

John O'Quigley

Survival Analysis

Proportional and Non-Proportional
Hazards Regression

Survival Analysis

John O'Quigley

Survival Analysis

Proportional and Non-Proportional Hazards
Regression



Springer

John O'Quigley
Department of Statistical Science
University College London
London WC1E 6BT, UK

ISBN 978-3-030-33438-3

ISBN 978-3-030-33439-0 (eBook)

<https://doi.org/10.1007/978-3-030-33439-0>

Mathematics Subject Classification: 62-07, 62B10, 62F03, 62F05, 62F07, 62F40, 62G30, 62G32, 62G10, 62H17, 62J02, 62J05, 62L05, 62L10, 62L12, 62N01, 62N02, 62N05, 62N86, 62P05, 62P10, 91B02, 92B15, 92C60, 92C50, 92D30, 97K80

© Springer Nature Switzerland AG 2021

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

In loving memory of my father, Arthur O'Quigley, whose guiding strategy when confronting the most challenging of problems – *use a sense of proportion* – would have baffled many a latter-day epidemiologist.

Preface

In common with most scientific texts this one is the result of several iterations. The seed for the iterations was the 2008 text *Proportional Hazards Regression* and the first step was in the direction of a second edition to that work. The next several iterations from there took us in a number of different directions before slowly converging to a text that appears to be more in harmony with modern approaches to data analysis, graphical techniques in particular. These are given greater emphasis in this book and, following recent work on the regression effect process, it seems clear that this process, first presented in O'Quigley (2003) ought be given a more prominent place. Just about everything we need, from complex estimation to simple tests, can be readily obtained from the regression effect process. Formal analysis can be backed up by intuitive graphical evaluation based directly on this process.

Proportional hazards models and their extensions (models with time-dependent covariates, models with time-dependent regression coefficients, stratified models, models with random coefficients, and any mixture of these) can be used to characterize just about any applied problem to which the techniques of survival analysis are appropriate. This simple observation enables us to find an elegant statistical expression for all plausible practical situations arising in the analysis of survival data. We have a single unifying framework. In consequence, a solid understanding of the framework itself offers the investigator the ability to tackle the thorniest of questions which may arise when dealing with survival data.

Our goal is not to present or review the very substantial amount of research that has been carried out on proportional hazards and related models. Rather, the goal is to consider the many questions which are of interest in a regression analysis of survival data, questions relating to prediction, goodness of fit, model construction, inference and interpretation in the presence of misspecified models. To address these questions the standpoint taken is that of the proportional hazards and the non-proportional hazards models.

This standpoint is essentially theoretical in that the aim is to put all of the inferential questions on a firm conceptual footing. However, unlike the commonly preferred approach among academic statisticians, based almost entirely on the construction of stochastic integrals and the associated conditions for the valid

application of the martingale central limit theorem for multivariate counting processes, we work with more classical and better known central limit theorems. In particular we appeal to theorems dealing with sums of independent but not necessarily identically distributed univariate random variables and of special interest is the classical functional central limit theorem establishing the Brownian motion limit for standardized univariate sums.

So, we avoid making use of Rebolledo's central limit theorem for multivariate counting processes, including the several related works required in order to validate this theorem, the Lenglart-Rebolledo inequality for example. Delicate measure-theoretic arguments, such as uniform integrability, borrowed from mathematical analysis can then be wholly avoided. In our view, these concepts have strayed from their natural habitat—modern probability theory—to find themselves in a less familiar environment—that of applied probability and applied statistics—where they are not well understood nor fully appreciated. While it could be advanced that this abstract theory affords greater generality, the author has yet to see an applied problem in which the arguably less general but more standard and well known central limit theorems for univariate sums fail to provide an equally adequate solution. None of the results presented here lean on Rebolledo's central limit theorem.

A central goal of this text is to shift the emphasis away from such abstract theory and to bring back the focus on those areas where, traditionally, we have done well—careful parsimonious modelling of medical, biological, physical and social phenomena. Our aim is to put the main spotlight on robust model building using analytical and graphical techniques. Powerful tests, including uniformly most powerful tests, can be obtained, or at least approximated, under given conditions. We discuss the theory at length but always looked at through the magnifying glass of real applied problems.

I would like to express my gratitude to several colleagues for their input over many years. The collaborative work and countless discussions I have had with Philippe Flandre, Alexia Iasonos, Janez Stare and Ronghui Xu in particular have had a major impact on this text. I have worked at several institutions, all of them having provided a steady support to my work. Specifically, I would like to acknowledge: the Department of Mathematics of the University of California at San Diego, U.S.A., the Fred Hutchinson Cancer Research Center, Seattle, U.S.A., the Department of Mathematics at the University of Leeds, U.K., the Department of Mathematics at Lancaster University, U.K., the Department of Biostatistics at the University of Washington, U.S.A., the Department of Biostatistics at Harvard University, Boston, U.S.A., the Division of Biostatistics at the University of Virginia School of Medicine, the Laboratory for Probability, Statistics and Modelling, University of Paris, Sorbonne and last (but not least), the Department of Statistical Science, University College London. I am very grateful to them all for this support.

A final word. At the heart of our endeavor is the goal of prediction, the idea of looking forward and making statements about the future. Perhaps somewhat paradoxically, this only makes real sense when looking backwards, that is when all the observations are in, allowing us to make some general summary statements.

Public health policy might use such statements to make planning decisions, clinical trials specialists may use them to increase the power of a randomized study. But, to use them to make individual predictions is unlikely to be of any help and could potentially be of harm. Even for the terminally ill patient it is not rare for those with the very worst prognosis to confound the predictions while others, looking much better in theory, can fare less well than anticipated. Such imprecision can only be greatly magnified when dealing with otherwise healthy subjects, the example of the BRCA1 and BRCA2, so-called, susceptibility genes being a striking one. We consider this question in the chapter dealing with epidemiology. While past history is indispensable to improving our understanding of given phenomena and how they can describe group behaviour, on an individual level, it can never be taken to be a reliable predictor of what lies ahead. Assigning probabilities to individuals, or to single events—however accurate the model—is not helpful.

When anyone asks me how I can best describe my experience in nearly 40 years at sea, I merely say ... uneventful. Captain Edward John Smith, RMS Titanic April, 1912.

London, UK
April 2021

John O'Quigley

Contents

1	Introduction	1
1.1	Chapter summary	1
1.2	Context and motivation	1
1.3	Some examples	2
1.4	Main objectives	6
1.5	Neglected and underdeveloped topics	7
1.6	Model-based prediction	12
1.7	Data sets	16
1.8	Use as a graduate text	16
1.9	Classwork and homework	17
2	Survival analysis methodology	19
2.1	Chapter summary	19
2.2	Context and motivation	19
2.3	Basic tools	20
2.4	Some potential models	28
2.5	Censoring	34
2.6	Competing risks	38
2.7	Classwork and homework	45
3	Survival without covariates	49
3.1	Chapter summary	49
3.2	Context and motivation	49
3.3	Parametric models for survival functions	50
3.4	Empirical estimate (no censoring)	56
3.5	Kaplan-Meier (empirical estimate with censoring)	58
3.6	Nelson-Aalen estimate of survival	68
3.7	Model verification using empirical estimate	69
3.8	Classwork and homework	70
3.9	Outline of proofs	72

4 Proportional hazards models	75
4.1 Chapter summary	75
4.2 Context and motivation	75
4.3 General or non-proportional hazards model.	77
4.4 Proportional hazards model	78
4.5 Cox regression model	78
4.6 Modeling multivariate problems	87
4.7 Classwork and homework	93
5 Proportional hazards models in epidemiology	97
5.1 Chapter summary	97
5.2 Context and motivation	97
5.3 Odds ratio, relative risk, and 2×2 tables	98
5.4 Logistic regression and proportional hazards	102
5.5 Survival in specific groups	107
5.6 Genetic epidemiology	110
5.7 Classwork and homework	117
6 Non-proportional hazards models	119
6.1 Chapter summary	119
6.2 Context and motivation	119
6.3 Partially proportional hazards models	120
6.4 Partitioning of the time axis	129
6.5 Time-dependent covariates	131
6.6 Linear and alternative model formulations	136
6.7 Classwork and homework	139
7 Model-based estimating equations	141
7.1 Chapter summary	141
7.2 Context and motivation	141
7.3 Likelihood solution for parametric models	143
7.4 Semi-parametric estimating equations	151
7.5 Estimating equations using moments	153
7.6 Incorrectly specified models	164
7.7 Estimating equations in small samples	176
7.8 Classwork and homework	186
7.9 Outline of proofs	189
8 Survival given covariate information	191
8.1 Chapter summary	191
8.2 Context and motivation	191
8.3 Probability that T_i is greater than T_j	193
8.4 Conditional survival given $Z \in H$	195

8.5	Other relative risk forms	201
8.6	Informative censoring	204
8.7	Classwork and homework	209
8.8	Outline of proofs.	211
9	Regression effect process	215
9.1	Chapter summary	215
9.2	Context and motivation	215
9.3	Elements of the regression effect process	217
9.4	Univariate regression effect process	224
9.5	Regression effect processes for several covariates	231
9.6	Iterated logarithm for effective sample size	235
9.7	Classwork and homework	237
9.8	Outline of proofs.	238
10	Model construction guided by regression effect process	261
10.1	Chapter summary	261
10.2	Context and motivation	261
10.3	Classical graphical methods	263
10.4	Confidence bands for regression effect process	267
10.5	Structured tests for time dependency	269
10.6	Predictive ability of a regression model	270
10.7	The R^2 estimate of Ω^2	275
10.8	Using R^2 and fit to build models	278
10.9	Some simulated situations	280
10.10	Illustrations from clinical studies	291
10.11	Classwork and homework	297
10.12	Outline of proofs.	298
11	Hypothesis tests based on regression effect process	301
11.1	Chapter summary	301
11.2	Context and motivation	301
11.3	Some commonly employed tests	302
11.4	Tests based on the regression effect process	308
11.5	Choosing the best test statistic	327
11.6	Relative efficiency of competing tests	334
11.7	Supremum tests over cutpoints	335
11.8	Some simulated comparisons	337
11.9	Illustrations	340
11.10	Some further thoughts	342
11.11	Classwork and homework	344
11.12	Outline of proofs.	346

Appendix A. Probability	351
A.1 Essential tools for survival problems	351
A.2 Integration and measure	351
A.3 Random variables and probability measure	354
A.4 Convergence for random variables	355
A.5 Topology and distance measures	356
A.6 Distributions and densities	358
A.7 Multivariate and copula models	362
A.8 Expectation	364
A.9 Order statistics and their expectations	366
A.10 Approximations	371
Appendix B. Stochastic processes	377
B.1 Broad overview	377
B.2 Brownian motion	378
B.3 Counting processes and martingales	385
B.4 Inference for martingales and stochastic integrals	392
Appendix C. Limit theorems	401
C.1 Empirical processes and central limit theorems	401
C.2 Limit theorems for sums of random variables	402
C.3 Functional central limit theorem	405
C.4 Brownian motion as limit process	408
C.5 Empirical distribution function	410
Appendix D. Inferential tools	413
D.1 Theory of estimating equations	413
D.2 Efficiency in estimation and in tests	420
D.3 Inference using resampling techniques	422
D.4 Conditional, marginal, and partial likelihood	425
Appendix E. Simulating data under the non-proportional hazards model	433
E.1 Method 1—Change-point models	433
E.2 Method 2—Non-proportional hazards models	437
Further exercises and proofs	439
Bibliography	449
Index	471

Summary of main notation

$\mathcal{W}, \mathcal{W}(\cup)$	Brownian motion on the interval $[0, 1]$
$B, B(t), W^0$	Brownian bridge on the interval $[0, 1]$
$\mathbb{I}(A), 1_A$	Indicator function for the set A
$\mathbb{P}(A), \Pr(A), P(A)$	Probability measure for the set A
$\xrightarrow{\mathcal{L}}$	Convergence in distribution (law)
$\xrightarrow{\mathbb{P}}$	Convergence in probability
$\xrightarrow{L^1}$	Convergence in mean
$a.s.$	Almost surely, <i>w.p.1</i>
$\mathcal{M}_{p \times p}(\mathbb{R})$	Matrices of dimension $p \times p$ with real coefficients
$\ a\ $	Supremum norm of the vector $a \in \mathbb{R}^p$
$\ A\ $	Supremum norm of the matrix $A \in \mathcal{M}_{p \times p}(\mathbb{R})$
\hat{F}	Kaplan-Meier estimator of F
n	Sample size
p	Dimension of covariable space
T_i	Time to event for individual i
X_i	Observed time under study for individual i
δ_i	Censoring indicator variable for individual i
Z_i	Vector of p covariables for individual i
$Z_i^{(l)}$	l ème covariable of individual i , $l = 1, \dots, p$
k_n	Number of effective (usable) failures
β	Regresson coefficient, log relative-risk
$\beta(t)$	Time-dependent regression coefficient
$\mathcal{Z}(t)$	Value of covariate for individual failing at time t
$\pi_i(\beta(t), t)$	Given t , probability that individual i is selected for failure under NPH model with parameter $\beta(t)$

$\mathcal{E}_{\beta(t)}(Z t)$	Expectation with respect to $\pi_i(\beta(t), t)$
$\mathcal{V}_{\beta(t)}(Z t)$	NPH model with coefficient $\beta(t)$ Variance with respect to $\pi_i(\beta(t), t)$
$U_n^*(\beta(t), t)$	NPH model with coefficient $\beta(t)$ Regression effect process evaluated at time t



Chapter 1

Introduction

1.1 Chapter summary

The chapters of this book correspond to a collection of broad themes. Taken together, these themes make up a sizeable component of what can be considered to be the topic of modern survival analysis. A feature that we maintain throughout the text is to summarize the salient features of each chapter in a brief section headed “Chapter summary”. Section 1.1 is the first such summary. Following these summaries comes a section headed “Context and motivation”, the purpose of which is to motivate our interest as well as to recall any key notions we may need as an aid in pursuing the topic of the chapter. In this introduction we consider a number of real-life examples in which the methods of survival analysis, proportional and non-proportional hazards models, in particular, have been used to advantage. Following this we highlight in a section entitled “Objectives” the way in which the described methodology can address many of the important questions that arise in the very diverse set of motivating examples. A large number of examples are worked throughout this book and we indicate where to find the relevant data for those readers interested in confirming the results for themselves before working on their own particular projects. For teaching purposes a section suggesting problems for classwork and homework is given. Formal proofs and proof outlines are removed from the main text into their own section at the end of each chapter.

1.2 Context and motivation

Applications of survival analysis methodology appear to be ever-widening. Recent areas of interest include unemployment data, lengths of time on and off welfare, for example, and some unusual applications such as interviewer bias

in sociological surveys. The more classic motivating examples behind the bulk of theoretical advances made in the area have come from reliability problems in engineering and, especially, clinical studies in chronic diseases such as cancer and AIDS. Many of the examples given in this text come from clinical research. Parallels for these examples from other disciplines are usually readily transposable. Proportional hazards regression is now widely appreciated as a very powerful technique for the statistical analysis of broad classes of problems. The ability to address issues relating to complex time-dependent effects and non-proportional hazards structures broadens this class yet further.

1.3 Some examples

Novel uses as well as innovative applications of the basic approach continue to grow. The following examples represent a small selection of problems for which the proportional hazards model has been shown to be useful as an analytic tool. Adding non-proportional hazards models to our toolkit provides us with an array of statistical techniques capable of tackling almost any problem in survival analysis.

