{\psi})$ by means of the EM-algorithm [DEM 77] illustrated in the following.

16.4.1. *Log-likelihood maximization*

To describe how the EM-algorithm may be implemented, we need to consider the complete data log-likelihood which is given by

$$\ell^*(\boldsymbol{\psi}) = \sum_{\boldsymbol{c}} \sum_{\boldsymbol{x}} m(\boldsymbol{c}, \boldsymbol{x}) \log[p(\boldsymbol{x}|\boldsymbol{c})p(\boldsymbol{c})],$$

with $m(\boldsymbol{c}, \boldsymbol{x})$ denoting the number of subjects with latent process configuration $\boldsymbol{c}$ and response configuration $\boldsymbol{x}$. After some algebra, this log-likelihood may be expressed as

$$\begin{aligned} \ell^*(\boldsymbol{\psi}) \quad &= \textstyle\sum_c f_c \log(\pi_c) + \sum_c \sum_d g_{cd} \log(\pi_{cd}) + \\ &+ \textstyle\sum_j \sum_c \{h_{jc1} \log(\lambda_{jc}) + h_{jc0} \log(1 - \lambda_{jc})\}, \end{aligned} \tag{16.5}$$

where f_c is the number of subjects that, at the first occasion, are in latent class c, g_{cd} is the number of transitions from class c to class d and h_{jcx} is the number of subjects that are in class c and respond x to item j.

Obviously, the above frequencies are not known and so $\ell^*(\boldsymbol{\psi})$ may not be calculated. However, this log-likelihood is exploited within the EM-algorithm to maximize the incomplete data log-likelihood $\ell(\boldsymbol{\psi})$ by alternating the following two steps until convergence is achieved:

– *E-step*: calculate the conditional expected value of the frequencies f_c, g_{cd} and h_{jcx} given the observed data $n(\boldsymbol{x})$ and the current estimate of $\boldsymbol{\psi}$;

– *M-step*: updates the estimate of $\boldsymbol{\psi}$ by maximizing the complete log-likelihood $\ell^*(\boldsymbol{\psi})$, as defined in [16.5], in which every frequency is substituted by the corresponding expected value calculated during the E-step.

These steps may be performed in a rather simple way. In particular, the E-step may be performed by using certain recursions similar to [16.2]. For the full LMR model, the M-step may be performed explicitly for what concerns the parameters π_c and π_{cd} and by using a standard algorithm for fitting generalized linear models for what concerns the parameters β_j and ξ_c. Note that, in order to ensure identifiability, we constrain β_1 to 0. The M-step is slightly more complex when constraints are put on the transition probabilities. For a more detailed description on how to implement these steps see [BAR 06].

Multimodality of the likelihood, which often arises when fitting latent variable models, can be detected by performing the above estimation algorithm with different sets of starting values and then choosing, as the ML estimate of the parameters, the value of $\boldsymbol{\psi}$ which at convergence gives the highest value of $\ell(\boldsymbol{\psi})$. This estimate is denoted by $\hat{\boldsymbol{\psi}}$.

16.4.2. *Likelihood ratio testing of hypotheses on the parameters*

In order to test a hypothesis H_0 on the parameters $\boldsymbol{\psi}$ of an LMR model with a certain number of latent classes and certain restrictions on the transition matrix, we suggest the use of the LR statistic

$$D = -2[\ell(\hat{\boldsymbol{\psi}}_0) - \ell(\hat{\boldsymbol{\psi}})],$$

where $\hat{\boldsymbol{\psi}}_0$ is the ML estimate of $\boldsymbol{\psi}$ under H_0; this estimate may also be calculated by using the EM-algorithm illustrated above. In particular, we are interested in testing hypotheses on the transition matrix, such as those formulated in [16.3] and [16.4]. As previously mentioned, standard asymptotic results on the distribution of LR test statistics may not be applied in this case. However, by using certain results of constrained statistical inference [SEL 87, SIL 04], it was shown by [BAR 06] that, under suitable regularity conditions, the null asymptotic distribution of D belongs to the chi-bar-squared family [SHA 88, SIL 04].

The chi-bar-squared distribution, denoted by $\bar{\chi}^2$, is the distribution of

$$Q = \boldsymbol{V}'\boldsymbol{\Sigma}^{-1}\boldsymbol{V} - \min_{\hat{\boldsymbol{V}}\in\mathcal{C}}(\hat{\boldsymbol{V}} - \boldsymbol{V})'\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{V}} - \boldsymbol{V}),$$

where $\boldsymbol{V}$ is a random vector of dimension t with distribution $N(\boldsymbol{0}, \boldsymbol{\Sigma})$ and $\mathcal{C}$ is a polyhedral convex cone in $\mathcal{R}^t$. It may be proved that the $\bar{\chi}^2$ distribution corresponds to a mixture of chi-squared distributions, so that

$$p(Q \geq q) = \sum_{i=0}^{t} w_i(\mathcal{C}, \boldsymbol{\Sigma})p(\chi_i^2 \geq q). \tag{16.6}$$

The weights $w_i(\mathcal{C}, \boldsymbol{\Sigma})$ may be explicitly calculated only in particular cases. However, these weights may always be estimated, with the required precision, by using a simple Monte Carlo algorithm. On the basis of [16.6], we can obtain an asymptotic p-value for the observed value of the LR statistic D, denoted by d.

A particular case of the above result is when we assume that all the off-diagonal elements of the transition matrix $\boldsymbol{\Pi}$ are equal to the same parameter α and we want to test the hypothesis $H_0 : \alpha = 0$. With $k = 3$ latent classes, for instance, we have

$$\boldsymbol{\Pi} = \begin{pmatrix} 1-2\alpha & \alpha & \alpha \\ \alpha & 1-2\alpha & \alpha \\ \alpha & \alpha & 1-2\alpha \end{pmatrix}.$$

In this case, the null asymptotic distribution of D is the mixture

$$0.5\chi_0^2 + 0.5\chi_1^2,$$

and so a p-value for d may be explicitly calculated as $0.5p(\chi_1^2 \geq d)$.

16.5. An application

We applied the approach presented above to a dataset derived from an epidemiologic survey carried out in Tuscany, Italy, in the 1990s. This dataset derives from the administration of $J = 7$ items measuring the ability in specific skills, such as motion, speech and drawing, to a sample of $n =$1,368 elderly people. An incorrect response to any of these items, the case in which the corresponding response variable is set equal to 0, is a symptom of dementia in the respondent. Our analysis was aimed, in particular, at testing if the LCR model is suitable for these data by comparing it with the LMR model. This kind of test makes sense since the items were administered in the same order to all the subjects.

The first step of our analysis was the choice of k, the number of latent classes. For this, we used a method based on the minimization of the BIC index [SCH 78] applied to the LCR model. For a model with k classes, this index is defined as

$$BIC_k = -2\hat{\ell}_k + r_k \log(n),$$

where $\hat{\ell}_k$ is the maximum log-likelihood of the model and r_k is the number of independent parameters.

The results obtained from the application of this method are reported in Table 16.1 which shows, for k between 1 and 4, the number of parameters of the LCR model, the maximum log-likelihood and the BIC index. According to these results, we found that a suitable number of latent classes for the LCR model is $k = 3$. The corresponding parameter estimates are reported in Table 16.2.

k	r_k	$\hat{\ell}_k$	BIC$_k$
1	7	-5,131.0	10,313.0
2	9	-4,702.0	9,469.1
3	11	-4,627.2	9,333.8
4	13	-4,624.9	9,343.7

Table 16.1. *Results obtained from the BIC-based method for the choice of the number of latent classes. k is the number of latent classes for the LCR model, r_k is the corresponding number of parameters, $\hat{\ell}_k$ the maximum log-likelihood and BIC$_k$ the value of the BIC index*

We then fitted the LMR model with $k = 3$ latent classes to compare it with the LCR model. The log-likelihood of this model is equal to -4,627.0 and the estimates of the parameters which are also present in the LCR model are reported in Table 16.2. Finally, the estimated transition matrix is

$$\begin{pmatrix} 1.000 & 0.000 & 0.000 \\ 0.000 & 1.000 & 0.000 \\ 0.000 & 0.013 & 0.987 \end{pmatrix}.$$

	LCR ($k = 3$)	LMR ($k = 3$)
β_2	-0.724	-0.746
β_3	-3.593	-3.638
β_4	1.127	1.089
β_5	0.940	0.882
β_6	1.382	1.310
β_7	0.875	0.783
ξ_1	-3.189	-3.249
ξ_2	0.622	0.542
ξ_3	2.714	2.676
π_1	0.048	0.047
π_2	0.425	0.397
π_3	0.527	0.556
$\hat{\ell}$	$-4,627.2$	$-4,627.0$

Table 16.2. *Estimates of the difficulty and ability parameters of the LCR and the LMR models with $k = 3$ latent classes. β_j is the difficulty level of item j ($\beta_1 \equiv 0$), ξ_c is the ability level of the subjects in class c, π_c denotes the corresponding probability (this is an initial probability for the LMR model) and $\hat{\ell}$ denotes the maximum log-likelihood of the model*

We can see that this transition matrix is very close to an identity matrix. The largest off-diagonal element is in fact equal to 0.013 and corresponds to the transition from the third to the second class. This is in accordance with the fact that the LR statistic between the LCR and the LMR model is equal to 0.35 with a p-value of 0.77. Therefore, we did not reject the first model in favor of the second and we concluded that there is no evidence of learning-through-training phenomena.

From the estimates of the LCR model we can observe that the first class, containing subjects with a high level of dementia, corresponds to 4.8% of the population under study. The other two classes, corresponding to a moderate level and an absence of dementia, represent respectively 42.5% and 52.7% of the population. Finally, among the items used in survey, there is a certain heterogenity for what concerns the difficulty levels. The most difficult item is the sixth and the easiest is the third.

16.6. Possible extensions

In the following, we briefly show how the LMR model may be extended to the case of discrete data with more than two categories and to the multivariate case in which we have more response variables for each time occasion.

16.6.1. *Discrete response variables*

Suppose that each response variable X_j has $s+1$ levels from 0 to s. The LMR model illustrated in section 16.3 may still be applied once a suitable parametrization of the conditional distribution of any X_j given Θ_j has been assumed; see also [SAM 96, BAR 06].

Let $\lambda_{jcx} = p(X_j = x|\Theta_j = \xi_c)$. When the response categories are not ordered, we can use a parametrization based on logits with the first category as baseline, i.e.

$$\log \frac{\lambda_{jcx}}{\lambda_{jc0}} = \xi_c - \beta_{jx},$$

with $j = 1, \ldots, J$, $c = 1, \ldots, k$ and $x = 1, \ldots, s$. When the response categories are naturally ordered, global logits are more appropriate and so we assume

$$\log \frac{\lambda_{jcx} + \ldots + \lambda_{jcs}}{\lambda_{jc0} + \ldots + \lambda_{jc,x-1}} = \xi_c - \beta_{jx}. \tag{16.7}$$

All the other assumptions of the LMR model remain unchanged and so it may be estimated by means of an EM-algorithm very similar to that described in section 16.4.1. Also, the asymptotic results described in section 16.4.2 continue to hold; for details see [BAR 06].

The approach described in this section was used by [BAR 06] to analyze a dataset derived from a longitudinal study on the use of marijuana among young people. This dataset concerns $n = 237$ respondents aged 13 years in 1976 who were followed until 1980. The use of marijuana is measured using $J = 5$ ordinal variables with 3 levels: 0 for never in the past year; 1 for no more than once a month in the past year; 2 for more than once a month in the past year. For a deeper description of the dataset see [ELL 89].

For the above dataset, an LM model with a parametrization based on global logits, $k = 3$ latent classes and a tridiagonal transition matrix was selected by using the asymptotic results here discussed. The conditional global logits, in particular, have been parameterized as in [16.7] under the constraint $\beta_{1x} = \cdots = \beta_{Jx}$ for $x = 1, \ldots, s$. In particular, the parameter ξ_c is interpreted as the tendency to use marijuana for the subjects in latent class c.

16.6.2. *Multivariate longitudinal data*

The LMR model may also be used to analyze data which are collected, for instance, by administering a certain set of items to the same subjects at different time occasions.

In this case, we have a multiple-indicator LMR model; see also [COL 92, LAN 94, VER 02].

Let $\boldsymbol{X}_j = \begin{pmatrix} X_{j1} & \cdots & X_{jv} \end{pmatrix}'$ denote the vector of binary response variables observable at the jth occasion, with $j = 1, \ldots, J$. The multiple-indicator LMR model assumes that $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_J$ are conditionally independent given $\boldsymbol{\Theta}$ and that the elements of $\boldsymbol{X}_j$ are conditionally independent given Θ_j. This implies that

$$p_j(\boldsymbol{x}|c) = p(\boldsymbol{X}_j = \boldsymbol{x}|\Theta_j = \xi_c) = \prod_u \lambda_{juc}^{x_u}(1-\lambda_{juc})^{1-x_u},$$

where x_u, with $u = 1, \ldots, v$, denotes the uth element of the vector $\boldsymbol{x}$ and $\lambda_{juc} = p(X_{ju} = 1|\Theta_j = \xi_c)$. The latter may be parameterized as in [16.1], with the difficulty parameters being occasion specific, i.e.

$$\text{logit}\,(\lambda_{juc}) = \xi_c - \beta_{ju}. \tag{16.8}$$

As shown in detail by [BAR 07], ML estimation of the above model may be performed by using an EM-algorithm similar to that illustrated in section 16.4.1. In particular, the log-likelihood of the model may be expressed as

$$\ell(\boldsymbol{\psi}) = \sum_{\boldsymbol{x}_1} \cdots \sum_{\boldsymbol{x}_J} n(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_J) \log[p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_J)],$$

where $p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_J) = p(\boldsymbol{X}_1 = \boldsymbol{x}_1, \ldots, \boldsymbol{X}_J = \boldsymbol{x}_J)$ may be calculated by means of a recursion similar to that introduced in section 16.3. Though the problem must be studied in more detail, the inferential results illustrated in section 16.4.2 can be applied, with minor adjustments, to this model.

The above model was used by [BAR 07] to analyze the conviction histories of a cohort of offenders who were born in England and Wales in 1953. These offenders were followed from the age of criminal responsibility, 10 years old, until the end of 1993. The offences were grouped into $v = 10$ major categories listed in Table 16.3 and gender was included in the model as an explanatory variable. The data are referred to six age bands (10-15, 16-20, 21-25, 26-30, 31-35 and 36-40), so that $J = 6$.

However, [BAR 07] found that the fit of the multivariate LMR model may be strongly improved by relaxing the assumption of homogenity of the latent Markov chain and substituting assumption [16.8] with the assumption that the conditional probabilities λ_{juc} do not depend on j and are equal to zero for $c = 1$. The model they selected under these assumptions has $k = 5$ latent classes, different initial probabilities between males and females (denoted by π_c^M and π_c^F), and common transition probabilities (denoted by $\pi_{cd}^{(j)}$), and conditional probabilities (denoted by λ_{uc}).

u	Category
1	*Violence against the person*
2	*Sexual offences*
3	*Burglary*
4	*Robbery*
5	*Theft and handling stolen goods*
6	*Fraud and forgery*
7	*Criminal damage*
8	*Drug offences*
9	*Motoring offences*
10	*Other offences*

Table 16.3. *Major categories of offences*

The estimates of the conditional probabilities, which are reported in Table 16.4, may be used to interpret the latent classes. For instance, the first class may be interpreted as that of the non-offenders, the second as that of the incidental offenders and so on. The estimated initial probabilities for males and females are reported in Table 16.5. We can note that about 50% of the males in age band 10-15 and about 96% of the females in the same age band are non-offenders. Finally, the estimated transition matrices are reported in Tables 16.6 and 16.7. Note that the first matrix refers to the transition from age band 10-15 to age band 16-20, whereas the second refers to the other transitions (from 16-20 to 21-25, from 21-25 to 26-30 and so on). The resulting model is then partially homogenous. On the basis of the estimated transition probabilities we can draw some conclusions about the evolution of the offending level of the subjects. For instance, we can observe that a high persistence is associated to the first class, that of the non-offenders; the same cannot be said about the other classes. For a more detailed description of these results see [BAR 07].

		Latent class (c)				
u	Category	1	2	3	4	5
1	*Violence against the person*	0.000	0.003	0.158	0.018	0.227
2	*Sexual offences*	0.000	0.003	0.029	0.003	0.026
3	*Burglary*	0.000	0.032	0.006	0.016	0.487
4	*Robbery*	0.000	0.000	0.005	0.002	0.039
5	*Theft and handling stolen goods*	0.000	0.096	0.067	0.546	0.777
6	*Fraud and forgery*	0.000	0.000	0.019	0.130	0.149
7	*Criminal damage*	0.000	0.016	0.091	0.010	0.233
8	*Drug offences*	0.000	0.000	0.075	0.016	0.099
9	*Motoring offences*	0.000	0.000	0.005	0.003	0.044
10	*Other offences*	0.000	0.000	0.060	0.039	0.347

Table 16.4. *Estimated conditional probabilities of conviction given the latent class*

c	π_c^M	π_c^F
1	0.496	0.963
2	0.472	0.020
3	0.000	0.000
4	0.000	0.016
5	0.033	0.000

Table 16.5. *Estimated initial probabilities for both males (π_c^M) and females (π_c^F)*

	d				
c	1	2	3	4	5
1	0.960	0.009	0.003	0.028	0.000
2	0.068	0.648	0.140	0.053	0.092
3	1.000	0.000	0.000	0.000	0.000
4	0.846	0.000	0.005	0.149	0.000
5	0.040	0.175	0.033	0.000	0.753

Table 16.6. *Estimated transition probabilities $\pi_{cd}^{(j)}$ from age band 10-15 to age band 16-20 for both males and females*

16.7. Conclusions

We illustrated a latent Markov version [WIG 73] of the Rasch model [RAS 61] which is referred to as the LMR model. With respect to the latent class version of the Rasch model, it has the advantage of allowing the subjects to move between latent classes during the survey according to a homogenous Markov chain which is not directly observable. We dealt, in particular, with ML estimation of the parameters of the model at issue and LR testing of hypotheses on the transition matrix of the latent process. ML estimation is carried out by means of an EM-algorithm [DEM 77] which is more sophisticated than that used to estimate the parameters of a latent class model. This algorithm is implemented through a series of MATLAB functions which are available from the authors upon request.

The LMR model can be extended by including a set of discrete or continuous covariates. In the presence of discrete covariates, which give rise to a stratification of the subjects in a small number of strata, a separate LMR model can be fitted for each stratum. A reduction in the number of parameters may possibly be achieved by imposing suitable restrictions between the models corresponding to the different strata. Continuous covariates, instead, can be included by adding a linear term depending on these covariates to expression [16.1]. Alternatively, we can allow the initial and the transition probabilities of the latent process to depend on the covariates by using a logit parametrization of these probabilities. In a similar context, this formulation was followed by [VER 99]. We have to clarify that the interpretation of the latent process

	d				
c	1	2	3	4	5
1	0.993	0.000	0.005	0.012	0.000
2	0.327	0.371	0.206	0.078	0.018
3	0.583	0.000	0.395	0.015	0.007
4	0.809	0.001	0.008	0.180	0.002
5	0.000	0.276	0.217	0.039	0.468

Table 16.7. *Estimated transition probabilities $\pi_{cd}^{(j)}$ from one age band to another for both males and females over 16*

is different in the two cases. In the second case, in particular, each response variable depends only on the corresponding latent variable and so this variable provides a synthetic measure (e.g. quality of life) of the individual characteristics measured by the response variables. This kind of interpretation is not reasonable when the covariates affect the conditional probabilities. In this case, the latent process allows us to take into account the heterogenity between subjects which is not explained by the observable covariates.

16.8. Bibliography

[BAR 06] BARTOLUCCI F., "Likelihood inference for a class of latent Markov models under linear hypotheses on the transition probabilities", *Journal of the Royal Statistical Society, Series B*, vol. 68, p. 155–178, 2006.

[BAR 07] BARTOLUCCI F., PENNONI F., FRANCIS B., "A latent Markov model for detecting patterns of criminal activity", *Journal of the Royal Statistical Society, series A*, vol. 170, p. 115-132, 2007.

[COL 92] COLLINS L. M., WUGALTER S. E., "Latent class models for stage-sequential dynamic latent variables", *Multivariate Behavioral Research*, vol. 27, p. 131-157, 1992.

[DEM 77] DEMPSTER A. P., LAIRD N. M., RUBIN D. B., "Maximum likelihood from incomplete data via the EM algorithm (with discussion)", *Journal of the Royal Statistical Society, Series B*, vol. 39, p. 1-38, 1977.

[ELL 89] ELLIOT D. S., HUIZINGA D., MENARD S., *Multiple Problem Youth: Delinquency, Substance Use and Mental Health Problems*, Springer-Verlag, New York, 1989.

[FOR 95] FORMANN A. K., "Linear logistic latent class analysis and the Rasch model", in: Rasch Models: Foundations, Recent Developments, and Applications, p. 239-255, Springer-Verlag, New York, 1995.

[GOO 74] GOODMAN L. A., "Exploratory latent structure analysis using both identifiable and unidentifiable models", *Biometrika*, vol. 61, p. 215-231, 1974.

[LAN 94] LANGEHEINE R., "Latent variables Markov models", in*: Latent Variables Analysis: Applications for Development Research*, p. 373–395, Sage, Thousand Oaks, 1994.

[LAZ 68] LAZARSFELD P. F., HENRY N. W., *Latent Structure Analysis*, Houghton Mifflin, Boston, 1968.

[LEE 86] DE LEEUW J., VERHELST N., "Maximum likelihood estimation in generalized Rasch models", *Journal of Educational Statistics*, vol. 11, p. 183-196, 1986.

[LIN 91] LINDSAY B., CLOGG C., GREGO J., "Semiparametric estimation in the Rash model and related exponential response models, including a simple latent class model for item analysis", *Journal of the American Statistical Association*, vol. 86, p. 96-107, 1991.

[MAC 97] MACDONALD I. L., ZUCCHINI W., *Hidden Markov and Other Models for Discrete-Valued Time Series*, Chapman and Hall, London, 1997.

[RAS 61] RASCH G., "On general laws and the meaning of measurement in psychology", *Proceedings of the IV Berkeley Symposium on Mathematical Statistics and Probability*, vol. 4, p. 321-333, 1961.

[SAM 96] SAMEJIMA F., "Evaluation of mathematical models for ordered polychotomous responses", *Behaviormetrika*, vol. 23, p. 17-35, 1996.

[SCH 78] SCHWARZ G., "Estimating the dimension of a model", *Annals of Statistics*, vol. 6, p. 461-464, 1978.

[SEL 87] SELF S. G., LIANG K.-Y., "Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions", *Journal of the American Statistical Association*, vol. 82, p. 605-610, 1987.

[SHA 88] SHAPIRO A., "Towards a unified theory of inequality constrained testing in multivariate analysis", *International Statistical Review*, vol. 56, p. 49-62, 1988.

[SIL 04] SILVAPULLE M. J., SEN P. K., *Constrained Statistical Inference: Inequality, Order, and Shape Restrictions*, Wiley, New York, 2004.

[STE 63] STERNBERG S., "Stochastic learning theory", in: *Handbook of Mathematical Psychology, vol. II*, p. 1-120, Wiley, New York, 1963.

[VER 99] VERMUNT J. K., LANGEHEINE R., BÖCKENHOLT U., "Discrete-time discrete-state latent Markov models with time-constant and time-varying covariates", *Journal of Educational and Behavioral Statistics*, vol. 24, p. 179-207, 1999.

[VER 02] VERMUNT J. K., GEORG W., "Longitudinal data analysis using log-linear path models with latent variables", *Metodología de las Ciencias del Comportamiento*, vol. 4, p. 37-53, 2002.

[WIG 73] WIGGINS L. M., *Panel Analysis: Latent Probability Models for Attitude and Behavior Processes*, Elsevier, Amsterdam, 1973.

This page intentionally left blank

Chapter 17

Selection of Items Fitting a Rasch Model

17.1. Introduction

Item Response Models (IRM) [FIS 95] [LIN 97] are models used in educational testing, psychology or health-related quality of life. These models consider that a latent trait (latent variable) explains the responses to the items. The latent trait is generally multidimensional, but IRMs generally only consider a unidimensional latent trait. Moreover, the relations between the items and the component of the latent trait often are unknown: they are assumed by suppositions of experts and the work of the statistician consists of validating them with psychometric models. The statistician generally is not part of the exploratory analysis that defines these links and helps the experts define links which are coherent with psychometric properties.

Some procedures (factor analysis, MSP, HCA/CCPROX) allow the links between the items and the latent traits to be defined, but none of them are based on the direct fit of an IRM to the data. [HAR 04] proposes a procedure which makes it possible to define these links and obtain scales with a good fit of a given IRM: the Multidimensional Marginally Sufficient Rasch Model, which is a multidimensional counterpart of the most famous IRM, the Rasch model.

However, the fit of a multidimensional IRM, evaluated by the likelihood of the model, is a long process to run if we consider the model as a Multidimensional Generalized Linear Mixed Model (GLMM): in this chapter, a fast way to carry out this procedure is proposed. It makes it possible to obtain correct results in a reasonable time.

Chapter written by Jean-Benoit HARDOUIN and Mounir MESBAH.

17.2. Notations and assumptions

17.2.1. *Notations*

Let $\Theta_q 1$ be the qth component of the multidimensional latent trait characterizing the individuals with $q = 1, \dots, Q$ and θ_{nq} the realization of this latent trait for the nth individual, $n = 1, \dots, N$. $\boldsymbol{\theta}_n$ is the vector of the values on the Q latent traits for the nth individual $(\theta_{n1}, \dots, \theta_{nq}, \dots, \theta_{nQ})$.

The jth item is characterized by a vector of parameters $\boldsymbol{\nu}_j$, $j = 1, \dots, J$ [FIS 95]. The response to this item is represented by the random variable X_j, while the realization for the nth individual is denoted x_{nj}.

We consider only dichotomous items, and for each of them, the more favorable response is named "positive response" and is coded 1, and the other is named "negative response" and is coded 0.

The Item Response Function (IRF) of the jth item is the probability that a given individual n positively responds to this item as a function of the value of the latent trait for the nth individual $\boldsymbol{\theta}_n$.

17.2.2. *Fundamental assumptions of the Item Response Theory (IRT)*

The IRT is the set of IRMs which verifies three fundamental assumptions [LIN 97]:

– Fixed dimension: the dimension Q of the latent trait is known. For the majority of IRM, the unidimensionality ($Q = 1$) is required.

– Local independency: the responses to the items are conditionally independent to the latent trait.

– Monotonicity: the IRF are non-decreasing functions of each component of the latent trait.

17.3. The Rasch model and the multidimensional marginally sufficient Rasch model

17.3.1. *The Rasch model*

The Rasch model [RAS 60] is a unidimensional IRM. The responses to the items are assumed to depend on a unidimensional latent trait: θ_n is a scalar. Moreover, each item j, $j = 1, \dots, J$, is defined by only one parameter δ_j: this parameter is interpreted as the difficulty of the jth item, because the higher its value, the more the probability will positively respond when item j is small. The IRF of the jth item is:

$$P(X_{nj} = x_{nj}/\theta_n; \delta_j) = \frac{\exp\left(x_{nj}\left(\theta_n - \delta_j\right)\right)}{1 + \exp\left(\theta_n - \delta_j\right)}. \tag{17.1}$$

The latent trait can be considered as a set of fixed parameters or as a random variable. In the fixed effects Rasch model, the classical maximum likelihood technique makes the estimations inconsistent.

The Rasch model is a famous IRM because this model has a specific property: the score $S_n = \sum_{j=1}^{J} X_{nj}$ is a sufficient statistic of the latent trait, that is to say, all the available information about the latent trait is contained in the score [AND 77]. Consequently, if the latent trait is considered as a set of fixed parameters, the conditional maximum likelihood can be used: the likelihood is maximized conditionally to the score calculated as the number of positive responses to all the items for each individual. These estimations are consistent [AND 70].

If the latent trait is considered a random variable, its distribution function $G(\theta)$ is assumed (generally as a centered Gaussian distribution of variance σ^2 [FIS 95]), and consistent estimations of the items parameters δ_j and of the parameters of this distribution (generally, only the variance σ^2) can be obtained by maximizing the marginal likelihood:

$$L_M(\sigma^2, \boldsymbol{\delta}/\boldsymbol{x}) = \prod_{n=1}^{N} \int \prod_{j=1}^{J} P(X_{nj} = x_{nj}/\theta; \delta_j) G(\theta/\sigma^2) d\theta, \tag{17.2}$$

with $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_j, \ldots, \delta_J)$.

17.3.2. *The multidimensional marginally sufficient Rasch model*

Hardouin and Mesbah [HAR 04] propose an extension of the Rasch model to the multidimensional case. In this model, the responses to the items are governed by a multidimensional latent trait (of dimension Q), but the response to a given item j is governed by only one component of the latent trait indexed q_j. Consequently, marginally to the other items and to the other components of the latent trait, all the items governed by the same component of the latent trait (that is to say, each scale) fit a classical Rasch model.

This model is a particular case of the "between items responses model" defined by Adams *et al.* [ADA 97].

In this model, the response function of the jth item is:

$$P(X_{nj} = x_{nj}/\boldsymbol{\theta}_n; \delta_j) = \frac{\exp\left(x_{nj}\left(\theta_{nq_j} - \delta_j\right)\right)}{1 + \exp\left(\theta_{nq_j} - \delta_j\right)} = P(X_{nj} = x_{nj}/\theta_{nq_j}; \delta_j). \tag{17.3}$$

As a consequence, the score $S_{nq} = \sum_{j=1/q_j=q}^{J} x_{nj}$ calculated with only the items associated with the qth component of the latent trait Θ_q is a sufficient statistic of Θ_q.

This model is named the Multidimensional Marginally Sufficient Rasch Model (MMSRM) for this reason. By considering the latent trait as a multivariate random variable (distributed by a multivariated centered Gaussian distribution with an unknown covariance matrix $\mathbf{\Sigma}$), the item parameters (and the elements of the $\mathbf{\Sigma}$ matrix) can be consistently estimated by a marginal maximum likelihood, by maximizing:

$$L_{M1}(\mathbf{\Sigma}, \boldsymbol{\delta}/\boldsymbol{x}) = \prod_{n=1}^{N} \int \prod_{j=1}^{J} P(X_{nj} = x_{nj}/\theta_{q_j}; \delta_j) G(\boldsymbol{\theta}/\mathbf{\Sigma}) d\boldsymbol{\theta}. \tag{17.4}$$

This method of estimation is a long process with traditional software such as SAS, Splus or Stata, because these integrals must be approximated at each step of the algorithm.

We can gain time by independently estimating the item parameters and the elements of the Σ matrix: as each marginal scale verifies a Rasch model, and the item parameters can be estimated scale by scale by the maximum marginal likelihood method. These estimations are denoted $\hat{\delta}_j$, $j = 1, \ldots, J$.

Then, the elements of the covariance matrix $\mathbf{\Sigma}$ can be estimated by assuming the item parameters as known parameters and by maximizing the new quantity:

$$L_{M2}(\mathbf{\Sigma}/\boldsymbol{x}, \hat{\boldsymbol{\delta}}) = \prod_{n=1}^{N} \int \prod_{j=1}^{J} P(X_{nj} = x_{nj}/\theta_{q_j}; \hat{\delta}_j) G(\boldsymbol{\theta}/\mathbf{\Sigma}) d\boldsymbol{\theta}, \tag{17.5}$$

which is an approximation of $L_{M1}(\mathbf{\Sigma}, \boldsymbol{\delta}/\boldsymbol{x})$.