CLINICAL TRIALS IN CHRONIC DISEASE

Patients, identified as having some chronic illness under study, are recruited to a clinical trial. A new, proposed, treatment, if effective, should prolong survival. The idea is that instead of aiming for a “cure” we aim to improve survival. In cancer research it has been common practice to equate cure with five years survival without relapse. Such a quantity may be difficult to estimate if some patients leave the study or die from unrelated causes during the five-year period. Survival methods address such difficulties. Proportional and non-proportional hazards models enable us to investigate the dependence of survival on controlled and uncontrolled sources of variation. Now, suppose that therapy A and therapy B have the same five-year survival rates but different survival curves. As an illustration consider curative treatments based on surgery having very high initial risks but, conditional upon having survived some length of time, a reduced long-term risk. Such differences in overall survival are important and would be considered relevant when weighing the relative advantages and disadvantages of any proposed new treatment. In actual studies patients may be lost to follow-up at any time. Proportional hazards and non-proportional hazards modeling can allow us to correctly use the partial information provided by such patients. These models provide the means to isolate as much of the treatment effect as we can from the several interrelated confounding effects.

APPLICATIONS IN EPIDEMIOLOGY STUDIES

The time-to-event variable in these studies is almost always age. A large number of individuals are studied across a relatively wide distribution of ages. Incidence of some target disease is measured. This incidence will typically vary with age and, in the majority of cases, the single most important source of variability in the observed incidence rates comes from age itself. Cancer is a particularly striking example. Here the variable age plays the role of the time variable in a survival model. Prolonging survival amounts to living to a longer age. How does incidence relate to risk variables, obvious examples being smoking, asbestos exposure or other industrial exposures. In practice the amount of time needed for such prospective studies may be too great and we may prefer to use a case-control study where, rather than compare the age distributions for different exposure categories, we compare the exposure distributions at different ages.

In doing this our purpose is to address some key feature of the conditional distribution of age given covariate information via a study of the conditional distribution of the covariates given age. The question arises as to the validity of this and to the efficiency of such a study design. Additional complications arise when the exposure categories change for some individuals. At any given age a particular individual can have a unique and complex exposure history. It would not be possible to study the relationship between any unique exposure history and disease outcome without assuming some structure, i.e., the potential ways in which the different components comprising the exposure history relate to one another and, more importantly, to outcome. The way to achieve this is via a statistical model, specifically, in the context of this work, the proportional hazards and non-proportional hazards regression model.

GENETIC SUSCEPTIBILITY TO DISEASE INCIDENCE

Recent years have seen a growing interest in identifying genetic mutations that may be linked to the occurrence of some particular disease. The use of registry data, such as the SEER data, in conjunction with a survival model allows us to quantify the risk associated with different classifications, in particular a classification based on genetic profiles and so-called susceptibility mutations. Family studies can be of great value in bringing to light potential risk factors although, unless the risk factors are particularly strong, it will be difficult to detect them without long follow-up times and careful modeling strategies. These can enable some degree of control over factors that may mask effects. Risk over several years including lifetime risk can be calculated using survival methods. That said, great care is needed in interpretation and we consider this in the chapter on epidemiology.

COMPLEX TIME-DEPENDENT EFFECTS IN CLINICAL STUDIES

In clinical research, not only can there be problems with patients going off study or being removed from the study due to causes believed to be unrelated to the main outcome under study, it is also not unusual for subjects to enter the study at different times. In chronic diseases such as AIDS, patients may leave the study because they are too sick to continue participation. Such loss to follow-up maybe of a different nature to that described above and may, of itself, provide information on the risk of being incident for the outcome. In technical terms we would describe the situation as being one where the censoring mechanism is not independent of the failure mechanism. In order to avoid potentially misleading biases it would be necessary to appeal to workable models to describe this dependence.

The complexity of potential applications can easily grow beyond the reach of available methods. These continue to promote and drive the interest for further methodological developments. One example, again from AIDS, where the methods presented here can provide insight concerns the situation where the time frame is lengthy. Patients may have very involved time-dependent exposure histories to the different treatments, the class of available treatments themselves having evolved through time as well as within any given patient. It is quite possible that no two patients would have exactly the same treatment history. Models enable us to put some kind of structure on this so that inferences can be made regarding treatment approaches.

In prognostic studies, many variables are measured on individuals and we would like to know how individually and collectively these variables influence survival (prognosis). It may be possible to use such information to stratify future clinical trials, the goal being to increase precision, or, possibly, to use the information in a context of clinical patient management. Further complexity arises with correlated data. For recurrent conditions such as asthma the same patient may be studied for more than one principal event. Cooperative clinical trial groups may use many centers to recruit patients, the center affects themselves not being of any intrinsic interest, but nonetheless contributing a correlation factor for within-center observations. Similar considerations arise in familial data in which the presence of genetic characteristics may influence the survival probabilities collectively. Proportional hazards models including random component coefficients, equivalently non-proportional hazards models, can be used to efficiently analyze such data structures.

INSURANCE, ECONOMICS, AND SOCIOLOGY

An area of application where the law of large numbers comes directly into play is that of insurance. The calculus of probability together with very large object populations enable effective prediction. We may wish to know how demographic and other variables influence insurance risk. Further examples arise in car insurance; how long on average does it take to be involved in a traffic accident and, more importantly, which variables influence this time. Such information will enable the insurer to correctly price policies for different driver profiles.

In economics there are many possibilities. How long does it take for a price index to increase to a certain value, to regain a previous value or to change by some given amount? In this and other examples the technical term "failure" may in a more common everyday sense correspond to success. Depending on the context we may wish to prolong or reduce the average time it takes to observe some kind of event. Failure models have been used to study, under differing circumstances, the average time taken for a welfare recipient to come off welfare, an event in social terms that we would regard as a success.

FINANCIAL ANALYSIS

Mathematical techniques of stochastic integration have seen wide application in recent years in the study of financial products such as derivatives, futures, and other pricing schemes for certain kinds of options. An alternative and potentially more flexible approach to many of these questions is via regression modeling. The increasing volatility of many of the markets has also given rise to the use of survival modeling as a tool to identify, among the large-scale borrowers, large industries, countries, and even financial institutions themselves the ones which are most likely to fail on their repayment schedule. Modeling techniques can make such analyses more precise, identifying just which factors are most strongly predictive.

MANUFACTURING AND RELIABILITY

One of the early applications of survival methods concerns the reliability of components as well as mechanical, electrical, or electronic systems. Careful modeling can enable us to use various sources of information to predict the probability of failure before some given time. Many consumers will have first-hand experience of such modeling whereby it is possible to keep small the number of manufactured products failing before the expiration of the guarantee while, at the same time, ensuring that the probability of survival well beyond the guarantee is not so large as to ultimately impact consumer demand. It can be of interest to model more

than one endpoint, tracing out the lifetime of a product. The first event may be something minor and subsequent events of increasing degrees of seriousness, or involving different components, until the product is finally deemed of no further use. The many different paths through these states which any product may follow in the course of a lifetime can be very complex. Models can usefully shed light on this.

1.4 Main objectives

The purpose of this book is to provide the structure necessary to the building of a coherent framework in which to view proportional and non-proportional hazards regression. The essential idea is that of prediction, a feature common to all regression models, but one that sometimes slips from sight amid the wealth of methodological innovations that has characterized research in this area. Our motivation should mostly derive from the need to obtain insights into the complex data sets that can arise in the survival setting, keeping in mind the key notion that the outcome time, however measured, and its dependence on other factors is at the center of our concerns.

The predictive power of a model, proportional and non-proportional hazards models, in particular, is an area that we pay a lot of attention to. It is necessary to investigate how we can obtain predictions in the absence of information (often referred to as explanatory variables or covariate information) which relate to survival and, subsequently, obtaining predictions in the presence of such information. How best such information can be summarized brings us into the whole area of model adequacy and model performance. Measures of explained variation and explained randomness can indicate to what extent our prediction accuracy improves when we include additional covariate information into any model. They also play a central role in helping identify the form of the unknown time-dependent regression effects model that is taken to generate the observations.

In order to give the reader, new to the area, a feeling as to how the Cox model fits into the general statistical literature we provide some discussion of the original paper of Professor Cox in 1972, some of the background leading up to that seminal paper, and some of the scientific discussions that ensued. The early successes of the model in characterizing and generalizing several classes of statistics are described.

As is true of much of science—statistics is no exception—important ideas can be understood on many levels. Some researchers, with a limited training in even the most basic statistical methods, can still appreciate the guiding principles behind proportional hazards regression. Others are mainly interested in some of the deeper inferential mathematical questions raised by the estimation techniques

employed. Hopefully both kinds of readers will find something in this book to their taste. The aim of the book is to achieve a good balance, a necessary compromise, between the theoretical and the applied. This will necessarily be too theoretical for some potential readers, not theoretical enough for others. Hopefully the average gap is not too great.

The essential techniques from probability and statistics that are needed throughout the text are gathered together in several appendices. These can be used for reference. Any of these key ideas called upon throughout the book can be found more fully developed in these appendices. The text can therefore serve as a graduate text for students with a relatively limited background in either probability or statistics. Advanced measure theory is not really necessary either in terms of understanding the proportional hazards model or for gaining insight into applied problems. It is not emphasized in this book and this is something of a departure from a number of other available texts which deal with these and related topics. Proofs of results are clearly of importance, partly to be reassured as to the validity of the techniques we apply, but also in their own right and of interest to those focusing mostly on the methods. In order not to interrupt the text with proofs, we give theorems, corollaries, and lemmas, but leave the proofs to be gathered together at the end of each chapter. The reader, less interested in the more formal presentation in terms of theorems and proofs, will nonetheless, it is hoped, find the style helpful in that, by omitting the proofs at the time of development, the necessary results are organized and brought out in a sharper focus. While our motivation always comes from real practical problems, the presentation aims to dig deep enough into the mathematics so that full confidence can be obtained in using the results as well as using the machinery behind the results to obtain new ones if needed.

1.5 Neglected and underdeveloped topics

The non-proportional hazards model, including as a special case the proportional hazards model, is so general in nature that essentially all relative risk models come under its umbrella. There are one or two exceptions, such as continuous covariates, but they can be approximated to as high a degree as we wish. This means that we have a full spectrum of all relative risk models where one extremity is an absence of any restrictive assumption while the other is that of proportional hazards. The true, generally unknown, model lies somewhere between the two. If our interest goes beyond that of just relative risk to, say, the underlying hazard functions, or intensity functions, then further restrictions can be made, an example being the use of parametric models such as the Weibull model. These are described in this text but little emphasis is made and the greatest part of our focus is on relative risk functions and their evolution through time.

It may be that certain parametrizations lead to simplification, one example arising in cardiovascular studies where an additive survival model allows a more parsimonious overall description than would a mathematically equivalent, non-proportional hazards model. Nevertheless, if only conceptually, it helps to keep in mind that models such as the additive and related models, at first glance appearing to lie outside of our general formulation, can indeed be expressed as non-proportional hazards one with a regression function that is time dependent.

This thinking guides our presentation. As a result certain topics that are often given an important place in survival analysis texts are neglected here. Some of these are briefly described below.

LEFT CENSORING AND LEFT TRUNCATION

Just as right censoring at time C means that the true, unobserved, time of interest is greater than C , left censoring at time C means that the true, unobserved, time of interest is lesser than C . Examples of left censoring in clinical trials are hard to find but, in epidemiology, it is conceivable that a study would include individuals for whom the event of interest has already occurred without knowing precisely when the event occurred; simply that it would be lesser than some known time point, often their time of entry into the study. In this text we do not study left censoring and the reader interested in analyzing problems in which left censoring arises might begin by looking at some specialist epidemiological texts that deal with longitudinal observations.

Note that observations that are only recorded to occur within some time interval may, in a formal way, be analyzed by combining methods that can simultaneously accommodate both left and right censoring. Again, we do not treat such problems in this book. Another type of observation that can arise may result from a phenomenon known as left truncation. If, having encountered the event of interest, before the study begins, the probability of being included in the study is small, possibly zero, then this is known as left truncation. It is easy to see how this phenomenon could lead to substantial bias if ignored. An illustration of left truncation can be seen in studies that focus on factors that might influence the rate of spontaneous abortion following conception. This rate is likely to be underestimated since the smallest times may often correspond to situations where the event is not recorded, the woman may not even be aware that she was pregnant.

Studying the impact of genetic and biomarkers on survival can run into the problem of left truncation. The decision to focus interest on some biomarker may not have been made at the beginning of the study so that, for those subjects who have left the study due to failure or other forms of censoring, it is no longer possible to ascertain to which of the different groups they may belong. As a result they are often ignored in analyses resulting in potentially strong biases. In this text we provide survival estimates that would overcome such biases but,

otherwise, we do not study left truncation in its own right. Readers looking for more on this topic are referred to Hyde (1977), Tsai et al. (1987), Keiding et al. (1987), Keiding and Gill (1990), Jiang et al. (2005), Shen (2006), Cain et al. (2011), and Geskus (2011).

LEARNING AND CLASSIFICATION

The ability, and the need, to deal with very large sets of data has been the motor for many developments of the past two decades in a discipline that is neither statistics nor computer science—we could describe it as a hybrid or an intersection of the two. Under this heading we might include machine learning, supervised and unsupervised, deep learning, classification and regression trees, random forests, and artificial intelligence. In the field of survival analysis the work of Breiman (2017) anticipated many of these developments. The bulk of the problems of machine learning, in the context of survival analysis, are directly addressed in this text albeit using terms that are more familiar to those from the less recent schools of statistics. While prediction, and the assessment of predictive strength, is a running theme in this work, we do not make an appeal to the techniques of classification. The literature has many examples of model construction based on classification techniques, for example, Ciampi et al. (1988), Ciampi et al. (1991), LeBlanc and Crowley (1992), Schumacher et al. (2003), Koziol et al. (2003), Loh (2011), and Loh (2014) and there are very many more. Linden and Yarnold (2017) provides a specific focus on clinical practice. This is a large topic and will be the focus of a future text but is not considered here in any great depth.

KERNEL METHODS AND LOCAL SMOOTHING

Local smoothing of the baseline hazard, or of the residuals, although an interesting topic, is not given any real space here. The same goes for kernel estimates of the density function or the use of polynomial splines. The view taken here is that, for the great majority of applications, cumulative quantities such as the empirical distribution function do “enough” smoothing. However, such work is valuable and certainly of interest. An absence of coverage in this text should not be taken as implying that such techniques have no importance. The choice of how much weight to afford different topics, including the assignment of a zero weight, is often a subjective one—a matter of taste. We have little familiarity with applied problems that would benefit from the use of these smoothing techniques. One way to view smoothing is to see it as a way to inject some structure into a model with too many moving parts, as can occur when there are a large number of potential parameters to estimate. The goal then would then be to bring some additional constraints to the model, in some sense tightening the overall structure. In this work, our approach to model building takes something of a contrary standpoint. We begin mostly with a quite rigid framework for the model and,

using the techniques of fit and predictive performance, gradually weaken certain restrictions.

There is more than one way to build more complex models from initially simpler ones. In the main we tend to do this by weakening the constraints around fixed parameters, for example, stepping up from a proportional hazards model to a piecewise proportional hazards model. A different approach focuses on the undetermined randomness and by broadening the probabilistic base from which observations can be made. Frailty models come under this heading so that unobserved quantities enter our inferential procedures indirectly. These can allow some simple parameters, often no more than the variance of the frailty distribution, to be replaced by data-driven estimates, thereby taking up some of the slack in order to provide a potentially better fit for the model. We give relatively little weight to the study of frailty models, outside the context of goodness of fit. Random effects models, although of the same flavor, do not come under this heading and, while these are described in the text, for frailty models, if we are to understand the term “frailty” as meaning a random effect for an individual (without repeated measurements), as opposed to a group (random effects models), then this is just another way of expressing lack of fit. In other words a proportional hazards model with a frailty term is equivalent to a non-proportional hazards model. Tests, described in the literature, for heterogeneity of individuals in this context can be misleading. This has been described in O’Quigley and Stare (2002).

PARTIAL LIKELIHOOD

The term “*partial likelihood*” is common in survival texts as a description of a particular approach to inference. It is not a very useful concept in our view, and perhaps it is time to put the term “*partial likelihood*” to a long deserved rest. In his introductory paper on proportional hazards models, Professor Sir David Cox (Cox, 1972) made no reference at all to partial likelihood. He presented a suitable likelihood with which to make inference, given the observations, and, no doubt unintentionally, generated some confusion by referring to that likelihood as a *conditional likelihood*. Some of the discussants of Cox’s paper picked up on that and were puzzled as to the sense of a conditional likelihood since the statistic upon which we would be conditioning is not very transparent. These discussants pointed out that there is no obvious statistic whose observed value is taken as fixed before we set about making inferences. Nonetheless, there is a lot of sequential conditioning that leads to the expression, first derived by Professor Cox, and, for want of a better term, “*conditional likelihood*” was not so bad. Cox’s choice for likelihood was a fully legitimate one. It was also a very good one, having among several properties that of leaving inference unaffected by increasing transformations on the time variable. More explanation may have been needed but, instead, regrettably, we all moved off in something of a wrong

direction: not dramatically wrong but wrong enough to cloud and confuse issues that ought, otherwise, have been quite plain.

These were very early days for proportional hazards regression and, for those concerned, it was not always easy to see clearly ahead. Efforts were made either to justify the likelihood given by Cox, as arising within a particular conditional sampling framework or to appeal to different techniques from probability that, by good fortune, led us to the same estimating equations like the ones derived by Cox. All of this had two unfortunate consequences that turned out to be very significant for the later development of the field. Those initial worries about the nature, and indeed the legitimacy, of the likelihood, took us on something of an arduous path over the next few years. In the late seventies, early eighties, a solution to the problem was believed to have been found. And in the place where we might least expect to find it: within the very abstract French school of probability. It was argued that the key to the solution lay in the central limit theorem of Rebolledo (1978) for multivariate counting processes.

Now, while stochastic processes, and counting processes, in particular, are at the very heart of survival analysis, this particular branch of the French school of probability, built around Lenglart's theorem, Rebolledo's multivariate central limit theorem for stochastic integrals, Jacod's formula, Brémaud's conditions, bracket processes, and locally square integrable martingales, is, and has remained, almost inaccessible to those lacking a strong background in real analysis. This development had the unfortunate effect of pushing the topic of survival analysis beyond the reach of most biostatisticians. Not only that but the field became for very many years too focused on the validity of Rebolledo's findings under different conditions. The beauty of the Cox model and the enormous range of possibilities it opened up took second place so that questions at the heart of model building: predictive strength, model validation, goodness of fit, and related issues did not attract the attention that they deserved.

Our goal is not to use the benefit of hindsight to try to steer the course of past history—we couldn't anyway—but we do give much less emphasis to the multivariate central limit theorem of Rebolledo in favor of more classical central limit theorems, in particular the functional central limit theorem. This appears to allow a very simple approach to inference. No more than elementary calculus is required to understand the underpinnings of this approach.

The concept of partial likelihood as a general technique of inference in its own right, and for problems other than inference for the proportional hazards model has never really been thoroughly developed. One difficulty with the concept of partial likelihood, as currently defined, is that, for given problems, it would not be unique. For general situations then it may not be clear as to the best way to proceed, for instance, which of many potential partial likelihoods ought we choose to work with. For these reasons we do not study partial likelihood as a tool for inference and the concept is not given particular weight. This is a departure from several other available texts on survival analysis. In this work, we do not

view partial likelihood as anything other than a regular likelihood consequent upon a particular sampling structure. As such, it inherits all of those desirable properties that we know to hold.

JOINT SURVIVAL-COVARIATE MODELS

An earlier version of this book in development included a chapter on joint survival-covariate models. It was decided to remove this chapter for the following reasons. Firstly, the contributions of Henderson et al. (2000), Tsiatis and Davidian (2004), as well as the text by Rizopoulos (2012), cover the field very thoroughly and, having not spent enough time in this area, I have nothing really to add in the way of commentary or insight. Secondly, in order to be coherent with the presentation here that hinges upon the regression effect process in transformed time, we would need to build a parallel setup for joint models. While this would not be conceptually difficult it remains to be carried out. Finally, the motivation to do this work is not strong since real applications of joint models are quite limited. There is an extensive literature on this topic and the interested reader is referred to; Hsieh et al. (2006), R. Brown and G. Ibrahim (2003), Huang and Liu (2007), Song and Wang (2008), Yu and Sandler (2008), Crowther et al. (2013), Chen et al. (2014) and Rizopoulos and Lesaffre (2017).

An important misconception that requires clarification is that joint modeling is necessary whenever the covariate—a common example being immunological status in HIV studies—changes through time. Our main goal is to study the dependence of survival on such measures and, for this, we need to study conditional distributions; survival time given the covariate or the covariate given survival time. Proportional hazards and non-proportional hazards models can handle time-dependent covariates very well. Modeling, and learning about, the covariate process itself through time, provides us with very little additional information in the best of cases and, in most cases, it does not provide us with predictive information at all. None of this is to say that the study of joint models has little value. Not at all. It is a very interesting activity but is more focused on determining plausible models for a whole collection of relevant processes taken together. In this book our focus is overwhelmingly on prediction, improvement of model fit, and the development of the most powerful tests in given situations. For these problems conditional distributions tell us all we need to know. Modeling the marginal covariate process itself does not help much, if at all, with this endeavor.

1.6 Model-based prediction

PREDICTIVE INDICES

The main purpose of model construction is ultimately that of prediction. The amount of variation explained is viewed naturally as a quantity that reflects the

predictive power of any model. This is a direct consequence of the Chebyshev inequality. And yet, considerable confusion on the topic of explained variation is prevalent among statisticians. This is to some extent surprising since the main ideas are very simple. The origin of the confusion for many was in a paper of Draper (1984), echoed subsequently by Healy (1984), which was strongly critical of the R^2 measure of explained variation for multi-linear models. Several authors, Kvalseth (1985), Scott and Wild (1991), and Willett and Singer (1988), followed in the wake of these two papers, adding further nails, it was argued, to the coffin of explained variation.

However, Draper's paper was in error. He was certainly the first to notice this and he wrote a retraction (Draper, 1985)—changing entirely the conclusion of his earlier paper from “ R^2 is misleading” to “ R^2 is a useful indicator” (his italics). By one of those curious twists governing the survival mechanisms of scientific ideas, few workers in the area, and apparently none of those building on Draper's 1984 paper, seem to be aware of the retraction. At the time of writing this text, the retraction has been cited three times, two of which were by this author and a colleague. The original contribution is cited 39 times while, of greater impact, the paper of Kvalseth has been cited over one thousand times. It continues to be frequently quoted and its erroneous conclusions are widely taken to be valid.

Indeed, there is little mystery to the idea of explained variation. Note that, by virtue of the Chebyshev-Bienaymé inequality, explained variation directly quantifies predictive ability. In O'Quigley (2008) a basic theory of explained variation was outlined. It helps to refer to that theory in order to have a good understanding of what needs to be thought through when we wish to consider the particular case of explained variation for survival models. In the light of this theory and the main results concerning inference it is relatively straightforward to develop suitable measures of explained variation (O'Quigley and Flandre (1994), O'Quigley and Xu (2001)). The related topic, explained randomness, in view of its direct connection to likelihood, is also discussed at length in O'Quigley (2008). In a way similar to explained variation, although leaning on the concepts of entropy rather than the Chebyshev-Bienaymé inequality, explained randomness will also translate predictive ability. For normal models these different population quantities happen to coincide. For non-normal models on the other hand, when it might be argued that measures of variance are less immediate as a means to quantify predictive effects, a case can be made for preferring explained randomness over explained variation. In this text we work with explained variation but this is not because of any perceived advantage over explained randomness. In practice we would anticipate a very high level of agreement in analyzing data whichever of the two measures are used.

Among the users of statistical models it can be argued that there are two quite different schools of thought; the first sees a model as an approximation to some, infinitely more complex, reality: that is the model is no more than a tool, taken as a more-or-less refined means to achieving some specific end,

usually that of prediction of some quantity of interest. Any model is then simply judged by its predictive performance. The second school sees the statistician's job as tracking down the "true" model that can be considered to have generated the data. The position taken in this work is very much closer to the first than the second. This means that certain well-studied concepts, such as efficiency, a concept which assumes our models are correct, are given less attention than is often the case. The regression parameter β in our model is typically taken to be some sort of average where the proportional hazards model corresponds to a summary representation of the broader non-proportional hazards model. We view the restricted model as a working model and not some attempt to represent the true situation. Let's remind ourselves that the proportional hazards model stipulates that effects, as quantified by β , do not change through time. In reality the effects must surely change, hopefully not too much, but absolute constancy of effects is perhaps too strong an assumption to hold up precisely. The working model enables us to estimate useful quantities, one of them being average regression effect, the average of β taken through time. Interestingly, the usual partial likelihood estimator in the situation of changing regression effects does not estimate an average effect, as is often believed. Even so, we can estimate an average effect but we do require an estimator different from that commonly used. Details are given in Xu and O'Quigley (2000), O'Quigley (2008) and recalled in Section 7.6.

INTERPRETING PREDICTIVE INDICES

Is 0.48 greater than 0.16? If so, then by how much. We might think that the answer to that is easy, Indeed, the answer is easy when the discussion concerns real numbers. If, on the other hand, the discussion concerns something more concrete, something that is hopefully quantified by these numbers then the question can only be answered by first identifying just what it is we are trying to quantify and, secondly, how that quantity relates to the real numbers used in its place. Several popular measures of predictive strength of a survival model fail this simple test. For example, for the values 0.48 and 0.16, obtained from the popular C -statistic, it is not even possible to state that the first of these two numbers translates greater predictive strength than the second, let alone consider that, in some well-defined sense, the predictive power of 0.48 is 3 times that of 0.16. As a result this statistic lacks any kind of interpretation. Studies that make use of estimated values of the C -statistic in their analyses, as a means to establish orders of importance on the variables in their investigation, need to carefully take account of the censoring as described by Gönen and Heller (2005). Otherwise, ordering the importance of these variables alphabetically, or even randomly, may work no less well. However, even if we follow Gonen and Heller's prescription to fix the censoring problem, the interpretation difficulty remains. Certainly, it is now possible to be sure that an ordering of the indices does translate an order-

ing in terms of predictive strength. But the scale is still not something that we can grasp on an intuitive level. The great majority of the so-called predictive indices reviewed by Choodari-Oskooei et al. (2012) offer no advantage over the C -statistic and fail to provide any useful statistical insight into the problem of prediction for a survival model.

Pepe et al. (2015) shows that some alternative measures, such as the net reclassification index (Pencina et al., 2008) do no better than the measures they set out to improve upon. The main problem is almost always that of interpretation and, in this context, improvised solutions that take care of the censoring have rarely been met with success. The way to proceed, described in Section 3.9 of O'Quigley (2008), is to start from first principles, considering the elementary definitions of explained variation, what they mean, and how they might be estimated in the presence of independent censoring. Some suggestions in the literature have no clear meaning and it is something of a puzzle that they were ever used at all, let alone that they continue to attract interest. The so-called proportion of treatment effect explained (Freedman et al., 1992), believed to quantify how much of the treatment effect can be explained by another variable, typically some kind of surrogate endpoint for treatment, enjoys no interpretation as a proportion. Flandre and Saidi (1999) finds values for this quantity, using data from published clinical trials, ranging from -13% to 249% . These are the point estimates and not the endpoints of the confidence intervals which reach as far as 416% . There was nothing unusual in the data from those examples and we can conclude that Freedman's measure is not any kind of a percentage which, necessarily, would lie between zero and one hundred percent. What it is we cannot say although there is no reason to believe that 249% corresponds to stronger predictive effects than -13% .

These measures are quite useless. Some measures, while not necessarily being useless, appear to be not very useful. Take, for example, the lifetime risk of getting breast cancer for carriers of the BRCA1 and BRCA2 genetic mutation. A few years back we learned that a celebrity was deemed to have an 87% lifetime risk of getting breast cancer on the basis of her BRCA status. Despite enjoying full health, as a result of that risk she chose to have a double mastectomy. This so-called risk prompting such drastic action became not only a topic of discussion in scholarly medical journals but also made it into the mainstream press where, almost universally, the celebrity was praised for her good judgment and her proactive stance, a stance taken to avoid the perils that appeared to be confronting her. As far as we are aware, there was little or no discussion on the obvious question ... what is meant by the number 87% . Lifetime risk is near impossible to interpret (O'Quigley, 2017) and, by any standards, cannot be considered a useful measure of anything. More interpretable measures are available and we take the view that lifetime risk is not only not useful but is in fact misleading. It does not measure any clinical or physiological characteristic and basing major treatment and surgical decisions on its calculated value makes no sense. We

return to this in the chapter covering applications in epidemiology, although not in great depth, noting, for now, that, without doubt, the lack of understanding of the advisors concerning the complex meaning of lifetime risk will have contributed in part to a lack of clarity that is crucial to the making of an informed decision before undertaking mutilating life-changing surgery.

1.7 Data sets

The importance of working through the calculations on actual data cannot be overemphasized. The reader is encouraged to use their own or any of the many available published data sets, as well as simulated data sets corresponding to specific conditions. In the text we often refer to the Freireich data, described by several authors including Kalbfleisch and Prentice (2002) and Professor Cox in his 1972 founding paper. These data arose in the context of a balanced comparison of two treatment groups. We also refer to breast cancer data which were gathered at the Institut Curie in Paris, France over a near thirty-year period. Finally, another important data set used here to illustrate many concepts was obtained in the context of a survival study in gastric cancer carried out at the St James Hospital, Leeds, U.K. (see, for example, Rashid et al. (1982). An interesting case of non-proportional hazards arose in a clinical trial studied by Stablein et al. (1981). The regression effect appeared to change direction during the study. We refer to this in the text as the Stablein data. All of the examples and illustrations in this text can be confirmed using available software and the original data the source of which is referenced.

1.8 Use as a graduate text

The text can be used as support for an introductory graduate course in survival analysis with particular emphasis on proportional and non-proportional hazards models. The approach to inference is more classical than often given in such courses, steering mostly away from the measure-theoretic difficulties associated with multivariate counting processes and stochastic integrals and focusing instead on the more classical, and well known, results of empirical processes. Brownian motion and functions of Brownian motion play a central role. Exercises are provided in order to reinforce the coursework. Their aim is not so much to help develop a facility with analytic calculation but more to build insight into the important features of models in this setting. Some emphasis then is given to practical work carried out using a computer. No particular knowledge of any specific software package is assumed.