17.4. The Raschfit procedure

In [HAR 04], a procedure of item selection in Rasch scales based on the fit of the items to an MMSRM is proposed. This procedure is referenced here by “Raschfit”.

At step k of this procedure, we assume to have a set of $J^{(k)}$ items, named kernel, which verifies a Rasch model, and we find if a Rasch model fits the data by adding a new item (indexed by $j = 0$) to the kernel. We compare the fit of a Rasch model, and the fit of the MMSRM with the items of the kernel relying on one component of the latent trait and the new item relying on another component.

At the first step of the Raschfit procedure, the initial kernel is composed of two or more items chosen by the user, or determined by a specific analysis. The order in which the other items will be introduced in the procedure can be freely determined but can have an impact in the final result. The authors propose to use the Mokken Scale Procedure (MSP) [HEM 95] to select the initial kernel and to order the other items.

The fit of the models is calculated by the Akaike Information Criterion (AIC) [HOI 97] with:

$$AIC_m = -2l_m + 2K_m, \tag{17.6}$$

where l_m is the value of the log-likelihood and K_m the number of parameters of the model m. At the kth step of the procedure, we have $K_1^{(k)} = J^{(k)} + 2$ for the Rasch model and $K_2^{(k)} = J^{(k)} + 4$ for the MMSRM.

The new item is selected at the kth step of the procedure in the scale if $AIC_1^{(k)} \leq AIC_2^{(k)}$. The procedure is stopped when there are no more items remaining.

This procedure is a very long process, because the log-likelihood of an MMSRM is long to approximate. The estimation can be implemented with the GLlamm program of Stata, the NLmixed procedure of SAS or the NLME library of Splus. These three programs approximate the multivariate integrals with (adaptive) Gaussian quadratures and necessitate a great amount of computer resources.

17.5. A fast version of Raschfit

We propose here an adaptation of the Raschfit procedure, referenced as "Raschfit-fast". This adaptation is based on the fixed effects Rasch model.

17.5.1. *Estimation of the parameters under the fixed effects Rasch model*

In the Rasch model, by considering the latent trait as a set of fixed parameters ($\theta_n,\ n = 1, \ldots, N$), the individual values of the latent trait $\theta_n,\ n = 1, \ldots, N$ cannot be consistently estimated by classical techniques [FIS 95].

Indeed, the Rasch model verifies the specific property of sufficiency of the unweighted score on the latent trait: this property signifies that, conditionally to the score $s_n = \sum_{j=1}^{J} x_{nj}$, the likelihood of the nth individual is independent of the latent trait θ_n.

$$L_{Cn}(\theta_n, \boldsymbol{\delta}/\boldsymbol{x_n}, s_n) = L_{Cn}(\boldsymbol{\delta}/\boldsymbol{x_n}, s_n). \tag{17.7}$$

In maximizing the quantity

$$L_C(\boldsymbol{\delta}/\boldsymbol{x}, \boldsymbol{s}) = \prod_{n=1}^{N} L_{Cn}(\boldsymbol{\delta}/\boldsymbol{x_n}, s_n), \tag{17.8}$$

we obtain a consistent estimation of the $\boldsymbol{\delta}$ vector of parameters.

The weighted maximum likelihood technique [FIS 95] consistently estimates the $\theta_s,\ s = 0, \ldots, J$ parameters by maximizing:

$$L_{WC}(\boldsymbol{\theta_S}/\hat{\boldsymbol{\delta}}, \boldsymbol{x}) = \prod_{n=1}^{N} L_n(\theta_{s_n}/\hat{\boldsymbol{\delta}}, \boldsymbol{x_n})g(\theta_{s_n}), \tag{17.9}$$

where $\boldsymbol{\theta_S} = (\theta_0, \ldots, \theta_s, \ldots, \theta_J)'$ and

$$g(x) = \prod_{j=1}^{J} \frac{\exp(x - \hat{\delta}_j)}{\left(1 + \exp(x - \hat{\delta}_j)\right)^2}. \tag{17.10}$$

17.5.2. *Principle of Raschfit-fast*

The principle of Raschfit-fast is globally the same as for Raschfit: at each step, a kernel fits a Rasch model, and a new item is added to this kernel if the new scale has a good fit to a Rasch model. The mixed Rasch model is replaced by a fixed effects Rasch model, and the MMSRM is replaced by a specific model built from the following consideration: if the set of items composed of the items of the kernel and the new item does not follow a Rasch model, then the responses to the new item are independent of the latent trait. The likelihood associated with these responses is estimated by a logistic form (as in the Rasch model) but only with an unknown parameter characterizing the item in the linear composant (which can be interpreted as a difficulty parameter).

At step k of the algorithm, let 0 be the index of the new item, the J_k items of the kernel being indexed from 1 to J_k.

17.5.3. *A model where the new item is explained by the same latent trait as the kernel*

If the new item is explained by the same latent trait as the kernel, a Rasch model can be used. At the kth step of the algorithm, the score $S_n^{(k*)}$ is calculated with the J_k items of the kernel and the new item.

The likelihood used to calculate the AIC is in this case similar to the one presented in the equation (17.9), and the number of parameters is $K_1^{(k)} = 2J_k + 3$.

17.5.4. *A model where the new item is not explained by the same latent trait as the kernel*

The likelihood of the responses to the J_k items of the kernel of the nth individual is similar to the one presented in the equation (17.9) and the corresponding log-likelihood is denoted $l_{kernel}^{(k)}$ at the kth step of the algorithm.

The likelihood of the response to the new item of the nth individual is estimated by:

$$P(X_{n0} = x_{n0}/\delta_0) = \frac{\exp\left[x_{n0}\left(-\delta_0\right)\right]}{1 + \exp\left(-\delta_0\right)}. \tag{17.11}$$

The estimation of the δ_0 parameter can be obtained by maximizing:

$$l_{C0}(\delta_0/\boldsymbol{x_0}) = \log \prod_{n=1}^{N} P(X_{n0} = x_{n0}/\delta_0). \tag{17.12}$$

We easily obtain $\hat{\delta}_0 = -\log\left(\frac{t_0}{N-t_0}\right)$ where $t_0 = \sum_{n=1}^{N} x_{n0}$.

The log-likelihood of the model at the kth step of the algorithm is evaluated by

$$l_2^{(k)} = l_{kernel}^{(k)} + l_{C0}(\delta_0/\boldsymbol{x_0}) \tag{17.13}$$

The number of parameters is in this case $K_2^{(k)} = 2J_k + 2$.

17.5.5. *Selection of the new item in the scale*

The new item 0 is selected in the scale at the end of the step if $AIC_1^{(k)} \leq AIC_2^{(k)}$.

17.6. A small set of simulations to compare Raschfit and Raschfit-fast

17.6.1. *Parameters of the simulation study*

To test the Raschfit procedure, Hardouin and Mesbah [HAR 04] have developed simulations based on a design proposed in [ABS 04]. Simulated data are unidimensional or bidimensional, and 2,000 individuals are used. The parameters used in the simulations are:

– the model used to simulate the data;

– the structure of the data;

– the correlation between the two latent traits;

– the number of items in each dimension;

– the discriminating power of the items of each dimension.

17.6.1.1. *Models*

The model used to simulate the data is a multidimensional counterpart of the five parameters accelerating model (5-PAM) which considers five parameters for each item: the difficulty (δ_j^*), the discriminating power (α_j^*), the random response to the item (γ_j^{low}), the maximal probability to respond to the item (γ_j^{up}), and an accelerating coefficient (ξ_j) which is a coefficient of dissymmetry of the IRF. The IRF of this model is:

$$\begin{aligned} P(X_{nj} = x_{nj}/\boldsymbol{\theta}_n; \delta_j^*, \boldsymbol{\alpha}_j^*, \gamma_j^{low}, \gamma_j^{up}, \xi_j) \\ = \gamma_j^{low} + \left(\gamma_j^{up} - \gamma_j^{low}\right) \left[\frac{\exp\left(1.7x_{nj}(\sum_{q=1}^{Q}(\alpha_{jq}^*\theta_{nq}) - \delta_j^*)\right)}{1+\exp\left(1.7\left(\sum_{q=1}^{Q}(\alpha_{jq}^*\theta_{nq}) - \delta_j^*\right)\right)} \right]^{\xi_j}, \end{aligned} \tag{17.14}$$

with $0 \leq \gamma_j^{low} < \gamma_j^{up} \leq 1$, $\alpha_{jq}^* > 0$ and $\xi_j > 0$. We name this model a M5-PAM (for multidimensional 5-PAM). If we have $\gamma_j^{low} = 0$, $\gamma_j^{up} = 1$ and $\xi_j = 1 \; \forall j$, the model is a multidimensional couterpart of the 2 parameter logistic model (noted M2-PLM).

The parameters $(\alpha_{jq}^*, \delta_j*)$ are calculated in order to obtain Item Characteristics Curves with the same maximal slope ($\frac{1.7\alpha_{jq}}{4}$) and the same localization of this maximal slope ($\frac{\delta_j}{\alpha_{jq}}$) on the qth component of the latent trait whatever the given value for (α_{jq}, δ_j).

In the simulations, we consider a bidimensional latent trait, so $Q = 2$. The components of the latent trait which influence the response to item j are indexed by q_j and $\overline{q_j}$.

The responses of a given item j can only be influenced by θ_{q_j}. In this case, $\alpha_{j\overline{q_j}} = 0$. This case corresponds to a simple structure (SS) [ZHA 99].

However, the responses of a given item j can be mainly influenced by one main component of the latent trait and weakly by the other ($0 < \alpha_{j\overline{q_j}} << \alpha_{jq}$). This case corresponds to an approximate simple structure (ASS) [ZHA 99]. In the simulations, we use $\alpha_{j\overline{q_j}} = 0.2$.

We consider four cases described in Table 17.1.

Case	$\alpha_{j\overline{q_j}}$	Structure	$(\gamma_j^{low}, \gamma_j^{up})$	ξ_j	Model
I	0	SS	(0, 1)	1	MMSRM
II	0.2	ASS	(0, 1)	1	M2-PLM
III	0	SS	(0.1, 0.9)	2	M5-PAM
IV	0.2	ASS	(0.1, 0.9)	2	M5-PAM

Table 17.1. *Values of the parameters used in the simulations*

We note that in cases I and II, we have $\alpha_{jq}^* = \alpha_{jq}, \; \forall q$ and $\delta_j^* = \delta_j$.

17.6.1.2. *Simulation of the multidimensional latent trait*

The two latent traits are simulated by a centered standardized multinormal distribution. The correlation coefficient between the two latent traits is denoted ρ and can take six different values: 0.0, 0.2, 0.4, 0.6, 0.8 and 1.0.

17.6.1.3. *The number of items in the two dimensions*

We use two different sizes for each dimension. These values correspond to a mean value of the number of items by dimension in quality of life questionnaires (seven items) and to a big value of this number (14 items). Three designs are used: seven items in each dimension, or seven items in one dimension and 14 in the other. The values used for the difficulty parameters in the simulations are chosen in the 2-PLM and in each dimension q as the $l/(J_q+1),\ l=1,\dots,J_q$ percentiles of a standardized centered Gaussian distribution.

17.6.1.4. *The discriminating powers of the items*

The same value α_q is used for the discriminating power of all the items mainly relying on the same dimension q. Three different values are used in the simulations for the parameters α_q, $q=1,\dots,Q$: a low value (0.4), a medium value (0.7) and a high value (1.7). Since we simulate 2 dimensions, we obtain six different designs by crossing two by two these three values.

17.6.1.5. *Description of the four main cases*

Data simulated in case I are equivalent to data simulated with an MMSRM. In this case, we can write the IRF of the jth item as:

$$\begin{aligned} P(X_{nj}=x_{nj}/\boldsymbol{\theta}_n;\delta_j,\boldsymbol{\alpha}_j) &= \frac{\exp\left(1.7x_{nj}(\alpha_{jq_j}\theta_{nq_j}-\delta_j)\right)}{1+\exp\left(1.7\left(\alpha_{jq_j}\theta_{nq_j}-\delta_j\right)\right)} \\ &= \frac{\exp\left(x_{nj}(\tilde{\theta}_{nq_j}-\tilde{\delta}_j)\right)}{1+\exp\left(\tilde{\theta}_{nq_j}-\tilde{\delta}_j\right)}, \end{aligned} \qquad (17.15)$$

with $\tilde{\theta}_{nq_j}=1.7\alpha_{jq_j}\theta_{nq_j}$ and $\tilde{\delta}_j=1.7\delta_j$. This expression is equivalent to equation (17.3). $\tilde{\boldsymbol{\theta}}_n=(\tilde{\theta}_{n1},\tilde{\theta}_{n2})'$ is in this case distributed as a centered multinormal distribution with a covariance matrix

$$\tilde{\Sigma}=\begin{pmatrix} (1.7\alpha_1)^2 & 1.7^2\alpha_1\alpha_2\rho \\ 1.7^2\alpha_1\alpha_2\rho & (1.7\alpha_2)^2 \end{pmatrix}. \qquad (17.16)$$

This case is interesting when we study the results of the procedure when the model underlying the data is an MMSRM, that is to say, the model used by the procedure.

Case II reveals the behavior of the procedures when the IRF of the items have the same general form as the MMSRM but when the structure is less particular (the SS is a very rare structure in practice, so the ASS is a likelier structure of real data). These cases make it possible to see if the introduction of a minor latent trait strongly affects in practice the notion of sufficiency of the score on the (main) latent trait.

Cases III and IV study the results of the procedure when the IRF are different from the supposed IRF. The results are more difficult to analyze because the underlying notion of sufficiency of the score on the latent trait is not verified in this model. These simulations consider the cases where the unidimensionality is the main link between the items.

17.6.1.6. *The number of simulations*

By crossing these five factors, we obtain 360 designs. Each of them is simulated one at a time.

17.6.2. ***Results and computing time***

17.6.2.1. *Tested procedures*

The simulated data are treated by four procedures:

– the Raschfit procedure;

– the Raschfit-fast procedure;

– a Mokken scale procedure (MSP) [HEM 95], which builds scales of items which verify the properties of the IRT by using a non-parametric IRM;

– a hierarchical cluster analysis on conditional measures of proximity (HCAC-CPROX), which clusters together the items having greater proximity (based on the conditional covariance between the items).

17.6.2.2. *MSP*

MSP is a procedure described in [HEM 95]. This procedure searches to build scales which verify a Mokken scale, that is to say, a scale verifying the fundamental assumptions of the IRT (unidimensionality, local independency and monotonicity).

The Mokken scale is a non-parametric model and necessitates fixing a threshold (c) as the minimum acceptable value for the indices used (the Loevinger H indices [LOE 48]). The authors of this procedure suggest choosing $c \geq 0.3$, this minimum value being used in the simulations.

17.6.2.3. *HCA/CCPROX*

HCA/CCPROX is defined in [ROU 98]. This method is based on the same methods as the classical HCA: a proximity matrix is defined and a each step, the two closest elements among all those defined at the preceding step are clustered together, until only one cluster is obtained. The authors defined, in the field of the IRT, three specific proximity matrices based on a weighted sum of the covariance, correlation or odds-ration calculated for each value of the score. They show with simulations that this method gives better results than classical measures of proximity with IRT items. The DETECT indice is used to chose the number of clusters of items (the partition which presents the minimal value for this indice is chosen). In the simulations, the distance based on the conditional covariances between the items and the WPGMA method of aggregation are used.

17.6.2.4. *Clustering of the results*

Let a major error in classification be defined as two items which have been simulated from two distinct dimensions and which are classified together. When we simulate two dimensions with a perfect correlation ($\rho = 1$: unidimensional case), a major error in classification is the classification which finds the original dimension of each item.

Each result is designated to a class among the following five:

– Class 1: the true classification of the items is found,

– Class 2: less than two items (for dimensions with $J_q = 7$) or three items (for dimensions with $J_q = 14$) are not classified in the two main dimensions,

– Class 3: the true classification of the items is not found but there is no major error of classement,

– Class 4: there is one or several major error(s) of classement,

– Class 5: unspecified results: at least two items (for dimensions with $J_q = 7$) or three items (for dimensions with $J_q = 14$) are unselected by the procedure (only for MSP).

17.6.2.5. *Computing time*

The average, the minimum and the maximum computing times for each procedure are presented in Table 17.2. The values depend on the number of items. The computer is cadenced at 950MHz with 512Mo of RAM.

Table 17.2 shows that Raschfit, in its original version, is a very long process and is unadapted in practice (on average, five hours to run the procedure with seven items in each dimension and 11 hours to run it with seven and 14 items in the two dimensions). Compared to Raschfit, Raschfit-fast reduces the computing time by a factor of 60 for seven items in each dimension, and by a factor of 30 for seven and 14 items in the two dimensions.

	Number		Computing time in seconds		
Procedure	of items	Average	Standard error	Minimum	Maximum
Raschfit	7;7	18,190	3,384	13,740	39,540
	7;14	38,829	9,291	23,460	64,800
Raschfit-fast	7;7	320	110	120	720
	14;7	1,216	530	240	2,640
MSP	7;7	20	5	5	187
	14;7	89	18	8	382
HCA/CCPROX	7;7	89	5	79	123
	14;7	353	81	287	905

Table 17.2. *Average, minimum and maximum computing time for one simulation for each procedure*

17.6.2.6. *Results of the data simulated by an MMSRM (case I)*

The results are influenced by two main factors: the correlation coefficient between the two components of the latent traits and the fact that the discriminating powers of the items in the two dimensions are either equal ($\alpha_1 = \alpha_2$) or not.

Concerning the correlation coefficient between the two component of the latent traits, we consider three main cases:

– the correlation is low ($\rho \leq 0.4$);

– the correlation is high ($0.6 \leq \rho \leq 0.8$);

– the two simulated latent traits are counfound ($\rho = 1$).

Table 17.3 presents the results of the four tested procedures in all these cases.

When the discriminating powers used in the simulations are the same in the two sets of items, results are good when the correlation coefficient is low ($\rho \leq 0.4$), and continue to be correct when it is medium ($0.6 \leq \rho \leq 0.8$) for Raschfit and HCA/CCPROX.

When the latent trait underlying the two sets of the items is the same ($\rho = 1$), the results are good for Raschfit and Raschfit-fast if the two sets of items have the same discriminating powers (and thus, if the global set of items can be considered as only one Rasch scale), and if the two sets of items have different discriminating powers, these two procedures tend to consider that the latent trait can be measured with two Rasch scales. HCA/CCPROX seems to be sensitive to the discriminating powers of the items, and produces few errors. MSP have success rates in these cases similar to those obtained with a lower value of the correlation coefficient.

Table 17.3. *Results with data simulated by an MMSRM*

Procedure	Results	Equal discriminating power between the two dimensions			Different discriminating power between the two dimensions		
		$\rho \leq 0.4$	$0.6 \leq \rho \leq 0.8$	$\rho = 1.0$	$\rho \leq 0.4$	$0.6 \leq \rho \leq 0.8$	$\rho = 1.0$
Number of simulations		18	12	6	27	18	9
Raschfit	good	14 (78%)	7 (58%)	4 (67%)	24 (89%)	13 (72%)	0
	bad	3 (17%)	5 (42%)	0	2 (7%)	4 (22%)	7 (78%)
Raschfit-fast	good	15 (83%)	0	6 (100%)	26 (96%)	15 (83%)	1 (11%)
	bad	3 (17%)	12 (100%)	0	1 (4%)	3 (17%)	3 (33%)
HCA/CCPROX	good	15 (83%)	8 (67%)	6 (100%)	26 (96%)	10 (56%)	0
	bad	1 (6%)	3 (25%)	0	1 (4%)	2 (11%)	1 (11%)
MSP	good	12 (67%)	3 (25%)	4 (67%)	8 (30%)	0	4 (44%)
($c = 0.3$)	bad	0	6 (50%)	0	0	9 (50%)	0
	unspecified	6 (33%)	2 (17%)	2 (33%)	18 (67%)	9 (50%)	5 (56%)

Procedure	Results	$\rho \leq 0.4$	$0.6 \leq \rho \leq 0.8$	$\rho = 1.0$
Number of simulations		45	30	15
Raschfit	good	27 (60%)	12 (40%)	4 (27%)
	bad	8 (18%)	18 (60%)	2 (13%)
Raschfit-fast	good	24 (53%)	8 (27%)	8 (53%)
	bad	21 (47%)	22 (73%)	1 (7%)
HCA/CCPROX	good	29 (64%)	9 (30%)	0
	bad	9 (20%)	10 (33%)	0
MSP	good	6 (13%)	2 (7%)	2 (13%)
($c = 0.3$)	bad	0	2 (7%)	0
	unspecified	39 (87%)	26 (87%)	8 (53%)

Table 17.4. *Results with data simulated by an M2-PLM*

Procedure	Results	$\rho \leq 0.4$	$0.6 \leq \rho \leq 0.8$	$\rho = 1.0$
Number of simulations		90	60	30
Raschfit	good	49 (54%)	23 (38%)	13 (43%)
	bad	35 (39%)	35 (58%)	6 (20%)
Raschfit-fast	good	59 (66%)	6 (10%)	12 (40%)
	bad	31 (34%)	54 (90%)	0
HCA/CCPROX	good	45 (50%)	17 (28%)	0
	bad	24 (27%)	28 (47%)	2 (7%)
MSP	good	16 (18%)	0	17 (57%)
($c = 0.3$)	bad	27 (30%)	45 (75%)	0
	unspecified	44 (49%)	14 (23%)	4 (13%)

Table 17.5. *Results with data simulated by an M5-PAM*

17.6.2.7. *Results of the data simulated by a M2-PLM (case II)*

When the data are simulated by an M2-PLM with an ASS (see Table 17.4), Raschfit, Raschfit-fast and HCA/CCPROX have similar rates of success when the correlation coefficient is different to 1: a high rate (53% to 64%) when the correlation is low, and a medium rate (27% to 40%) when the correlation is high). MSP have poor rates of success and an important rate of unspecified results (87%).

When the two components of the latent traits are confounded, Raschfit-fast is the procedure which produces the best rate of success (53%), but HCA/CCPROX has the advantage of not producing bad results.

17.6.2.8. *Results of the data simulated by an M5-PAM (cases III and IV)*

Table 17.5 presents the results obtained with data simulated by an M5-PAM.

When data are simulated from an M5-PAM, results are comparable to those obtained with the M2-PLM with a better rate for bad results for Raschfit, Raschfit-fast and HCA/CCPROX. The rate of unspecified results for MSP is lower but unspecified results "become" bad results. The relatively good results obtained with Raschfit and a model very different from the MMSRM can be attributed to the fact that, in the simulations, the parameters are fixed in order to obtain Item Characteristic Curves (ICC) with the same location of the maximum slope and the same value of the maximum slope, so this model cannot be far enough from the MMSRM to create a large disturbance of the algorithm.

17.7. A large set of simulations to compare Raschfit-fast, MSP and HCA/CCPROX

Raschtest-fast seems to give similar results to Raschfit, but the slowness of Raschfit does not fit a large set of simulations with this procedure. In this section, we propose a simulation study with a large set of simulations, to compare Raschfit-fast, MSP and HCA/CCPROX. The simulated cases involve a unidimensional scale (is the procedure able to detect a unidimensional scale? – case A) and disturbance created by the addition of a disturbing item (are the procedures able to detect a bad item? – cases B-E).

17.7.1. *Parameters of the simulations*

We simulate uni or bidimensional data.

We define four cases indexed from A to E. In each case, the correlation between the two components of the latent trait is fixed ($\rho = 0.0$ for B, $\rho = 0.2$ for C, $\rho = 0.4$ for D, $\rho = 0.6$ for E). Case A is the unidimensional case: no item relies on the second component of the latent trait.

In each case, seven items are simulated relative to the first component of the latent trait, and one item is simulated relative to the second component (except in case A where there is no such item). The aim is to determine if the procedures can detect one bad item in a set of items (in case A, the aim is to see if the procedures detects bad items in a set of unidimensional items).

We define four scenarios, numbered I to IV, following the average of the discriminating power of the items: $\mu_{\alpha_I} = .5$, $\mu_{\alpha_{II}} = 1$, $\mu_{\alpha_{III}} = 2$, and $\mu_{\alpha_{IV}} = 3$. The discriminating power of each item is randomly defined among a normal distribution of standard error 0.2.

The difficulty of the items is taken as the percentiles of the standardized Gaussian distribution. The latent trait is simulated by a multinormal centered distribution, with the identity matrix as covariance matrix.

800 replications are simulated in each case.

17.7.2. *Discussion*

Raschfit-fast obtains good results as soon as the correlation between the two components of the latent trait is low, and/or the discriminating power of the items is high. When the set of items is unidimensional, Raschfit-fast detects that there is no disturbing item. Indeed, the quality of the results strongly decreases with the increasing correlation between the two components of the latent trait.

MSP produces many unspecified results, notably when the discriminating powers of the items are low (less than or equal to 1). In the others cases, the results are correct, except if the discriminating powers of the items are too high (greater than 1) and if the correlation between the two components of the latent trait are too correlated (coefficient greater than 0.4).

HCA/CCPROX produces good results when the discriminating powers of the items are high (greater than 1), otherwise, medium results.

Raschfit-fast, in the cases where the real model is close to MMSRM, is the more powerful procedure among these three ones, notably when the conditions are less favorable: high correlation between the components of the latent traits and/or low discriminating power of the items.

17.8. The Stata module "Raschfit"

We propose a Stata module named -raschfit- to carry out the Raschfit procedure. By default, this module runs Raschfit-fast. The Stata module -raschfit- can be downloaded from the FreeIRT Project.

The syntax of -raschfit- is simple. The user indicates the names of the items used. By default, MSP is run under the items to order in a negative order, and the two first items selected by MSP are considered as the initial kernel of the scale. The other items are ordered with MSP from the last item selected by MSP to the first one (except the kernel).

It is possible to modify the method to order the items with the "itemsorder" option which can be "msp" (by default), "mspinv" (the kernel is also selected by default, but the other items are taken in the inverse order) and "order" which orders the items in the same order as the one defined by the user.

The number of sub-scales to build is defined by the "nbscales" option (1 by default). The size of the kernel of the first sub-scale can be defined by the "kernel" option (2 by default). Finally, it is possible to run the original version of Raschfit with the “nofast” option.

With the syntax of the Stata manual, the syntax of the -raschfit- module is:

.**raschfit** *varlist* [, **kernel**(*#*) **nbscales**(*#*) **itemsorder**(*keyword*) **nofast**]

17.9. Conclusion

A new procedure named Raschfit had been proposed in another work. It selects the items which fit a Rasch model. This procedure is based on the fit of the data to a multidimensional IRM, instead of on the correlations between items (as in the factor analysis) or on the properties of the items (as in the MSP).

Raschfit performs better than the existing procedures which are based on the unidimensionality of the items, especially when the multidimensional model underlying the data is close to one used in this procedure. The main drawback of Raschfit is the computing time (several hours, even when the number of items is small). A new version of this procedure, named Raschfit-fast, is proposed in this chapter. Raschfit-fast estimates more quickly the likelihood of the models, and considerably reduces the computing time, even if already existing procedures (MSP for example) are still faster. This adaptation of Raschfit gives similar results compared to the former version if the latent traits underlying each set of items have a low correlation.

Although these are encouraging results for this type of procedure, based on the fit to IRM, new improvements are necessary to reduce the rates of bad results.

Case/		Raschfit-fast			MSP				HCA/CCPROX		
scenario		Correct	Medium	Bad	Correct	Medium	Bad	Unspecified	Good	Medium	Bad
A	I	781	19	-	0	0	-	800	0	800	-
	II	800	0	-	26	221	-	553	0	800	-
	III	800	0	-	800	0	-	0	800	0	-
	IV	800	0	-	800	0	-	0	800	0	-
B	I	625	0	175	0	0	0	800	0	775	25
	II	800	0	0	23	73	0	704	2	789	9
	III	800	0	0	800	0	0	0	799	1	0
	IV	800	0	0	800	0	0	0	800	0	0
C	I	562	0	238	0	0	0	800	0	767	33
	II	799	0	1	29	97	0	674	1	786	13
	III	800	0	0	800	0	0	0	800	0	0
	IV	800	0	0	800	0	0	0	800	0	0
D	I	469	0	331	0	0	0	800	0	755	45
	II	712	0	88	22	107	0	671	0	770	30
	III	800	0	0	800	0	0	0	795	5	0
	IV	800	0	0	780	0	20	0	800	0	0
E	I	259	0	541	0	0	0	800	0	733	67
	II	186	0	614	30	63	0	0	0	753	47
	III	187	0	613	495	0	305	0	767	32	1
	IV	702	0	98	1	0	799	0	800	0	0

Table 17.6. *Results of the simulations concerning the comparison of Raschfit-fast, MSP and HCA/CCPROX*

17.10. Bibliography

[ABS 04] VAN ABSWOUDE A. A. H., VAN DEN ARK L. A., SITJSMA K., “A comparative study on test dimensionality assessment procedures under non-parametric IRT models”, *Applied Psychological Measurement*, vol. 28, p. 3–24, 2004.

[ADA 97] ADAMS R. J., WILSON M. R., WANG W., “The multidimensional random coefficient multinomial logit model”, *Applied Psychological Measurement*, vol. 21, p. 1–23, 1997.

[AND 70] ANDERSEN E. B., “Asymptotic properties of conditional maximum likelihood estimators”, *Journal of the Royal Statistical Society, Series B*, vol. 32, p. 283–301, 1970.

[AND 77] ANDERSEN E. B., “Sufficient statistics and latent trait models”, *Psychometrika*, vol. 42, num. 1, p. 69–81, 1977.

[FIS 95] FISHER G. H., MOLENAAR I. W., *Rasch Models. Foundations, Recent Developments, and Applications*, New York: Springer edition, 1995.

[HAR 04] HARDOUIN J. B., MESBAH M., “Clustering binary variables in subscales using an extended Rasch model and Akaike Information Criterion”, *Communications in Statistics – Theory and Methods*, vol. 33, num. 6, p. 1277–1294, 2004.

[HEM 95] HEMKER B. T., SITJSMA K., MOLENAAR I. W., “Selection of unidimensional scales from a multidimensional item bank in the polytomous Mokken IRT model”, *Applied Psychological Measurement*, vol. 19, p. 337–352, 1995.

[HOI 97] HOIJTINK H., MOLENAAR I. W., “A multidimensional item response model : constrained latent class analysis using the Gibbs sampler and posterior predictive checks”, *Psychometrika*, vol. 62, num. 2, p. 171–189, 1997.

[LIN 97] VAN DEN LINDEN W. J., HAMBLETON R. K., *Handbook of Modern Item Response Theory*, New York: Springer-Verlag edition, 1997.

[LOE 48] LOEVINGER J., “The technique of homogeneous tests compared with some aspects of scale analysis and factor analysis”, *Psychological Bulletin*, vol. 45, p. 507–530, 1948.

[RAS 60] RASCH G., *Probabilistic Models for Some Intelligence and Attainment Tests*, Copenhagen: Nielsen & Lydiche edition, 1960.

[ROU 98] ROUSSOS L. A., STOUT . F., MARDEN J. I., “Using new proximity measures with hierarchical cluster analysis to detect unidimensionality”, *Journal of Educational Measurement*, vol. 35, num. 1, p. 1–30, 1998.