1.9 Classwork and homework

1. For the several examples described in Section 1.2 write down those features of the data which appear common to all examples. Which features are distinctive?
2. In Section 1.5 it is claimed that cumulative quantities may do enough smoothing in practice. How do you understand this idea of “enough smoothing?” What does a statistician aim to achieve by smoothing?
3. Suppose that the methods of survival analysis were not available to us. Suggest how we might analyze a randomized clinical trial using (i) multiple linear regression, (ii) multiple logistic regression.
4. For Exercise 3, describe the shortcomings that we expect to be associated with either type of analysis. How are these shortcomings amplified by the presence of increasing censoring, and in what way do we anticipate the techniques of survival analysis to address these shortcomings?
5. Consider a long-term survival study in which the outcome of interest is survival itself and we wish to study any possible dependence of this upon two other variables, one binary and one continuous. The marginal distribution of survival is unknown but suspected to be very skew. For the data at hand there is no censoring, i.e., all the actual survival times have been observed. Describe at least one possible approach, aside from those of survival analysis, to analyzing such data. Do you think more could be obtained from the data using survival techniques? Explain.
6. How would you describe to a non-scientist the purpose of a statistical model, explaining the issues involved with misspecified models. How would you describe to a physicist the difference between statistical models which may be employed in epidemiology and statistical models that he or she may be used to elaborate theories in quantum mechanics.
7. Consider a joint model for survival and an individual's covariate path through time. Suppose that the parametrization is such that the model for the covariate path and the model for survival do not share any parameters. Will knowledge, provided by data, on the parameters that model the covariate path, help improve inference for the regression coefficients of the survival model? If not, why not?
8. In the context of joint modeling, find a way to make use of information on the individual's covariate process to improve inference on the survival process without appealing to any structure that postulates shared parameters between models. Otherwise, how might we decide the plausibility of such a structure?

9. Check out the definitions for explained variation and explained randomness for bivariate continuous distributions. Show that, in the case of a multi-normal model, explained randomness and explained variation coincide.
10. Suppose that T_1 and T_2 are two survival variates considered in a competing risks setting. We suppose independence. Imagine though that the observations of T_2 can be modeled by treating T_1 as a time-dependent covariate in a Cox model. When the regression coefficient is non-zero describe how this would impact the Kaplan-Meier curves obtained for a set of observations; T_{21}, \dots, T_{2n} . What can be said when the model is an NPH one and the regression coefficient a function of time?



Chapter 2

Survival analysis methodology

2.1 Chapter summary

We recall some elementary definitions concerning the probability distributions, putting an emphasis toward one minus the usual cumulative distribution function, i.e., the survival function. This is also sometimes called the survivorship function. The closely related hazard function has, traditionally, been the most popular function around which to construct models. For multistate models it can be helpful to work with intensity functions, rather than hazard functions since these allow the possibility of moving in and out of states. This is facilitated by the very important function, $Y(t)$, the “at-risk” indicator. A number of special parametric cases of proportional hazards models are presented. The issue of censoring and the different kinds of censoring is discussed. The “at-risk” indicator $Y_i(w, t)$, taking the value one when the subject i is at risk of making a transition of a certain kind, indicated by w , makes it particularly simple to address more complex issues in survival such as repeated events, competing risks, and multistate modeling. We consider some tractable parametric models, the exponential model in particular.

2.2 Context and motivation

Survival time T will be a positive random variable, typically right skewed and with a non-negligible probability of sampling large values, far above the mean. The fact that an ordering, $T_1 > T_2$, corresponds to a solid physical interpretation has led some authors to argue that time is somehow different from other continuous random variables, reminiscent of discussion among early twentieth-century physicists about the nature of time “flowing inexorably in and of itself”. These characteristics are sometimes put forward as a reason for considering techniques other than the classic techniques of linear regression. From a purely statistical

viewpoint, this reasoning is incorrect. Elementary transformations fix the skewness problems which, in consequence, reveal themselves as quite superficial. Nor is there any worthwhile, statistical, distinction between time and, say, height or weight. The reason for considering particular techniques, outside of the classical ones of linear regression, is the presence of censoring. In early work, censoring came to be viewed as a nuisance feature of the data collection, hampering our efforts to study the main relationships of interest. A great breakthrough occurred when this feature of the data, the censoring, was modeled by the “at-risk” function. Almost immediately it became clear that all sorts of much more involved problems; competing risks, repeated events, correlated outcomes, could all be handled with almost no extra work. Careful use of the “at-risk” indicator was all that would be required. At the heart then of survival analysis is the idea of being at risk for some event of interest taking place in a short time frame (for theoretical study this short time will be made arbitrarily small). Transition rates are then very natural quantities to consider. In epidemiology these ideas have been well rooted for a half-century where age-dependent rates of disease incidence have been the main objects under investigation.

2.3 Basic tools

TIME AND RISK

The insurance example in the introduction highlights an obvious, but an important, issue. If driver A, on average, has a higher daily risk than driver B, then his or her mean time to be involved in an accident will be shorter. Conversely, if driver B has a longer mean time to accident, then he has, on average, a lower daily risk. For many examples we may tend to have in mind the variable time and how it is affected by other variables. But we can think equally well in terms of risk over short time periods, a viewpoint that we will see generalizes more readily to be able to deal with complicated situations. The connection between time and risk is outlined more formally below.

CONDITIONAL DISTRIBUTIONS

Our goal is to investigate dependence, its presence, and, if present, the degree of dependence. The information on this is conveyed to us via the conditional distribution of time given the covariate, keeping in mind that this covariate may be a combination of several covariates. The conditional distribution of the covariate given time may not immediately strike us as particularly relevant. However, it is very relevant because: (1) the marginal distribution of time and the marginal distribution of the covariate contain little or no information on dependency so that the two conditional distributions can be viewed, in some sense, as being equivalent and, (2) problems relating to censoring are near intractable when dealing with the conditional distribution of time given the covariate whereas they become

somewhat straightforward when we consider the conditional distribution of the covariate given time. This is a very central idea and runs throughout this text.

HAZARD AND RELATED FUNCTIONS

The purpose here is to continue the introduction of preliminary notions and some basic concepts. Before discussing data and estimation we consider the problem in its most simplified form as that of the study of the pair of random variables (T, Z) , T being the response variable “survival” of principal interest and Z an associated “explanatory” variable. There would be little difficulty in applying the host of techniques from linear regression to attacking this problem were it not for the presence of a “censoring” variable C . The particularity of C is that, when observed, i.e., $C = c$, we are no longer able to observe values of T for which $T > c$. Also, in most cases, when T is observed, we are no longer able to observe C . Nonetheless an observation on one tells us something about the other, in particular, that it must assume some greater value.

Although the joint distribution of (T, Z) can be of interest, we are particularly interested in the conditional distribution of T given Z . First let us consider T alone. The probability density function of T is defined as

$$f(t) = \lim_{\Delta t \rightarrow 0^+} \frac{1}{\Delta t} \Pr(t < T < t + \Delta t), \quad (2.1)$$

where $\lim_{\Delta t \rightarrow 0^+}$ means that Δt goes to 0 only through positive values. We define as usual $F(t) = \int_0^t f(u) du$. The survivorship function is written as $S(t) = 1 - F(t)$. If we view the density as the unconditional failure rate, we can define a conditional failure rate as being the same quantity after having accounted for the fact that the individual has already survived until the time point t . We call this $\lambda(t)$ and we define

$$\lambda(t) = \lim_{\Delta t \rightarrow 0^+} \frac{1}{\Delta t} \Pr(t < T < t + \Delta t | T > t). \quad (2.2)$$

It helps understanding to contrast Equation (2.2) and (2.1) where we see that $\lambda(t)$ and $f(t)$ are closely related quantities. In a sense, the function $f(t)$ for all values of t is seen from the standpoint of an observer sitting at $T = 0$, whereas, for the function $\lambda(t)$, the observer moves along with time looking at the same quantity but viewed from the position $T = t$. Analogous to a density, conditioned by some event, we can define

$$\lambda(t|C > t) = \lim_{\Delta t \rightarrow 0^+} \frac{1}{\Delta t} \Pr(t < T < t + \Delta t | T > t, C > t). \quad (2.3)$$

The conditioning event $C > t$ is of great interest since, in practical investigations, all our observations at time t have necessarily been conditioned by the event. All associated probabilities are also necessarily conditional. But note that, under an

independent censoring mechanism, we have that $\lambda(t|C > t) = \lambda(t)$. This result underlies the great importance of certain assumptions, in this case that of independence between C and T . The conditional failure rate, $\lambda(t)$, is also sometimes referred to as the hazard function, the force of mortality, the instantaneous failure rate or the age-specific failure rate. If we consider a small interval then $\lambda(t) \times \Delta t$ closely approximates the probability of failing in a small interval for those aged t , the approximation improving as Δt goes to zero. If units are one year then these are yearly death rates. The cumulative hazard function is also of interest and this is defined as $\Lambda(t) = \int_0^t \lambda(u)du$. For continuous $\lambda(t)$, using elementary calculus we can see that:

$$\lambda(t) = f(t)/S(t), \quad S(t) = \exp\{-\Lambda(t)\}, \quad f(t) = \lambda(t) \exp\{-\Lambda(t)\}.$$

Although mathematically equivalent, we may prefer to focus attention on one function rather than another. The survival function, $S(t)$, is the function displaying most clearly the information the majority of applied workers are seeking. The hazard function, $\lambda(t)$, of central concern in much theoretical work, provides the most telling visual representation of time effects. An important function, of theoretical and practical interest, is the conditional survivorship function,

$$S(t, u) = \Pr(T > t | T > u) = \exp\{\Lambda(u) - \Lambda(t)\}, \quad (u < t).$$

From this it is clear that $S(t, u) = S(t)/S(u)$ and that $S(u, u) = 1$ so that it is as though the process had been restarted at time $t = u$. Other quantities that may be of interest in some particular contexts are the mean residual lifetime, $m(t)$, and the mean time lived in the interval $[0, t]$, $\mu(t)$, defined as

$$m(t) = E(T - t | T \geq t), \quad \mu(t) = \int_0^t S(u)du. \quad (2.4)$$

Like the hazard itself, these functions provide a more direct reflection on the impact of having survived until time t . The mean residual lifetime provides a very interpretable measure of how much more time we can expect to survive, given that we have already reached the timepoint t . This can be useful in actuarial applications. The mean time lived in the interval $[0, t]$ is not so readily interpretable, requiring a little more thought (it is not the same as the expected lifetime given that $T < t$). It has one strong advantage in that it can be readily estimated from right-censored data in which, without additional assumptions, we may not even be able to estimate the mean itself. The functions $m(t)$ and $\mu(t)$ are mathematically equivalent to one another as well as the three described above and, for example, a straightforward integration by parts shows that $m(t) = S^{-1}(t) \int_t^\infty S(u)du$ and that $\mu(\infty) = E(T)$. If needed, it follows that the survivorship function can be expressed in terms of the mean residual lifetime by

$$S(t) = m^{-1}(t)m(0) \exp\left(-\int_0^t m^{-1}(u)du\right).$$

We may wish to model directly in terms of $m(t)$, allowing this function to depend on some vector of parameters θ . If the expression for $m(t)$ is not too intractable then, using $f(t) = -S'(t)$ and the above relationship between $m(t)$ and $S(t)$, we can write down a likelihood for estimation purposes in the situation of independent censoring. An interesting and insightful relationship (see, for instance, the Kaplan-Meier estimator) between $S(t)$ and $S(t,u)$ follows from considering some discrete number of time points of interest. Thus, for any partition of the time axis, $0 = a_0 < a_1 < \dots, a_n = \infty$, we see that

$$S(a_j) = S(a_{j-1})S(a_j, a_{j-1}) = \prod_{\ell \leq j} S(a_\ell, a_{\ell-1}). \quad (2.5)$$

The implication of this is that the survival function $S(t)$ can always be viewed as the product of a sequence of conditional survival functions, $S(t,u)$. This simple observation often provides a foundation that helps develop our intuition, one example being the case of competing risks looked at later in this chapter. Although more cumbersome, a theory could equally well be constructed for the discrete case whereby $f(t_i) = \Pr(T = t_i)$ and $S(t_i) = \sum_{\ell \geq i} f(t_\ell)$. We do not explore this here.

INTENSITY FUNCTIONS AND COMPARTMENT MODELS

Modern treatment of survival analysis tends to focus more on intensity than hazard functions. This leads to great flexibility, enabling, for example, the construction of simple models to address questions in complex situations such as repeated events (Andersen and Gill, 1982). We believe that both concepts can be useful and we will move back and forth between them according to the application. Intensity functions find their setting in the framework of stochastic processes where the random nature of T is suppressed, t being taken simply as an index to some stochastic process. The *counting process* $N(t)$, takes the value 0 at



Figure 2.1: An elementary alive/dead compartment model with two states

$t = 0$, remaining at this same value until some time point, say $T = u$, at which the event under study occurs and then $N(t) = \mathbf{1}_{t \geq u}$. We can then define, in an infinitesimal sense, i.e., the equality only holds precisely in the limit as dt goes

to zero through positive values

$$\Pr\{(N(t) - N(t - dt)) = 1 | \mathcal{F}_{t-dt}\} = \alpha(t)dt \quad (2.6)$$

where \mathcal{F}_{t-dt} , written as \mathcal{F}_{t-} when we allow $dt > 0$ to be arbitrarily close to zero, is the accumulated information, on all processes under consideration, observed up until time $t - dt$ (Figure 2.1).

The observed set \mathcal{F}_{t-} is referred to as the history at time t . The set is necessarily non-decreasing in size as t increases, translating the fact that more is being observed or becoming known about the process. The Kolmogorov axioms of probability, in particular sigma additivity, may not hold for certain noncountable infinite sets. For this reason probabilists take great care, and use considerable mathematical sophistication, to ensure, in broad terms, that the size of the set \mathcal{F}_{t-} does not increase too quickly with t . The idea is to ensure that we remain within the Kolmogorov axiomatic framework, in particular that we do not violate sigma additivity. Much of these concerns have spilled over into the applied statistical literature where they do not have their place. No difficulties will arise in applications, with the possible exception of theoretical physics, and the practitioner, unfamiliar with measure theory, ought not to be deterred from applying the techniques of stochastic processes simply because he or she lacks a firm grasp of concepts such as filtrations. It is hard to imagine an application in which a lack of understanding of the term “filtration” could have led to the error. On the other hand, the more accessible notions of history, stochastic process, and conditioning sets are central and of great importance both to understanding and to deriving creative structures around which applied problems can be solved. Viewing t as an index to a stochastic process rather than simply the realization of a random variable T , and defining the intensity process $\alpha(t)$ as above, will enable great flexibility and the possibility to model events dynamically as they unfold.