[ZHA 99] ZHANG J., STOUT W. F., “The theorical DETECT index of dimensionality and its application to approximate simple structure”, *Psychometrika*, vol. 64, num. 2, p. 213–249, 1999.

This page intentionally left blank

Chapter 18

Analysis of Longitudinal HrQoL using Latent Regression in the Context of Rasch Modeling

18.1. Introduction

For the last ten years, both in clinical tests and in epidemiological studies, the health-related quality of life (HrQoL) has been considered a very important element in evaluating the state of health and, therefore, in defining the most suitable treatment. In particular, in the context of palliative care for terminal cancer patients, HrQoL is no longer a secondary end-point of the treatment (survival having been the primary end-point), but is now the primary end-point. Therefore, the detection of good measurement and analysis tools is important to provide information at clinicians' disposal to facilitate the therapy decisional process.

From a methodological point of view, decisions on two different but connected aspects have to be taken into account for the study of a latent variable, such as HrQoL. The first is the definition of a suitable method of measurement to translate the qualitative information coming from a set of observable items partial indicators of the latent variable, into quantitative information. The second is the detection of a suitable statistical methodology to explain the latent variable.

The analysis presented in this chapter is developed in the framework of the Rasch-family models ([RAS 60]; [FIS 95b]) in relation to the former aspect and in the framework of multilevel models for repeated measures [SNI 99] with regards to latter aspect.

Chapter written by Silvia BACCI.

In the most common approach, the latent variable is estimated by means of a measurement model (for instance, a Rasch model) and then the estimates are used in a regression model as observed values of the response variable. However, some elements suggest caution toward a two separate steps approach: the bias and the inconstistency of the latent variable estimates ([GOL 80]; [LOR 84]), the underestimation of the true association between latent variable and covariates [MES 04] and, more generally, the lack of flexibility and integration between applied psychometricians and statisticians [WIL 04]. A possible solution is represented by a global approach that integrates the measurement phase and the analysis phase into the same model: in this case we talk of latent regression [AND 77]. The linear latent regression Rasch model has already been studied for several years [ZWI 91], but longitudinal data has developed recent interest.

In this chapter, a longitudinal latent regression model that uses the Rasch analysis as a measurement tool is proposed for the first time. It consists of a random intercept and slopes logistic model with covariates at the second aggregation level. The implementation of the model is performed in SAS, through the Nlmixed procedure. In such a way, it is possible to achieve better flexibility for the utilization by users of general statistical softwares.

In section 18.2 a brief review about global models for longitudinal data is presented, whereas in section 18.3 the latent regression model for repeated measures is developed. In particular, attention is focused on the model structure, on the correlation matrix, on the estimation and on the computational implementation. In section 18.4 the results of a case study on HrQoL of terminal cancer patients under palliative care are shown. The applied analysis is specially focused on the time effect and on the significance of the baseline condition and of other variables on HrQoL. Finally, some concluding remarks are made.

18.2. Global models for longitudinal data analysis

Submitting the same questionnaire to the same set of people repetitively over time is a rather common situation. The objective is to monitor the trend of a latent variable to understand if significant changes occur; then, it raises the problem to detect variables that explain these changes. In a context of this kind, the data structure is formed by observations for every person on more than one item, and, for every item, on more than one time point. In the global approach framework, methodological developments are concerned with three different types of models, though case studies are rather infrequent and related to rather simple examples: the linear logistic model with relaxed assumptions (LLRA), the multidimensional Rasch model and the three-level regression model.

– *Linear logistic test model with relaxed assumption* (LLRA)

Fischer [FIS 95a] formalized the measurement of change by means of the concept of "virtual" items. Any change of the latent variable occurring between two time points can be described for every person as a change of the item parameters. An item I_j given to the same person at two different time points, T_1 and T_2, can be considered as a pair of "virtual" items, I_a^* and I_b^*, associated with a pair of "virtual" difficulty parameters, β_a^* and β_b^*. If the amount of change between T_1 and T_2 is equal to a constant δ (overall people), then the pair of "virtual" items generated by real item I_j with difficulty β_j is characterized by the two parameters: $\beta_a^* = \beta_j$ and $\beta_b^* = \beta_j + \delta$. Thus, the item parameters are a linear combination of the real item parameter and the change effect. This kind of model belongs to linear logistic test model (LLTM) family. Substituting $(\theta_i - \beta_j)$ with only one parameter θ_{ij} a generalization of LLTM is obtained: it is named the linear logistic model with relaxed assumptions and it is a LLTM with two (in the case of two repeated measures) items for every person.

– *Multi-dimensional Rasch model*

Wang and Chyi-In [WAN 04] model the latent variable change over time using a multi-dimensional Rasch model. They discern an initial ability parameter and a modifiability parameter for each occasion following the first one: all of these parameters are considered as different latent dimensions which describe the change among adjacent time points.

– *Three-level regression model*

Pastor and Beretvas [PAS 06] start the Rasch model formulation as a two-level model, where the item responses are the first-level units and individuals are the second-level units. Then, they put beside this model a longitudinal regression model: it is also a two-level model, where measurement occasions are the first-level units and individuals are still the second-level units. Putting together the two models, a three-level model results with item responses aggregated in measurement occasions and measurement occasions aggregated in individuals. The output is a model that explains the change, where what is changing is the latent variable and, under every second-level unit, there is a measurement model.

Among the three mentioned models, LLRA and the multi-dimensional Rasch models are not very suitable for complex data structure. Indeed, for both of them the number of parameters increases by increasing the number of measurement occasions for every person. In detail:

– LLRA: for each real item there are as many virtual items as measurement occasions; moreover, every virtual parameter is a linear combination of the real item parameter and of a change effect. Thus, if change effect is constant for all items, the number of parameters is equal to the number of measurement occasions; otherwise, if the change effect varies among items, the number of parameters is equal to the product of the number of measurement occasions and the number of real items.

– Multi-dimensional Rasch model: the latent ability for every measurement occasion is a dimension; moreover, for every pair of dimensions three parameters have to be estimated, i.e. two variances and one correlation coefficient. Therefore, every time point added to the model causes new parameters: a variance and a correlation coefficient between the new dimension and each of the previous ones.

The multilevel model is, then, the most flexible, though the three-level structure is computationally heavy for the estimation process. Moreover, though the model does not become more complicated when the measurement occasions increase and it allows different people to have a different number of time points, an insignificant number of observations for every person (that is an insignificant number of second-level units) may cause inaccurate estimates and high standard errors.

18.3. A latent regression Rasch model for longitudinal data analysis

18.3.1. *Model structure*

To remedy the above stated problems, an alternative model is proposed to analyze longitudinal data in the context of Rasch measurement models. The framework of reference is the latent regression: the aim is to model the latent variable θ as a function of the time and of other covariates. Given the multilevel structure of longitudinal data (measurement occasions are first-level units and individuals are second-level units), a (latent) regression model with random intercept is estimated to take into account the variability among individuals. Moreover, it is reasonable to also assume that the time effect – in the clinical context it is often the effect of therapy – is variable among people. In the end, the dependent variable, i.e. the latent ability, is assumed to be continuously (in particular, normally) distributed.

Let us denote by $i = 1, 2, \ldots, N$ the individuals; t the measurement occasion; θ_{it} the "ability" parameter (i.e. the HrQoL level) for i-th person at t time; $f(t)$ any function of t (e.g. a linear or a quadratic function); z_i the value of covariate z assumed by the i-th person; α the regression coefficient of z_i; δ_{0i} and δ_{1i} the random intercept and random coefficient, respectively; γ_{00} and γ_{11} the fixed components of δ_{0i} and δ_{1i}, respectively; u_{0i} and u_{1i} the random components of δ_{0i} and δ_{1i}, respectively. In more detail, the random intercept u_{0i} explains how the initial level (i.e. for $t = 0$) of θ_{it} of each patient differs from the average population value. Additionally, the random coefficient u_{1i} represents how the individual time effect differs from the average population time effect. A linear multilevel model is described by the following equations:

$$\left\{\begin{array}{lcl} \theta_{it} & = & \delta_{0i} + \delta_{1i} \cdot f(t) + \alpha \cdot z_i, \\ \delta_{0i} & = & \gamma_{00} + u_{0i}, \\ \delta_{1i} & = & \gamma_{11} + u_{1i}, \end{array}\right. \tag{18.1}$$

$$\Updownarrow$$

$$\theta_{it} = (\gamma_{00} + \gamma_{11} \cdot f(t) + \alpha \cdot z_i) + (u_{0i} + u_{1i} \cdot f(t)).$$

With reference to the second member of equation (18.1), the terms in the first brackets define the fixed part of the model, that is, the mean μ_θ of θ_{it}, whereas the terms in the second brackets define the random part of the model. The structure of the random effects is:

$$\begin{bmatrix} u_{0i} \\ u_{1i} \end{bmatrix} \sim Normal \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{u0} & \sigma_{u01} \\ \sigma_{u01} & \sigma^2_{u1} \end{pmatrix} \right] \tag{18.2}$$

With respect to other multilevel linear models, this is different because the response variable is not observed: hence, it has to be estimated by means of a Rasch model. For dichotomous items, the Rasch model is given by the following equation [FIS 95b], where the subscript t is introduced here for coherence with equation (18.1):

$$\begin{array}{c} P(X_{jti} = 1|\theta_{it}; \beta_j) = \frac{exp[\theta_{it} - \beta_j]}{1 + exp[\theta_{it} - \beta_j]} \\ \Updownarrow \\ logit[P(X_{jti} = 1|\theta_{it}; \beta_j)] = \log\left[\frac{P(X_{jti}=1|\theta_{it}, \beta_j)}{1 - P(X_{jti}=1|\theta_{it}, \beta_j)}\right] = \theta_{it} - \beta_j, \end{array} \tag{18.3}$$

where: i is the i-th person ($i = 1, 2, \ldots, N$); j is the j-th item ($j = 1, 2, \ldots, J$); x_{jti} is the response category of the j-th item, i.e. 0 or 1, chosen by the i-th person at t-th time; and β_j is the difficulty parameter of j-th item.

If θ_{it} is substituted in the Rasch model with the structural model of equation (18.1), the following *longitudinal latent regression model* (LLRM) is obtained:

$$\begin{aligned} &logit[P(X_{jti} = 1|z_i, u_{0i}, u_{1i}; \gamma_{11}, \alpha, \mathbf{\Sigma}, \beta_j)] \\ &\qquad = (\gamma_{00} + \gamma_{11} \cdot f(t) + \alpha \cdot z_i - \beta_j) + (u_{0i} + u_{1i} \cdot f(t)), \end{aligned} \tag{18.4}$$

where $\mathbf{\Sigma}$ is the variance and covariance matrix of the latent variable. In order for the model to be identifiable, the intercept γ_{00} of the structural model has to be fixed at 0; alternatively, one of the difficulty parameters may be fixed and so γ_{00} is free.

The proposed model is still a random intercept and slope model; it is not linear, but logistic, and it allows a direct estimation of parameters from observed item responses. Several generalizations are possible. First of all, more than one covariate can be added and the computational heaviness does not increase substantially. Similarly, for polytomous items, the partial credit model[1] [MAS 82] is simple to use by means of

[1] The formula of partial credit model is

$$P(X_{itj} = x_{ij}|\theta_{it}, \beta_{jx}) = \frac{exp[\sum_{k=0}^{x_{ij}} (\theta_{it} - \beta_{jk})]}{\sum_{h=0}^{H_j} exp \sum_{k=0}^{h} (\theta_{it} - \beta_{jk})}, \tag{18.5}$$

where $x_{ij} = 0, 1, \ldots, h, \ldots, H_j$ and β_{jk} means the difficulty level of the k-th threshold of the j-th item.

substituting equation (18.1) in equation (18.5), that is, substituting β_j with β_{jk} parameters. Moreover, the model can be used to estimate the differential item functioning (DIF) effect on one or more items, which describes a different functioning of items among different groups of individuals. It is sufficient to add an interaction effect to items suspected of DIF: i.e. β_j has to be substituted by $\beta_j \cdot z_{0i}$, where z_{0i} is a dummy variable that indicates the belonging of the i-th person to a group of people (e.g. males vs females). Moreover, by substituting β_j with a random coefficient β_{ji}, we can test the hypothesis that a two-parameter logistic model [BAK 04] describes the data better: β_{ji} has a fixed component, which denotes the average difficulty of item for the population, and a random component, which denotes how different the difficulty of item is for the i-th person. Still, if an interaction effect between a difficulty parameter β_j and the time variable is added, the time stability of questionnaire can be evaluated. It is important to remind that, in order for a questionnaire to be valid, it must be stable, that is, each item must give a contribution to measure the same latent variable in every time point.

Output from the LLRM model is similar to output from "classical" longitudinal models. If the main purpose of analysis is an evaluation of time effect on the latent variable, attention is on the second level residual estimates (the $u_{1i}s$). The $u_{1i}s$ show how time effect on the latent variable for a specific person diverges from the average effect for the whole population. For example, in a clinical study, monitoring the impact of therapy on patients could be interesting, so the time effect describes the effect of the therapy (if the questionnaire is filled out repeatedly during therapy) and an $u_{1i}s$ residuals analysis detects individual characteristics of patients with positive or negative reactions. This kind of information can be useful to produce a classification of patients based on individual characteristics to anticipate (before therapy begins) the impact therapy will have on the HrQoL of every patient.

18.3.2. *Correlation structure*

To facilitate the analysis and the understanding of the longitudinal latent regression model, it is useful to know the correlation structure of the latent variable θ_{it}. By recalling equation (18.1), the variances and covariances matrix elements are promptly defined:

$$\begin{aligned}
&Var(\theta_{it}|z_i) = \\
&\qquad\qquad = Var[\gamma_{00} + \gamma_{11}\cdot f(t) + \alpha\cdot z_i) + (u_{0i} + u_{1i}\cdot f(t)] = \\
&\qquad\qquad = Var[u_{0i} + u_{1i}\cdot f(t)] = \\
&\qquad\qquad = \sigma^2_{u0} + f^2(t)\cdot\sigma^2_{u1} + 2\cdot f(t)\cdot\sigma_{u01}; \\
&Cov(\theta_{it}, \theta_{i't'}|z_i) = \\
&\qquad\qquad = Cov[(u_{0i} + u_{1i}\cdot f(t); (u_{0i'} + u_{1i'}\cdot f(t')] = \\
&\qquad\qquad = \{\sigma^2_{u0} + f(t)f(t')\sigma^2_{u1} + \sigma_{u01}[f(t) + f(t')]\}\cdot I_{(i=i')}.
\end{aligned}$$

These formulae show that the variance and covariances (and, therefore, the correlations) of the latent variable depend only on the measurement occasion: this means that the correlation between the measurements of the i-th person's latent variable at two different time points changes with the change of the time lag and, in particular, it decreases when the time lag increases (as is shown in the applied analysis in Table 18.2). Naturally, coherently with the assumptions of the model, the correlation between the measurements of the latent variable for two different individuals is equal to 0.

18.3.3. *Estimation*

There are three main approaches to estimate a Rasch model [BAK 04]: joint maximum likelihood (JML), conditional maximum likelihood (CML) and marginal maximum likelihood (MML). In JML and CML person parameters are considered fixed effects, whereas in MML they are assumed random and independent variables drawn from a density distribution that describes the population. The latent regression point of view considers the latent ability as a random variable: thus, the most suitable approach is the MML. The log-likelihood function is obtained integrating out the conditional log-likelihood on the random variable; then, the log-likelihood for the longitudinal latent regression model is the following one:

$$\log L(\boldsymbol{\beta}, \gamma_{11}, \alpha, \boldsymbol{\Sigma}) = \sum_{i=1}^{N} \log \int_{\Theta} P(\mathbf{x_{it}}|\theta_{it}; \beta) \cdot \phi(\theta_{it}|\gamma_{11}, \alpha, \boldsymbol{\Sigma}) d\theta_{it}, \qquad (18.6)$$

where $\mathbf{x_{it}}$ is the response pattern for the i-th person at t time, $\boldsymbol{\beta}$ is the difficulty parameters vector and $P(\mathbf{x_i}|\theta_{it}; \boldsymbol{\beta})$ is the probability of $\mathbf{x_{it}}$:

$$P(\mathbf{x_{it}}|\theta_{it}; \boldsymbol{\beta}) = \prod_{j=1}^{J} \frac{exp[x_{jti}(\theta_{it} - \beta_j)]}{1 + exp(\theta_{it} - \beta_j)}.$$

Because the integral in the marginal log-likelihood does not have a closed solution, a numerical integration method has to be used. In this chapter the adaptive Gaussian quadrature was adopted. Finally, the estimation of random effects is usually based on the posterior distribution of the latent variable: the mean or the mode of this distribution is used as point estimates of u_{0i} and u_{1i} (the empirical Bayes estimates).

18.3.4. *Implementation with SAS*

One of the main problems about the utilization of measurement models by common-users is that on one side traditional computational tools do not have specialized routines to estimate these kinds of models, and on the other side devoted software are hardly configurable. To remedy these problems, the most recent research is addressed to study the potential of generic statistical softwares, primarily SAS (for instance, [DOR 03]; [WIL 04]; [HAR 07]) and Stata [RAB 04]. So, it is possible to estimate Rasch models without using compulsorily devoted software, and to estimate generalized Rasch models, such as latent regression models, that otherwise are not applicable. This chapter discusses current studies intending to implement in SAS the longitudinal latent regression model proposed in the previous section.

The multilevel structure of the Rasch model – item responses as the first-level and persons as the second-level – makes it possible to use the Nlmixed procedure of SAS [WIL 04]. In comparison with a simple Rasch model, the longitudinal latent regression model is still a multilevel logistic model with two aggregation levels and with a random intercept, but also with a random coefficient to explain the time effect. The random component of the model is not θ, but it is formed by the two random effects u_0 and u_1, that appear in the regression model for θ.

For a dichotomous Rasch model, let time be the linear time effect (more in general, any kind of time effect can be introduced), let z be a covariate with fixed effect, let beta1, . . . , betaJ be the difficulty item parameters, let alpha be the regression fixed coefficient of z, let Response be the response vector, let theta be the latent variable, let Individual be the second-level unit, let I1, . . . , IJ be the indicators of items, let s2u0, cu01, s2u1 be the first-level variance, the covariance, the second-level variance, respectively, and let u0 and u1 be the random effects. Then the SAS code to estimate the LLRM is the following:

```
proc nlmixed data = filename qpoints = number;
parms beta1 = num1 beta2 = num2 ... betaJ = numJ alpha=numAlpha s2u0 =
numS2u0 s2u1 = numS2u1 cu01 = numCu01;
meantheta = gamma11 · time + alpha · z;
theta = meantheta + u0 + u1 · time;
eta = theta - (beta1 · I1 + beta2 · I2 + ... + betaJ · IJ);
expeta = exp(eta);
p = expeta / (1 + expeta);
model Response ~ binomial (1, p);
random u0 u1 ~ normal ([0, 0], [s2u0, cu01, s2u1]) subject = Individual out =
residual;
run;
```

Data have to be organized in a matrix so that there is not one line for each person, but rather one line for each person-item-measurement occasion combination. Note that the number of measurement occasions can change for different people.

The model developed in this section can also be implemented in Stata by means of the Gllamm routine [RAB 04]. The analysis in the next section has been implemented with both SAS and Stata: outcomes are the same, but Stata's computational times are longer (about 45 minutes in SAS and 4 hours and 30 minutes in Stata).

18.4. Case study: longitudinal HrQoL of terminal cancer patients

This section presents an application of the LLRM. The analyzed data are supplied by an Italian multicentric study called "Staging", and it concerns 485 terminal cancer patients under domiciliary palliative care. A questionnaire for measuring HrQoL was submitted to every patient both before the beginning of the treatment (at baseline) and afterwards once a week until death or, in most fortunate cases, until the end of the study period[2]. At baseline, numerous (about 40) individual characteristics have been surveyed about the patient, his/her family, his/her house, his/her disease, and his/her clinical situation at the moment of the first visit, in the previous week, in the previous month and in the previous year. The questionnaire is called TIQ (Therapy Impact Questionnaire [TAM 92]). It was implemented in 1987 at Pain Therapy and Palliative Care division of National Cancer Institute in Milano, Italy, to measure HrQoL in cancer patients. The questionnaire is composed of 36 items with four ordinal response categories ("not at all" = high HrQoL, "some", "a lot", "very much" = very low HrQoL). The following analysis is concerned with only a subset of 8 items describing the psychological component of HrQoL: difficulty in performing usual free time activities (diffree), fatigue (fatigue), illness (illness), sad or depressed feelings (sad), difficulty in concentrating or paying attention (difconce), nervousness (nervousness), insecurity (insecurity) and confusion (confusion).

As already mentioned in section 18.1, when the disease is in a terminal phase, the most important end point of the pain therapy is no longer the survival, but the quality of life of the patient. The aim of the survey carried out by the "Staging" study is twofold:

– evaluating the time effect on HrQoL trend during the palliative care;

– understanding if the measurement of HrQoL at baseline gives sufficient information to predict HrQoL during the therapy or, on the contrary, if the estimation at baseline has to be completed with other information on individual characteristics.

The latent regression model was estimated on 285 patients with at least three measurements on quality of life (plus the measurement at baseline). The selection model occurred in subsequent steps, taking into account a different number of random effects and performing several explorative analysis to select the significant covariates.

[2] The number of compiled questionnaires is very variable for different people: the minimum is 1 (only survey at baseline) and maximum is 32 (survey at baseline and 31 weeks under palliative care).

Finally, the following model has been selected:

$$
\begin{aligned}
&logit[P(X_{jti} = 1|u_{0i}, u_{1i}; \mu_\theta, \mathbf{\Sigma}, \boldsymbol{\beta})]\\
&\quad = (0.271 \cdot \sqrt{t} + 0.825 \cdot HrQoL_{base_i} + 0.546 \cdot \textsf{difconce} - 0.308 \cdot \textsf{confusion}\\
&\qquad + 1.997 \cdot \textsf{diffree} + 1.077 \cdot \textsf{illness} + 0.302 \cdot \textsf{insecurity} + 2.487 \cdot \textsf{fatigue}\\
&\qquad\qquad + 1.275 \cdot \textsf{sad} + 0.632 \cdot \textsf{nervousness}) + (u_{0i} + u_{1i} \cdot \sqrt{t})
\end{aligned}
$$

To fully understand the estimated regression coefficients, the dichotomization applied for the item response categories has to be taken into account: 0 for the category "not at all", that is, absence of symptom or disorder, 1 for the categories "some", "a lot" or "very much", that is, the presence of a symptom or disorder. Thus, the probability of category 1 rather than 0 is higher the lower the HrQoL level; and vice versa, patients with a good HrQoL are inclined to choose category 0. Therefore, a high numerical value (both at baseline and after) is related with a low level of HrQoL.

For further details, Table 18.1 shows the estimated values of parameters with standard errors, p-values and 95% confidence intervals; Table 18.2 shows the correlation matrix (for the first 8 measurement occasions). The correlation between the HrQoL of the same patient in two different time points is always highly positive, though it decreases when the time lag increases.

	Estimate	Stand.error	p-value	Lower limit	Upper limit
difconce	-0.546	0.088	$< .0001$	-0.719	-0.373
confusion	0.308	0.873	0.0005	0.137	0.480
diffree	-1.997	0.096	$< .0001$	-2.187	-1.808
insecurity	-0.302	0.088	$< .0001$	-0.474	-0.129
illness	-1.077	0.090	0.0007	-1.253	-0.901
nervousness	-0.632	0.088	$< .0001$	-0.805	-0.458
fatigue	-2.487	0.102	$< .0001$	-2.687	-2.287
sad	-1.275	0.091	$< .0001$	-1.454	-1.097
$\sqrt{t}$	0.271	0.044	$< .0001$	0.186	0.357
$HrQoL_{base}$	0.825	0.062	$< .0001$	0.703	0.947
σ^2_{u0}	0.971	0.138	$< .0001$	0.699	1.242
σ^2_{u1}	0.272	0.041	$< .0001$	0.190	0.351
σ_{u01}	-0.139	0.060	0.0222	-0.257	-0.020

Table 18.1. *LLRM: estimates, standard errors, p-values, confidence intervals* ($\alpha = 0.05$)

The outcomes from the estimated model lead to several considerations. As regards the main objective of analysis (detecting determinants of HrQoL during palliative care), only quality of life at baseline is significant with a positive effect (low

	0	1	2	3	4	5	6	7	8
0	1.000	0.860	0.743	0.650	0.575	0.514	0.463	0.420	0.383
1		1.000	0.978	0.945	0.910	0.878	0.849	0.823	0.800
2			1.000	0.992	0.975	0.956	0.937	0.919	0.903
3				1.000	0.996	0.986	0.975	0.963	0.951
4					1.000	0.997	0.991	0.984	0.976
5						1.000	0.998	0.994	0.989
6							1.000	0.999	0.996
7								1.000	0.999
8									1.000

Table 18.2. *LLRM: correlation matrix (first 8 measurement occasions)*

values of HrQoL at baseline are related with low values of HrQoL during the treatment and vice versa). None of the 40 individual characteristics observed at baseline give significant information on the future trend of HrQoL. Then, about the initial question on what variables have to be gathered at baseline to predict the HrQoL trend, the model shows that the questionnaire filled at the beginning by patient can explain some variability of future HrQoL, whereas the knowledge of further characteristics does not add any relevant information.

The other significant variable is time effect: time affects the numerical value of HrQoL with a quadratic positive effect. In particular, the value of 0.271 for the regression coefficient means that when the time lag increases the numerical value of HrQoL increases (i.e. its level gets worse), with a marginally decreasing rate. The deterioration of HrQoL is explicable with the approach to death. The time effect can be considered as the result of two different, but hardly separable effects: the approach to death and palliative care. Understanding whether palliative care has a positive effect on HrQoL should be an interesting question. Unfortunately, the dataset does not give any aid to answer to this question: the study should be repeated with a control group, in order to compare patients under palliative care and patients under a different kind of therapy.

The fixed time coefficient equal to 0.271 defines the *average* time effect on the population. Second-level residuals (u_{1i}), on the other hand, show how much time effect for the i-th person diverges from the average value of population. Positive values mean an individual time effect greater than average, that is, a more negative impact on quality of life; negative values mean a less negative individual effect and, particularly, values smaller than -0.271 cause an improvement in HrQoL over the time. Figure 18.1 shows a classification of patients based on u_{1i} residuals. For every patient, the residual value and the corresponding confidence interval (based on suggestions of Goldstein and Healy [GOL 95]) are shown: two patients can be considered significantly different when the respective intervals do not overlap. As the graph shows, only individuals with extreme residuals are significantly different from the remaining population. In particular, only 15 patients (called Group 1) show a confidence interval with values smaller than -0.271: in other words, at a confidence level of 95%, HrQoL of these people improves over the time. On the other hand, the confidence interval of 77 patients (called Group 2) contains only values greater than -0.271, so the individual time impact on HrQoL is negative at 95% level. Table 18.3 shows the individual characteristics that have a different distribution between the two groups of patients. From a descriptive point of view, patients from Group 1 are distinguished by a lower percentage of people that live alone and, consequently, by a higher percentage of people that have a partner; more than $1/3$ of them do not have any metastasis versus the 11% of patients from Group 2; moreover, some physical symptoms, such as break-through pain, endocranial hypertension and decubitus lesions, are more rare. Finally, only 36% (versus 59%) are confined to bed and 60% (versus 74%) depend on someone else for the daily living activities. The differences between the two groups of patients are statistically significant only in relation with the absence of metastasis (1% significativity level) and the condition to stay in bed (5% significativity level).

	Group 1		Group 2	
	%	Total	%	Total
Lives alone	0.07	1	0.16	12
Has a partner	0.80	3	0.58	32
No metastasis**	0.36	5	0.11	8
Break-through pain	0.07	1	0.19	13
Endocranial hypertension	0.00	0	0.08	6
Decubitus lesion	0.00	0	0.10	8
Confined to bed*	0.36	5	0.59	44
Dependent for adl	0.60	9	0.74	57
Total		15		77

Table 18.3. *Comparison between patients with a positive individual time effect on HrQoL (Group 1) and patients with a negative individual time effect on HrQoL (Group 2) (** = significativity at 1% level; * = significativity at 5% level)*

The other second-level residual component, u_{0i}, explains the alteration of intercept for the average patient in $t = 0$, that is, after the first week of therapy. The interpretation is similar to $u_{1i}s$: positive values mean an initial HrQoL level worse than the average population and negative values mean a level better than average. The covariance between second-level residuals is negative; the correlation coefficient between u_{0i} and u_{1i} is equal to -0.270. This means that patients with an HrQoL level better than average after the first visit ($t = 0$) have a worse reaction during the continuation of therapy; vice versa, the therapy should be more effective for people in a more dire situation at the beginning. A possible interpretation of negative correlation can be ascribed to the characteristics of palliative care, whose effectiveness level appears better in more crucial situations.

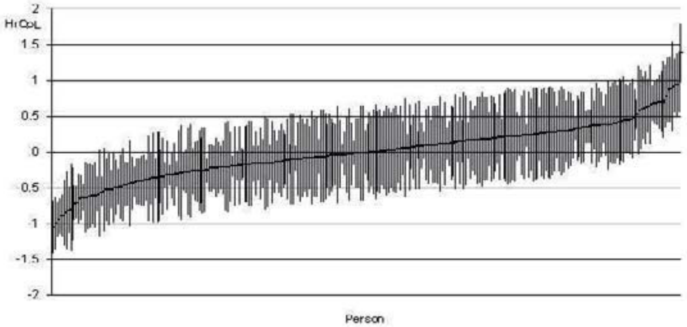


Figure 18.1. *Second-level residuals u_{1i}: estimates and confidence intervals*

In the end, the estimation of the LLRM gives information about the difficulty of items: diffree and fatigue items are the "easiest", that is, they are seen in the majority of patients, and patients that have a good HrQoL level; in contrast, confusion and insecurity items are the most "difficult", that is, only people with a low HrQoL level show them.

To conclude, the analysis, a two step approach was applied, separating the measurement model from the structural model. First of all, the HrQoL of every person at different time points has been estimated, then, these values were used as a dependent variable in a multilevel linear model (see equation (18.1)), with the same covariates of LLRM (without items, obviously):

$$Y_i = (\gamma_{00} + \gamma_{11} \cdot \sqrt{t} + \alpha \cdot HrQoL_{base_i}) + (u_{0i} + u_{1i} \cdot \sqrt{t} + \epsilon_{ji}).$$

Output from the two step approach is similar with outputs coming from other analyses ([ZWI 91] and [ADA 97]): as Table 18.4 shows, both time effect and HrQoL at baseline coefficient estimated by the two steps approach are smaller than the corresponding values estimated by the latent regression model.

18.5. Concluding remarks

This chapter is concerned with the relationship between measurement models and structural models when the analysis is focused on a latent variable. A global approach that integrates measurement and explanation of a latent variable is more suitable than an approach that separates the two aspects. Following this idea, a latent regression

	LLRRM	Long.linear model
$\sqrt{t}$	0.271	0.035
$HrQoL_{base}$	0.825	0.542
σ^2_{u0}	0.971	0.026
σ^2_{u1}	0.271	0.006
σ_{u01}	-0.139	-0.005

Table 18.4. *Global approach and two steps approach: comparison*

model for longitudinal data is developed by using a Rasch model as a measurement tool: it consists of a logistic model (or ordinal logistic in the case of polytomous responses) with a random intercept and a random coefficient, which relates explanatory covariates of the latent variable *directly* with item responses. The model was implemented by means of the Nlmixed procedure of SAS.