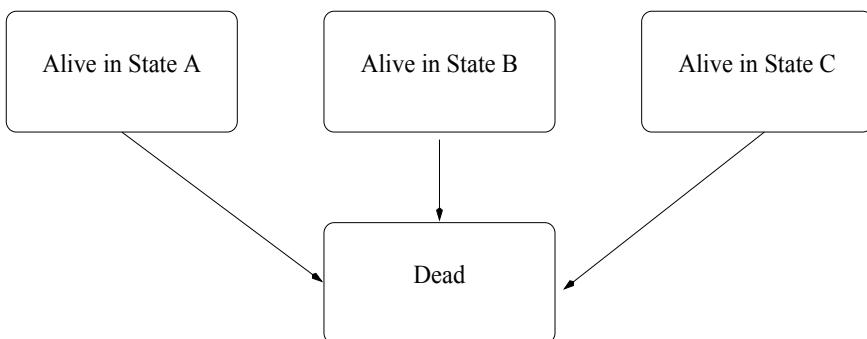


Figure 2.2: A compartment model with 3 covariate levels and an absorbing state. Model is fully specified by the 3 transition rates from states A, B, or C to the state Dead

AT-RISK FUNCTIONS $Y(t)$, $Y(w,t)$ AND MULTISTATE MODELS

The simplest case we can consider occurs when following a randomly chosen subject through time. The information in \mathcal{F}_{t-} tells us whether or not the event has yet occurred and if the subject is still at risk, i.e., the set \mathcal{F}_{t-} is providing the same information as an observation on the function $Y(t)$ where we take $Y(t)$ to be left continuous, assuming the value one until the occurrence of an event, or removal from observation, at which time it assumes the value zero. If the simple fact of not having been removed from the study, the event ($C > t$) is independent of the event ($t < T < t + dt$), then conditioning on $Y(t) = 1$ is the same as conditioning on $T > t$. Referring then to Equation 2.2, it is clear that if $Y(t) = 0$ then $\alpha(t) = 0$ and, if $Y(t) = 1$ then $\alpha(t) = \lambda(t)$. Putting these two results together we have

$$\alpha(t) = Y(t)\lambda(t). \quad (2.7)$$

This relation is important in that, under the above condition, referred to as the independent censoring condition, the link between the intensity function and the hazard function is clear. Note that the intensity function is random since Y is random when looking forward in time. Having reached some time point, t say, then $\alpha(t)$ is fixed and known since the function $Y(u)$, $0 < u < t$ is known and $Y(t)$ is left continuous.

We call $Y(\cdot)$ the “at risk” function (left continuous specifically so that at time t the intensity function $\alpha(t)$ is not random). The idea generalizes readily and in order to cover a wide range of situations we also allow Y to have an argument w where w takes integer values counting the possible changes of state. For the i th subject in any study we will typically define $Y_i(w,t)$ to take the value 1 if this subject, at time t , is at risk of making a transition of type w , and 0 otherwise. Figure 2.2 summarizes a situation in which there are four states of interest, an absorbing state, death, and three states from which an individual is able to make a transition into the death state. Transitions among the three non-death states themselves cannot occur. Later we will consider different ways of modeling such a situation, depending upon further assumptions we may wish or not wish to make.

In Figure 2.3 there is one absorbing state, the death state, and two non-absorbing states between which an individual can make transitions. We can define $w = 1$ to indicate transitions from state 1 to state 2, $w = 2$ to indicate transitions from state 2 to state 1, $w = 3$ to indicate transitions from state 1 to state 3 and, finally, $w = 4$ to indicate transitions from state 2 to state 3. Note that such an enumeration only deals with whether or not a subject is at risk for making the transition, the transition probabilities (intensities) themselves could depend on the path taken to get to the current state.

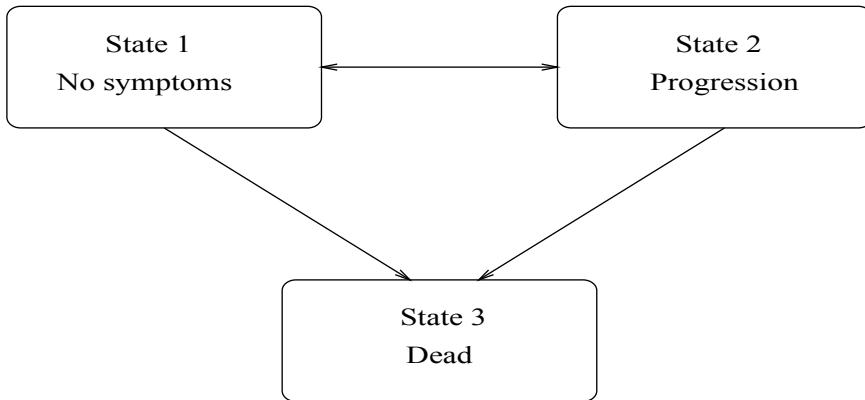


Figure 2.3: Binary time-dependent covariate and single absorbing state. Probabilistic structure is fully specified by the transition rates at time t , $\alpha_{ij}(t)$ for leaving state i for state j . Fix and Neyman's illness-death model, also known as the semi-competing risks model, is a special case when $\alpha_{21}(t) = 0, \forall t$.

We can then appreciate why it can be helpful to frame certain questions in terms of compartment models, intensity functions and the risk function. Rather complex situations can be dealt with quite straightforwardly, the figures illustrating simple cases where we can use the argument w in $Y_i(w, t)$ to indicate, at any t , which kinds of transition any given subject i is available to make. In Figure 2.4 there are two absorbing states, one of which can only be reached from state 2. The transition rate between state 2 and state 4 may or may not depend on the number of times a subject moves between states 1 and 2. Allowing for transitions between states greatly adds to the flexibility of any model so that, in Figure 2.2, although the explanatory variable (state) has three levels, the model is, in principle, much simpler than that described in Figure 2.3 where the explanatory variable maybe characterized by the history as well as the current state.

AT-RISK INDICATOR $Y(w, t)$ AND REPEATED EVENTS

Some studies have the particularity that an occurrence of the event of interest does not remove the subject from further observation. Additional events, of the same or of different types, may happen. An example is benign breast disease, potentially followed by malignant disease. A patient may have several incidences of benign breast disease at different intervals of time. Following any one of these incidences, or even before such an incidence takes place the subject may become an incident for malignant disease. If our interest is essentially focussed on the incidence of malignant disease then we would treat the time-dependent history of benign breast disease as a potential explanatory variable for incidence of malignant disease. However, we may also be interested in modeling directly

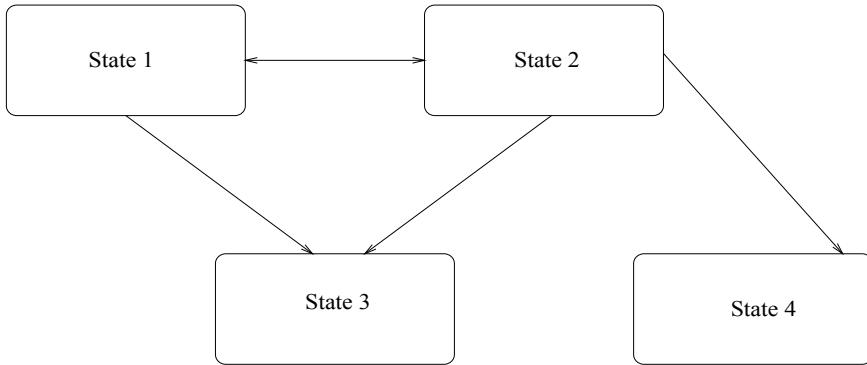


Figure 2.4: Binary time-dependent covariate and two absorbing states. Transitions to state 4 pass through state 2, i.e., $\alpha_{j4}(t) = 0, j \neq 2, \forall t$. Note also that $\alpha_{4j}(t)=0$.

the repeated incidence of benign breast disease in its own right. Clearly a patient can only be at risk of having a third incident of benign breast disease if she has already suffered two earlier incidents. We can model the rate of incidence for the j th occurrence of benign disease as,

$$\alpha_j(t) = Y(j,t)\lambda_j(t - t_{j-1}), \quad (2.8)$$

where $t_0 = 0$ and t_j is the observed occurrence of the j th event. Different options may be considered for modeling $\lambda_j(t)$. Usually there will be at least one covariate, Z , indicating two distinct prognostic groups, possibly established on the basis of different treatments. The model will involve coefficients multiplying Z and thereby quantifying treatment effect. Allowing these coefficients to also depend upon j provides the broadest generality and is equivalent to analyzing separate studies for each of the occurrences. Stronger modeling, imposing greater structure, might assume that the coefficients do not depend upon j , in which case the information provided by a subject having three incident cases is comparable to that of three independent subjects each providing information on a single incident. So-called marginal models have been proposed in this context. Here, it would be as though the subject, after an event, starts the clock from zero and, aside from covariate information, is deemed to be in the same position as another subject who has just entered the study without having yet suffered a single event. A lot of information would appear to be thereby gained but the setup seems rather artificial and implausible. Starting the clock from zero, after each event, is sensible but it is more realistic to assume that the underlying hazard rates, i.e., those not adjusted by covariate information, would change with the number of prior incidents. In other words a more plausible model would condition on this information allowing the baseline hazard rate to change according to the

number of events counted so far. Stratified proportional hazards models have been successfully used in order to analyze such problems.

2.4 Some potential models

SIMPLE EXPONENTIAL

The simple exponential model is fully specified by a single parameter λ . The hazard function, viewed as a function of time, does not in fact depend upon time so that $\lambda(t) = \lambda$. By simple calculation we find that $\Pr(T > t) = \exp(-\lambda t)$. Note that $E(T) = 1/\lambda$ and, indeed, the exponential model is often parameterized directly in terms of the mean $\theta = E(T) = 1/\lambda$. Also $\text{Var}(t) = 1/\lambda^2$. This model expresses the physical phenomenon of no aging or wearing out since, by elementary calculations, we obtain $S(t+u, u) = S(t)$; the probability of surviving a further t units of time, having already survived until time u , is the same as that associated with surviving the initial t units of time. The property is sometimes referred to as the lack of memory property of the exponential model.

For practical application the exponential model may suggest itself in view of its simplicity or sometimes when the constant hazard assumption appears realistic. A good example is that of a light bulb which may only fail following a sudden surge in voltage. The fact that no such surge has yet occurred may provide no information about the chances for such a surge to take place in the next given time period. If T has an exponential distribution with parameter λ then λT has the so-called standard exponential distribution, i.e., mean and variance are equal to one.

Recall that for a random variable Y having normal distribution $\mathcal{N}(\mu, \sigma^2)$ it is useful to think in terms of a simple linear model $Y = \mu + \sigma\epsilon$, where ϵ has the standard distribution $\mathcal{N}(0, 1)$. As implied above, scale changes for the exponential model lead to a model still within the exponential class. However, this is no longer so for location changes so that, unlike the normal model in which linear transformations lead to other normal models, a linear formulation for the exponential model is necessarily less straightforward. It is nonetheless of interest to consider the closest analogous structure and we can write

$$Y = \log T = \alpha + bW, \quad (2.9)$$

where W has the standard extreme value density $f(w) = \exp\{w - \exp(w)\}$. When $\alpha = 0$ we recover an exponential model for T with parameter b , values other than zero for α pushing the variable T out of the restricted exponential class into the broader Weibull class discussed below.

PROPORTIONAL HAZARDS EXPONENTIAL

In anticipation of the central topic of this book (that of heterogeneity among the subjects under study) imagine that we have two groups, indicated by a binary variable $Z = 0$ or $Z = 1$. For $Z = 0$ the subjects follow an exponential law with parameter λ_0 . For $Z = 1$ the subjects follow an exponential law with parameter λ_1 . It is clear that for the hazard functions there exists real β ($= \log \lambda_1 - \log \lambda_0$) such that

$$\lambda(t|Z) = \lambda(t|Z=0) \exp(\beta Z) = \lambda_0 \exp(\beta Z). \quad (2.10)$$

The important point to note here is that the ratio of the hazards, $\lambda(t|Z=1)/\lambda(t|Z=0)$ does not involve t . It also follows that $S(t|Z=1) = S(t|Z=0)^\alpha$ where $\alpha = \exp(\beta)$. The survival curves are power transformations of one another. This is an appealing parameterization since, unlike a linear parameterization, whatever the true value of β , the constraints that we impose upon $S(t|Z=1)$ and $S(t|Z=0)$ in order to be well-defined probabilities, i.e., remaining between 0 and 1, are always respected. Such a model is called a proportional hazards model. For three groups we can employ two indicator variables, Z_1 and Z_2 , such that, for group 1 in which the hazard rate is equal to λ_0 , $Z_1 = 0$ and $Z_2 = 0$, for group 2, $Z_1 = 1$ and $Z_2 = 0$ whereas for group 3, $Z_1 = 0$ and $Z_2 = 1$. We can then write;

$$\lambda(t|Z) = \lambda_0 \exp(\beta_1 Z_1 + \beta_2 Z_2), \quad (2.11)$$

where $\lambda_0 = \lambda(t|Z_1 = Z_2 = 0)$. It is worthwhile bringing the reader's attention to just where the constraints of the model express themselves here. They concern the hazard rates for all groups, which are assumed to be constant. Given this constraint there are no further constraints concerning the relationship between the groups. Suppose, though, that we were to consider a further group, group 4, defined by $Z_1 = 1$ and $Z_2 = 1$. In order to add a fourth group without introducing a further binary coding variable Z_3 , we introduce the constraint that the hazard for group 4 is simply expressed in terms of the hazards for groups 2 and 3. Such assumptions are commonly made in routine data analysis but, nonetheless, ought to come under critical scrutiny. We return to this issue in later chapters. The extension to many groups follows in the same way. For this we take Z to be a p -dimensional vector of indicator variables and β a vector of parameters having the same dimension as Z , the product βZ in Equation 2.10 now implying an inner product, i.e., $\beta Z = \sum_{i=1}^p \beta_i Z_i$. In this case, the proportional hazards exponential model (2.10) implies that every group follows some simple exponential law, a consequence being that the survivorship function for any group can be expressed as a power transformation of any other group. Once again, it is important to keep in mind just which assumptions are being made, the potential impact of such assumptions on conclusions, as well as techniques for bringing under scrutiny

these assumptions. The proportional hazards constraint then appears as a very natural one in which we ensure that the probabilities $S(t|z)$ and subsequent estimates always remain between 0 and 1. A linear shift added to $S(t|0)$ would not allow for this. We do nonetheless have a linear shift although on a different, and thereby more appropriate, scale and we can write

$$\log -\log S(t|Z) = \log -\log S(t|0) + \sum_{i=1}^p \beta_i Z_i.$$

This formulation is the same as the proportional hazards formulation. Noting that $-\log S(T|Z = z)$ is an exponential variate some authors prefer to write a model down as a linear expression in the transformed random variable itself with an exponential error term. This then provides a different link to the more standard linear models we are familiar with.

PIECEWISE EXPONENTIAL

The lack of flexibility of the exponential model will often rule it out as a potential candidate for application. Many other models, only one or two of which are mentioned here, are more tractable, a property stemming from the inclusion of at least one additional parameter. Even so, it is possible to maintain the advantages of the exponential model's simplicity while simultaneously gaining flexibility. One way to achieve this is to construct a partition of the time axis $0 = a_0 < a_1 < \dots < a_k = \infty$. Within the j th interval (a_{j-1}, a_j) , $(j = 1, \dots, k)$ the hazard function is given by $\lambda(t) = \lambda_j$. We can imagine that this may provide quite a satisfactory approximation to a more involved smoothly changing hazard model in which the hazard function changes through time. We use $S(t) = \exp\{-\Lambda(t)\}$ to obtain the survival function where

$$\Lambda(t) = \sum_{j=1}^k I(t \geq a_j) \lambda_j (a_j - a_{j-1}) + \sum_{j=1}^k I(a_{j-1} \leq t < a_j) \lambda_j (t - a_{j-1}). \quad (2.12)$$

Properties such as the lack of memory property of the simple exponential have analogs here by restricting ourselves to remaining within an interval. Another attractive property of the simple exponential is that the calculations are straightforward and can be done by hand and, again, there are ready analogues for the piecewise case. Although the ready availability of sophisticated computer packages tends to eliminate the need for hand calculation, it is still useful to be able to work by hand if for no other purposes than those of teaching. Students gain invaluable insight by doing these kinds of calculations the long way.