The LLRM has several advantages. First of all, it overcomes the drawbacks of the two steps approach. Secondly, it is developed in a framework (multilevel modeling) well-known in the field: this allows us to estimate the model by means of generical statistical tools, supporting the utilization of latent regression by common-users. Moreover, it can be easily interpreted on the basis of random and fixed coefficients: in particular, the estimation of random components gives information on how each individual differs from the average population. In comparison with other global models for longitudinal data, its complexity does not increase with the number of measurement occasions; moreover, the treatment of a different number of observations for every individual does not represent any particular problem. Finally, as outlined in section 18.3.1, it can be extended to take into account more general data structures.

An important question to study in a following analysis is concerned with the components of the proposed model to the Rasch family. In other words, it is interesting to understand whether the LLRM verifies the main properties of Rasch models, i.e. the sufficiency of the raw scores and the specific objectivity.

Moreover, the complexity of the analyzed data set illustrates other problems that will be dealt with in the future. First of all, the LLRM should be extended to also consider one or more latent covariates, in order to take into account their random nature, and to consider the multi-dimensional nature of the questionnaire. Secondly, another problem is the presence of informative drop out due to the death of patients during the therapy.

Acknowledgements

Thanks are expressed to Professor Mounir Mesbah for his support in the development of the original idea. Thanks also to Professor Leonardo Grilli for his advices on the applied aspects of data analysis and to Dr. Guido Miccinesi who supplied the

data set. Special thanks are finally expressed to Professor Bruno Chiandotto for his continuous encouragement.

18.6. Bibliography

[ADA 97] ADAMS R., WILSON M., WU M., “Multilevel item response models: an approach to errors in variable regression”, *Journal of Educational and Behavioral Statistics*, vol. 22, num. 1, p. 47–76, 1997.

[AND 77] ANDERSEN E., MADSEN M., “Estimating the parameters of the latent population distribution”, *Psychometrika*, vol. 42, p. 357–374, 1977.

[BAK 04] BAKER F., KIM S., *Item Response Theory: Parameter Estimation Techniques*, Dekker, 2004.

[DOR 03] DORANGE C., CHWALOW, MESBAH M., “Analysing Quality of Life data with the ordinal Rasch model and NLMIXED SAS procedure”, 2003.

[FIS 95a] FISCHER G., “Linear logistic models for changes”, FISCHER G., MOLENAAR I., Eds., in: *Rasch Models. Foundations, Recent Developments, and Applications*, p. 157–180, Springer-Verlag, 1995.

[FIS 95b] FISCHER G., MOLENAAR I., *Rasch Models: Foundations, Recent Developments and Applications*, Springer-Verlag, 1995.

[GOL 80] GOLDSTEIN H., “Dimensionality, bias, independence and measurement scale problems in latent trait test score models”, *British Journal of Mathematical and Statistical Psychology*, vol. 33, p. 234–260, 1980.

[GOL 95] GOLDSTEIN H., HEALY M., “The graphical presentation of a colection of means”, *Journal of the Royal Statistical Society series A*, vol. 158, p. 175–177, 1995.

[HAR 07] HARDOUIN J., MESBAH M., “The SAS macro-program %ANAQOL to estimate the parameters of Item Response Theory models”, *Communications in Statistics. Theory and Methods*, vol. 36, num. 2, p. 437–453, 2007.

[LOR 84] LORD F., *Maximum Likelihood and Bayesian Parameter Estimation in IRT*, Educational Testing Service, 1984.

[MAS 82] MASTERS G., “A Rasch model for partial credit scoring”, *Psychometrika*, vol. 47, p. 149–174, 1982.

[MES 04] MESBAH M., “Measurement and analysis of health related quality of life and environmental data”, *Environmetrics*, vol. 15, num. 5, p. 473–481, 2004.

[PAS 06] PASTOR D., BERETVAS S., “An illustration of longitudinal Rasch modeling in the context of psychotherapy outcomes assessment”, *Applied Psychological Measurement*, vol. 30, p. 100–120, 2006.

[RAB 04] RABE-HESKETH S., SKRONDAL A., PICKLES A., “GLLAMM Manual”, http://www.gllamm.orgh, 2004.

[RAS 60] RASCH G., *Probabilistic Models for Some Intelligence and Attainment Tests*, Danish Institute for Educational Research, 1960.

[SNI 99] SNIJDERS T., BOSKER R., *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*, Educational Testing Service, 1999.

[TAM 92] TAMBURINI M., ROSSO S., GAMBA A., MENCAGLIA E., DE CONNO F., VENTAFRIDDA V., “A therapy impact questionnaire for quality-of-life assessment in advanced cancer research”, *Annals of Oncology*, vol. 3, p. 565–570, 1992.

[WAN 04] WANG W., CHYI-IN W., “Gain score in item response theory as an effect size measure”, *Educational and Psychological Measurement*, vol. 64(5), p. 758–780, 2004.

[WIL 04] WILSON M., DE BOECK P., *Explanatory Item Response Models: a Generalized Linear and Nonlinear Approach*, Springer-Verlag, 2004.

[ZWI 91] ZWINDERMAN A., “A generalized Rasch model for manifest predictor”, *Psychometrika*, vol. 56, num. 4, p. 589–600, 1991.

Chapter 19

Empirical Internal Validation and Analysis of a Quality of Life Instrument in French Diabetic Patients during an Educational Intervention

19.1. Introduction

Non-insulin dependent diabetes mellitus (NIDDM) is a major public health problem in France, not only because of the prevalence of the disease but also due to the severity of its complications. Current estimates of the prevalence range from 1.8-2.5%. This estimate is rising as the population ages (Detourney *et al.* [1]; Deschaseaux and Detourney [2], UKPDS [3]).

The Federation of the *Association des Structures d'Aide à la Vie et à l'Education du Diabétique* (FASAVED) developed an educational program for health care professionals (general practitioners and visiting nurses) in three pilot sites: the Sarthe, the Jura and the Pas de Calais. This study was conducted using a prospective, randomized, two armed, clinical study design (Vray and Schwartz [4], Vray [5]). It studied the impact of different types of educational strategies for health care providers on the management of their NIDD patients seen in ambulatory care. The objective of this overall study was to evaluate the impact of these two different types of educational programs on patient outcomes at T0, year one and year two (Berger and Jorgens [6]; Rose *et al.* [7]; Pipernik-Okanovic *et al.* [8]; Pouwer *et al.* [9]; Landy *et al.* [10];

Chapter written by Judith CHWALOW, Keith MEADOWS, Mounir MESBAH, Vincent COLICHE and Étienne MOLLET.

Hanefeld *et al.* [11]). Included in these outcome measures was patient reported quality of life. As diabetes is a life long, evolving chronic illness, the measurement of disease-specific health-related quality of life has become an important outcome measure (Anderson *et al.* [12, 13]; Coffrey *et al.* [14]; Goddijn *et al.* [15]; Déclaration de Saint-Vincent [16]). Quality of life was measured using two questionnaires: the SF36 (Ware *et al.* [17], Leplège *et al.* [18], Clouet *et al.* [19]) and the disease specific diabetes health profile (DHP) (Meadows *et al.* [20], Goddijn *et al.* [15a]). The DHP was originally developed for use with insulin dependent diabetic patients but data from this study, as well as a study by the original authors (Meadows *et al.* [21], Goddijn *et al.* [22b]) enabled its validation in a population of non-insulin dependent diabetic patients.

The objectives of this sub-study were to validate the psychometric properties of the DHP in French for a population of NIDDM patients and, if valid, to evaluate the effect of the two educational programs on the quality of life of these French patients.

Abbreviations:

– DHP – Diabetes Health Profile

– FASAVED - Federation of the Association des Structures d'Aide à la Vie et à l'Education du Diabétique

– IRT – Item Response Theory

– NIDDM - Non-insulin dependent diabetes mellitus

– SF-36 – Short Form 36

– UKPDS – United Kingdom Prospective Diabetes Study

– WHO – World Health Organization

19.2. Material and methods

19.2.1. *Health care providers and patients*

All physicians in the pilot regions were contacted by telephone (N=1074). In order to participate, physicians had to 1) accept to be randomized into the experimental or control group without knowing their status at the time of acceptance; 2) identify a visiting nurse with whom they agreed to constitute a working pair; and 3) select three patients (NIDDM) who were willing to give written consent to participate in a two year study before being told their study status. 109 physicians met the criteria and agreed to participate. A total of 214 patients were enrolled in the study. Following the patient enrollment, the physicians were randomized into experimental and control groups. Patient reported quality of life was assessed at inclusion, year one and year two using French versions of the SF36 and the DHP. Self-administered questionnaires were sent to patients with an addressed and stamped envelope so that they could be returned directly to the research-coordinating center. All patient data were kept confidential and never individually reported to clinicians. These procedures are in accordance with standard ethical practice and the appropriate ethics committees in France who had approved the study.

19.2.2. *Psychometric validation of the DHP*

The DHP is a diabetes-specific quality of life scale that was developed in the mid-1980s to assess behavioral and psychological dysfunction among insulin dependent diabetic (IDD) patients. The original English version of the questionnaire is cited in Meadows *et al.* [20]. The French version used in this study is shown in section 19.8, Appendix 1. The original authors of the DHP (Meadows *et al.* [20]) do not define the scale as a quality of life scale. Strictly speaking, the DHP is a measure of psychosocial and behavioral dysfunction associated with diabetes rather than a measure of all the multidimensional concepts associated with quality of life. Nevertheless, the dimensions investigated by the DHP are those dimensions of quality of life most associated with its deterioration process among diabetic patients of type 1 or 2. Obviously, any remaining dimensions can be easily assessed using any generic quality of life instrument such as the SF36.

The DHP measures three dimensions: *barriers to activity* (13 items), *psychological distress* (14 items) and *disinhibited eating* (5 items). The sub-scale, *barriers to activity,* reflects the deterioration of social activities and behavioral limits related to anxiety (e.g. “difficult to go out for the evening”, “fear of entering a crowded store”). The sub-scale, *psychological distress,* is concerned with specific and non-specific aspects of psychological dysfunction related to diabetes. These items refer to feelings of vulnerability in the presence of depression and a tendency towards emotional instability (e.g. feelings of despair, irritability or hostility). *Disinhibited eating*, the third sub-scale, reflects inhibitions about nutritional behavior (“binge when you are bored”) as well as about food (“it is hard not to eat”).

All the items in the scale are polytomic (four categories of ordinal responses). At T0, 201 patients completed the questionnaire. The psychometric analyses were done using the data from T0.

19.2.3. *Psychometric methods: from construct validity and reliability to item response theory*

Measuring individual quality of life is frequently done by calculating a score. This approach assumes that the set of items being considered represent a single dimension. One of the methods allowing for the assessment of the unidimensionality of a scale is factor analysis with Varimax rotation. This method allows for the restitution of items into coherent sub-scales and for the identification of isolated items. In this study, as we were using an already validated scale, the construct validity had been defined *a priori*, but we were able to choose the number of factors we would use later on i.e. three. We were then able to compare the classification of items we had obtained to that of the existing scale.

In the measurement of quality of life it is important to know if a questionnaire is adapted to the population being studied. The standard measurement test for reliability of the questionnaire was used: Chronbach's α. Chronbach's α is mainly known as a reliability coefficient. It has also been proven that, under the assumption of a parallel model, this coefficient can be easily interpreted as a coefficient of unidimensionnality (Moret, *et al.* [23]). Moreover, it is easy to show (using the Spearman-Brown formula), that under an assumption of a parallel model, Chronbach's α is an increasing function of the number of items in the scale, so a Backward Chronbach's alpha curve (Curt *et al.* [24]), more precise than the Chronbach's α value, can be used as a validation criterion of the parallel model and its unidimensionality.

The trace of an item is the graph of the proportion of people having responded positively to the item (item response) as a function of their score S_i. Such graphical methods are related to *item response theory* (IRT) models. Modern psychometric theory demands a more rigorous statistical methodology that allows for the modeling of latent traits. These models assume that the probability of an item response depends on the person's aptitude and the parameters associated with the question. These models are, therefore, valid for qualitative as well as quantitative responses. Among these IRT models, the Rasch model for dichotomous responses (Hamon [25], Hamon and Mesbah [26], Hardouin and Mesbah [27], Samb [28]), as well as the Mokken model (Mokken [29]), were used in this study. More details about these models are given in section 19.8, Appendices 2 (Rasch) and 3 (Mokken).

All psychometric analyses were performed using RSP, MSP and SAS (Proc Calis, Proc Corr, Proc Factor, Proc NLmixed) softwares (Mokken [29], Rasch [30], Saporta [31], SASa [32], Friendly [33], Littell *et al.* [34], SASb [35]).

19.2.4. *Comparative analysis of quality of life by treatment group*

Quality of life was measured at three data points, inclusion, year one and year two, in the two treatment groups using the DHP and the SF36. The descriptive statistics (mean±SD) were calculated for each of the different quality of life scores. In order to compare the two treatment groups, a number of analyses were conducted:

– *Univariate analysis of variance* was used to compare quality of life scores in each treatment group. These comparisons were done using a Fisher's exact test.

– *Multivariate analysis of variance* allowed us to study the global differences between the two treatment groups at each time period, while controlling the information contributed by the different quality of life scores. These analyses were done using a Hotelling-Lawley test.

– *Analysis of repeated measures* was the longitudinal analysis used. The use of general linear models for repeated measures was the last type of analysis done in this study. These models control the group effects of treatment, time and the group interaction of treatment*time, as well as the correlations among the responses of the same individual.

All comparative analyses were done using SAS software (SASa [32], SASb [35]).

19.3. Results

19.3.1. *Internal validation of the DHP*

Questionnaire structure

A three-factor factor analysis with Varimax rotation was used. Results are shown in Table 19.1 below.

The cumulative percentage of variance of the first three factors reached 38.91. The sub-scale, *psychological distress* came out as the first factor (22.37% of the variance) *barriers to activity* (9.74% of the variance) as the second and *disinhibited eating* (6.80% of the variance) as the third. For the sub-scale *barriers to activity*, three items loaded on a different factor: items 6, 20 and 26. For the sub-scale, *psychological distress*, item 4 was the only item that shifted to another factor. Finally, in the third sub-scale, *disinhibited eating,* all items seem to measure the same dimension as they loaded on the same factor. A confirmatory factor analysis using SAS Proc Calis and testing the null hypothesis that the clustering of the items around factors is the same as the one found with the original questionnaire was done. The result was the rejection of the null hypothesis with a high significance ($p < 0.0001$). One of the main goals of the following statistical analyses was to identify those departures from the original item clustering.

From reliability of the questionnaire to unidimensional set of items

The stepwise method used for the calculation of the Chronbach's alpha coefficient allowed for a graphic criterion for the selection of items. The alpha coefficient was calculated for each sub-scale. Then, at each iteration, the item that gave the scale the strongest reliability coefficient was removed. The iterations were repeated until only two items were left. If the curve obtained did not increase, we concluded that one or more items were suspect.

These results seem satisfactory, with regard to the final Cronbach alpha value even if no items are discarded in the sub-scales. In the sub-scales *barriers to activity* and

Item	Factor 1	Factor 2	Factor 3
Barriers to activity			
1 Avoid going out if sugars on low side		66	
6 Foods control life			39
20 Edgy when out and nowhere to eat			43
24 Worry about colds or flu		53	
26 Frightened in busy/crowed shops			42
41 Days tied to meal times		51	
44 Difficult staying out late		71	
5 Worry about doing too much and going hypo		64	
9 Avoid going too far in case of hypo		62	
21 Worry about going into diabetic coma		63	
23 Nagging fear of hypos		60	
42 Difficult doing things		63	
43 Plan day around injections		58	
Psychological distress			
19 Lose temper/shout due to diabetes	72		
8 Lose temper over testing/diet	58		
11 Lose temper over small things	72		
13 Touchy/moody about diabetes	70		
14 Because of diabetes get depressed	50		
37 More arguments at home	69		
22 Looks forward to the future	27		
2 Throw things when upset/lose temper	60		
12 Tension headaches	25		
3 Hurt self when upset	49		
4 Wish diabetes would just go away		43	
10 Because of diabetes cries/feels like crying	57		
15 Wished dead	58		
16 Wished never born	62		
Disinhibited eating			
32 Wished not so many nice things to eat			32
34 Eat something extra when bored			53
36 Not easy to stop eating			60
38 Eat to cheer self up			66
39 Hard saying no to food			77
Variance explained by each factor	4.98	4.79	2.68

Table 19.1. *Results of the factor analysis*

* Item labels are condensed. See section 19.7, Appendix 1 for the exact item label.

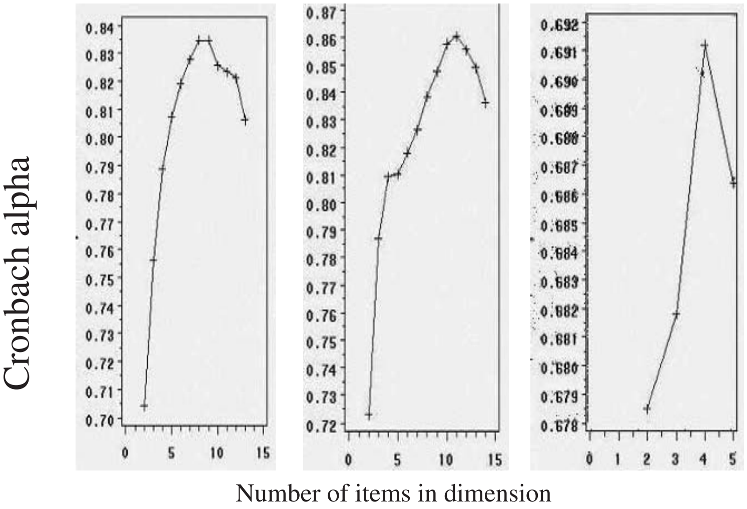


Figure 19.1. *Stepwise curves for Cronbach's alpha coefficients for the sub-scales: barriers to activity, psychological distress and disinhibited eating*

psychological distress, the values of the alpha coefficients exceed 0.70 (0.8 and 0.83, respectively). *Disinhibited eating* is at 0.68. These values are satisfactory, when Nunnally's rule is applied (Nunnally and Bernstein [36]). This requires a minimal Cronbach's alpha value around 0.7. When the increasing monotonicity of the curve is considered, we need to remove those items from the dimension that do not allow for an increasing curve. The "Nunnally rule" is an empirical rule without any scientific justification. Table 19.2 shows the items in each sub-scale that, if removed, would guarantee a perfectly increasing curve, and so, a *unidimensional scale*.

Sub-scale	Suspect items	Maximum reliability
Barriers to activity	20, 6, 26, 41	0.83
Psychological distress	22, 4, 12	0.86
Disinhibited eating	32	0.69

Table 19.2. *Recapitulative of suspect items by sub-scale*

Item traces

Figure 19.2 gives the traces of the items of the sub-scale *disinhibited eating* using the first recode.

The Dichotomous Rasch Model

Methodology: the DHP is composed of 32 polytomic, ordinal items with four response options (coded from 0 to 3). In order to be able to apply the dichotomous

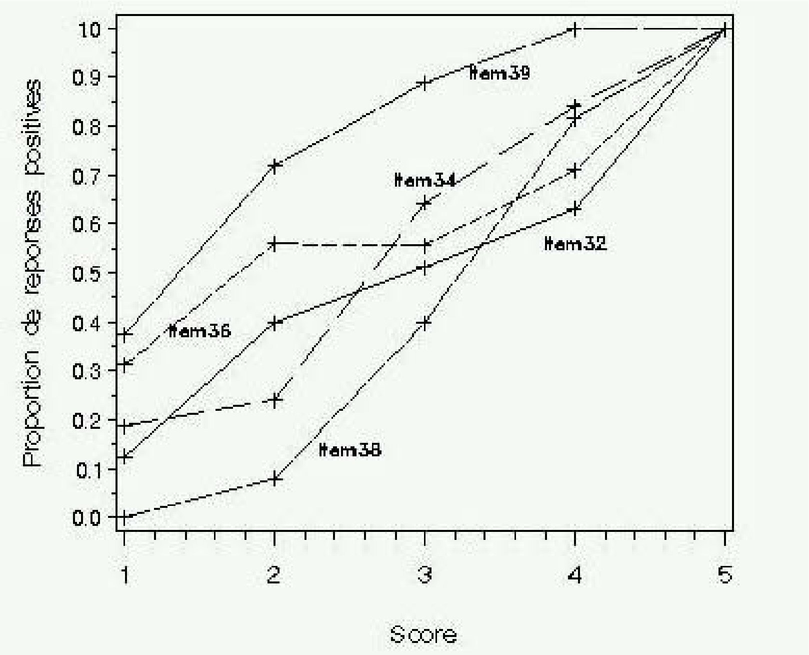


Figure 19.2. *Item traces for disinhibited eating*

Rasch model (responses 0 or 1) and keep the ordinal nature of the response set, several different categorical reclassifications were used. Here, three different reclassifications were done:

	Coding 0	**Coding 1**
1st recode	0	1, 2 and 3
2nd recode	0 and 1	2 and 3
3rd recode	0, 1 and 2	3

Table 19.3. *Dichotomous recoding of responses to items*

The estimations of the parameters of difficulty for the sub-scale *disinhibited eating* using the first classification are presented in Table 19.4. All of these various estimations are very similar. Even if the obtained estimation values are not exactly the same, the ordering of item parameters given by any of these methods is always unchanged. Here, therefore, we used RSP software only, in order to perform specific goodness-of-fit tests not yet available in SAS Proc NLmixed.

	Rasch scaling program		**SAS NLMIXED**
Item	**CML**	**MML**	**MML**
32	0.418	0.417	0.402
34	0.071	0.065	0.015
36	0.038	0.032	0.032
38	0.664	0.663	0.694
39	-1.190	-1.176	-1.145
Variance	-	1.530	2.360

Table 19.4. *Estimations of parameters of difficulty for the dimension disinhibited eating, first recode*

	Barriers to activity	**Psychological distress**	**Dysfunctional eating**
$\mathbf{R}_{1c}$	70.85, 13.11, -	105.69, 43.19, -	21.44, 5.25, 3.26
DF	24, 12, -	39, 26, -	8, 4, 4
Prob($\mathbf{R}_{1c}$**)**	**<0.001,** 0.3600, -	**<0.0010, 0.0167, -**	**0.0061,** 0.2600, 0.5100
$\mathbf{R}_{2c}$	183.34, 113.08, 76.61	202.88, 163.92, 148.07	18.67, 27.60, 16.73
DF	72, 72, 72	84, 84, 84	8, 8, 8
Prob($\mathbf{R}_{2c}$**)**	**<0.001, 0.0014,** 0.3300	**<0.001, <0.001, <0.001**	**0.0167, <0.001, 0.033**

Table 19.5. *Results of the tests of adjustment with the three recodes (first, second and third)*

Goodness-of-fit tests

Table 19.5 shows the results of these tests for the three sub-scales when successively using the three recodes. With this regrouping of the data, the above results show the presence of poor items in the three sub-scales.

Tables 19.6, 19.7, and 19.8 show the method of selection using the U statistic for the three sub-scales.

	Item 20	**Item 26**	**Item 6**	**Item 9**
Max |U|	3.44	2.18	3.7	1.62
Prob($\mathbf{R}_{1c}$**)**	<**0.001**	<**0.001**	<**0.001**	0.120
Prob($\mathbf{R}_{2c}$**)**	<**0.001**	<**0.001**	<**0.001**	<**0.001**

Table 19.6. *Sub-scale barriers to activity (first recode)*

If we remove items 20, 26 and 6, the R_{1c} test is no longer significant. On the other hand, the R_{2c} test is always significant.

If we remove items 22, 4, 12, 37 and 14, the R_{1c} test is no longer significant. On the other hand, the R_{2c} test is always significant.

	Item 22	**Item 4**	**Item 12**	**Item 37**	**Item 14**	**Item 8**
Max |U|	2.70	1.79	2.08	1.03	1.34	2.01
Prob(R_{1c})	**<0.0010**	**0.0029**	**0.0016**	**0.0047**	**0.0248**	0.0805
Prob(R_{2c})	**<0.001**	**<0.001**	**0.0004**	**0.011**	**0.0271**	**0.0003**

Table 19.7. *Sub-scale psychological distress (first recode)*

	Item 36	**Item 32**	**Item 38**
Max |U|	2.15	2.52	1.34
Prob(R_{1c})	**0.0061**	**0.0014**	0.0900
Prob(R_{2c})	**0.0167**	**0.0130**	0.0900

Table 19.8. *Sub-scale disinhibited eating (first recode)*

If we remove items 36 and 32, the tests are no longer significant.

	Item 22	**Item 4**
Max |U|	2.20	1.65
Prob(R_{1c})	**0.018**	0.100
Prob(R_{2c})	**<0.001**	**0.007**

Table 19.9. *Sub-scale psychological distress (first recode)*

If we remove item 22, the R_{1c} test is no longer significant.

The tests of adjustment for the first recode showed many more suspicious items than the subsequent ones. Let us consider that the recoding corresponded to 0 versus 1, 2, 3 and that “0” meant “no dysfunction”. This manner of recoding increased the difficulty of the items as the most difficult response to obtain was “0”. Inversely, the third recode makes the items easier which explains the absence of items rejected by this recoding. In another previous work (Dorange, Chwalow and Mesbah, [37]) the ordinal Rasch model (the partial credit model) was used to analyze the same data. Such a model does not need any recoding of the original items. Results from that analysis are summarized in Table 19.11.

These results are interesting as we find items that have also been rejected in previous studies (Meadows *et al.* [21]).

Mokken’s polytomic model

This is an IRT model for polytomic response items. First, we used the procedure using all the items without indicating the number of sub-scales expected. This gave the following results.

Psychological distress is principally loaded on the first sub-scale with items 22 and 12 excluded. On the second sub-scale, *barriers to activity*, most of the original items load on the original sub-scale except for item 23, which loads on the first sub-scale, item 6 on the fourth and items 20 and 26, which were excluded. Finally, the third sub-scale, which characterizes the dimension *disinhibited eating*, has only one item, 32, that was loaded on the fourth sub-scale (see Table 19.10).

Item	1st	2nd	3rd	4th	Items
	scale	scale	scale	scale	exc.
Barriers to activity					
1 Avoid going out if sugars on low side		X			
6 Foods control life				X	
20 Edgy when out and nowhere to eat					X
24 Worry about colds or flu		X			
26 Frightened in busy/crowed shops					X
41 Days tied to meal times		X			
44 Difficult staying out late		X			
5 Worry about doing too much and going hypo		X			
9 Avoid going too far in case of hypo		X			
21 Worry about going into diabetic coma		X			
23 Nagging fear of hypos	X				
42 Difficult doing things		X			
43 Plan day around injections		X			
Psychological distress					
Psychological distress					
19 Lose temper/shout due to diabetes	X				
8 Lose temper over testing/diet	X				
11 Lose temper over small things	X				
13 Touchy/moody about diabetes	X				
14 Because of diabetes get depressed	X				
37 More arguments at home	X				
22 Looks forward to the future					X
2 Throw things when upset/lose temper	X				
12 Tension headaches					X
3 Hurt self when upset	X				
4 Wish diabetes would just go away	X				
10 Because of diabetes cries/feels like crying	X				
15 Wished dead	X				
16 Wished never born	X				
Disinhibited eating					
32 Wished not so many nice things to eat				X	
34 Eat something extra when bored			X		
36 Not easy to stop eating			X		
38 Eat to cheer self up			X		
39 Hard saying no to food			X		

Table 19.10. *First results using the Mokken model*

A second analysis consisted of analyzing the scales one by one, and removing items step by step as a function of their H coefficients for the entire scale and the individual H for each item.

Table 19.11 shows the H coefficient for each sub-scale for the total of the sub-scale, as well as items to remove as a function of their desired quality (weak, moderate or strong).

	Barriers to activity	Psychological distress	Disinhibited eating
Original H	0.29	0.35	0.34
Weak scale	20	-	-
Moderate scale	6, 26, 41	22,12	32
Strong scale	42, 43,44, 24, 9	4, 3, 15, 16, 14	36

Table 19.11. *Items removed as a function of their desired quality*

We can see that in order to obtain a moderate scale all three dimensions included, we must remove items 20, 6, 26 and 41 (*barriers to activity*), 22 and 12 (*psychological distress*), and 32 (*disinhibited eating*). We had already found these items to be loaded on inappropriate factors, except for item 23.

Summary of the validation results

Table 19.12 summarizes all of the items that had been considered “doubtful” in the diverse analyses that were conducted in order to validate the DHP in French.

	Barriers to activity	Psychological distress	Disinhibited eating
Factor analysis	20, 6	4	-
Rasch (1st recode)	20, 26, 6	22, 4, 12, 37, 14	36, 32
Rasch (2nd recode)	-	22	-
Partial credit	1, 43, 44	-	34
Mokken model	20, 6, 26, 41	22, 12	32

Table 19.12. *Summary of the results by type of analysis*

It is evident that for each sub-scale we find similar items no matter what type of analysis was used. For *barriers to activity* items 20, 6 and 26 are present in most of the types of analyses conducted. We find items 22, 12 and 4, repeated in all analyses of *psychological distress*. For the sub-scale *disinhibited eating*, item 32 is repeated. The scale had already been used in a real-life clinical setting from which we obtained our database. Details about partial credit model can be found in Dorange, Chwalow and Mesbah [3]).

19.3.2. *Comparative analysis of quality of life by treatment group*

Among the 214 questionnaires sent, 201 responded at T0, making the response rate 93.9%. Among these patients, there was one case missing data for age and sex, but no missing data in the treatment group.

Treatment group	**n**	**Age mean (SD)**	**Men(%)**	**Women(%)**
Control	83	60.8(7.4)	57.3	42.7
Education	118	57.8(8.4)	49.1	50.9

Table 19.13. *Characteristics of patients responding to the questionnaire at T0*

On average, the patients in the control group are slightly older and more likely to be male.

Descriptive statistics of quality of life scores by treatment

In the two treatment groups the mean values of the DHP sub-scales are not significantly different. For the sub-scales *barriers to activity* and *psychological distress* the means are relatively low (between 17% and 22%). This is positive, as the lower the score the better the quality of life. For these dimensions patients report a relatively good quality of life. For the sub-scale *disinhibited eating*, the two groups differ at year one. The education group reports a better quality of life. The mean scores for this sub-scale are generally higher than for the other two scores.

For the SF36, the mean scores are higher than 50% and stay stable over the two-year period. With the SF36, the higher the score the better the quality of life. The two treatment groups are not significantly different over time or by sub-scale.

	Barriers to activity	**Psychological distress**	**Disinhibited eating**
Inclusion (T0)			
Education	17.7 (14.7)	22.4 (16.5)	37.4 (22.2)
Control	18.5 (14.4)	19.0 (10.9)	36.5 (22.6)
1 yr (T1)			
Education	21.0 (16.5)	20.3 (16.4)	32.1 (20.2)
Control	17.3 (12.5)	20.3 (11.6)	39.5 (22.1)
2 yrs (T2)			
Education	22.0 (13.8)	22.4 (16.2)	36.5 (22.3)
Control	20.7 (12.5)	20.9 (13.0)	38.5 (19.4)

Table 19.14. *DHP means ±SD of treatment groups by time*

	Baseline (T0)		**Year 1 (T1)**		**Year 2 (T2)**	
	Educ	Control	Educ	Control	Educ	Control
Physical activity	53.4 (19.7)	55.5 (18.7)	55.2 (19.3)	54.4 (15.9)	53.7 (18.3)	51.3 (18.8)
Feeling about HIPA*	75.8 (22.9)	77.9 (19.8)	77.5 (21.9)	77.3 (21.9)	75 (21.5)	75.8 (24.1)
Pain	58.6 (42.4)	68.5 (39.0)	68.3 (42.2)	63.1 (41.8)	63.1 (42.0)	63.4 (42.5)
Perceived health	60.9 (21.9)	65.7 (19.6)	63.4 (22.2)	63.7 (17.1)	61.2 (20.8)	60.9 (21.9)
Vitality	72.0 (23.9)	72.7 (22.9)	65.5 (39.9)	63.2 (38.1)	70.3 (25.8)	71.8 (23.8)
Social activity	37.4 (22.2)	36.5 (22.6)	65.5 (39.9)	63.2 (38.1)	60.2 (39.3)	62.2 (39.9)
Feeling about HIDA**	64.3 (25.8)	66.2 (26.1)	66.1 (25.5)	67.8 (25.2)	64.3 (23.5)	62.0 (26.2)
Emotional health	58.9 (20.6)	59.7 (19.2)	58.3 (17.6)	57.4 (18.7)	56.2 (17.6)	56.8 (18.9)
Evolution of health status	55.7 (19.6)	51.8 (17.3)	57.5 (17.6)	54.7 (17.3)	55.2 (20.4)	50.4 (18.2)

Table 19.15. *SF36 means ±SD of treatment groups by time*

*HIPA= Health Impact on Physical Activity **HIDA= Health Impact on Daily Activity

Comparison of the two scores

Using SAS Proc GLM, univariate analyses of variances were performed to show a significant difference in the means between the two treatment groups for *disinhibited eating* (p=0.032) at T1. The group who saw clinicians and who had been educated had a better quality of life. The difference of scores between consecutive times was also studied, using a t-test, but showed no significant differences.

Finally, the multivariate analysis (*SAS Proc GLM* procedure *MANOVA* option) allowed us to study the global differences between the two treatment groups over time, while controlling the different quality of life scores. There were no lasting differences between the two groups.

Longitudinal analysis

This analysis attempted to see if there was a group effect on the quality of life scores over time. A mixed model (using SAS Proc Mixed) was used which allowed us to analyze whether there was, for each quality of life score, a group, time or the interaction of group-time effect. In this analysis, time was introduced as fixed as well as random effect. The covariance structure type was chosen as compound symmetry. This analysis allowed us to conclude whether the three sources of variation could explain the differences in the quality of life scores. There were no statistically significant differences between the two groups.

19.4. Discussion

This study showed no differences between the two treatment groups except for *disinhibited eating* at T0, where the experimental group did have a significantly better quality of life. It also showed almost no differences between measures taken using the DHP or the SF36. The SF36 is a well-known, validated, generic quality of life questionnaire that has been in existence for over 10 years. The lack of difference in the results obtained by the two scales, one generic and the other specific to the illness, shows concurrent validity for the French version of the DHP.

This does not answer the question: "why were there no differences between the treatment groups?" There are a number of possible explanations:

1) There truly is no difference! The educational strategy used was not effective.

2) In the educational strategy, there may have been more emphasis on eating rather than depression, which would explain why differences in the eating scale were identified.

3) The size of the sample was relatively small (214 individuals), which contributed to a lack of power in the analyses.

4) The instruments might not have been sensitive enough to detect changes in the sample or the intervention might not have been strong enough.

5) We believe that the most likely reason was that this was a randomized clinical trial conducted in the field rather than in a clinic or laboratory setting which engendered a "contamination" or "Hawthorne" effect. As such, the individual interactions of the clinicians after being randomized, then educated and then sent back to practice in the community, could not be controlled.

6) There is also a problem with missing data. In fact, psychometrically, in order to establish the different quality of life scores, patients must respond to all questions or their scores are not counted. For the DHP, 185 people responded to the entire questionnaire at T0 as opposed to 130 at T2. For the SF36, there were 199 at T0 versus 142 at T2. Missing data was a real problem in this study and is the object of a separate sub-study (Chavance [38]).

7) We can probably find significant differences between groups if we use more powerful statistical methods combining validation and comparative analysis parts. Modern latent regression methods, perform comparisons on the latent value θ_i, not on its surrogate value S_i. In the case of a true difference, we can improve the "true" significance (Mesbah [39]).

19.5. Conclusion

It is rare to use multiple methods to validate cross-cultural psychometric validations of quality of life scales, and in statistical science, it is common to use multiple and complementary methods when there is no unique method to answer the scientific question. As questions are being asked about "what are we measuring?" (Hendry, McVittie [40], Sac Guidelines [41]) we must be at least capable of answering the question "whatever it is, are we measuring the same thing". Construct and concurrent validity should not be taken lightly. The main goal of this chapter, was, by using various concurrent models, to prove that all theses different analyses find very similar subscales. Deleting only a few items gave a strong consensus!

Seven of 32 items in this scale consistently showed difficulty in being loaded on their original scales no matter what type of analysis was used. While the factor structure may be questioned, the overall construct validity of the scale is good. Use of this scale in comparative, clinical studies suggests that an overall score should be given. Our results, reported above, indicate that this would seem to be a good idea. Concurrent validity is another concept and the DHP shows that it gives the same results as another, older, widely used scale, the SF36. The DHP is specific to patients with diabetes mellitus and much easier to analyze than the SF36. If a choice had to be made between the two scales, by clinicians, the DHP would be easier to use and just as sensitive as the SF36. It is possible that the number of cases on which the scale was validated (n=214) is not sufficient to show differences. We are currently analyzing the psychometric properties of this scale with responses from approximately 3,500 French NIDDM patients. This analysis should provide definitive answers to questions raised in this pilot study about the methodological properties of this scale in the French population.

19.6. Bibliography

[1] Detourney B, Vauzelle-Kervroedan F, Charles MA, Forhan A, Fagnani F, Fender P, Eschwege E. Epidémiologie, prise en charge et coût du diabète de type 2 en France en 1998. Diabetes Metab, 1999; 25: 356–365.

[2] Deschaseaux C, Detournay B. Analyse des données économiques recueillies lors de l'étude ENTRED: Étude réalisée pour l'ANCRED. Cemka, 2004; 17–37.

[3] UKPDS 16. Overview of 6 years' therapy of type II diabetes: a progressive disease. Diabetes, 1995; 44, 1249-1258.

[4] Vray M, Schwartz D. Comments on a pragmatic trial. J. Clin Epidemio 1996; 49: 949–50.

[5] Vray M. Pragmatic and explanatory trials: ask and answer different questions. Applied Clinical trials 1999; 42–50.

[6] Berger M, Jorgens V: Therapeutical effects of diabetes education: evaluation of diabetes teaching programs. In *Diabetes Education, How to Improve Patient Education. Excerpta Medica*, 1983; pp37–50.

[7] Rose M, Fliege H, Hildebrandt M, Schirop T, Klapp BF. The network of psychological variables in patients with diabetes and their importance for quality of life and metabolic control. Diabetes Care. 2002; Jan: 25(1): 35–42.

[8] Pibernik-Okanovic M, Prasek M, Poljicanin-Filipovic T, Pavlic-Renar I, Metelko Z. Effects of an empowerment-based psychosocial intervention on quality of life and metabolic control in type 2 diabetic patients. Patient Educ Couns, 2004; Feb: 52(2): 193–9.

[9] Pouwer F, Snoek FJ, van der Ploeg HM, Ader HJ, Heine RJ. Monitoring of psychological well-being in outpatients with diabetes. Effects on mood, HbA1c, and the patient's evaluation of the quality of diabetes care: a randomized controlled trial. Diabetes Care, 2001; 24: 1929–1935.

[10] Landy J, Stein J, Brown MM, Brown GC, Sharma S. Patient, community and clinician perceptions of the quality of life associated with diabetes mellitus. Med Sci Monit, 2002; 8(8): CR543-8.

[11] Hanefeld M, Julius U, Fischer S, Schulze J and the DIS Group. Continuous health education is needed to achieve long-term improvement in quality of diabetes control (Abstract). Diabetologia, 1995; 38, suppl 1: A24.

[12] Anderson RM, Funnell MM, Butler PM, Arnold MA, Fitzgerald JT, Feste CC. Patient empowerment. Results of a randomized controlled trial. Diabetes Care, 1995a: 18: 943–949.

[13] Anderson RM, Fitzgerald JT, Wisdom K, Davis WK, Hiss RG. A comparison of global versus disease-specific quality-of-life measures in patients with NIDDM. Diabetes Care, 1997b Mar: 20(3): 299–305.

[14] Coffey JT, Brandle M, Zhou H, Marriott D, Burke R, Tabaei BP, Engelgau MM, Kaplan RM, Herman WH. Valuing health-related quality of life in diabetes. Diabetes Care. 2002; Dec: 25(12): 2238–43.

[15] Goddijn P, Bilo HJG, Meadows KA *et al.* Longitudinal study on glycemic control and quality of life in patients with type 2 diabetes mellitus referred for intensified control. Diabetic Medicine, 1999a: 16: 23–30.

[16] Prise en charge, traitement et recherche en Europe. Le texte de la Déclaration de Saint-Vincent. Diabetic Metab, 1992; 18: 335–337.

[17] Ware JF, Snow KK, Kosinski M, Gandek B. SF-36 health survey: manual and interpretation guide. Boston, MA: The Health Institute, New England Medical Center, 1993.

[18] Leplège A, Ecosse E, Verdier A, Perneger TV. The French SF-36 health survey: translation, cultural adaptation and preliminary psychometric evaluation. J. Clin Epidemio, 1998; 51 11: 1013–1023.

[19] Clouet F, Excler-Cavailher G, Christophe B, Masson F, Fasquel D. Type 2 Diabetes and Short Form 36-items Health Survey. Diabetes Metab. 2001; Dec: 27(6): 711–717.

[20] Meadows KA, Steen N, McColl E *et al.* The Diabetes Health Profile (DHP): A new instrument for assessing the psychosocial profile of insulin requiring patients: development and psychometric evaluation. Qual Life Res, 1996; 5: 255–264.

[21] Meadows KA, Abrams C. Sandbaek A. The adaptation of the diabetes health profile (DHP-1) for use with type-2 DM patients: psychometric evaluation and cross-cultural comparison. Diabet Med. 2000; 17: 572–580.

[22] Goddijn P, Bilo HJG, Meadows KA *et al.*. The validity and reliability of the diabetes health profile (DHP) in NIDDM patients referred for insulin therapy. Qual Life Res, 1996b: 5: 433–442.

[23] Moret L, Mesbah M, Chwalow J, Lellouch J. Validation interne d'une échelle de mesure: relation entre analyse en composantes principales, coefficient alpha de Cronbach et coefficient de corrélation intra-classe. RESP. 1993; vol. 41 (2), 179–186.

[24] Curt F, Mesbah M, Lellouch J, Dellatolas G. Handedness scale: how many and which items? Laterality, 1997; 2(2), 137–154.

[25] Hamon A. Modèle de Rasch et validation d'un questionnaire de qualité de vie. Thesis, 2000.

[26] Hamon A. and Mesbah M. Validation statistique interne d'un questionnaire de qualité de vie. RESP. 1999; 47, 571–583.

[27] Hardouin JB, Mesbah M. Processus de sélection d'items unidimentionnels en phase exploratoire dans le cadre du modèle "Objectif" de Rasch multidimentionnel. 2001; SABRES Technical Laboratory Report.

[28] Samb T. Les modèles de Rasch polytomiques pour l'analyse des échelles de qualité de vie. 1998; SABRES Technical Laboratory Report.

[29] Mokken Scale Program 5 for Windows: User's Manual, Pro Gamma, 2000.

[30] Rasch Scaling Program: User's Manual, Pro Gamma, 1993.

[31] Saporta G. *Probabilités Analyse des données et Statistiques*, Technip, 1990.

[32] SAS Institute Inc. SAS/STAT Sofware: changes and enhancements through release 6.12, Cary, NC: SAS Institute Inc., 1997a.

[33] Friendly M. SASSystem for Statistical Graphics, First Edition, Cary, NC: SAS Institute Inc. 1991.

[34] Littell RC, Milliken GA, Stroup WW, Wolfinger RD. SASSystem for Mixed Models, Cary, NC: SAS Institute Inc., 1996.

[35] SAS Institute Inc. SAS/STAT User's Guide, Version 8, Cary, NC: SAS Institute Inc, 1999b.

[36] Nunnally JC, Bernstein I. *Psychometric Theory*, McGraw-Hill College, January 1994.

[37] Dorange, C., Chwalow, J., Mesbah, M. (2003) Analyzing Quality of Life data with the ordinal Rasch model and NLMixed SAS procedure. In *Proc. of the International conference on Advance in Statistical Inferential Methods "ASIM2003"*. Almaty: KIMED Ed. 2003.

[38] Chavance M. "Handling missing items in quality of life studies", *Communication and Statistics. Theory and Methods*, 2004; 33(6), 1277–1294.

[39] Mesbah M. Measurement and analysis of health related quality of life and environmental data. Environmetrics. 2004; vol. 15 (5), 471–481.

[40] Hendry F, McVittie C. Is Quality of Life a Healthy Concept? Measuring and Understanding Life Experiences of Older People. Qualitative Health Res, 2004; 14:7:961–975.

[41] Scientific Advisory Committee of the Medical Outcomes Trust. Key Guidelines: Assessing health status and quality of life instruments: attributes and review criteria. Qual Life Res, 2002; May: 11 (3): 193–205.

[42] Hamon A, Dupuy J.F. and Mesbah M.(2002) Validation of Model Assumptions in Quality Of Life Measurements. In *Goodness of Fit tests and Model Validity*. Eds : Huber, C., Nikulin, M., Balakrishnan, N. and Mesbah, M. Birkhauser, Boston. pp. 371–386.

[43] Hemker, B. T., Sijtsma K. and Molenaar I. W. Selection of unidimensional scales from a multidimensional item bank in the polytomous Mokken IRT model. Applied Psychological Measurement, 19(4) : 337–352,1995.

[44] Masters, G.N. (1982). A Rasch model for partial credit scoring. Psychometrika, 47, 149–174.

[45] Meijer R. R. The number of Guttman errors as a simple and powerful person-fit. Applied psychological measurements. 1994, vol. 18 (4), pp. 311–314.

19.7. Appendices

Appendix 1: French version of the diabetes health profile

1 Evitez-vous de sortir si vos glycémies sont basses ?
6 Est-ce que la nourriture “ gouverne ” votre vie ?
20 Etes-vous à cran quand vous sortez et qu’il n’y a nulle part où manger ?
24 A cause de votre diabète vous tracassez-vous à l’idée d’attraper des rhumes ou des grippes ?
26 Trouvez-vous effrayant ou inquiétant d’entrer dans des magasins animés ou bondés ?
41 Avoir un diabète signifie-t-il que vos journées sont régulées par vos heures de repas ?
44 Avoir un diabète signifie-t-il qu’il est difficile de sortir le soir ?
5 Avez-vous peur de vous surmener et de vous mettre en hypoglycémie ?
9 Évitez-vous d’aller trop loin tout seul par crainte de faire une hypoglycémie ?
21 Vous tracassez-vous à l’idée de faire un coma diabétique ?
23 Avez-vous une peur tenace des hypoglycémies ?
42 Avoir un diabète signifie-t-il qu’il est difficile de faire des choses au moment où vous le voulez ?
43 Avoir un diabète signifie-t-il que vous devez planifier vos journées autour de vos injections ?
Sub-scale psychological distress
19 Est ce que votre diabète vous fait mettre en colère ou crier ?
8 Vous-vous mettez en colère si on vous harcèle à propos de votre autosurveillance glycémique ou de votre régime?
11 Trouvez-vous que vous vous mettez en colère pour des petits rien ?
13 Etes-vous susceptible ou de mauvaise humeur à propos du diabète ?
14 Est-ce que votre diabète arrive à vous déprimer ?
37 Ya-t-il plus de disputes ou de bouleversements à la maison qu’il n’y aurait si vous n’aviez pas de diabète ?
22 Etes-vous tourné vers l’avenir ?
2 Envoyez-vous balader les choses si vous êtes contrarié ou si vous vous mettez en colère ?
12 Avez-vous des maux de tête (dus à la tension nerveuse) ?
3 Vous faites-vous du mal ou avez-vous de vous faire du mal quand vous êtes contrarié ?
4 Souhaitez-vous que votre diabète disparaisse quand vous ne savez plus où donner de la tête ?
10 A cause de votre diabète pleurez-vous ou avez-vous envie de pleurer ?
15 Le fait de souhaiter être mort vous effleure-t-il l’esprit ?
16 Auriez-vous aimé ne jamais être né ?
Sub-scale disinhibited eating
32 Souhaitez-vous qu’il n’y ait pas autant de bonnes choses à manger ?
34 Risquez-vous de faire un extra alimentaire quand vous vous ennuyez ou quand vous en avez marre ?
36 Quand vous commencez à manger trouvez-vous qu’il soit facile d’arrêter ?
38 Avez-vous des difficultés à suivre votre régime parce que vous mangez pour vous remonter le moral ?
39 Avez-vous des difficultés à suivre votre régime parce que vous trouvez dur de renoncer à la nourriture que vous aimez?

Table 19.16. *Sub-scale barriers to activity*

Appendix 2: the Rasch model

Let us consider a group of n individuals who have responded to a questionnaire constructed of k items. Each question is associated with a positive response (or "correct") graded "1" and a negative response graded "0". Each person i has a probability π_{ij} to give the positive response to the question j. Therefore, X_{ij} is the response of the individual, i, to the question j. As we are using questions with dichotomous responses, X$_{ij}$ can take 2 forms: 0 or 1. This is a Bernoulli variable.

The parameter of the person's aptitude i is noted as θ_i. This might be psychological distress, for example. The parameter of the difficulty associated with the item j is noted as β_j. This is the aptitude that an individual must have in order that his probability to positively respond is equal to 0.5. θ_i and β_j are scalar.

The assumptions of the model are as follows:

1) The latent trait θ_i and the difficulty parameter β_j are scalar.

2) The individuals are independent.

3) The probability of a response x_{ij} of a person i to the question j is a logistic function of the difference $\theta_i - \beta_j$:

$$P(X_{ij} = x_{ij}/\theta_i; \beta_j) = \frac{\text{exp}((x_{ij}(\theta_i - \beta_j))}{1 + \exp(\theta_i - \beta_j)}.$$

The item responses are locally independent, i.e. if the latent trait fixed, the responses to the questions are independent.

$$P(X_{ij} = x_{ij}, X_{ik} = x_{ik})/\theta_i; \beta_j, \beta_k) = \frac{\text{exp}((x_{ij}(\theta_i - \beta_j))}{1 + \exp(\theta_i - \beta_j)} \times \frac{\text{exp}((x_{ik}(\theta_i - \beta_k))}{1 + \exp(\theta_i - \beta_k)}.$$

This model describes the probability of a response to an item as a logistic function of the difference between a parameter of aptitude and a parameter of difficulty. In addition, this model presents a nice property: the total score of the individual i, S$_i = \sum_{j=1}^{k} X_{ij}$, is a *sufficient statistic* for the latent trait. This means that, for the individual i, all the information on its latent trait θ_i, that we can get from the observations X$_{ij}$, j=1 to n, can be recovered using only S$_i$. *Estimation of the parameters*. The easiest way to estimate the parameters is to use the method of maximum likelihood. We therefore estimate the parameters of difficulty of the items and the parameters of aptitude of the individuals, at the same time. Unfortunately this method does not give consistent estimators for the θ_i. In addition, when we are validating a scale, only the estimations of the difficulties are of interest. We therefore find two types of methods that allow us to avoid the estimation of the aptitudes. These are the estimations using conditional maximum likelihood or the marginal likelihood estimation method of the difficulties.

The conditional maximum likelihood method depends on the sufficient property of the score S_i. The marginal likelihood estimation of the difficulties is, therefore, consistent. Using the marginal likelihood estimation leads us to consider the Rasch model as a mixed generalized linear model. The θ parameter is then considered as a random variable most frequently chosen as normally distributed.

There is a specific software program that analyzes the dichotomic Rasch Model. RSP (Rasch scaling program) enables the application of these two estimation methods. Using the NLMixed procedure in SAS also allows the use of the second method.

Goodness-of-fit tests. The goodness-of-fit tests allow us to know the nature of discrepancy to the model and, on that basis, to remove deviant items. Glas (1988) proposed goodness of fit tests for the two types of estimation methods. When a questionnaire is being validated, however, it is the items and their parameters of difficulty that is of interest. We therefore chose to adapt a model to fixed effects.

For this model, the first test is based on the number of people having a score s and who had responded to item j. This was a first order test that we called R_{1c}. The second test was based on the number of people having a score s and who had responded to item j. This was a second order test that we called R_{2c}. They tested the null hypothesis, H_0, that the Rasch model was correct.

The R_{1c} statistic allowed us to test H_0 against the alternative that the response curves were not parallel. If one assumes that for an item the parallel hypothesis is not true, it means that the theoretical curve is far from the observed response curve. The R_{2c} statistic allowed us to test H_0 against the alternative that the unidimensionality and the local independence were not respected.

As the Rasch model is based on strong hypotheses, the present tests frequently show differences with the model. Except in various instances, we saw that only a few items seemed to provide a poor adjustment. We removed the deviant items, one by one until we obtained a sub-group of items adjusted correctly.

In practice, if the R_{1c} test is significant, it means that one or more items do not respect the parallel hypothesis. In order to identify these items, we used the U_j statistic that allowed each item j to test the null parallel hypothesis of the response curve against the alternative that the curve was not parallel. As a general rule, we considered that an item was “dubious” if the absolute value of U exceeded 1.96.

More details on goodness of fit tests for Rasch models can be found in Hamon, Dupuy and Mesbah [42].

Appendix 3: the Mokken model

The Hemker, Sitjma and Molenaar methods [43] are based on Mokken's monotonic homogenous models. These models are part of the family of IRT models for polychotomous items. These models are non-parametric as they do not assume a parametric definition of the item traces or of the distribution of the latent trait as it relates to individuals. This model only produces an ordinal measure and is known for its ability to adapt to the data.

Let X_j be a random variable for the score of the item $j, j = 1 \ldots J$ with r_j categories of arranged responses. An item step is the imaginary plateau between two categories of adjacent, arranged responses. We make the hypothesis that each item is based on r_j – 1 and hypothetical, dichotomous item steps called Y_{jl} with l = 1...r_{j-1}. Going from one response to a higher, adjacent response gives 1 point; if not, the score 0 is given. Within an item, an item step cannot pass if the precedent item step has not been done. Therefore, $X_j = \sum_{l=1}^{rj-1} Y_{jl}$ and $Y_{jl-1} = 0 \Rightarrow$ Y_{jl}=0 and Y_{jl+1}= 1 $\Rightarrow Y_{jl}$ = 1. Note that P(Y_{jl} = 1/θ) = P($X_j \geq$ 1 /θ) = $\pi_{jl}(\theta)$. We call this probability the function of the response to the item step.

The monotonic homogeneous Mokken model is defined by three hypotheses:

1) unidimentionality;

2) local Independence;

3) monotonicity in θ (implying that if $\theta_A < \theta_B$ then $\pi_{jl}(\theta_A) < \pi_{jl}(\theta_B)$).

The monotonic homogenicity of Mokken's models are the non-parametric versions of the models in response to definitive gradations defined by Masters [44]. Then α_j is the parameter of discrimination of the item j and λ_{jl} the first order parameter of the steps l of the item I, the graded response models place:

$$\pi_{jl} = \frac{\exp[\alpha_j(\theta_{-\lambda jl})]}{1 + \exp[\alpha_j(\theta_{-\lambda jl})]},$$

with λ_{jl} = δ_j + τ_{jl} or $\sum_{l=1}^{r_j-1} \tau_{jl} = 0$, and $\delta_j = \frac{\sum_{l=1}^{r_j-1} \lambda jl}{r_{j-1}}$. δ_j is the parameter of difficulty of the step l of the item j. In addition, we find a "polytomic" Rasch model by placing α_j = 1 $\forall j$ = 1..J and the Rasch model by placing r_j= 2. Using the monotonic homogenicity of Mokken's model, we can easily prove that the covariance σ_{jj}' between the items j and j' is not negative.

Selection of items: the method is based on the utilization of the Loevinger coefficient H. His calculation is based on the utilization of Guttman's errors (Meijer R. R. [45]). In the context of this study, we based the analysis on Mokken's criteria that suggest:

– 0.30 $\leq$ H $<$ 0.40: poor scale;

– $0.40 \leq \mathrm{H} < 0.50$: moderate scale;

– $0.50 \leq \mathrm{H}$: good scale.

The Mokken scale procedure: this is a process of item selection using an ascending stepwise procedure. It is available in Mokken's Scale Procedure software (MSP) .

Appendix 4: complementary results. Estimation of the parameters of difficulty for the Rasch model

	First recode			Second recode			Third recode		
	RSP		**SAS**	**RSP**		**SAS**	**RSP**		**SAS**
Item	**CML**	**MML**	**MML**	**CML**	**MML**	**MML**	**CML**	**MML**	**MML**
1	0.684	0.684	0.751	0.783	0.793	0.850	-0.217	-0.222	-0.120
6	-2.347	-2.359	-2.335	-1.682	-1.694	-1.819	-1.616	-1.633	-1.523
20	-0.066	-0.064	-0.099	0.181	0.180	-0.016	1.021	1.038	0.463
24	0.224	0.226	0.227	0.181	0.180	0.288	1.021	1.038	1.053
26	-0.605	-0.603	-0.695	-0.684	-0.691	-0.877	-0.859	-0.877	-0.727
41	-0.605	-0.603	-0.667	-1.251	-1.262	-1.376	-1.106	-1.126	-1.095
44	0.684	0.684	0.687	1.117	1.133	1.121	1.021	1.038	1.042
5	0.356	0.357	0.454	0.284	0.285	0.442	-0.404	-0.413	-0.270
9	0.635	0.635	0.670	0.284	0.285	0.471	0.257	0.262	0.256
21	0.311	0.313	0.224	0.394	0.397	0.206	0.580	0.590	0.055
23	-0.642	-0.64	-0.603	-0.180	-0.185	0.071	-0.859	-0.877	-0.510
42	0.139	0.141	0.207	0.394	0.397	0.637	0.580	0.590	1.044
43	1.231	1.230	1.177	0.181	0.180	0.001	0.580	0.590	0.333
Variance		1.356	1.906		1.465	1.827		1.763	2.867

Table 19.17. *Estimation of parameters of difficulty for the sub-scale barriers to activity*

	First recode			Second recode			Third recode		
	RSP		**SAS**	**RSP**		**SAS**	**RSP**		**SAS**
Item	**CML**	**MML**	**MML**	**CML**	**MML**	**MML**	**CML**	**MML**	**MML**
19	0.426	0.434	0.409	0.490	0.497	0.605	-0.189	-0.161	-0.188
8	-0.092	-0.850	-0.036	0.128	0.145	0.358	-0.531	-0.492	-0.465
11	-1.802	-1.815	-1.815	-1.324	-1.282	-1.062	-1.882	-1.833	-1.724
13	0.178	0.185	0.105	0.391	0.401	0.609	-0.370	-0.336	-0.329
14	-0.062	-0.055	-0.161	0.596	0.599	0.900	1.436	1.375	1.483
37	1.528	1.526	1.546	0.391	0.401	0.562	0.019	0.038	0.144
22	-2.690	-2.717	-2.684	-2.955	-3.077	-2.318	-1.430	-1.379	-1.364
2	-0.857	-0.854	-0.773	0.391	0.401	-2.495	-0.531	-0.492	-0.478
12	-0.649	-0.645	-0.551	-0.364	-0.334	-0.248	-0.370	-0.336	-0.331
3	1.117	1.119	1.146	1.695	1.651	1.932	2.245	2.155	2.246
4	-0.827	-0.824	-0.858	-1.437	-1.396	-1.280	-2.634	-2.619	-2.567
10	1.274	1.275	1.205	0.490	0.497	0.614	2.245	2.155	1.478
15	0.79	0.796	0.804	0.391	0.401	0.600	0.555	0.547	0.648
16	1.664	1.661	1.662	1.116	1.099	1.220	1.436	1.375	1.449
Variance		1.368	1.941		1.343	1.712		1.750	2.617

Table 19.18. *Estimation of parameters of difficulty for the sub-scale psychological distress*

	Second recode			Third recode		
	RSP		**SAS**	**RSP**		**SAS**
Item	**CML**	**MML**	**MML**	**CML**	**MML**	**MML**
32	0.603	0.627	0.622	1.163	1.133	1.032
34	-1.210	-1.219	-1.287	-1.768	-1.789	-1.754
36	-0.996	-0.999	-0.966	-0.124	-0.090	-0.058
38	1.361	1.311	1.339	0.917	0.900	0.935
39	0.242	0.280	0.292	-0.188	-0.154	-0.155
Variance		1.238	1.437		1.884	3.125

Table 19.19. *Estimation of parameters of difficulty for the sub-scale disinhibited eating*

PART IV

Related Topics

This page intentionally left blank

Chapter 20

Deterministic Modeling of the Size of the HIV/AIDS Epidemic in Cuba

20.1. Introduction

The first Acquired Immunodeficiency Syndrome (AIDS) case in Cuba was diagnosed in April of 1986. This signaled the starting point of the Human Immunodeficiency virus (HIV/AIDS) epidemic in the country, although some HIV-positive persons had been detected at the end of 1985. In 1983, the Ministry of Health in Cuba set up a national committee on AIDS and this committee started taking several preventive measures to try to contain the possible outbreak of the epidemic [GRA 95, PER 96, SWA 95]. Among these measures was a total ban on the import of blood and blood byproducts. Once the first cases were confirmed, a program based on the experience with the spread of other sexually transmitted diseases was started, and by June 1990 eight million tests for HIV had been performed. At that time all provinces in Cuba started building AIDS sanatoriums. By 1993, day care hospitals started treating patients replacing the sanatorium system [KOI 02]. The AIDS program also had, among other measures, traced sexual contacts of known HIV-positive (HIV+) person, to prevent the spreading of HIV. When a person is detected as living with HIV, an epidemiological interview is carried out by the Epidemiology Department of his municipality or by his family doctor as part of the Partner Notification Program. After this interview the Epidemiology Department tries to locate the sexual partners of the person through the network of the health system. The person living with HIV usually does not participate in this process, though they normally help in notifying their present partners. Trying to locate the sexual partners is a very complex job and one

Chapter written by Rachid LOUNES, Héctor DE ARAZOZA, Y.H. HSIEH and Jose JOANES.

that in some cases takes a lot of time. This task is one of high level of priority for the health system, and it is something that is under constant supervision to try to determine how effective it is in the prevention of the spread of HIV. All data used in this chapter is from the time period of 1986-2002.

The number of AIDS cases in Cuba at the end of 2005 was 2,848 with 575 females and 2,300 males. Of the males 86.5% are men that have sex with other men (MSM). Through the health system HIV/AIDS program a total of 6,967 HIV-positive individuals have been found, including 1,361 females and 5,606 males. Of the HIV-positive males, 87.6% are MSM.