PROPORTIONAL HAZARDS PIECEWISE EXPONENTIAL

In the same way as for the simple exponential model, for two groups, indicated by a binary variable $Z = 0$ or $Z = 1$, each having constant piecewise rates on the same intervals, it is clear that there exists β_j such that, for $t \in [a_{j-1}, a_j)$,

$$\lambda(t|Z) = \lambda(t|Z=0) \exp(\beta_j Z) = \lambda_0(t) \exp\{\beta(t)Z\}, \quad (2.13)$$

where we now have a function $\beta(t) = \sum_{j=1}^k \beta_j I(a_{j-1} \leq t < a_j)$. This can be described as a nonproportional hazards model and, if, under a further restriction that $\beta(t)$ is a constant function of time, i.e., $\beta_1 = \beta_2 = \dots = \beta_k = \beta$, then, as for the simple exponential model, we have $S(t|Z=1) = S(t|Z=0)^\alpha$ where $\alpha = \exp(\beta)$ and, once again, such a model is called a proportional hazards model. The model can once more be described in terms of a linear translation on $\log - \log S(t|z)$.

WEIBULL MODEL

Another way to generalize the exponential model to a wider class is to consider a power transformation of the random variable T . For any positive γ , if the distribution of T^γ is exponential with parameter λ , then the distribution of T itself is said to follow a Weibull model whereby

$$f(t) = \lambda \gamma (\lambda t)^{\gamma-1} \exp\{-(\lambda t)^\gamma\}$$

and $S(t) = \exp\{-(\lambda t)^\gamma\}$. The hazard function follows immediately from this and we see, as expected, that when $\gamma = 1$ an exponential model with parameter λ is recovered. It is of interest to trace out the possible forms of the hazard function for any given λ . It is monotonic, increasing for values of γ greater than 1 and decreasing for values less than 1. This property, if believed to be reflected in some given physical situation, may suggest the appropriateness of the model for that same situation. An example might be the time taken to fall over for a novice one-wheel skateboard enthusiast—the initial hazard may be high, initially decreasing somewhat rapidly as learning sets in and thereafter continuing to decrease to zero, albeit more slowly.

The Weibull model, containing the exponential model as a special case, is an obvious candidate structure for framing questions of the sort—is the hazard decreasing to zero or is it remaining at some constant level? A null hypothesis would express this as $H_0 : \gamma = 1$. Straightforward integration shows that $E(T^r) = \lambda^{-r} \Gamma(1+r/\gamma)$ where $\Gamma(\cdot)$ is the gamma function,

$$\Gamma(p) = \int_0^\infty u^{p-1} e^{-u} du \quad p > 0.$$

For p integer $\Gamma(p) = (p - 1)!$ The mean and the variance are $\lambda^{-1}\Gamma(1 + 1/\gamma)$ and $\lambda^{-2}\Gamma(1 + 2/\gamma) - E^2$, respectively. The Weibull model can be motivated by the theory of statistics of extremes. The distribution coincides with the limiting distribution of the smallest of a collection of random variables, under broad conditions on the random variables in question (Kalbfleisch and Prentice, 2002, page 48).

PROPORTIONAL HAZARDS WEIBULL

Once again, for two groups indicated by a binary variable $Z = 0$ or $Z = 1$, sharing a common γ but different values of λ , then there exists a β such that $\lambda(t|Z)/\lambda(t|Z = 0) = \exp(\beta Z)$. Since, as above, the right-hand side of the equation does not depend on t , then we have a proportional hazards model. This situation and the other two described above are the only common parametric models that come under the heading proportional hazards models by simply expressing the logarithm of the location parameter linearly in terms of the covariates. The situation for more than two groups follows as before. Consider however a model such as

$$\lambda(t|Z) = \lambda\gamma(\lambda t)^{\gamma-1} \exp(\beta Z), \quad (2.14)$$

in which Z indicates three groups by assuming the values $Z = 1, 2, 3$.

Unlike the model just above in which three groups were represented by two distinct binary covariates, Z_1 and Z_2 , we have only one covariate. In the context of estimation and a given set of data we will almost invariably achieve greater precision in our estimates when there are less parameters to estimate. We would then appear to gain by using such a model. As always though, any such gain comes at a price and the price here is that we have made much stronger assumptions. We are assuming that the signed “distance” between groups 1 and 2, as measured by the logarithm of the hazard, is the same as the signed distance between groups 2 and 3. If this is not the case in reality then we are estimating some sort of compromise, the exact nature of which is determined by our estimating equations. In an extreme case in which the distances are the same but the signs are opposite we might erroneously conclude that there is no effect at all. At the risk of being repetitive, it cannot be stressed too much just how important it is to identify the assumptions we are making and how they may influence our conclusions. Here the assumptions concern both the parametric form of the underlying risk as well as the nature of how the different groups are related. Allowing a shape parameter γ to be other than one provides a more flexible model for the underlying risk than that furnished by the simple exponential model. The choice of covariate coding, on the other hand, is more restrictive than the earlier choice. All of this needs to be studied in applications. An interesting point is that, for the three group cases defined as above, the “underlying” hazard, $\lambda(t|Z = 0) = \lambda\gamma(\lambda t)^{\gamma-1}$ does not correspond to the hazard for any of the three groups under study. It is

common in practice to consider a recoding of Z , a simple one being $Z - \bar{Z}$, so that the underlying hazard will correspond to some kind of average across the groups. For the case just outlined, another simple recoding is to rewrite Z as $Z - 2$, in which case the underlying hazard corresponds to the middle group, the other two groups having hazard rates lower and greater than this, respectively.

LOG-MINUS-LOG TRANSFORMATION

As a first step to constructing a model for $S(t|Z)$ we may think of a linear shift, based upon the value of Z , the amount of the shift to be estimated from data. However, the function $S(t|Z)$ is constrained, becoming severely restricted for both $t = 0$ and for large t where it approaches one and zero respectively. Any model would need to accommodate these natural constraints. It is usually easiest to do this by eliminating the constraints themselves during the initial steps of model construction. Thus, $\log S(t|Z) = -\Lambda(t)$ is a better starting point for modeling, weakening the hold the constraints have on us. However, $\log -\log S(t|Z) = \log \Lambda(t)$ is better still. This is because $\log \Lambda(t)$ can take any value between $-\infty$ and $+\infty$, whereas $\Lambda(t)$ itself is constrained to be positive. The transformation $\log -\log S(t|Z)$ is widely used and is called the log-minus-log transformation. The above cases of the exponential and Weibull proportional hazards models, as already seen, fall readily under this heading.

OTHER MODELS

The exponential, piecewise exponential, and Weibull models are of particular interest to us because they are especially simple and of the proportional hazards form. Nonetheless there are many other models which have found use in practical applications. Some are directly related to the above, such as the extreme value model in which

$$S(t) = \exp \left(-\exp \left(\frac{t-\mu}{\sigma} \right) \right),$$

since, if T is Weibull, then $\log T$ is extreme value with $\sigma = 1/\gamma$ and $\mu = \log \lambda$. These models may also be simple when viewed from some particular angle. For instance, if $M(s)$ is the moment-generating function for the extreme value density then we can readily see that $M(s) = \Gamma(1+s)$. A distribution, closely related to the extreme value distribution (Balakrishnan and Johnson, 1994), and which has found wide application in actuarial work is the Gompertz where

$$S(t) = \exp (\beta \alpha^{-1} (1 - e^{\alpha t})) .$$

The hazard rates for these distributions increase with time, and, for actuarial work, in which time corresponds to age, such a constraint makes sense for studying disease occurrence or death. The normal distribution is not a natural candidate in view of the tendency for survival data to exhibit large skewness, not

forgetting that times themselves are constrained to be positive. The log-normal distribution has seen some use but is most often replaced by the log-logistic, similar in shape apart from the extreme tails, and much easier to work with. The form is particularly simple for this model and we have

$$S(t) = (1 + (\alpha t)^\gamma)^{-1}.$$

For two groups, sharing a common γ but different values of α it is interesting to note that the hazard ratio declines monotonically with time t to its asymptotic value of one. Such a model may be appropriate when considering group effects which gradually wane as we move away from some initial time point.

PARAMETRIC PROPORTIONAL HAZARDS MODELS

In principle, for any parametric form, the above providing just a very few examples, we can make a straightforward extension to two or more groups via a proportional hazards representation. For example, if the survivorship functions of two groups are $S(t|Z = 1)$ and $S(t|Z = 0)$ then we can introduce the parameter α to model one group as a power transform of the other. Rewriting α to include Z via $\alpha = \exp(\beta Z)$ then we have an expression involving the regressors,

$$\log -\log S(t|Z) = \log -\log S(t|Z = 0) + \beta Z. \quad (2.15)$$

All parameters, including β , can be estimated using standard techniques, maximum likelihood, in particular, the only restriction being that we require some conditions on the censoring variable C . In practice, standard techniques are rarely used, most likely as a consequence of the attractive proposal of Cox (1972) whereby we can estimate β without having to consider the form of $S(t|Z = 1)$ or $S(t|Z = 0)$. As attractive as the Cox approach is though, we should not overlook the fact that, in exchange for generality concerning the possible parametric forms of functions of interest, such as $S(t|Z)$, making inferences on these population quantities become that much more involved. Parametric proportional hazards models may be an area that merits renewed interest in applications.

2.5 Censoring

The most important particularity of survival data is the presence of censoring. Other aspects such as the positivity and skewness of the main random variable under study, time T , and other complex situations such as repeated measures or random effects, are not of themselves reasons for seeking methods other than linear regression. Using transformations and paying careful attention to the structure of the error, linear models are perfectly adequate for dealing with almost any situation in which censoring does not arise. It is the censoring that forces us to consider other techniques. Censoring can arise in different ways.

We typically view the censoring as a nuisance feature of the data, and not of direct interest in its own right, essentially something that hinders us from estimating what it is we would like to estimate. In order for our endeavors to succeed we have to make some assumptions about the nature of the censoring mechanism. The assumptions may often be motivated by convenience, in which case it is necessary to give consideration as to how well grounded the assumptions appear to be as well as to how robust are the procedures to depart from any such assumptions. In other cases the assumptions may appear natural given the physical context of interest, a common case being the uniform recruitment into a clinical trial over some predetermined time interval. When the study closes patients for whom the outcome of interest has not been observed are censored at study close and until that point occurs it may be reasonable to assume that patients are included in the study at a steady rate.

It is helpful to think of a randomly chosen subject being associated with a pair of random variables (T, C) , an observation on one of the pair impeding observation on the other, while at the same time indicating that the unobserved member of the pair must be greater than the observed member. This idea is made more succinct by saying that only the random variable $X = \min(T, C)$ can be fully observed. Clearly $\Pr(X > x) = \Pr(T > x, C > x)$ and we describe censoring as being independent whenever

$$\Pr(X > x) = \Pr(T > x, C > x) = \Pr(T > x)\Pr(C > x). \quad (2.16)$$

TYPE I CENSORING

Such censoring most often occurs in industrial or animal experimentation. Items or animals are put on test and observed until failure. The study is stopped at some time T^* . If any subject does not fail it will have observed survival time at least equal to T^* . The censoring times for all those individuals being censored are then equal to T^* . Equation (2.16) is satisfied and so this is a special case of independent censoring, although not very interesting since all subjects, from any random sample, have the same censoring time.

TYPE II CENSORING

The proportion of censoring is determined in advance. So if we wish to study 100 individuals and observe half of them as failures we determine the number of failures to be 50. Again all censored observations have the same value T^* although, in this case, this value is not known in advance. This is another special case of independent censoring.

TYPE III CENSORING

In a clinical trial patients enter randomly. A model for entry is often assumed to be uniform over a fixed study period, anywhere from a few months to several years but determined in advance. Survival time is the time from entry until the event of interest. Subjects can be censored because (1) the end of the study period is reached, (2) they are lost to follow-up (3) the subject fails due to something unrelated to the event of interest. This is called random censoring. So, unlike for *Type I* or *Type II* censoring, for a random sample C_1, \dots, C_n , the C_i could all be distinct.

For a random sample of pairs (T_i, C_i) , $i = 1, \dots, n$, we are only able to observe $X_i = \min(T_i, C_i)$. A fundamental result in this context was discovered by Tsiatis (1975). The result says that, for such data, we are unable to estimate the joint distribution of the pair (T, C) . Only the marginal distributions can be estimated under the independent censoring assumption, the assumption itself not being testable from such data. It is common then to make the assumption of independent censoring, sometimes referred to as non-informative censoring, by stipulating that

$$\Pr(X_i > x) = \Pr(T_i > x, C_i > x) = \Pr(T_i > x) \Pr(C_i > x). \quad (2.17)$$

The assumption is strong but not entirely arbitrary. For the example of the clinical trial with a fixed closing date for recruitment it seems reasonable to take the length of time from entry up until this date as not being associated with the mechanism generating the failures. For loss to follow-up due to an automobile accident or due to leaving the area, again the assumption may be reasonable, or, at least, a good first approximation to a much more complex, unknown, and almost certainly unknowable, reality.

INFORMATIVE CENSORING, KOZIOL-GREEN MODEL

When censoring is informative, which we can take to be the negation of non-informative, then it is no longer possible to estimate the main quantities of interest without explicitly introducing some model for the censoring. The number of potential models relating C and T is infinite and, in the absence of special knowledge, it can be helpful to postulate some simple relationship between the two, the proportional hazards model itself having been used in this context (CsÖrgÖ and Horvath, 1981; Slud and Rubinstein, 1983). This model was first introduced by Koziol and Green (1976) and has a wide range of applicability (Gaddah and Braekers, 2009; Gather and Pawlitschko, 1998; Stute, 1992). Obvious examples might be surrogate endpoints in the study of the evolution of AIDS following treatment, where, for falling CD4 cell counts, below a certain point patients can be withdrawn from the study. Censoring here is clearly informative. This will be the case whenever the fact of removing a subject, yet to experience the event of

interest, from the study implies a change in risk. Informative censoring is necessarily more involved than non-informative censoring and we have to resort to more elaborate models for the censoring itself in order to make progress. If, as might be the case for a clinical trial where the only form of censoring would be the termination of the study, we know for each subject, in advance, their censoring time C , we might then postulate that

$$\log - \log S(t) = \log - \log(S(t|C < t)) + \beta I(C > t).$$

This would be a proportional hazards model for a dependent censoring mechanism. More generally we would not know C in advance of making observations on T , but we could write down a similar model in terms of intensity functions, viewing the censoring indicator as a predictable stochastic process. For the purposes of estimation we may require empirical quantities indicating how the risk changes once censoring is observed, and for this we need to be able to compare rates between those censored at some point and those who are not. Mostly, once censoring has occurred, it is no longer possible to observe the main event under study so that, for data of this nature, we are not able to estimate parameters of interest without further assumptions. These assumptions are usually that the censoring is independent of the failure process or that it is conditionally independent given covariate values. The paper of Tsiatis (1975) demonstrates this intuitive observation formally.