YEAR	HIV+	AIDS	DEATH DUE TO AIDS
1986	99	5	2
1987	75	11	4
1988	93	14	6
1989	121	13	5
1990	140	28	23
1991	183	37	17
1992	175	71	33
1993	102	82	59
1994	122	102	62
1995	123	116	82
1996	235	99	93
1997	363	129	99
1998	362	150	99
1999	493	177	123
2000	545	258	143
2001	642	392	118
2002	644	447	92
2003	740	280	75
2004	768	226	109
2005	942	238	117
Total	6967	2875	1361

Table 20.1. *New HIV+, AIDS cases and deaths due to AIDS by year Cuba: 1986–2002*

From the table we can see the epidemic is very low-prevalent. Indeed, with a population of around 11 million, Cuba has a cumulative incidence rate for AIDS of 261 per million (13 per million per year). One of the characteristics of the Cuban program for the HIV/AIDS epidemic is that there is an active search of HIV-positive persons through the sexual contacts of known HIV-infected persons. As a result, for the period 2000-2005, 21.7% of all HIV-positive persons detected have been found through contact tracing. The rest of the infected persons are found through a “blind” screening, a search of HIV-positive individuals by serotesting of blood donors , pregnant women, persons with other sexually transmitted diseases, etc. at clinics. In recent years, through the community doctors, a new source of detection has been introduced that is having an impact in the detection system. This form of detection is done by the family doctors that, through counseling, suggest to persons that may be at risk of having the virus to take the HIV test. For the period 2000–2005, 27% of those detected have been detected by the family doctors. Non-parametric estimation of the mean time for the health authority to find a sexual partner notified by a detected HIV-positive person through the Contact Tracing Program has been found to be 54.3 months, with a standard deviation of 0.631 (Figure 20.1).

Contact tracing has been used as a method to control endemic contagious diseases [HET 82, HET 84]. While there is still debate about contact tracing for the HIV infection [APR 95, RUT 88], the resurgence of infectious tuberculosis and outbreaks of drug-resistant tuberculosis secondary to HIV-induced inmunodepression is forcing many public health departments to reexamine this policy [ALT 97, CDC 91]. The introduction of anti-retroviral therapy (ART) is also changing the way that health authorities will look in the future at the epidemic. A model of the HIV epidemic allowing for contact tracing would help evaluate the effect of this method of control on the size of the HIV epidemic, and give some idea as to the effectiveness of the health system in finding them.

Our objective is to model the contact tracing aspect of the HIV detection system, to try to obtain some information that could be useful to the Health System in Cuba in evaluating the way the program is working, and to ascertain its usefulness in terms of intervention and treatment of HIV. Other models have been used to study the effect of contact tracing with this objective in mind [LOU 99, ARA 00]. However, these were essentially linear models. We will now introduce non-linearity to model contact tracing. We will also discuss the implications of our results for the purpose of intervention and treatment of HIV/AIDS in Cuba, and to estimate the size of the epidemic in Cuba.

20.2. The models

As we noted, the Cuban program to control the HIV/AIDS epidemic is based on the active search for persons infected with HIV, long before they show any signs of AIDS. Our objective is not to model how new infections by HIV are generated, but how the HIV-infected persons are detected. We will consider the following variables:

1) $X(t)$ is the number of HIV-infected persons that do not know they are infected at time t,

2) $Y(t)$ is the number of HIV-infected persons that know they are infected at time t,

3) $Z(t)$ is the number of persons with AIDS at time t.

As the detection system has several search methods, we will separate the individuals in $Y(t)$ into two classes:

- $Y_1(t)$ is the number of HIV-infected persons that know they are infected at time t and were detected in a random type search,
- $Y_2(t)$ the number of HIV-infected persons that know they are infected at time t and were detected using contact tracing.

Evidently, $Y(t) = Y_1(t) + Y_2(t)$ for all t. With the following constant coefficients:

1) N the sexually active population,

2) α the rate of recruitment of new HIV-infected persons infected by X,

3) α' the rate of recruitment of new HIV-infected persons infected by Y,

4) k_1 the rate at which the unknown HIV-infected persons are detected by the system, independently of other HIV-positive persons (through "random" screening),

5) k_2 the rate at which unknown HIV-infected persons are detected by the system through other sources, independent of contact tracing (random screening).

6) β the rate at which the undetected HIV-positive persons develop AIDS, reciprocal of the mean incubation,

7) β' the rate at which the detected HIV-positive persons develop AIDS, the reciprocal of the mean time it takes to go from Y to Z,

8) μ the mortality rate of the sexually active population,

9) μ' the mortality rate of the population with AIDS.

The model dynamics is described by the following system:

$$\begin{aligned}
\frac{dX}{dt} &= \alpha N X + \alpha' N Y - (k_1 + \mu + \beta)\, X - f(k_2, X, Y),\\
\frac{dY_1}{dt} &= k_1 X - (\mu + \beta')\, Y_1,\\
\frac{dY_2}{dt} &= f(k_2, X, Y) - (\mu + \beta')\, Y_2,\\
\frac{dZ}{dt} &= \beta X + \beta' Y - \mu' Z.
\end{aligned} \tag{20.1}$$

We consider the system only in the region $\mathcal{D} = \{X \geq 0,\ Y \geq 0,\ Z \geq 0\}$. It is clear that $\mathcal{D}$ is positively invariant under the flow induced by (20.1). First, we make the following three remarks regarding the system in (20.1):

1) In (20.1) there are two ways individuals can move from the unknown HIV infected class (X) to the known HIV infected class (Y). One way is through the term $f(k_2, X, Y)$. This is the term we use to model contact tracing. In other words, the individual is found through his contact with persons that are known to live with HIV. The other way they can be detected is through the screening term $k_1 X$ which models all the other "random" ways of searching for HIV-positives. For other HIV models using constant and nonlinear screening terms, see [HSI 91, VEL 94, ARA 03]. It is important to note that $1/k_1$ can be viewed as the mean time from infection to detection for the persons found through means other than contact tracing.

2) The term $f(k_2, X, Y)$ models contact tracing. The way it is given in the model indicates that the process is one that goes on for a long time, because it involves all the individuals in the class Y and this is confirmed numerically by the result that the mean time to finding a contact is 54.3 months (Figure 20.1). If we consider the numerical result that the mean time from detection to AIDS is 86.8 months (Figure 20.2) we can see that, in the mean, infected sexual partners of an HIV-positive person are found more than four years after the person was detected as HIV-positive.

3) We assume that the known HIV infected persons are infectious, but at a much lower rate than those who do not know they are infected. In this sense α' will be taken as a fraction of α.

4) The passage to AIDS is modeled in a linear way. This could be modeled in a more general way, but for the Cuban case the best fit to an incubation curve is still an exponential. This can be seen in Figure 20.3 which gives us the cumulative hazard function for the time to AIDS which is a straight line and this corresponds to an exponential model.

Several possibilities arise for the contact tracing term $f(k_2, X, Y)$ [HSI 05a]. In this chapter, we will consider the following four models:

1) $k_2 X$

2) $k_2 Y$

3) $k_2 XY$

4) $k_2 \frac{XY}{X+Y}$

The first two models are linear models, while the last two are non-linear. In order to make full comparison, we will give some analytical results for each of the models and calculate the basic reproduction number for each one. Furthermore, we will fit the models to the Cuban contact tracing data to see which model explains the data most satisfactorily.

Let $k = k_1 + k_2$, $\lambda = \alpha N$ and $\lambda' = \alpha' N$.

20.2.1. *The k_2X model*

In this case the system is:

$$\begin{aligned} \frac{dX}{dt} &= (\lambda - k - \mu - \beta)X + \lambda' Y, \\ \frac{dY}{dt} &= kX - (\mu + \beta')\, Y, \\ \frac{dZ}{dt} &= \qquad\qquad\qquad\qquad \beta X + \beta' Y - \mu' Z. \end{aligned} \tag{20.2}$$

It is a linear model and the basic reproduction number is:

$$\mathcal{R}_0 = \frac{\lambda}{k + \mu + \beta} + \frac{\lambda'}{\mu + \beta'}\,\frac{k}{k + \mu + \beta}.$$

If $\mathcal{R}_0 > 1$, then all trajectories go to infinity, otherwise the disease-free equilibrium (DFE) at $(0, 0)$ is the only equilibrium and is globally asymptotically stable.

20.2.2. *The k_2Y model*

In this case the system is:

$$\begin{aligned} \frac{dX}{dt} &= \lambda X + \lambda' Y - (k_1 + \mu + \beta)X - k_2 Y, \\ \frac{dY}{dt} &= k_1 X + k_2 Y - (\mu + \beta')\, Y, \\ \frac{dZ}{dt} &= \qquad\qquad\qquad\qquad \beta X + \beta' Y - \mu' Z. \end{aligned} \tag{20.3}$$

It is also a linear model and the basic reproduction number is:

$$\mathcal{R}_0 = \frac{\lambda}{k_1 + \mu + \beta} + \frac{k_1 \lambda'}{(k_1 + \mu + \beta)(\mu + \beta')}\,\frac{k_2}{k_2 + \mu + \beta}\,\frac{\mu + \beta - \lambda}{\mu + \beta'}.$$

Again, if $\mathcal{R}_0 > 1$, then all trajectories go to infinity, otherwise $(0, 0)$ is unique and globally asymptotically stable.

20.2.3. *The k_2XY model*

This model is similar to one studied in [ARA 02], but only considers the variable Y. The system is

$$\begin{aligned}
\frac{dX}{dt} &= \lambda X + \lambda' Y - (k_1 + \mu + \beta)X - k_2XY,\\
\frac{dY}{dt} &= k_1X + k_2XY - (\mu + \beta')\,Y, \qquad (20.4)\\
\frac{dZ}{dt} &= betaX + \beta'Y - \mu'Z,
\end{aligned}$$

where $\lambda = \alpha N$ and $\lambda' = \alpha' N$.

In the region $\mathcal{D} = \{X \geq 0,\ Y \geq 0,\ Z \geq 0\}$, the system has two equilibria: one is the DFE at $P_0 = (0,0,0)$, and the other is a unique endemic point $P^* = (X^*, Y^*, Z^*)$ at

$$X^* = \frac{\sigma\,\gamma + \lambda' k_1}{k_2(\sigma + k_1)}, \quad Y^* = \frac{\sigma\,\gamma + \lambda k_1}{k_2(\gamma - \lambda')}, \quad Z^* = \frac{\beta X^* + \beta' Y^*}{\mu'}, \qquad (20.5)$$

with $\sigma = \lambda - k_1 - \beta - \mu, \quad \gamma = \beta' + \mu$.

The basic reproduction number for the system is

$$\mathcal{R}_0 = \frac{\lambda}{k_1 + \mu + \beta} + \frac{\lambda'}{\mu + \beta'}\,\frac{k_1}{k_1 + \mu + \beta}.$$

If $\mathcal{R}_0 < 1$ and the endemic equilibrium P^* is feasible (i.e. $\gamma > \lambda'$), then the DFE P_0 is stable and its basin of attraction consists of a triangle formed by the axes and the line of slope

$$\frac{\sigma + k_1}{\sigma(\lambda' - \gamma)}\left\{\frac{\lambda' k_1}{\gamma - \lambda'} + \lambda_1\right\},$$

that passes through P^*, where λ_1 is the negative eigenvalue of the Jacobian matrix at P^*.

If $\mathcal{R}_0 > 1$, then P_0 is unstable and P^* is globally asymptotically stable in the region $\mathcal{D}$.

If $\mathcal{R}_0 = 1$, P^* and P_0 coincide, 0 is a simple eigenvalue for the Jacobian at P_0 and the other eigenvalue is $\sigma - \gamma$ which is negative. Therefore, P_0 is globally asymptotically stable in region $\mathcal{D}$.

20.2.4. *The* $k_2\frac{XY}{X+Y}$ *model*

The system considered here is:

$$\begin{aligned}\frac{dX}{dt} &= \lambda X + \lambda' Y - (k_1 + \mu + \beta)X - k_2\frac{XY}{X+Y},\\ \frac{dY}{dt} &= k_1 X + k_2\frac{XY}{X+Y} - (\mu + \beta')\,Y,\\ \frac{dZ}{dt} &= \beta X + \beta' Y - \mu' Z.\end{aligned} \tag{20.6}$$

As before, we will consider the system formed by the first two equations in (20.6). Let $x = \frac{X}{X+Y}$, $y = \frac{Y}{X+Y}$ be the respective proportions of unknown and known HIV-positives in the HIV-positive population. Since $x + y = 1$, we have

$$x' = (\lambda' - \lambda + k_2 + \beta - \beta')x^2 + (\lambda - 2\lambda' - k_1 - k_2 + \beta' - \beta)x + \lambda'.$$

Here $x' = 0$unknown$-(\lambda x + \lambda'(1-x) - (\beta+\mu)x)\,(1-x) + k_1x + k_2x(1-x) - (\beta+\mu)x(1-x) = 0$. Equivalently $(\lambda - \gamma\lambda - k_2)x^2 - (\lambda - 2\gamma\lambda - k_1 - k_2)x - \gamma\lambda = 0$ has a unique positive solution x^* between 0 and 1. Let y^* be the corresponding value for y. x^* is globally asymptotically stable on the one-dimensional line between 0 and 1, which means that the straight line $\frac{X(t)}{Y(t)} = \frac{x^*}{y^*}$ on the positive XY-quadrant is an asymptotic line for all trajectories, either going to $(0,0)$ or infinity.

Let $\sigma = k_2y^*(1 + x^*)$. Then, the basic reproduction number of the system is:

$$\mathcal{R}_0 = \frac{\lambda}{\mu+\beta+k_1+\sigma}\,\frac{k_1}{\mu+\beta+k_1+\sigma}\,\frac{\lambda}{\mu+\beta'} - \sigma\,\frac{\lambda + k_1 - (\beta+\mu)}{(\mu+\beta')(\mu+\beta+k_1+\sigma)}.$$

20.3. The underreporting rate

In every detection system, there is always underreporting, that is, cases that were detected later. With the variables that we have defined in section 20.2, X gives the number of cases that were not reported at each instant t. The ratio of underreporting is defined as X/Y. In this section we will give a way of estimating this value using one of the linear models. We will use a model that is a variation of the one represented by Equation (20.2). We will consider $\lambda' = 0$.

$$\begin{aligned}\frac{dX}{dt} &= (\lambda - k - \mu - \beta)X,\\ \frac{dY}{dt} &= kX - (\mu + \beta')\,Y,\\ \frac{dZ}{dt} &= \beta X + \beta' Y - \mu' Z.\end{aligned} \tag{20.7}$$

From the solution of this system we have: $\frac{X(t)}{Y(t)} = \left[\left(\frac{1}{r_0} - \frac{1}{r^*}\right) e^{-(kr^*)t} + \frac{1}{r^*}\right]^{-1}$, where $r_0 = X_0/Y_0$ is the initial ratio of living undetected HIV-positives to the number of known HIV-positives at time $t = 0$, and

$$r^* = \frac{\lambda - k + (\beta' - \beta)}{k}.$$

This value will be called the *approximate rate of underreporting* [HSI 05b].

If $\lambda - k + \beta' - \beta > 0$, then we have:

$$r^* = \lim_{t \to \infty} \frac{X(t)}{Y(t)}.$$

Clearly, if the parameter values remain stable over time, then the ratio of the number of living undetected HIV-positives to the number of known living HIV-positives approaches r^* in time, hence the name “approximate rate of underreporting”.

Observe that when $(\lambda - k + \beta' - \beta)t_1$ is sufficiently large for a specific time t_1, $\frac{X(t_1)}{Y(t_1)} \approx r^*$. That is, r^* gives an approximation to the underreporting within the population at time t_1.

The expression for r^* is very intuitive. First, if r^* is negative, then the removal from the unknown HIV-infective class is so fast that there is no underreporting. Moreover, $(\beta' - \beta)$ gives the difference in the removal rates of the known HIV-positives and the undetected HIV-positives due to progression to AIDS, while $\lambda - k$ measures the rate of change of the undetected HIV-positives disregarding such removal. If either

1) β' is decreased (i.e. longer time from detection to AIDS),

2) the rate of infection λ is decreased, or

3) detection k is increased,

the resulting approximate ratio of underreporting will be smaller.

Note that an increase in k produces the most dramatic effect, as it also appears in the denominator of r^*. Thus, an intensive detection strategy is the most effective way to combat underreporting at the early stages of the epidemic, or when HIV prevalence is low.

20.4. Fitting the models to Cuban data

We fit the models to the data for the known HIV-positives and AIDS cases in Cuba. We will use the following values for the parameters which were estimated from the HIV data in Cuba:

– $X(0) \in [200, 230]$, the number of unknown HIV-positives in 1986 estimated from the number of HIV-positives who were detected after 1986 but were found to be already infected in 1986,

– $Y(0) = 94$, the number of HIV-positives who were known to be alive at the end of 1986,

– $Z(0) = 3$, the number of AIDS cases who were alive at the end of 1986,

– $\mu = 0.0053$, the yearly mortality rate for the HIV-positive cases in 1991–2000, ($SD = 0.0030$), calculated from the number of deaths for HIV-infected persons not related to AIDS,

– $\mu' \in [0.66, 0.85]$, obtained from the 95% confidence interval for the median of the survival time of AIDS (1987–2000),

– $\lambda = \alpha N = 0.5744$, the infection rate of the undetected HIV-positives obtained from the value of the parameter λ in the model developed in [ARA 00], (S.D.=0.0096),

– $\beta = 0.1135$, from the incubation period ($1/\beta$) estimated from 1,218 persons whose probable date of infection has been determined during the observation period 1987–2000, (S.D.=0.0031),

– $\beta' = 0.1350$, from the mean time from detection to AIDS ($1/\beta'$) (1987–2000), (S.D. = 0.0026).

– We take λ' the infection rate of the known HIV-positives, to be a fraction of λ ; $\lambda' = r\lambda$ and consider $r \in (0, 0.1)$.

We fit the models to the data to obtain values for k_1 and k_2 by minimizing a relative error function. As traditional optimization methods failed to work properly we used a genetic algorithm approach and a random search for local minima [ARA 00]. To calculate standard errors for the parameters, 300 fitting runs were made using different values for the known parameters taken randomly form their confidence interval.

Table 20.2 gives the mean values found for k_1 and k_2.

Model	mean k_1	sd k_1	mean k_2	sd k_2	mean error	sd error
k_2X	0.2423	0.0229	0.1232	0.0110	16.6823	1.2322
k_2Y	0.2786	0.0280	0.0984	0.0036	18.5417	1.0827
$k_2\frac{XY}{X+Y}$	0.2547	0.0242	0.2457	0.0133	17.1199	1.1790
k_2XY	0.3031	0.0254	0.00024	0.000038	20.3743	0.9922

Table 20.2. *Parameters*

20.5. Discussion and concluding remarks

The question now is which model is the best one in the sense of best fit for the data.

In other words, which of the four models offers the best model for contact tracing? By comparing the mean errors of the four models in Table 20.2, the order of suitability of the four models in the sense of the smallest mean errors is:

1) k_2X
2) $k_2\frac{XY}{X+Y}$
3) k_2Y
4) k_2XY

However, we note that the contact tracing term $k_2\frac{XY}{X+Y}$ in model 4 can be approximated by model 1 (i.e. k_2X) when $Y >> X$ and by model 2 (i.e., k_2Y) when $X >> Y$. In other words, model 4 approximates whichever that is the smaller of the two linear models whenever one contact tracing term is much larger than the other. Moreover, $k_2\frac{XY}{X+Y}$ is smaller than the corresponding contact tracing terms in either model k_2X or model k_2Y. Thus, model 4 ($k_2\frac{XY}{X+Y}$), as a model for the contact tracing, offers a good compromise between the two extremes of contact tracing in the linear models and should be the best from the theoretical point of view. Estimates of the unknown HIV-positive population in Cuba [ARA 03, HSI 02, HSI 01], though not a negligible number, have shown that, in recent years, approximately two-thirds of the HIV-positive persons in Cuba have been detected. Thus, realistically model 4 is probably more appropriate than either models 1 or 2. The simple “mass action” contact tracing term in model 3 (k_2XY) gives the “worst” fit and should be discarded.

To further use our result, we calculate the mean detection time for random screening to detect an HIV-positive, $(1-p)/k_1$, for models 1, 2 and 4, and the mean detection time for contact tracing to detect one HIV-positive person, p/k_2, for model 1 only. Here $p = 0.291$ is the proportion of known HIV-positive persons detected through contact tracing. The results are given in Table 20.3. The result indicates that using model 1, the contacting tracing program shortens the time of detection for an HIV-positive person by 6.8 months. The other models cannot be used to draw any similar conclusions.

model	mean detection time by random screening (A)	mean detection time by contact tracing (B)	difference (A-B)
k_2X	35.1 months	28.3 months	6.8 months
$k_2\frac{XY}{X+Y}$	33.4 months	NA	NA
k_2Y	30.1 months	NA	NA

Table 20.3. *Mean detection time by random screening and contact tracing, when applicable*

For the underreporting ratio we fitted model (20.7) to the data (this was done again because there is no λ') and obtained an average value for r^* of 0.3058 (IQR: 0.2372 - 0.3888). We can use this value to estimate the percentage of the HIV/AIDS epidemic that has been detected by the health system. We need the proportion $\frac{Y(t)}{X(t)+Y(t)}$. With the mean value for r^* we get that the system has detected 76.6%. If we take the extremes of the IQR we have values between 71.0% and 81.0%.

The detection system has, according to the models, an impact on the size of the epidemic. However, if only around 25% of the infected population is infective, this means that the size of the epidemic does not grow very fast. Typically, once a person knows his or her HIV sero-positive status, change in sexual behavior occurs even without having a good educational background. Only a minority of people keep on indulging in risky behavior after knowing his/her positive serological status. Thus, early diagnosis through random screening, contact tracing, and, more recently, anonymous testing has been instrumental in keeping the HIV prevalence in Cuba at a low level [JOA 03]. The detection of HIV-positive persons subsequently made their treatment possible. The first therapeutic methods employed in Cuba, appearing first in 1986, consisted of the use of domestically produced immune-modulators. Treatments making use of transfer factor and the recombinant interferon alpha were conducted with satisfactory results. In 1987, AZT therapy was introduced into Cuba's health care system. Donations made by individuals and international organizations made it possible in 1996 to offer triple AIDS therapy to 100 Cubans afflicted with the disease. Though more substantial donations were made in the years to follow, due the recent increase in the HIV-infected population size [HSI 01] the needs of the population were, at that point, not entirely met [ABR 03]. In 1997, when the domestic production of these pharmaceuticals entered its research phase, the Cuban government paid international prices to acquire all of the medication needed to offer triple therapy (HAART) to mothers and children who were HIV-positive. From 2001 onwards, a wider variety of domestically manufactured anti-retroviral agents became increasingly available in Cuba, resulting in a 100% level of coverage by the end of 2002. With the advances in therapeutic treatment, the early detection and diagnosis of HIV-positive persons through contact tracing, as evident from our modeling, has taken an increasingly important role in improving the quality of life for those living with HIV/AIDS.

Acknowledgment

R. Lounes and H. de Arazoza would like to thank "Épidémiologie de l'infection du VIH à Cuba: Modélisation stochastique et prévision" for financial support.

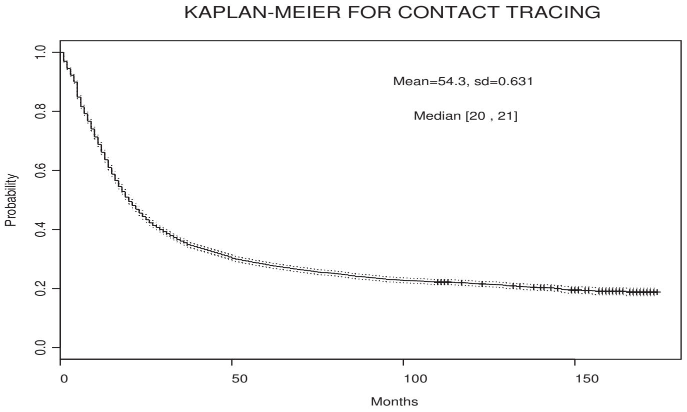


Figure 20.1. *Estimated time for contact tracing using declared sexual partners from 1986-2001, using the Kaplan-Meier method*

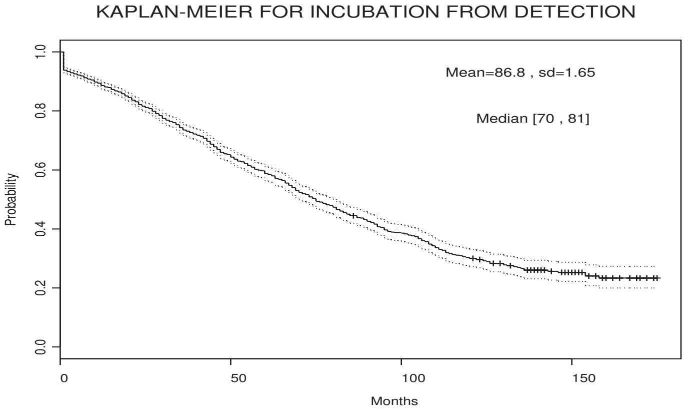


Figure 20.2. *Estimated time from detection to AIDS-defined illness (ADI) for 4,517 HIV/AIDS patients in Cuba from 1986-2002 using the Kaplan-Meier method*

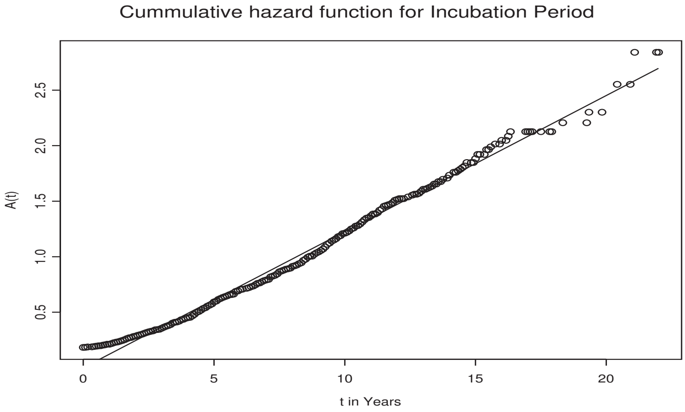


Figure 20.3. *Cumulative hazard function for time to AIDS in Cuba, 1986-2002*

20.6. Bibliography

[ABR 03] ABREU M. I. L., Head of National Aids Programme, personal communication, Ministry of Public Health, Cuba, 2003.

[ALT 97] ALTMAN L., "Sex, privacy and tracking the HIV infection.", *New York Times*, 11/4/1997.

[APR 95] APRIL K., THÉVOZ F., "Le contrôle de l'entourage (contact tracing) a été négligé dans le cas des infections par le VIH", *Revue Médicale de la Suisse Romande*, vol. 115, p. 337–340, 1995.

[ARA 00] ARAZOZA H., LOUNES R., HOANG T., INTERIAN Y., "Modeling HIV epidemic under contact tracing – the Cuban case", *Journal of Theoretical Medicine*, vol. 2, p. 267–274, 2000.

[ARA 02] DE ARAZOZA H., LOUNES R., "A non linear model for a sexually transmitted disease with contact tracing", *IMA. J. Math. Appl. Med. Biol*, vol. 19, num. 3, p. 30–37, 2002.

[ARA 03] DE ARAZOZA H., LOUNES R., PÉREZ J., HOANG T., "What percentage of the Cuban HIV-AIDS epidemic is known?", *Revista del Instituto de Medicina Tropical de Cuba*, vol. 55, num. 1, p. 30–37, 2003.

[CDC 91] CDCP, "Transmission of multidrug resistant tuberculosis from an HIV positive client in a residential substance-abuse treatment facility — Michigan", *MMWR*, vol. 40, num. 8, p. 129, 1991.

[GRA 95] GRANICH R., JACOBS B., MERMIN J., PONT A., "Cuba's national AIDS program – the first decade", *West J Med.*, vol. 163, num. 2, p. 139-144, 1995.

[HET 82] HETHCOTE H., YORKE J., NOLD A., "Gonorrhea modeling: comparison of control methods", *Math. Biosci.*, vol. 58, p. 93–109, 1982.

[HET 84] HETHCOTE H., YORKE J., *Gonorrhea Transmission Dynamics and Control*, Springer Verlag, lecture notes in Biomathematics 56, 1984.

[HSI 91] HSIEH Y. H., "Modeling the effect of screening in HIV transmission dynamics", *Proceedings of International Conference on Differential Equations and its Applications to Mathematical Biology, Claremont*, Berlin Heidelberg, New York, Springer-Verlag, Lecture Notes Biomath., V.92, p. 99–120, 1991.

[HSI 01] HSIEH Y. H., LEE S. M., CHEN C. W. S., ARAZOZA H., "On the recent sharp increase of HIV infections in Cuba", *AIDS*, vol. 15, num. 3, p. 426–428, 2001.

[HSI 02] HSIEH Y. H., ARAZOZA H., LEE S. M., CHEN C. W. S., "Estimating the number of HIV-infected Cubans by sexual contact", *Int. J. of Epidemiology*, vol. 31, p. 679–683, 2002.

[HSI 05a] HSIEH Y. H., DE ARAZOZA H., LOUNES R., JOANES J., "A class of methods for HIV contact tracing in Cuba: implications for intervention and treatment", *Deterministic and Stochastic Models for AIDS Epidemics and HIV Infection with Interventions*, Singapore, World Scientific, Ed. W.Y. Tan, p. 77–92, 2005.

[HSI 05b] HSIEH Y.-H., WANG H.-C., DE ARAZOZA H., LOUNES R., TWU S.-J., HSU H.-M., "Ascertaining HIV underreporting in low prevalence countries using the approximate ratio for underreporting", *Journal of Biological Systems*, vol. 13, num. 4, p. 441–454, 2005.

[JOA 03] JOANES-FIOL J., National Aids Programme, personal communication, Ministry of Public Health, Cuba, 2003.

[KOI 02] KOIKE S., "Assessment of the Cuban approach to AIDS and HIV", *Nippon Koshu Eisei Zasshi*, vol. 49, num. 12, p. 1268–77, 2002.

[LOU 99] LOUNES R., DE ARAZOZA H., "A two-type model for the Cuban national programme on HIV/AIDS", *IMA. J. Math. Appl. Med. Biol*, vol. 16, p. 143–154, 1999.

[PER 96] PEREZ J., TORRES R., JOANES J., M. LANTERO H. D. A., "HIV control in Cuba", *Biomed. and Pharmacother*, vol. 50, p. 216–219, 1996.

[RUT 88] RUTHERFORD G., WOO J., "Contact Tracing and the Control of Human Inmunodeficiency Virus", *J. Amer. Med. Assoc.*, vol. 259, p. 3609–3610, 1988.

[SWA 95] SWANSON J. M., GILL A. E., WALD K., SWANSON K. A., "Comprehensive care and the sanatorium: Cuba's response to HIV/AIDS", *JANAC*, vol. 6, num. 1, p. 33–41, 1995.

[VEL 94] VELASCO-HERNANDEZ J. X., HSIEH Y. H., "Modeling the effect of treatment and behavioral change in HIV transmission dynamics", *J. Math. Biol.*, vol. 32, p. 233–249, 1994.

This page intentionally left blank

Chapter 21

Some Probabilistic Models Useful in Sport Sciences

21.1. Introduction

Research in sport sciences is increasingly complex, and necessities support increasingly specific statistics. In this chapter, we present five fields of research in sport sciences which are subjects of studies and statistical modelings by the group “Statistics in sport sciences” of the laboratory “Statistique Mathématique et ses Applications – University of Bordeaux 2”. In the first part, we tackle the problem of the evaluation of sportsmen by a jury. Here, we present an application in figure skating modeled by the Gauss-Markov model. Next, we tackle the problem of the planning and optimization of performance. Within this framework, a new formulation of the Banister model is proposed and simulations are done.

We continue with industrial problems: the evaluation of individual sport equipment. Here, the sensorial analysis and its treatment by fuzzy subset theory are at the heart of the statistical processing controlled experiment.

The statistical analysis of the result of the duel type sports match is approached in the fourth part of this chapter. We propose a modeling by logistic regression, followed by sequential numerical simulations which create a new perspective on the analysis of the result. Lastly, we tackle the problem of modeling in sporting epidemiology. We

Chapter written by Léo GERVILLE-RÉACHE, Mikhail NIKULIN, Sébastien ORAZIO, Nicolas PARIS and Virginie ROSA.

propose using degradation models (initiated reliability) to model osteoarthritis of the knee.

21.2. Sport jury analysis: the Gauss-Markov approach

The use of a jury to evaluate sports is a natural and historical procedure and impossible to avoid today. This evaluation is a group decision consisting of voting or any other decision-making that uses several parameters (see [GN 99], [GN 96] and [ZUE 97] for example). Thus, it is essential according to game theory that the mode of taking into account individual opinions determines the final decision. For sporting juries, compilations of the notes of the judges are varied. Are these procedures optimal? Up to what point do they allow an undeniable classification? The statistical methods presented here bring partial answers, but more than that show the limits of the existing methods using an example of the results of a figure skating world championship by a couple in the 1970s.

21.2.1. *Gauss-Markov model*

Suppose that I skaters are evaluated by J judges. Each judge evaluates each skater independently with the same scale.

Let x_{ij} be the numeric evaluation of judge j for skater i, $i = 1, ..., I$; $j = 1, ..., J$. Suppose that x_{ij} is the realization of the following random variable:

$$X_{ij} = a_i + b_j + \epsilon_{ij}, \quad i = 1, ..., I; \quad j = 1, ..., J,$$

where a_i is the "true value" of the skater i, b_j is the "systematic error" (or bias) of the judge j and ϵ_{ij} is a random error.

We suppose that ϵ_{ij} are iid with normal probability $\mathcal{N}(0,\sigma^2)$. Values σ^2, a_i and b_j are unknown. Finally, we assume that $b_1 + b_2 + ... + b_J = 0$.

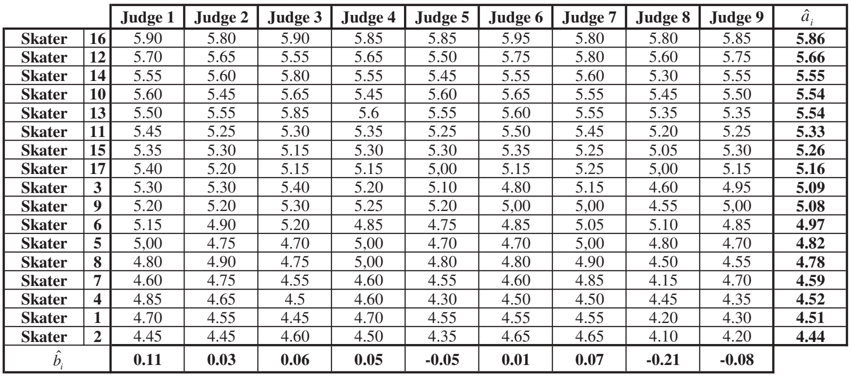
Then it is easy to obtain MSE estimators of a_i and b_j. We have:

$$\widehat{a}_i = \frac{1}{J}\sum_{j=1}^{J} x_{ij} \qquad \text{and} \qquad \widehat{b}_j = \frac{1}{I}\sum_{i=1}^{I} x_{ij} - \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} x_{ij}.$$

21.2.2. *Test for non-objectivity of a variable*

We wish to test the hypothesis H_0 that at least one variable is an outlier. Let

$$\Lambda = \frac{IJ}{(J-1)(I-1)} \frac{\max_{ij} \delta_{ij}}{S^2} \quad \text{where} \quad \delta_{ij} = X_{ij} - \widehat{a}_i - \widehat{b}_j, S^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \delta_{ij}^2.$$



		Judge 1	Judge 2	Judge 3	Judge 4	Judge 5	Judge 6	Judge 7	Judge 8	Judge 9	$\hat{a}_i$
Skater	**16**	5.90	5.80	5.90	5.85	5.85	5.95	5.80	5.80	5.85	**5.86**
Skater	**12**	5.70	5.65	5.55	5.65	5.50	5.75	5.80	5.60	5.75	**5.66**
Skater	**14**	5.55	5.60	5.80	5.55	5.45	5.55	5.60	5.30	5.55	**5.55**
Skater	**10**	5.60	5.45	5.65	5.45	5.60	5.65	5.55	5.45	5.50	**5.54**
Skater	**13**	5.50	5.55	5.85	5.6	5.55	5.60	5.55	5.35	5.35	**5.54**
Skater	**11**	5.45	5.25	5.30	5.35	5.25	5.50	5.45	5.20	5.25	**5.33**
Skater	**15**	5.35	5.30	5.15	5.30	5.30	5.35	5.25	5.05	5.30	**5.26**
Skater	**17**	5.40	5.20	5.15	5.15	5,00	5.15	5.25	5,00	5.15	**5.16**
Skater	**3**	5.30	5.30	5.40	5.20	5.10	4.80	5.15	4.60	4.95	**5.09**
Skater	**9**	5.20	5.20	5.30	5.25	5.20	5,00	5,00	4.55	5,00	**5.08**
Skater	**6**	5.15	4.90	5.20	4.85	4.75	4.85	5.05	5.10	4.85	**4.97**
Skater	**5**	5,00	4.75	4.70	5,00	4.70	4.70	5,00	4.80	4.70	**4.82**
Skater	**8**	4.80	4.90	4.75	5,00	4.80	4.80	4.90	4.50	4.55	**4.78**
Skater	**7**	4.60	4.75	4.55	4.60	4.55	4.60	4.85	4.15	4.70	**4.59**
Skater	**4**	4.85	4.65	4.5	4.60	4.30	4.50	4.50	4.45	4.35	**4.52**
Skater	**1**	4.70	4.55	4.45	4.70	4.55	4.55	4.55	4.20	4.30	**4.51**
Skater	**2**	4.45	4.45	4.60	4.50	4.35	4.65	4.65	4.10	4.20	**4.44**
$\hat{b}_i$		**0.11**	**0.03**	**0.06**	**0.05**	**-0.05**	**0.01**	**0.07**	**-0.21**	**-0.08**	

Figure 21.1. *Evaluation of judges and estimations of* $\widehat{a}_i$ *and* $\widehat{b}_j$

Using the Chauvenet test (see [GN 96]), we show that under H_0 for all $z \geq 0$,

$$\mathbf{P}\{\Lambda \geq z\} \leq IJ\left[1 - I_z\left(\frac{1}{2}, \frac{1}{2}(IJ - I - J)\right)\right],$$

where $I_z\,(x, y)$ is the incomplete beta Euler function. We can test the hypothesis H_0 according to which $x_{i_0 j_0}$ is an objective variable (where i_0 and j_0 are indices for which δ_{ij} is maximal) against the hypothesis H_1 that this variable is non-objective (an outlier).

If we reject H_0, we can correct $x_{i_0 j_0}$ as following:

$$\widetilde{x}_{i_0 j_0} = x_{i_0 j_0} + \frac{IJ}{(I-1)(J-1)}\delta_{i_0 j_0}$$

Example: $\Lambda = 0.198$ and $\mathbf{P}\{\Lambda \geq 0.198\,|H_0\} \leq 0.000015$. We reject H_0 and we deduce that the variable given by judge 8 to skater 6 is too great. We can correct it with the value: $\widetilde{x}_{6;8} = 4.7$ (instead of $x_{6;8} = 5.1$).

21.2.3. *Test of difference between skaters*

We wish now to test the hypothesis that the "true values" of k skaters (with indices $(i_1, ..., i_k) = \Phi$) are the same i.e.: $H_0(\Phi)$:$a_{i_1} = ... = a_{i_k}$. Let

$$X^2(\Phi) = \sum_{i \in \Phi}\left(a_i - \frac{1}{k}\sum_{i \in \Phi} a_i\right)^2.$$

It is easy to show, using ANOVA techniques, that under $H_0(\Phi)$:

$$Z = \frac{J(J-1)(I-1)}{(k-1)S^2} X^2(\Phi) \sim F_{k-1,(J-1)(I-1)}.$$

We can test the hypothesis $H_0(\Phi)$ that k skaters are indistinguishable against the hypothesis that at least two of the k skaters have different "true values".

Example: We wish to test:

a) "Skaters 14, 13 and 10 have the same true value". $H_0(\Phi)$:$a_{10} = a_{13} = a_{14}$.

$$Z = 0.00577 \text{ and } \mathbf{P}\{Z \geq 0.00577\,|H_0(\Phi)\} = 0.99$$

$H_0(\Phi)$ is not rejected i.e. the fact that those skaters are indistinguishable.

b) "Skaters 16 and 12 have the same true value". $H_0(\Phi)$:$a_{12} = a_{16}$.

$$Z = 10.6 \text{ and } \mathbf{P}\{Z \geq 10.6\,|H_0(\Phi)\} = 0.0015$$

We reject $H_0(\Phi)$ at the significance level 0.05 and we conclude that there is a significant difference between those two skaters.

21.2.4. *Test for the less precise judge*

This last test can detect if one judge j_0 is significantly less precise than the others. We test the hypothesis:

$$H_0 : \mathbf{Var}(Y_{ij}) = \sigma^2, i = 1, ..., I\ ;\ j = 1, ..., J,$$

against

$$H_1 : \begin{cases} \mathbf{Var}(Y_{ij_0}) > \sigma^2,\ i = 1, ..., I \\ \mathbf{Var}(Y_{ij}) = \sigma^2,\ i = 1, ..., I\ ;\ j = 1, ..., J\ ; j \neq j_0. \end{cases}$$

Let

$$\Lambda_j = \frac{J}{(J-1)S} \sum_{i=1}^{I} \delta_{ij}^2.$$

Using one more time the Chauvenet test techniques, we show that, under H_0, for all $z \geq 0$,

$$\mathbf{P}\{\max \Lambda_j \geq z\} \leq J \left[1 - I_z\left(\frac{I-1}{2}, \frac{(I-1)(J-2)}{2}\right)\right]$$

Example: $\max \Lambda_j = \Lambda_8 = 0.279$ (we find again the judge number 8). $\mathbf{P}\{\max \Lambda_j \geq 0.279\,|H_0\} \leq 0.01$. We reject then H_0 at the significance level 0.05, and then the judge 8 is significantly less precise than the others.

21.3. Sport performance analysis: the fatigue and fitness approach

It has become common to apply systems theory to describe the effects of physical training on physical performances ([BCS 75], [BAN 91], [BCZ 99], [BBB 02], [BUS 03], [GOA 03], [AG 04]). In such an approach, the subject is characterized by a system in which training loads represent the input and performance the output. Therefore, understanding how the system works requires compiling daily training loads and recording as many performances as possible over a defined time period. From training loads and recorded performances it is thus possible to simulate individual profiles of fitness and fatigue, both of which provide interesting perspectives of physiological testing in athletes and other populations. How fitness and fatigue actually decline during a period of time where training loads are either absent or significantly reduced is of great importance in producing training programs. Successive periods of training and taper have to be accurately controlled in order to reach the maximal performance at the right moment of the sport season (see [BCZ 99] and [BBB 02]). In 1975, Banister [BCS 75] proposed the first mathematical function to model the relations between training and performance. This model assumes that the difference between two functions accordingly termed fitness, $g(t)$ and fatigue, $h(t)$ is a good estimator of the athlete performance at the time t. Indeed, once training loads and the corresponding series of performances have been compiled, the mathematical treatment is difficult, because at a specific time t, $g(t)$ and $h(t)$ are the cumulative result of a series of training doses (w_j).

The main aim of this chapter was to propose alternative formulations of fitness, $g(t)$ and fatigue, $f(t)$ to be able to implement the model with any datasheet program. A second issue faced by this study was to estimate the accuracy level required for the constant parameters included in the fitness and fatigue functions. The errors associated with experimental application of the systems model are difficult to assess and a sufficient statistical analysis is still lacking to assess the quality of the model. It is worth noting that the quality of the fit of modeled performance against recorded performance, as often assessed by the R2 or adjusted R2 (see [BUS 03]), does not provide the quality of the estimates of fitness and fatigue. Therefore, although the measurement itself is straightforward, there is a need for methods for the estimation of variability and confidence intervals of the estimates in order to evaluate the statistical significance of the information obtained, particularly since obtaining repeated measurements is often not practical. An approach that economists commonly use to test the reliability of predictions is the Monte Carlo simulation.

21.3.1. *Model characteristics*

Original formulation of Banister's model [BCS 75]

Briefly, the theoretical performance $p(t)$ generated by the Banister model at any instant t is such that:

$$p(t) = p_0 + g(t) - h(t) + \epsilon(t),$$

where p_0 is the initial level of performance before the considered training period, $g(t)$ and $h(t)$ are evaluated as iterated functions of discrete training impulses, and $\epsilon(t)$ is the distributed residual error. At a specific time t, $g(t)$ and $h(t)$ are the cumulative result of a series of training doses (w_j, with j varying from 1 to $t-1$):

$$g(t) = k_1 \sum_{j=1}^{t-1} w_j e^{-(t-j)/\tau_1}, \quad h(t) = k_2 \sum_{j=1}^{t-1} w_j e^{-(t-j)/\tau_2}.$$

Thus, once the modeled performance $p(t)$ has been fitted to real performances measured serially throughout the period of interest, fitness and fatigue are characterized by the value taken by their respective weighting factors (k_1 and k_2) and time constants (τ_1 and τ_2).

Alternative formulation of Banister's model

We proposed presently that the Banister model can be formulated as a system of auto-recursive functions ARX, modeling fitness and fatigue (see [GOA 03], [AG 04]) We discover that $g(t)$ and $h(t)$ satisfy:

$$g(t) = M_1 g(t-1) + I_1 w_{t-1}, \quad h(t) = M_2 g(t-1) + I_2 w_{t-1},$$

with $M_i = e^{-1/\tau_i}$, $I_i = k_i e^{-1/\tau_i}$, $i = 1, \ldots, 2$.

As in the original Banister model, fitness and fatigue are characterized by the value taken by:

(i) their respective weighting factors presently termed Impact (I_1 and I_2) of training to avoid confusion with the original term;

(ii) the value taken by the parameters indicating their rate of decrease presently termed Memories (M_1 and M_2) because they indicate the amount of fitness (respectively fatigue) that can be retained in the system at time t as a function of $t-1$. For instance, $M_1 = 0.9$ in the above equation of $g(t)$ would indicate that 90% of the amount of fitness at $t-1$ has been retained at t.

The model becomes a system of ARX, thus providing an alternative formulation of the Banister model that is easily programmable in any worksheet.

21.3.2. *Monte Carlo simulation*

A pattern of $p(t)$ was generated from an hypothetical series of daily training loads and a set of known parameters whose arbitrary values were presently: $p_0 = 100$, $M_1 = 0.8$, $I_1 = 0.01$, $M_2 = 0.6$, $I_2 = 0.02$.

A series of random Gaussian-distributed errors with mean zero and variance $\sigma^2 = 9$, $N(0; \sigma^2)$ were imposed on the 30 simulated performances before performing each fit. The procedure was repeated 1,000 times so that 1,000 estimates of each parameter were obtained and analyzed in terms of distribution and dependence.

21.3.3. *Results*

Estimates are not normally distributed except for p_0 (Figure 21.2) and relations between parameters indicated unclear independence.

It has been recently attempted to improve the original Banister model by introducing a gain term for the fatigue component which is a state variable depending on the amount of past training ([AG 04]). Although this is an interesting idea taking into account the possible increase in the fatigue effect resulting from repeated training sessions, this unfortunately requires the estimation of one additional free floating parameter in the original model. As we hope to have demonstrated in the present analysis of the statistical significance of modeled responses, there exists an obvious risk that many sets of parameters would allow us to fit modeled performances to real performances, thus making the actual individual responses finally unknown.

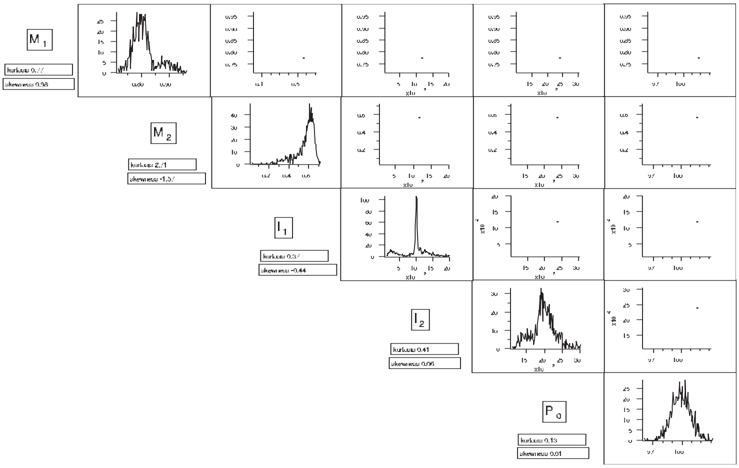


Figure 21.2. *Parameter estimates of the Banister model*

21.4. Sport equipment analysis: the fuzzy subset approach

Sport performance is considered physical, technical and tactical entertainment. Nevertheless, sports equipment has a non-negligible option to perform. Industries create perfectly fitted equipment to respond to athletes' needs. However, the industry gets into difficulties producing sports equipment for everybody. Indeed, every human is unique. So, physiological, biomechanical and sensorial responses to the same stimulus are specific to every being. In this chapter, we focus our research on sensorial analysis. We know that sensorial capacities change among subjects for three reasons:

1) every human has his/her own physiological state of the observation [MNS 02],

2) the same observation could be described by different terms depending if they are considered in one category or another [AB 97],

3) the same input could be perceived differently according to subject assessment, environment, etc. [GUI 54].

So, when we carry out sensorial analysis, we can not consider that every subject can give an unique answer for the same perceived stimulus. Also, adapted from [SHH 85], the terms of a semantic structure scale are not equidistant. Statistic analysis must take these comments into account. The sensorial analysis of sport equipment needs to try a method which improves results. A method for evaluating the psychological response has been elaborated by introducing the concept of fuzzy probability [ZAD 65], [HU A97], in particular on food (Urdapilleta *et al.*, 1999). This method consists of diffusing the sensorial answer according to the representation of this observation for the group of subjects. Moreover, sensorial analysis on sports equipment necessitates an analysis during physical effort exercise [RJH 01]. In this chapter, we choose to illustrate this sensorial method with the study of the free shoulder movement in surf wetsuits during an oar spring race with ten subjects. The aim of this study is essentially to investigate the advantage of the application of the fuzzy subset theory in ergonomic sport equipment research.

21.4.1. *Statistical model used*

In classic set theory, a term can either belong to a set or not. It is a binary relation:

$$A \cup \bar{A} = U, \quad A \cap \bar{A} = 0.$$

A fuzzy subset is described by a characteristic function μ_A, defined on a set U and a value on $[0,1]$, which expresses the degree of belonging of U elements to A:

$$\mu_A : U \rightarrow [0,1].$$

The intersection between A and B is the fuzzy subset C such that:

$$\forall u \in U, \quad \mu_C(u) = min(\mu_A(u), \mu_B(u))$$

When the diffusion principle is applied, every answer must have the same importance. Then, the cardinality of fuzzy subset must be equal to one. The cardinality of a fuzzy subset is defined by:

$$|A| = \int_{u \in U} \mu_A(u).$$

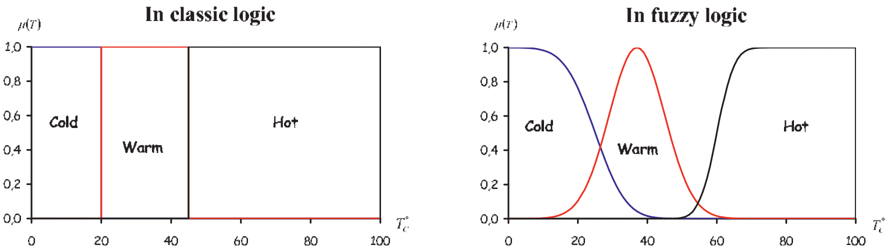


Figure 21.3. *Left: classic subset theory. Right: fuzzy subset theory*

The left graph of Figure 21.3 is a sensorial representation according to classic subset theory with the example of the assessment with three items of water temperature: absolutely cold, absolutely warm or hot.

The right graph of Figure 21.3 is a sensorial representation according to fuzzy subset theory with the same example. On a defined range temperature, fuzzy subset theory superposes the cold term with the warm term and hot term. It seems similar to human functioning.

21.4.2. *Sensorial analysis step*

Urdapilleta *et al.* (2001) proposed a quotation scale of 8 numbers (from 0 to 7) to apply fuzzy subset theory. Two types of questionnaires are necessary to apply this method. In the first, subjects must define one or more number(s) for each term of each descriptor in accordance with their particular representation. The second questionnaire with item scales has been elaborated to qualify the free shoulder movements on wetsuits. After each race, the subjects might fill in these two questionnaires.

21.4.2.1. *Evaluation of the products by subjects*

Each subject i, $i = 1, \ldots, n$ evaluates surf wetsuits on a descriptor (free shoulders movements). At the end of each oar spring race, subjects evaluated the wetsuits tested by allotting a verbal item j, $j = 1, \ldots, k$. Let J_i be a verbal item chosen by individual i during the test.

21.4.2.2. *Measures of observations of the subjects*

The subjects are invited to fill in the calibration questionnaire in which they must change a verbal item$j$ into notes l, $l = 0, \ldots, m$ Let L_{ij} be the note allotted by individual i to a verbal item j. Let

$$C_{jl} = \frac{1}{n} \sum_{i=1}^{n} 1\{L_{ij} = L\}$$

be the association coefficient of note l with the item verbal j.

21.4.2.3. *Construction of the representative filters of each verbal item*

To build the filters with the distributions of the answers of the calibration questionnaire, the intersections of each item on the structured scale are calculated. Let $[F_{jj'}]_{k \times 1}$ be the filter of the verbal item j, ($j' = 1, \ldots, k$), with

$$F_{jj'} = \sum_{l=1}^{m} min(C_{jl}, C_{j'l}), \quad \forall j = 1, \ldots, k.$$

21.4.2.4. *Principle of diffusion*

Filters are cardinalized in order, each verbal item having the same weight. This consists of reducing the total of each item to 1. Let

$$D_{jj'} = \frac{F_{jj'}}{\sum_{j'=1}^{m} F_{jj'}},$$

be the cardinalized filters, and the diffused result of the item verbal j is:

$$R_j = \sum_{j'=1}^{m} \sum_{i=1}^{n} D_{j'j} 1\{J_i = j'\}$$

Data analysis will be carried out by applying fuzzy subsets theory.

21.4.3. *Results*

The two graphs show the heterogenity of subjects' feelings: the answers of each product are spread out the verbal scale. We can see that the artifacts present in the graph on the left concerning the curves of the products F evolved in the graph on the right. However, for wetsuit B, the artifact persists. Whereas it seems difficult to classify all wetsuits in the graph on the left, we can easily distinguish three groups in the graph on the right: wetsuits E seems to be the best, the second group we could

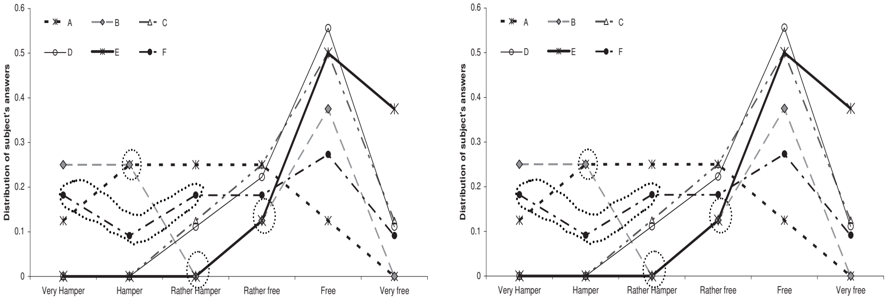


Figure 21.4. *Left: result of sensorial analysis without fuzzy subset theory. Right: result of sensorial analysis with fuzzy subset theory*

regard as average wetsuits consisting of (D and C), and the third group with the worst wetsuits consisting of A, B and F.

In our study and in agreement with [AB 97], semantic items have, on average, a membership of 3 notes on a scale of 8 numbers. Thus, the definition of the concept of threshold between two items is vague. This result consolidates the interest to smooth the data collected during a sensory analysis of sport equipment. The application of this theory during a sensory analysis rationalizes measurement. In this study, the extent of the non-smoothed sensory answers is probably related 1) to the differences in individual morphologies (even if this group were relatively homogenous), 2) to the row of passage of wetsuits for each subject and 3) to the differences in observations.

Essentially, using fuzzy subsets theory on sensory analysis of sport products decreases the assessment subjectivity and takes into account individual specifications. This method might allow industries to increase their equipment quality based on better subjective observations.

21.5. Sport duel issue analysis: the logistic simulation approach

A tennis match can be considered as a sequence of points, which are won or lost by opposing players. The confrontation between players can be represented as a sequence of Bernoulli random variables. For example, [KM 01a] worked on the distribution of points during a match and showed that the independence of points and homogenous distribution hypothesis can, in some circumstances, lead to a good rough estimate of the probability of winning the match. These studies led to an estimation of the probability of winning a match, at the beginning of the match and also while the match is being played, using a computing program called “Tennisprob”. However, this program assumes independence and an identical distribution of points. The authors again suppose that the probability of winning each point, irrespective of the other points, is a Bernoulli random variable.

It is easy to see the difficulty implicit in defending this hypothesis. Accepting this approach would imply that the length of a match has no effect on the gap in the players’ levels. It would clearly be wrong to postulate that neither of the two players feels tiredness, for example, or that this tiredness has exactly the same effect on both players. Furthermore, it appears relatively clear that the identity of the server plays an essential role in the probability of winning a point. The same holds for the success or otherwise of the first serve.

These remarks naturally lead us to propose an alternative method for modeling the match. Since it is a matter of modeling a sequence of binary variables ($Y = 1$ if the point is won by player A and otherwise 0), taking into account explanatory variables (identity of the server, success at the service, current score, etc.) the logistic model appears to be a sensible method. Moreover, with logistic regression we can calculate an individual probability of winning each point in the match. Through numerical simulations, it is then also possible to estimate the probability of winning the match.

The construction of a logistic model depends on the observation of real data; the match chosen for this study is the final of the Roland Garros tournament in 2000, won by Gustavo Kuerten against Magnus Norman.

21.5.1. *Modeling by logistic regression*

Modeling a match with logistic regression assumes that the probability of winning the jth point of the match can be represented by the following logistic function:

$$P\{y_j = 1|x_j\} = \frac{1}{1 + exp(-\alpha_0 - \sum_{i=1}^{k} \alpha_i x_{ij})},$$

where y_j is a Bernoulli variable worth 1 if the j-th point is won by the reference player and otherwise 0, and $x_j = (x_{1j}, \dots, x_{kj})^T$ is the vector of covariables determined by the position of the jth point of the match (server, success first ball and score). Supposing that the random variables y_j (knowing x_j) are independent, the coefficients α_i can be estimated by maximization of the following likelihood:

$$L(\alpha, x, y) = \prod_{j=1}^{n} P\{y_j = 1|x_j\}^{Y_j}(1 - P\{y_j = 1|x_j\})^{1-Y_j}$$

The quality of the model is assessed with a cross-validation method. The model is built on a base sample including 75% of the points of the final determined by an equiprobable random draw. Remaining points constitute a test sample. The rate of correct forecasts measured on the sample enables us to assess the quality of the model.

21.5.2. *Numerical simulations*

The main idea is to simulate the match a great number of times in order to deduce the probability, after the event, of each player winning it. To simulate a match, we only need to be able to simulate a point. The logistic model enables us to calculate the probability of winning each point knowing the vector of covariables associated with it.

The first stage involves calculating the probability of winning the first point in the match. At that time, there is obviously no score, and $P(Y = 1)$ depends only on the server's parameters and first ball. To determine the server, an equiprobable drawing of lots is performed as usual. The success of the first serve ball is simulated from the frequency of the first ball observed during the referring match. The interaction is calculated from both variables. In this way, $P(Y = 1)$ can be calculated and the winning of the point simulated. The result of the point is determined by drawing a uniform pseudo-random variable on $[0, 1]$. If the realization is under $P(Y = 1)$, the point will be won by Kuerten, otherwise it will be won by his opponent. The program calculates the winner of the point – the score before the second point is known. All that remains for us to do is simulate the success of the first ball to calculate $P(Y = 1)$ for the second point. When a game is over, the program automatically changes the server. To calculate $P(Y = 1)$ for every point of the match, we simply have to repeat the operations until one of the players gets 3 sets. The global probability of winning the match is estimated from 1,000 iterations of a match. The simulation of 1,000 matches, allocating the constant value of the frequency of points won by Kuerten (0.526) to $P(Y = 1)$, was also performed.

21.5.3. *Results*

Variables	Coefficients	p-value	Variables	Coefficients	p-value
Server	-0.289	0.4508	***Games Kuerten***	-0.242	0.4508
1st serve	-0.659	0.0678	***Games Norman***	0.232	0.0678
Interaction	2.034	0.0002	***Sets Kuerten***	-0.718	0.0002
Points Kuerten	-0.273	0.0366	***Sets Norman***	0.786	0.0366
Points Norman	0.234	0.0706	***Constante***	1.058	0.0706

Table 21.1. *Results of the logistic regression*

During this match, the most important variable is the interaction ($p < 0.001$). Moreover, coefficient a is very high (2.03). That means that when Kuerten is serving, and when he succeeds in his first ball, the probability of winning the point for Kuerten is higher than in other cases. The variable “server” does not seem significant. This conveys the fact that when a point is played on the second ball, both players neutralize themselves. The number of Kuerten’s sets seems too important with a high negative coefficient. This can be understood: the more sets Kuerten progresses to, the lower his probability of winning a point. A third set of this match has been led by Kuerten, that way it’s not surprising that we find back this kind of coefficient. On the contrary, the sets won by Norman increases Kuerten’s probability. When Norman wins a set, Kuerten leads less and his probability rises again. So, this coefficient goes in the same way as the number of Kuerten’s set coefficient. Finally, the constant, with a coefficient slightly over 1 and a p-value around 0.05 gives an advantage to Kuerten.

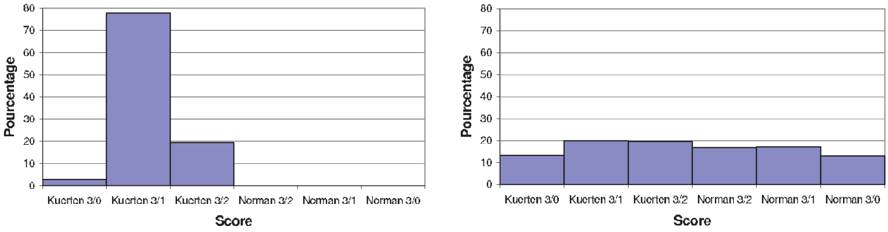


Figure 21.5. *Left: repartition of results with logistic probabilities. Right: repartition of results with the observed frequency* $p = 0.516$

The histograms presented above show the repartition of the results obtained allocating to each point a probability of winning calculated from the parameters of the regression (left graph) and the one obtained allocating the frequency observed as probability (right graph). The right graph clearly shows that with a constant probability ($P(Y = 1) = 0.51$) of winning each point, the probability of winning the match for Kuerten is of 0.53 with a repartition relatively uniform between the possible different scores. On the other hand, in the left graph, the results are not really different. Attributing to each point a probability of winning the point calculated from coefficients of the logistic regression, Kuerten will now win the match with a probability of nearly 1 and on the score of 3 sets to 1 for the essential matches. This vision is consistent with the appearance of the studied match.