MARGINAL AND CONDITIONALLY INDEPENDENT CENSORING

When considering many groups, defined by some covariate value Z , there are essentially two types of independence commonly needed. The stronger assumption is that of marginal independence in which the variables T and C as well as C and Z are pairwise independent. The censoring distribution for C is the same for different values of Z . A weaker assumption that is often made is that of conditional independence. Here, the pair (T, C) are independent given Z . In other words, for each possible value of Z , the pair (T, C) is independent, but the censoring distribution C can be different for different values of Z . A nice illustration of conditional independence was shown by Cox (1972) in his analysis of the Freireich data. The assumption appeared to be a reasonable one. The stronger assumption of marginal independence is often needed in more complex situations, an example being the consistent estimation of R^2 , and needs to be borne in mind and examined critically. We return to this later in the text.

FINITE CENSORING SUPPORT

Many mathematical issues simplify immediately when the failure variable T is continuous, as we generally suppose, but that the censoring variable is restricted to having support on some finite subset. We can imagine that censoring times

are only allowed to take place on the set $\{a_0, a_1, \dots, a_k\}$. This is not a practical restriction since we can make the division (a_j, a_{j-1}) as fine as we wish. We will frequently need to consider the empirical distribution function and analogues (Kaplan-Meier estimate, Nelson-Aalen estimate) in the presence of censoring. If we adopt this particular censoring setup of finite censoring support, then generalization from the empirical distribution function to an analogue incorporating censoring is very straightforward. We consider this in greater detail when we discuss the estimation of marginal survival.

COMPETING RISKS AND CENSORING

Recalling the “at-risk” indicator function, $Y_i(w, t)$, which takes the value one if, at time t , the i th subject is at risk of making a transition of type w , and is zero otherwise, we can imagine a simple situation in which w takes only one of two values. Calling these $w = 1$ and $w = 2$, consider a constraint whereby $Y_i(1, t) = Y_i(2, t)$. In other words, if the i th subject is at risk of one kind of transition, then he or she is also at risk of the other kind. If the subject is no longer at risk then this means that they are not at risk for either kind of transition. Thus, if a subject suffers an event of type $w = 1$ then he or she is no longer considered at risk of suffering an event of type $w = 2$, and conversely.

This is the situation of competing risks. As long as the subject is at risk, then either of the event types can occur. Once one type of event has occurred, then it is no longer possible to observe an occurrence of an event of the other type. Such a construction fits in immediately with the above models for survival involving censoring. If at time $t = t_1$ an event of type $w = 1$ takes place, then, as far as events of type $w = 2$ are concerned, the subject is simply censored at $t = t_1$. A subject may be at risk of death from stroke or at risk from either stroke or cirrhosis of the liver. Once one of the types of death has occurred, then the other type of event can no longer be observed. We will assume that the subject is censored at this point, in as much as our attention focuses on the second type of event, and the above discussion on the different censoring models applies in the same way. We will need to make some assumptions, most often that of independent censoring or that of independent censoring conditional on covariate information in order to make progress.

2.6 Competing risks

A number of models for competing risks are in use and we consider these briefly without going into technical detail. We might view the simplest approach to this question as one, seen from a somewhat abstract theoretical angle, where everyone will ultimately suffer any and all events given enough time. We only observe the first such event and all others are censored at that time. This corresponds to the competing risks model based on latent events and with an independence

assumption (CRM-I). An alternative theoretical construct, CRM-J, requires us to break down the event space into a partition, i.e., only one of the event types can be observed and, for a failure, one of the event types must be observed. If we use the random variable D to denote the type of failure then we can write

$$\lambda_k(t) = \lim_{\Delta t \rightarrow 0^+} \frac{1}{\Delta t} \Pr(t < T < t + \Delta t, D = k | T > t), \quad (2.18)$$

described as the cause-specific hazard rate. This is straightforward enough, notwithstanding the strong assumptions relating to the partition, so that, given data in the form of risk sets and events of type k taking place within relatively short intervals, we can estimate $\Lambda_k(t) = \int_0^t \lambda(u) du$. From this we have the survivorship function, $S_k(t) = \exp\{-\Lambda_k(t)\}$. The interpretation of this is not so easy. It is certainly not accessible to non-statisticians if only for the fact that it does not stand alone and depends crucially on what other causes of failure are under consideration. This is immediately apparent from the expression for the survivorship function of T , the first observed failure, as, $S(t) = \exp\{-\sum_k \Lambda_k(t)\}$.

The authors on this topic will often point out that the interpretation of $S_k(t)$ is difficult, in particular since as t increases without bound $S_k(t)$ will not tend to zero. While this is clear from its definition it is not our biggest problem here. Our main problem is one of interpretation. As far as epidemiologists are concerned, the CRM-J model will be challenging to fit since we are generally unable to estimate the rates specified by this model using available registry data. These data take the form of numbers of individuals at the beginning of some age interval “at risk” of the event in question. The number of such events occurring in the interval can be counted. While the ratio of these provides valid estimates of short-term rates, this validity does lean on a strong untestable assumption which is that, for those subjects lost to follow-up during any interval, had we somehow managed to not lose them to follow-up then the subsequent rates we are trying to estimate remain unaffected.

This model for a joint outcome (CRM-J) contrasts with CRM-I for competing risks where everyone suffers everything ultimately. While any event will, in a similar way, censor all others, our outcome variable is now a joint one, the time to the event together with the type of event observed. A third model (CRM-ID) is a competing risks model built around the illness-death (hence the ID) model of Fix and Neyman. This model is also referred to as the semi-competing risks model and has been used to advantage in Fine and Gray (1999), Fine et al. (2001), Ghosh (2006) and Haneuse and Lee (2016). Lee et al. (2015) considered inference for such models from a Bayesian viewpoint while some of the difficulties in likelihood construction were studied in Hsieh and Huang (2012). The impact that competing risks can have on the confounding of the main treatment effect has recently been discussed in Huang R. and Dulai (2020).

Models CRM-I and CRM-J have their advantages and drawbacks. The model based on a joint outcome may seem the most appealing, in some ways the more

realistic, but there are many technical issues associated with this model and there is considerable practical difficulty in outlining an appropriate partition of the event space. The survival probabilities of main interest are confounded with all of the other survival probabilities and, necessarily, dependent on the chosen partition. This can be problematic from the point of view of interpretation. It is also very difficult in terms of finding suitable data that allow estimation of the main quantities of interest. Competing risks models have seen greater application in epidemiology than in clinical research and we consider this again in the context of genetic epidemiology (Chapter 5).

For the simplest case, that of just two competing risks, say failure and censoring, a quite fundamental result was derived by Tsiatis (1975). The result indicates that the marginal distributions of failure or censoring—viewed as competing risks, i.e., the observation on one impedes the observation on the other—cannot, under general conditions, be estimated. These distributions are not identifiable. In order to find consistent estimates of these distributions from observations arising under such competing risks, it is necessary to add further restrictive conditions. The most common condition used in this context is that of independence. Tsiatis (1975), in a paper that is something of a cornerstone to subsequent theory, showed that we can obtain consistent estimates using, for example, the Kaplan-Meier estimator, under the assumption of independence.

Consider a clinical trial where patients enter the study according to a uniform law over some time interval. This appears to be a somewhat reasonable model. Their time to incidence of some event of interest is then recorded. At some point the study is closed and the available data analyzed. Not only will we consider those patients yet to experience the event to be censored but such censoring will be taken to be independent. The assumption seems reasonable. It would break down of course if the recruitment over the interval was subject to change, in particular, the more robust patients being recruited early on. So, even in such a simple case, the needed independence assumption ought to come under scrutiny. We can encounter other situations where the independence assumption requires something of a stretch of the imagination. For example, if patients who are responding poorly to treatment were to be removed from the study.

Next, let's consider a more realistic but greatly more complex situation in which there are several competing risks. We cannot reasonably assume independence. The occurrence of any particular outcome will prevent us from observing all of the other outcomes. In real-life situations this is what happens of course, when dealing with an absorbing state such as death. If a patient were to die of heart failure then he or she will no longer be in a position to die of anything else. In order to make progress in analyzing such data it is necessary to make a large number of, essentially, unverifiable assumptions. In making these assumptions we need to be guided by a basic principle in all statistical modeling, the principle of parsimony. The situation we wish to model is, almost always, infinitely more complex than that we can realistically hope to model accurately. Even if we

could find a plausible model deemed to describe the mechanism that generates the observations, the lack of precision of our parameter estimates will be such as to drown in noise any worthwhile signal we are aiming to identify. Modeling requires something of the artist's ability to sense when, in trying too hard, their work can fall short of the goal in mind. The dilemma confronting the statistical modeler is not altogether different. Competing risks models are particularly complex and great care is called for in order to remain fully confident in our conclusions. The field is very large and several published texts are devoted solely to the topic. We do not have the space here to do justice to this extensive body of work apart from to make some observations on the 3 main approaches that have been widely adopted. For those new to the field two clear overviews are given by Putter et al. (2007) and Xu and Tai (2010). There is a large body of literature on these topics and suggested reading would include Crowder (1994), Tsiatis (2005), Moeschberger and Kocher (2007), Andersen and Rosthoj (2002), Bakoyannis and Touloumi (2012) and Austin and Fine (2017). The impact of competing risks on clinical trials is looked at in Satagopan and Auerbach (2004) and Freidlin and Korn (2005).

In order to assess the impact of model choice on estimation we can consider the development more formally. For CRM-I, the independence assumption for all causes is made together with the idea of latent event times. In this setup every individual is taken, initially, to be at risk of an event of all causes. If we have m competing risks and times to event, T_1, T_2, \dots, T_m , then we only get to observe $X = \min(T_1, T_2, \dots, T_m)$ and all event times greater than the minimum remain unobserved and recorded as censored at X . This structure is simple and appealing. Cancer registries such as the SEER registries compiling data from several millions of individuals provide information of a form that can be used by CRM-I. Very large numbers of individuals, at given ages, are taken to be at risk from one of the possible outcome events under study. The number of occurrences of any particular cause over some small age interval divided by the number at risk, either at the beginning of the interval or some average over the interval, provide an estimate of the age-specific incidence rate for that particular cause. This is then repeated for the next adjacent interval. Cumulative incidence follows from just adding together the age-specific incidence rates. Using the elementary formulas introduced in Section 2.3 we can obtain a survival probability. This probability is not easily interpreted, leaning as it does on many near impossibly restrictive assumptions and, more bothersome still, an idealized characterization of reality that is so far removed from human existence as to make its direct applicability problematic.

The implications of choosing CRM-J rather than CRM-I are highlighted by consideration of Equation 2.5. At some time point, let's say a_j , we have that

$$S(a_j) = S_I(a_{j-1})S(a_j, a_{j-1}) \quad (2.19)$$

where $S_I(t)$ is the survival function for the event of interest under an assumption of independence of the competing events, i.e., $S_I(t) = S(t)$ for all t . In Equation 2.5 there was only a single competing event, the censoring itself, taken to be independent, but let us extend that to all competing events. The left-hand side of the equation gives the probability of surviving beyond time point a_j and this is the simple product of the probabilities of reaching time point a_{j-1} together with the probability of surviving the interval (a_{j-1}, a_j) . These probabilities are not affected by an independent censoring mechanism, the working assumption that we are making. So, when it comes to estimation we can use convergent estimators of survival, such as the Kaplan-Meier or the Nelson-Aalen formulas together with the observed rate estimated from the interval (a_{j-1}, a_j) . From a theoretical viewpoint any subject censored before time a_{j-1} remains in an abstract sense at risk for the event under study albeit, given actual data, for this subject, the event can not be observed. The event for this subject is considered a latent variable, it exists but is not observable.

For the competing risk model CRM-J, this is no longer true. Once one type of event has been observed then all other types of event are considered to not exist, we do not have a latent variable structure. In consequence, in the above formula, $S(a_j)$ describes the probability of the event of interest occurring after time point a_j , and this is the product of the incidence rate for this event between a_{j-1} and a_j and the probability, $S_J(a_{j-1})$ that all events—the one of interest together with all competing events—occur some time later than a_{j-1} . In this case we have:

$$S(a_j) = S_J(a_{j-1})S(a_j, a_{j-1}). \quad (2.20)$$

Clearly, $S_I(t)$ and $S_J(t)$ in Equations 2.19 and 2.20 may differ so that in turn, the cumulative incidence rates, $1 - S(a_j)$, depending on whether model CRM-I or CRM-J is assumed, can also differ. For a breast cancer study, Satagopan and Auerbach (2004) found the cumulative incidence rates based on CRM-I or CRM-J to all but coincide while, for an example in haematologic malignancy, the two working models led to quite different estimates.

The working assumption behind the characterization of CRM-I in terms of latent variables is that, ultimately, if given enough time, then everyone dies from everything. Of course this is a hypothetical construct and while of value in the appropriate context, it has little role to play in individual prediction. Such prediction can give an entirely false impression on the non-specialist. For any cause, say, for example, breast cancer, the only reason registry data will furnish a probability that grows to anything less than one (100%) is a lack of observations, a limitation on the overall time span. This is a somewhat abstract, theoretical observation—its mathematical expression comes from the divergence of the harmonic series—that does require we think carefully about how we interpret any cumulative risks. The age-specific risks are calculated by constantly updating the risk sets, having removed deaths from all other causes in the denominator.

Essentially, in this calculation of age-specific risk, via this constant updating, we are not allowing a subject to fail through any cause other than that under study.

Only a severe lack of observations on highest age groups prevents the lifetime risk—the probability of being incident or of dying from the specified cause—from gradually approaching 100%. Lifetime risk will typically be provided over intervals not going much beyond 75 or 80 years. This does not make it any more meaningful and clinical decisions, such as prophylactic mastectomies, need to be based on considerations other than such probabilities. These probabilities are near impossible to give a clear interpretation to, and this is by experts in probability theory. They should be considered impossible to interpret by non-specialists and ought not to be used to advise individuals on their chances of being incident for some disease. We return to the problems of lifetime risk in the chapter dealing with epidemiology and, in particular, we review the assessments of breast cancer risk and its relation to the genetic markers, BRCA1 and BRCA2.

The several difficulties associated with CRM-I, in particular the interpretation problems have led to more emphasis being given to CRM-J. The structure of CRM-J has attracted a lot of theoretical interest and there now exists a large body of work that we will not be able to review here. The main advantage of CRM-J is it is more straightforward interpretability although, even here, this is not without problems. The main disadvantage is that there exists very little data suitable for analysis. The many data banks held in the registries cannot be readily analyzed via CRM-J. For whatever disease is in question, the registries are based on large risk sets and the process of counting the incident cases over some relatively short age interval. All other events during the interval are removed from further assessment without having identified the nature or reason for their removal. This fits in well with model CRM-I. The information needed to estimate risks within the structure of CRM-J is not generally available. That said, there are several examples in the literature where a solution is cobbled together by taking death as a competing risk for incidence of some outcome. All other risks are assumed independent and rates of death, obtained from life tables, can be used as a rough approximation to the unknown survivorship function of all events.