The method suggested shows clearly that to model a match of tennis, we cannot accept the assumption that the probabilities of gaining each point independently are all Bernoulli random variables. The program “Tennisprob” initiated by Magnus and Klaassen [KM 01b] could thus be refined by integrating covariables highlighted by the logistic model. This research shows the interest of the logistic regression in the analysis *a posteriori* of a match of tennis.

21.6. Sport epidemiology analysis: the accelerated degradation approach

In this section, we deal with the problem of modeling degenerative osteoarthritis. There are two types of degenerative osteoarthritis: one connected to age and to heredity (where the aetiology is not clear); and one which is caused by other pathologies like obesity, joint trauma or repetitive joint use [FCC 98]. We are particularly interested in knee osteoarthritis which seems to be a particularly frequent pathology. An American study in 1987 [FNA 87] showed the dimension of this pathology: 33% of Americans from 63 to 94 years old are affected by knee degenerative osteoarthritis. Currently, the degenerative osteoarthritis which is a degenerative inflammatory pathology can only be slowed down. To our knowledge, we cannot find a cure for a degenerative knee. More interestingly, there is a possibility of quantitatively measuring the degree of knee degradation. For that, we can use the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) scores (see [BEL 95]). This score makes it possible to model with a mathematical function the promotion degree of the disease in time. Furthermore, this knee degeneration is able to evolve to a functional incapacity which we are also capable of measuring. In this case, the six minutes walk test (see [GST 85] and [GNR 00]) is the most commonly used. Thus, the functional incapacity could constitute the breakdown process necessary to model the knee degradation process with our model. Beyond the simple modeling of the degradation of the degenerative knee, we are able to estimate the effect of the covariable on the degradation functions obtained. Long-term exposure to the physical practice is a variable that will be associated with the degradation function. Many studies showed that the exposure to sport (cross-country, skiing, soccer, ice hockey) would be a factor of risk for the occurrence of osteoarthritis (2,9, IC 1.3-6.5).

21.6.1. *Principle of degradation in reliability analysis*

In degradation models, an item is regarded as failed when degradation reaches a critical level or when a censoring traumatic event occurs, and the degradation itself is modeled by a stochastic process $Z(t), t > 0$, with some properties, depending on the phenomena in consideration. The degradation process $Z(t)$ can be considered as an additional time-dependent covariable, which describes the process of wear or the usage history up to time t.

The ageing process is due to cumulated wear and fatigue. As failure time data can be very rare, an alternative is to measure some parameters characterizing degradation of the subject (the product) in time. Then a unit is considered as failed when its degradation (internal wear) reaches a critical level or when a traumatic event occurs. Intensity of traumatic failures is an increasing function of degradation value.

Degradation and failure time data in survival analysis, reliability and medical statistics are usually used to model the relation between the occurrence of the traumatic event at time T and the degradation process $Z(t)$. Having the joint probability model of the random element $(T, Z(t)$ we can estimate it and calculate the predictive probabilities of any events of interest (see [NBB 06], [NGO 06] and [BN 04]).

Denote by T^0 the moment at which degradation attains some critical value z_0. The failure time T_0 is sometimes called *soft failure* (or failure directly due to wear) because in most industrial applications, z_0 is fixed and the experiment is voluntarily ceased at the time the degradation process reaches level z_0 or just after this time. The moment of the unit's failure is $\tau = T^0 \wedge T$.

21.6.2. *Accelerated degradation model*

To model the degradation-failure time process, we suppose that the real degradation process is modeled by the *general path model* (see [ME 98]):

$$Z_r(t) = g(t, A)$$

where $A = (A_1, \ldots, A_r)$ is a positive random vector, and g is a continuously differentiable increasing in t function. The typical form of the degradation curves in the case $r = 2$ is

$$g(t, a) = e^{a_1}(1 + t)^{a_2},$$

where (a_1, a_2) is a realization of $A = (A_1, A_2)$. In the particular case of linear degradation, $a_2 = 0$.

Denote by h the function inverse to g with respect to the first argument. Evidently, it is continuously differentiable and increasing in t. Moreover,

$$T^0 = h(z_0, A).$$

The real degradation process $Z_r(t)$ often is not observed, and we have to measure (estimate) it. In this case the observed degradation process $Z(t)$ is different from the real degradation process $Z_r(t)$. For example, if we suppose that the values of the real degradation process are measured at time moments $t_1, \ldots, t_m$, we may consider a *degradation model with measurement errors*. According to this model the *observed degradation values* are

$$Z(t_j) = Z_r(t_j)\, U(t_j),$$

where $e_j = \ln U(t_j)$ are iid random variables, $e_j \sim N(0, \sigma^2)$. This model was studied in [BN 02],[BN 04], [NGC 07] and [NGO 06], for example.

In the continuous case we may consider the so called *degradation model with noise*. According to this model the *observed degradation process* is

$$Z(t) = Z_r(t)\, U(t),$$

where

$$V(t) = \ln U(t) = \sigma W(c(t)),$$

W being the *standard Wiener process* independent on A, and $c : [0, \infty) \to [0, \infty)$, $c(0) = 0$, being a specified continuous and increasing time function, $c(0) = 0$.

For any $t > 0$ the median of the random variable $U(t)$ is 1.

Using the second model, for example, it is easy to construct an interesting joint model.

Bagdonavicius and Nikulin [BN 02] proposed a model in terms of the conditional survival function of T given the realization of real degradation process:

$$S_T(t|A) = P\{T > t|g(s, A), 0 \leq s \leq t\} = \exp\left\{-\int_0^t \lambda_0(s, \theta)\lambda(g(s, A))ds\right\},$$

where λ is the unknown intensity function, and $\lambda_0(s, \theta)$ is from parametric family of hazard functions. The distribution of A is not specified. This model states that the conditional hazard rate $\lambda_T(t|A)$ at moment t given the degradation $g(s, A), 0 \leq s \leq t$, has the multiplicative form as in the famous Cox model:

$$\lambda_T(t|A) = \lambda_0(s, \theta)\lambda(g(s, A)).$$

If, for example, $\lambda_0(s, \theta) = (1 + t)^\theta$ or $\lambda_0(s, \theta) = e^{t\theta}$, then $\theta = 0$ corresponds to the case when the hazard rate at any moment t is a function of the degradation level at this moment. Here, the function λ, characterizing the influence of degradation on the hazard rate, is non-parametric.

It is clear that such a general approach can be used to analyze degenerative osteoarthritis, for example. We note that ten years ago Wulfsohn and Tsiatis (1997) considered another joint model for survival and longitudinal data measured with error, given by

$$\lambda_T(t \mid A) = \lambda_0(t)e^{\beta(A_1 + A_2 t)},$$

with bivariate normal distribution of (A_1, A_2). We can see the difference between two considered models: in the *Wulfsohn-Tsiatis model* the function λ, characterizing the influence of degradation on the hazard rate, is parametric, and in the Bagdonavičius-Nikulin model the function λ is non-parametric, which seems more natural in many practical cases, since in general we do not know in advance how degradation influences the hazard rate. On the other hand, the baseline hazard rate λ_0 (it is proportional to the hazard rate which should be observed if degradation is absent) is non-parametric in Wulfsohn-Tsiatis model and parametric in Bagdonavičius-Nikulin model.

Such general approaches can be used to analyze degenerative osteoarthritis. Such applications are in process in the “Statistique mathématique et ses applications” laboratory of University of Bordeaux 2.

21.7. Conclusion

This chapter tries to show how probabilistic approaches can be useful in sport sciences. Most of the time, statistical analysis in sport sciences is reduced to Student's test or Pearson correlation. We proposed, in some classical examples, new statistical approaches which show how data can be treated more precisely.

Of course, statistical solutions proposed here are not unique. For example, in the cause of the evaluation of sportsmen by a jury, item response theory used in quality of life analysis could produce interesting alternative models.

Sport scientists need statisticians to help them with their data analysis. We hope that this chapter will contribute to the development of statistics in sport sciences.

21.8. Bibliography

[AB 97] Aladenise N., Bouchon-Meunier B. (1997). *Acquisition de Connaissances Imparfaites: mise en Évidence d'une Fonction d'Appartenance.* Revue internationale de systémique, 11(1), 109–127.

[AG 04] Arsac L. Gerville-Réache L. (2004). *Statistical significance of fitness and fatigue modeled from training effects on performance*, 1st International French-Russian Workshop "Longevity, Aging and Degradation Models in Reliability, Public Health, Medicine and Biology", St Petersbourg, Russia 52–63.

[BAN 91] Banister, EW. *Modeling Elite Athletic Performance.* Physiological Testing of Elite Athletes, edited by H. J. Green, J. D. McDougal, and H. Wenger. Champaign, IL: Human Kinetics, 1991, 406–424.

[BBB 02] Busso T, Benoit H, Bonnefoy R, Feasson L, Lacour JR. (2002). *Effects of training frequency on the dynamics of performance response to a single training bout*. J Appl Physiol. 92(2): 572–580.

[BCS 75] Banister, EW, Calvert TW, Savage MV, Bach TM. (1975). *A systems model of training for athletic performance*. Aust. J. Sports Med. 7(3): 57–61.

[BCZ 99] Banister EW, Carter JB, Zarkadas PC. (1999). *Training theory and taper: validation in triathlon athletes*. Eur J Appl Physiol Occup Physiol. 79(2): 182–191.

[BEL 95] Bellamy N. (1995). *WOMAC Ostheoarthritis Index: A User's Guide*, London, Ontario.

[BER 93] Berg S. (1993). *Condorcet's jury theorem, dependency among jurors.* Social Choice and Welfare, (10) 87–95.

[BN 02] Bagdonavičius V, Nikulin M. (2002). *Accelerated life models*. Chapman & Hall/CRC, Boca Raton.

[BN 04] Bagdonavičius V, Nikulin M. (2004). Semiparametric analysis of degradation and failure time data with covariates, in *Parametric and Semiparametric Models with Applications to Reliability, Survival Analysis, and Quality of Life*, (Eds: Nikulin M., Balakrishnan N., Mesbah M., Limnios N.). Birkauser, Boston: 41–64.

[BUS 03] Busso T. (2003). *Variable Dose-response Relationship between Exercise Training and Performance*. Med Sci Sports Exerc. 35(7): 1188–1195.

[FCC 98] Felson DT, Couropmitree NN, Chaisson CE, Hannan M, Zhang Y, Mc Alindon TE, La Valley M, Levy D, Myers RH. (1998). *Evidence for a mendelin gene in a segregation analysis of generalized radiographic osteoarthrisis.* Arthritis Rheum 41, 1064–71.

[FNA 87] Felson DT, Naimark A, Anderson J, Kazis L, Castelli W, Meenan RF. (1987). *The Prevalence of Knee Osteoarthrisis in the Elderly. The Framingham Osteoarthrisis Study.* Arthrisis Rheum. 30, 914–8.

[GN 96] Greenwood P.E., Nikulin M. (1996). *A Guide to Chi-squared Testing.* John Wiley and Sons.

[GN 99] Gerville-Réache L., Nikulin M.S. (1999). *Analyse statistique de l'évaluation des sportifs par un jury.* Proceedings "XXXIème Journées de statistique", Grenoble, France, 51–54.

[GNR 00] Gail D, Nancy E, Robert L, Michael G, Matthew B, Stephen C. (2000). *Effectiveness of Manual Physical Therapy and Exercise in Osteoarthritis of the Knee.* Annals of Internal Medicine.132, 173–181.

[GOA 03] Gerville-Réache L., Orazio S., Arsac L. (2003). *Relations entraînement – performance : Modélisation et analyse statistique par la méthode de Monte-Carlo*. Congrès International des Chercheurs en Activités Physiques et Sportives, Toulouse, France, 268–269.

[GST 85] Guyatt GH, Sullivan MJ, Thompson PJ, Fallen EL, Pugsley SO, Taylor DW, *et al.* (1985). *The 6-minute walk: a new measure of exercice capacity in patients with chronic heart failure.* Can Med Assoc J. 132, 919–23.

[GUI 54] Guilford J.P. (1954). *Psychometric Methods*, New York, McGraw-Hill.

[HB 71] His, B.P. and Burych, D.M. (1971). *Games of two players*, Applied Statistics, 20, 86–92.

[HU A97] Huang (1997). *Principle of information diffusion.* Fuzzy Sets and Systems, 91, 69–90.

[KM 01a] Klaassen, F.J.G.M., Magnus, J.R. (2001a). *Are points in tennis independent and identically distributed? Evidence from a dynamic binary panel data model.* Journal of the American Statistical Association, 96, 500–512.

[KM 01b] Klaassen, F.J.G.M., Magnus, J.R. (2001b). *Forecasting the winner of a tennis match.* Center Discussion Paper (Int. r. no. 2001–38). Econometrics, 20 pp.

[ME 98] Meeker, W., Escobar, L. (1998) *Statistical Methods for Reliability Data*. Wiley, New York.

[MNS 02] Mündermann A, Nigg BM, Stefanyshyn DJ, Neil Humble R (2002) *Development of a reliable method to assess footwear comfort during running.* Gait and Posture 16, 38–45.

[NBB 06] Nikulin M, Barberger-Gateau P, Bagdonavičius V. (2006) Accelerated degradation model and its applications to statistical analysis of the role played by dementia and sex in the loss of functional autonomy in elderly patients. *Advances in Gerontology*,19, 44–54.

[NGO 06] Nikulin M, Gerville-Réache L, Orazio S (2006) About one parametric degradation model used in reliability, survival analysis and quality of life. In: *Advances in statistical methods for health sciences*, (Eds: Balakrishnan N, Aujet J.-L, Mesbah M, Molenberghs), Birkhauser, 131–142.

[NGC 07] Nikulin M, Gerville-Réache L, Couallier V (2007) *Statistique des essais accélérés*, Hermes: London.

[RJH 01] Roberts J., Jones R., Harwwod C., Mitchell S., Rothberg S. (2001) *Human Perceptions of Sports Equipment Under Playing Conditions.* J. Sports Sciences, 19, 485–497.

[SHH 85] Shand P.J., Hawrysh Z.J., Hardin R.T., Jeremiah L.E. (1985). *Descriptive sensory analysis of beef steaks by category scaling, line scaling and magnitude estimation.* Journal of Food Science, 50, 495–500.

[WZ97] Wulfson M, Tsiatis A (1997) A joint model for survival and longitudinal data measured with error. *Biometrics*, 53, 330–339.

[ZAD 65] Zadeh L.A (1965). *Fuzzy Sets.* Information Control, 8, 338–353.

[ZUE 97] Zuev Y.U. (1997). *On the estimation of efficiency of voting procedures.* Theory Probab. Appl. 42 (1) 73–81.

Appendix A

European Seminar: Some Figures

A.1. Former international speakers invited to the European Seminar

B. Abdous (Canada), V.N. Anisimov (Russia), V. Bagdonavičius (Lituania), N. Balakrishnan (Canada), F. Bartolucci (Italy), C. Ceci (Italy), D.R. Cox (United Kingdom), B. Droge (Germany), M. Finkelstein (South Africa), I. Gertsbakh (Israel), S. Gulati (USA), W. Härdle (Germany), H. Hojtink (Holland), J. Janssen (Belgium), U. Jensen (Germany), W. Kahle (Germany), V.S. Korolyuk (Ukrain), S. Kreiner (Denmark), S. Lagakos (USA), H. Läuter (Germany), C. Lefevre (Belgium), A. Lemann (Germany), H. Liero (Germany), G. Martynov (Russia), V. Nair (USA), K. Sijtsma (Holland), N.D. Singpurwalla (USA), V. Solev (Russia), V. Spokoiny (Germany), W. Stute (Germany), M.-L. Ting Lee (USA), F. Vonta (Cyprus), N. Wermuth (Germany), Z. Ying (USA), M. Zelen (USA).

A.2. Former meetings supported by the European Seminar

– International conference on “Goodness-of-fit Tests and Validity Models”, GOF’ 2000, 2000, Paris, France,

– International conference on “Mathematical Methods in Reliability”, MMR2000, 2000, Bordeaux, France,

– International workshop on “Semiparametric Models and Its Applications”, Mont Saint Michel, 2003, France,

– International conference on “Advances in Statistical Inferential Methods”, 2003, Almaty, Kazakhstan.

– Franco-Russian Conference “Modèles de Longévité, de Vieillissement et de Dégradation en Fiabilité, Santé Publique, Médecine et Biologie”, 2004, Saint Petersburg, Russia.

– International meeting on “Statistical Modelling in Industry and Life Sciences”, 2005, Potsdam, Germany.

– International conference “Degradation, Damage, Fatigue and Accelerated Life Models in Reliability Testing”, ALT’ 2006, Angers, France.

– International conference on “Statistical Models in Biomedical and Technical Systems”, Biostat 2006, Limassol, Cyprus.

A.3. Books edited by the organizers of the European Seminar

Many books written or edited by authors or editors who were the organizers of this seminar are related to the subjects of the seminar and are mainly published by Birkhäuser, Boston, Springer, Chapman & Hall/CRC, and Hermes/Lavoisier, Paris/London.

Editor: N. Limnios and M. Nikulin

1) “Recent Advances in Reliability Theory: Methodology, Practice and Inference”, Boston: Birkhäuser, 2000, 514 p.

2) Abstracts of “The Second International Conference on Mathematical Methods in Reliability”, University of Bordeaux 2 Victor Segalen, 4-7 July, 2000, Bordeaux, France, v.1, p. 1-536.

3) Abstracts of “The Second International Conference on Mathematical Methods in Reliability”, University of Bordeaux 2 Victor Segalen, 4-7 July, 2000, Bordeaux, France, v.2, p.537-1070.

Editors: C. Huber-Carol, N. Balakrishnan, M. Nikulin and M. Mesbah “Goodness-of-fit Tests and Model Validity”, Birkhäuser: Boston, 2002.

Editors: M. Nikulin, N. Balakrishnan, N. Limnios and M. Mesbah “Parametric and Semiparametric Models with Applications to Reliability, Survival Analysis and Quality of Life”, Birkhäuser: Boston, 2004.

Editors: V. Antonov, C. Huber, M. Nikulin, V. Polischook. “Longevity, Aging and Degradation Models”, v.1, Saint Petersburg, 2004.

Editors: V. Antonov, C. Huber, M. Nikulin, V. Polischook. “Longevity, Aging and Degradation Models”, v.2, Saint Petersburg, 2004

Editors: M. Nikulin, D. Commenges, C. Huber “Probability, Statistics and Modelling in Public Health”, Springer: New York, 2006.

Editors: F. Vonta, M. Nikulin, N. Limnios, C. Huber “Statistical Models and Methods for Biomedical and Technical Systems”, Birkhäuser, 2007.

A.4. Institutions supporting the European Seminar (names of colleagues)

– René Descartes University, Paris 5 (C. Huber)
– University of Technology of Compiègne (N. Limnios)
– Victor Segalen University, Bordeaux 2 (M. Nikulin)
– Potsdam University, Germany (H. Lauter)
– Otto-von-Guericke University, Germany (W. Kahle)
– McMaster University, Hamilton, Canada (N. Balakrishnan)
– Steklov Mathematical Institute, Saint Petersburg (V. Solev)
– Vilnius State University, Lithuania (V. Bagdonavicius)
– University of Rome la Sapienza (F. Spitziccino)
– French Academy of Sciences, Paris (P. Deheuvels)
– Pierre and Marie Curie University, Paris 6 (M. Mesbah)
– Polytechnical University, Saint Petersburg (V. Antonov)
– KIMEP, Almaty (V. Voinov)
– Ohio State University, Columbus (M.L.T. Lee)
– Université Libre de Bruxelles (C. Lefèvre)
– University of Padova (G. Celant)
– University of Cyprus (F. Vonta)
– French Atomic Energy Commission, Saclay (M. Eid)
– University of Rouen (V. Barbu)
– University of Caen, IUT de Lisieux (N. Heute)
– StatXpert, Bordeaux (L. Denis)

More information about the European Seminar can be found on the webpage below:

`http://www.lmac.utc.fr/~nlimnios/SEMINAIRE/`

This page intentionally left blank

Appendix B

Contributors

Belkacem ABDOUS
Laval University
Département de médecine sociale et préventive
Pavillon de l'Est, local 1138A
Ste-Foy, Quebec, G1K 7P4, Canada
E-mail: belkacem.abdous@msp.ulaval.ca

Roza ALLOYAROVA
Halyk Savings Bank of Kazakhstan
Almaty, Kazakhstan
E-mail: RozaAl@halykbank.kz

Héctor DE ARAZOZA
Facultad de Matemática y Computación
Universidad de la Habana
San Lázaro y L Vedado
CP 10400 C.Habana, Cuba
E-mail: arazoza@matcom.uh.cu

Silvia BACCI
Department of Statistics "G. Parenti"
University of Florence
Viale Morgagni 59,
50134 Florence, Italy
E-mail: s.bacci@ds.unifi.it

Vilijandas BAGDONAVIČIUS
Dep. of Mathematical statistics
Faculty of Mathematics and Informatics
Vilnius University
Naugarduko 24
Vilnius, Lithuania
E-mail: `Vilijandas.bagdonavicius@mif.vu.lt`

N. BALAKRISHNAN
Department of Mathematics and Statistics
McMaster University
Hamilton, Ontario, Canada L8S 4K1
E-mail: `bala@mcmaster.ca`

Francesco BARTOLUCCI
Dipartimento di Economia, Finanza e Statistica
University of Perugia
Via A. Pascoli, 20
06123 Perugia, Italy
E-mail: `bart@stat.unipg.it`

Michel BRONIATOWSKI
Laboratory of Theoretical and Applied Statistics (LSTA)
University of Paris 6 – Pierre and Marie Curie
175 rue du Chevaleret
75013 Paris, France
E-mail: `mbr@ccr.jussieu.fr`

Élodie BRUNEL
MAP 5, UMR 8145 CNRS
Paris Descartes University
45, rue des Saints-Pères
75270 Paris cedex 06, France
E-mail: `elodie.brunel@univ-paris5.fr`

Julien CHIQUET
Laboratoire “Statistique et Génome”
UMR CNRS 8071, INRA 1152
La genopole Tour Évry 2
523 place des Terrasses
91000 Evry, France
E-mail: `julien.chiquet@genopole.cnrs.fr`

Judith CHWALOW
INSERM
National Federation of the Blind
Baltimore, USA
E-mail: `JChwalow@nfb.org`

Jacqueline CLAVEL
INSERM U754, IFR69
University of Paris 11
16 av. Paul Vaillant Couturier
94807 Villejuif, France
E-mail: `clavel@vjf.inserm.fr`

Luc CLERJAUD
UFR, “Sciences et Modélisation”
University of Bordeaux 2 – Victor Segalen
33076, Bordeaux, France
E-mail: `Lucclerjaud@aol.com`

Vincent COLICHE
Service de Diabétologie
Hôpital de Montgeron, France
E-mail: `vincent.coliche@uni-medecine.fr`

Fabienne COMTE
MAP 5, UMR 8145 CNRS
Paris Descartes University
45, rue des Saints-Pères
75270 Paris cedex 06, France
E-mail: `fabienne.comte@univ-paris5.fr`

Jean-Pierre DAURÈS
IURC, Laboratoire de Biostatistique
641 Av. du Doyen Gaston Giraud
34093 Montpellier, France
E-mail: `daures@iurc.montp.inserm.fr`

Alexandre DEPIRE
Laboratory of Theoretical and Applied Statistics (LSTA)
University of Paris 6 – Pierre and Marie Curie
175 rue du Chevaleret
75013 Paris, France
E-mail: `depire@gmail.com`

Jean-François DUPUY
Laboratoire de Statistique et Probabilités
Institut de Mathématiques de Toulouse, UMR 5219
Paul Sabatier University
31062 Toulouse cedex 9, France
E-mail: dupuy@math.ups-tlse.fr

Léa FORTUNATO
INSERM U754, IFR69
University of Paris 11
16 av. Paul Vaillant Couturier
94807 Villejuif, France
E-mail: fortunato@vjf.inserm.fr

Léo GERVILLE-RÉACHE
University of Bordeaux 2 – Victor Segalen
Institut de Mathématiques de Bordeaux
UFR des Sciences du Sport
12, avenue Camille Jullian
33607 Pessac cedex, France
E-mail: gerville@u-bordeaux2.fr

Claudine GRAS-AYGON
Registre des Tumeurs de l'Hérault
Bâtiment Recherche
Rue des Apothicaires
34091 Montpellier, France
E-mail: registre-tumeur@wanadoo.fr

Chantal GUIHENNEUC-JOUYAUX
INSERM U754, IFR69
University of Paris 11
16 av. Paul Vaillant Couturier
94807 Villejuif, France

&

CNRS UMR 8145 UFR Biomédicale
University of Paris 5
45 rue des saints pères
75006 Paris, France
E-mail: chantal.guihenneuc@univ-paris5.fr

Jean-Benoit HARDOUIN
Team of Biostatistics, Clinical Research and Subjective Measures in Health Sciences
Department of Biomathematics and Biostatistics
University of Nantes
1, rue Gaston Veil
BP 53508 – 44035 Nantes cedex 1, France
E-mail: jean-benoit.hardouin@univ-nantes.fr

Denis HÉMON
INSERM U754, IFR69
University of Paris 11
16 av. Paul Vaillant Couturier
94807 Villejuif, France
E-mail: hemon@vjf.inserm.fr

Y.H. HSIEH
Department of Applied Mathematics
National Chung-Hsing University
Taichung, Taiwan
E-mail: hsieh@amath.nchu.edu.tw

Yen-Lung HUANG
Department of Mathematics
Tamkang University
Tamsui 251, Taiwan
E-mail: carl0918@yahoo.com.tw

Catherine HUBER
MAP 5, UMR 8145 CNRS
Paris Descartes University
45, rue des Saints-Pères
75270 Paris cedex 06, France
E-mail: catherine.huber@univ-paris5.fr

Jose JOANES
Department of Epidemiology
Ministry of Public Health
La Habana, Cuba
E-mail: ssida@infomed.sld.cu

Dominique LAURIER
Institut de Radioprotection et de Sûreté Nucléaire
IRSN/DRPH/SRBE/LEPID
BP17 – F-92262 Fontenay-aux-Roses cedex, France
E-mail: dominique.laurier@irsn.fr

Henning LÄUTER
Institute of Mathematics
University of Potsdam
Am Neuen Palais 10
D-14469 Potsdam, Germany
E-mail: laeuter@uni-potsdam.de

Eve LECONTE
GREMAQ
Manufacture des Tabacs 21, allée de Brienne
31000 Toulouse, France
E-mail: leconte@cict.fr

James LEDOUX
Institut National des Sciences Appliquées de Rennes
20 avenue des buttes de Coesmes
CS 14315
35043 Rennes cedex, France
E-mail: james.ledoux@insa-rennes.fr

Nikolaos LIMNIOS
Université de Technologie de Compiègne
Laboratoire de Mathématiques Appliquées
Centre de Recherche de Royallieu
BP 20529 – 60205 Compiègne cedex, France
E-mail: nikolaos.limnios@utc.fr

Rachid LOUNES
Paris Descartes University
MAP5 UMR CNRS 8145
45 rue des Saints Pères
75270 Paris cedex 06, France
E-mail: lounes@math-info.univ-paris5.fr

Chien-Tai Lin
Department of Mathematics
Tamkang University
Tamsui 251, Taiwan
E-mail: chien@math.tku.edu.tw

Monia Lupparelli
Department of Statistics
University of Milan-Bicocca
Milan, Italy
E-mail: mlupparelli@stat.unipg.it

Inga Masiulaityte
Department of Mathematical Statistics
Faculty of Mathematics and Informatics
Vilnius University
Naugarduko 24, Vilnius, Lithuania
E-mail: inmagik@yahoo.com

Ève Mathieu-Dupas
SysDiag, Unité CNRS-BIO-RAD
1682 rue de la Valsière
34790 Grabels, France
E-mail: eve.dupas@sysdiag.cnrs.fr

Keith Meadows
RD Tower Hamlets PCT, North East London
Consortium for Research and Development (NELCRAD)
Mile End Hospital
London, UK
E-mail: keith.meadows@thpct.nhs.uk

Mounir Mesbah
Laboratory of Theoretical and Applied Statistics (LSTA)
University of Paris 6 – Pierre and Marie Curie
175 rue du Chevaleret
75013 Paris, France
E-mail: mesbah@ccr.jussieu.fr

Étienne Mollet
Service de Diabétologie
Hôpital de Montgeron, France
E-mail: molleteml@aol.com

Mikhail NIKULIN
University of Bordeaux 2 – Victor Segalen
Institut de Mathématiques de Bordeaux
UFR Sciences et Modélisation
146, rue Léo Saignat
33076 Bordeaux cedex, France
E-mail: nikou@sm.u-bordeaux2.fr

Sébastien ORAZIO
University of Bordeaux 2 – Victor Segalen
Institut de Mathématiques de Bordeaux
UFR des Sciences du Sport
12, avenue Camille Jullian
33607 Pessac cedex, France
E-mail: sorazio2003@yahoo.fr

Nicolas PARIS
University of Bordeaux 2 – Victor Segalen
Institut de Mathématiques de Bordeaux
UFR des Sciences du Sport
12, avenue Camille Jullian
33607 Pessac cedex, France
E-mail: nicolas-paris@wanadoo.fr

Fulvia PENNONI
Department of Statistics
University of Milan-Bicocca
Via Bicocca degli Arcimboldi 8
20126 Milan, Italy
E-mail: fulvia.pennoni@unimib.it

Odile PONS
INRA Mathématiques
78352 Jouy-en-Josas cedex, France
E-mail: Odile.Pons@jouy.inra.fr

Natalie PYA
Operations Management and Information Systems Department
Kazakhstan Institute of Management, Economics and Strategic Research
Almaty, Kazakhstan
E-mail: pya@kimep.kz

Ya'acov RITOV
Hebrew University of Jerusalem
Mount Scopus Campus
Department of Statistics
Jerusalem, Israel
E-mail: `yaacov.ritov@gmail.com`

Virginie ROSA
University of Bordeaux 2 – Victor Segalen
UFR des Sciences du Sport
12, avenue Camille Jullian
33607 Pessac cedex, France
E-mail: `rosavirginie@hotmail.com`

Margot TIRMARCHE
Institut de Radioprotection et de Sûreté Nucléaire
IRSN/DRPH/SRBE/LEPID
BP17
F92262 Fontenay-aux-Roses cedex, France
E-mail: `margot.tirmarche@irsn.fr`

Vassilly VOINOV
Operations Management and Information Systems Department
Kazakhstan Institute of Management, Economics and Strategic Research
Almaty, Kazakhstan
E-mail: `voinovv@kimep.kz`

Filia VONTA
Department of Mathematics and Statistics
P.O. Box 20537
CY-1678, Nicosia, Cyprus
E-mail: `vonta@ucy.ac.cy`

This page intentionally left blank

Index