There are many more things to worry about with competing risks models. We open up a huge topic simply by aiming to relax independence assumptions. We also need to give thought to the competing causes we wish to consider. There is an almost unlimited spectrum of different partitions of the space of causes we might wish to analyze and, for any individual cause, associated probabilities of future transitions will depend not only on its own status, and related risk factors, but also on the status of those other causes we wish to include in the analysis. It all sounds very complex. It is very complex and yet the tools can be invaluable to the investigator who does not have to rely on only one of CRM-I or CRM-J. They are at liberty to employ them both depending on the context. They can both throw some light on the infinitely complex array of factors that relate to prognosis. Interpreting survival probabilities is almost impossibly difficult and,

fortunately, this is not the goal. The goal is to provide no more than some rough and ready tools to help the investigators gain better insight into the myriad of factors, and their approximate relative importance, that may be implicated in the outcomes on which we are focused.

A slightly modified structure, that is well described within the framework of the compartment models described earlier, gives rise to the so-called semi-competing risks models. For these models, one of the competing causes is not an absorbing state so that, having arrived in the non-absorbing state, subjects are still free to leave this state in order to transition to the absorbing state. Seen from the viewpoint of a compartment model it is clear that semi-competing risks models coincide with the classical illness-death model (Fix and Neyman, 1951). For this reason we refer to the model as CRM-ID. Some of the underlying assumptions can differ however, in particular the existence and behavior of latent times, and these assumptions will impact the task of estimation. There is an extensive literature on such models and the methods of this book can be adapted with more or less effort in order to deal with inferential questions that arise in this context. In this text our central focus is on regression and while we do not study in depth the question of regression with competing risks models there are several papers that do, among which Larson and Dinse (1985), Lunn and McNeil (1995), Kim (2007), Scheike and Zhang (2008), Lau and Gange (2009), Dignam and Kocherginsky (2012), and Haller et al. (2013).

Which of the 3 main models, CRM-I, CRM-J, and CRM-ID, we may use for analysis will mostly be guided by the application. The different underlying assumptions can have a more or less strong impact on inferences. The most useful focus is regression modeling, which can involve assessing the impact of certain risk factors after having adjusted for others, the joint impact of these risk factors and the relative importance of different risk factors. In this, arguably the most important, setting the model choice tends to be relatively robust. If one model provides a hierarchy of importance of risk factors then, in the main, the same or a very similar hierarchy obtains with other models. The main differences due to model choice will occur when we consider absolute rather than relative quantities such as the probability of surviving so many years without being incident of some particular disease. A great deal of circumspection is recommended when describing absolute rather than relative assessments, especially when any such assessments can be heavily model dependent. The author's own limited experiment in asking fellow statisticians—mostly not in the particular field—to interpret the 87% chance of breast cancer risk given to some carriers of a mutated gene was revealing. Not one person got it right—the author is no exception, at least initially—leading to the inescapable conclusion that this number, 87%, is almost impossibly difficult to interpret. For non-specialists, and most particularly the carriers themselves, it is hard to understand the wisdom behind asking them to make life-changing decisions on the basis of any such number. We return to this in the chapter on epidemiology and, here, we point this out as a way to

underline an important fact: no less effort ought to go into the purpose of our competing risks models than in their construction.

The main focus of this text is on regression and we do not investigate the specifics of this in the competing risks setting. Many ideas carry though. Many do not. An introduction to regression modeling of the sub-distribution function for models CRM-J and CRM-ID can be found in Fine and Gray (1999) and Zhang and Fine (2011).

2.7 Classwork and homework

1. Using the definition for $\lambda(t) = f(t)/S(t)$, show that $S(t) = \exp\{-\Lambda(t)\}$ and that $f(t) = \lambda(t) \exp\{-\Lambda(t)\}$.
2. For a Weibull variate with parameters λ and k , derive an expression for the conditional survivorship function $S(t+u, u)$. How does this function vary with t for fixed u ? With u for fixed t ?
3. Use numerical integration to calculate the mean residual lifetime $m(t)$ and the mean time lived in the interval $[0, t]$, $\mu(t)$ for the Weibull with parameters 2 and 1.5. With parameters 2 and 0.7. Plot these as functions of time t .
4. Consider two groups in which $f(t) = \lambda\gamma(\lambda t)^{\gamma-1} \exp\{-(\lambda t)^\gamma\}$, i.e., each group follows a Weibull distribution where, for the first group, $\lambda = \lambda_1$, $\gamma = \gamma_1$ and, for the second, $\lambda = \lambda_2$, $\gamma = \gamma_2$. Under which conditions will this situations be described by proportional hazards?
5. Undertake a numerical and graphical study of the conditional survivorship function, $S(t+u, u)$, for the Weibull model, the extreme value model, the Gompertz model and the log-logistic model. What conclusions can be drawn from this?
6. Repeat the previous class project, focusing this time on the mean residual lifetime. Again what conclusions can be drawn from the graphs.
7. Consider a disease with three states of gravity (state 1, state 2, and state 3), the severity corresponding to the size of the number. State 4 corresponds to death and is assumed to follow state 3. New treatments offer the hope of prolonged survival. The first treatment, if it is effective, is anticipated to slow down the rate of transition from state 2 to state 3. Write down a compartmental model and a survival model, involving a treatment indicator, for this situation. A second treatment, if effective, is anticipated to slow down all transition rates. Write down the model for this. Write down the relevant null and alternative hypotheses for the two situations.

8. Consider a non-degenerative disease with several states; $1, 2, \dots$, counting the occurrence of these together with a disease state indicating a progression to something more serious, e.g., benign and malignant tumors or episodes of mild asthma with the possibility of progression to a more serious respiratory ailment. Write down possible models for this and how you might formulate tests of hypotheses of interest under varying assumptions on the role of the less serious states.
9. Suppose we have data; T_1, \dots, T_n , from a Weibull distribution in which the shape parameter γ is known to be equal to 1.3. Use the delta-method to find an estimate for the variance of the estimated median (transform to a standard form).
10. For a proportional hazards Weibull model describe the relationship between the respective medians.
11. Investigate the function $S(t, u)$ for different parametric models described in this chapter. Draw conclusions from the form of this two-dimensional function and suggest how we might make use of these properties in order to choose suitable parametric models when faced with actual data.
12. Consider two possible structures for a parametric proportional hazards model;

$$\log S(t|Z) = \log\{S[t|E(Z)]\} \exp(\beta Z)$$
$$\log S(t|Z) = \log\{ES[t|Z]\} \exp(\beta Z).$$

How do the interpretations differ and what difficulties are likely to be encountered in fitting either of the models?

13. Consider a clinical trial comparing two treatments in which patients enter sequentially. Identify situations in which an assumption of an independent censoring mechanism may seem a little shaky.
14. On the basis of a single data set, fit the exponential, the Weibull, the Gompertz, and the log-normal models. On the basis of each model estimate the mean survival. On the basis of each model estimate the 90th percentile. What conclusions would you draw from this?
15. Suppose our focus of interest is on the median. Can you write down a model directly in terms of the median. Would there be any advantage/drawback to modeling in this way rather than modeling the hazard and then obtaining the median via transformations of the hazard function?
16. A cure model will typically suppose two populations; one for which the probability of an event follows some distribution and another for which

the hazard rate can be taken to be zero. The ratio of the sizes of the populations is $\theta/(1 - \theta)$. Write down such a model and indicate how we could go about estimating the mixing parameter, θ . How might a Bayesian tackle such a problem.

17. An alternative to a cure model is based on an approach akin to that used in relative survival. For some reference population the hazard rate is given by $\lambda_0(t)$ whereas, for the population of interest, it is given by $\lambda_1(t)$. For $t < \tau$, we will assume that the rate $\lambda_1(t)$ is strictly greater than $\lambda_0(t)$ as a result of some risk factor, most often treatment, but that for $t > \tau$, the two rates are so close as to be essentially indistinguishable. Discuss the advantages and drawbacks, in terms of estimation and interpretability, of this approach when compared to that described in the previous question.



Chapter 3

Survival without covariates

3.1 Chapter summary

The marginal survival function is of central interest even when dealing with covariates. We need to find good estimates of this function and we use those estimates in several different contexts. Good estimates only become possible under certain assumptions on the censoring mechanism. Attention is paid to the exponential and piecewise exponential models, both of which are particularly transparent. The exponential model, fully characterized by its mean, can appear over restrictive. However, via the probability integral transform and empirical estimates of marginal survival, it can be used in more general situations. The piecewise exponential is seen, in some sense, to lie somewhere between the simple exponential and the empirical estimate. Particular attention is paid to empirical processes and how the Kaplan-Meier estimator, very commonly employed in survival-type problems, can be seen to be a natural generalization of the empirical distribution function. In the presence of parametric assumptions, it is also straightforward to derive suitable estimating equations. The equations for the exponential model are very simple.

3.2 Context and motivation

Our interest is mostly in the survival function $S(t)$. Later we will focus on how $S(t)$, written as $S(t|Z)$, depends on covariates Z . Even though such studies of dependence are more readily structured around the hazard function $\lambda(t|Z)$, the most interpretable quantity we often would like to be able to say something about is the survival function itself. In order to distinguish the study of the influence of Z on $S(t|Z)$ from the less ambitious goal of studying $S(t)$, we refer to the former as conditional survival and the latter as marginal survival.

Since we will almost always have in mind some subset Z from the set of all possible covariates, and some distribution for this subset, we should remind ourselves that, although Z has been “integrated out” of the quantity $S(t)$, the distribution of Z does impact $S(t)$. Different experimental designs will generally correspond to different $S(t)$. Marginal survival, $S(t)$, corresponds to two situations: (i) the subjects that are considered as *i.i.d.* replicates from a single population or (ii) the subjects can be distinct, from many, and potentially an infinite number of populations, each population being indexed by a value of some covariate Z . It may also be that we have no information on the covariates Z that might distinguish these populations. In case (ii), $S(t)$ is an average over these several populations, not necessarily representing any particular population of interest in itself. It is important to appreciate that, in the absence of distributional assumptions, and the absence of observable Z , it is not possible, on the basis of data, to distinguish case (i) from case (ii). The homogeneous case then corresponds to either case (i) or case (ii) and it is not generally useful to speculate on which of the cases we might be dealing with. They are not, in the absence of observable Z , identifiable from data. We refer to $S(t|Z)$ as the conditional survival function given the covariate Z . This whole area, the central focus of this work, is studied in the following chapters. First, we need to consider the simpler case of a single homogeneous group.

3.3 Parametric models for survival functions

Let us suppose that the survival distribution can be completely specified via some parametric model, the parameter vector being, say, θ . We take θ to be a scalar in most cases in order to facilitate the presentation. The higher-dimensional generalization is, in most cases, very straightforward.

MAXIMUM LIKELIHOOD ESTIMATION

The data will consist of the n i.i.d. pairs $(x_i, \delta_i); i = 1, \dots, n$. We assume an independent censoring mechanism. This leads to the important lemma:

Lemma 3.1. *Under an independent censoring mechanism the log-likelihood can be written as $\log L(\theta) = \sum_{i=1}^n \log L_i(\theta)$, where*

$$\log L_i(\theta) = \delta_i \log f(x_i; \theta) + (1 - \delta_i) \log S(x_i; \theta) \quad (3.1)$$

This covers the majority of cases in which parametric models are used. Later, when we focus on conditional survival involving covariates Z , rather than marginal survival, the same arguments follow through. In this latter case the common assumption, leading to an analogous expression for the log-likelihood, is that of conditional independence of the pair (T, C) given Z .

ESTIMATING EQUATION

The maximum likelihood estimate is obtained as the value of θ , denoted $\hat{\theta}$, which maximizes $L(\theta)$ over the parameter space. Such a value also maximizes $\log L(\theta)$ (by monotonicity) and, in the usual case where $\log L(\theta)$ is a continuous function of θ this value is then the solution to the estimating equation (see Appendix D.1),

$$U(\theta) = \partial \log L(\theta) / \partial \theta = \sum_i \partial \log L_i(\theta) / \partial \theta = 0.$$

Next, notice that at the true value of θ , denoted θ_0 , we have $\text{Var}\{U(\theta_0)\} = EU^2(\theta_0) = EI(\theta_0)$ where

$$I(\theta) = \sum_{i=1}^n I_i(\theta) = -\partial^2 \log L(\theta) / \partial \theta^2 = -\sum_{i=1}^n \partial^2 \log L_i(\theta) / \partial \theta^2.$$

As for likelihood in general, some care is needed in thinking about the meaning of these expressions and the fact that the operators $E(\cdot)$ and $\text{Var}(\cdot)$ are taken with respect to the distribution of the pairs (x_i, δ_i) but with θ_0 fixed. The score equation is $U(\hat{\theta}) = 0$ and the large sample variance is approximated by $\text{Var}(\hat{\theta}) \approx 1/I(\hat{\theta})$. It is usually preferable to base calculations on $I(\hat{\theta})$ rather than $EI(\hat{\theta})$, the former being, in any event, a consistent estimate of the latter (after dividing both sides of the equation by n). The expectation itself would be complicated to evaluate, involving the distribution of the censoring, and unlikely, in view of the study by Efron and Hinkley (1978) to be rewarded by more accurate inference. Newton-Raphson iteration is set up from

$$\hat{\theta}_{j+1} = \hat{\theta}_j + I(\hat{\theta}_j)^{-1} U(\hat{\theta}_j), \quad j \geq 1, \tag{3.2}$$

where $\hat{\theta}_1$ is some starting value, often zero, to the iterative cycle. The Newton-Raphson formula arises as an immediate application of the mean value theorem (Appendix A). The iteration is brought to a halt once we achieve some desired level of precision.

Large sample inference can be based on any one of the three tests based on the likelihood function; the score test, the likelihood ratio test, or the Wald test. For the score test there is no need to estimate the unknown parameters. Many well-established tests can be derived in this way. In exponential families, also the so-called curved exponential families (Efron et al., 1978), such tests reduce to contrasting some observed value to its expected value under the model. Confidence intervals with optimal properties (Cox and Hinkley, 1979) can be constructed from uniformly most powerful tests. For the exponential family class of distributions the likelihood ratio forms a uniformly most powerful test and, as such, allows us to obtain confidence intervals with optimal properties. The other tests are asymptotically equivalent so that confidence intervals based on the above test procedures will agree as sample size increases. Also we can use

such intervals for other quantities of interest such as the survivorship function which depends on these unknown parameters.

ESTIMATING THE SURVIVAL FUNCTION

We can estimate the survival function as $S(t; \hat{\theta})$. If Θ_α provides a $100(1 - \alpha)\%$ confidence region for the vector θ then we can obtain a $100(1 - \alpha)\%$ confidence region for $S(t; \theta)$ in the following way. For each t let

$$S_\alpha^+(t; \hat{\theta}) = \sup_{\theta \in \Theta_\alpha} S(t; \theta), \quad S_\alpha^-(t; \hat{\theta}) = \inf_{\theta \in \Theta_\alpha} S(t; \theta), \quad (3.3)$$

then $S_\alpha^+(t; \hat{\theta})$ and $S_\alpha^-(t; \hat{\theta})$ form the endpoints of the $100(1 - \alpha)\%$ confidence interval for $S(t; \theta)$. Such a quantity may not be so easy to calculate in general, simulating from Θ_α or subdividing the space being an effective way to approximate the interval. Some situations nonetheless simplify such as the following example, for scalar θ , based on the exponential model in which $S(t; \theta)$ is monotonic in θ . For such cases it is only necessary to invert any interval for θ to obtain an interval with the same coverage properties for $S(t; \theta)$.

EXPONENTIAL SURVIVAL

For this model we only need estimate a single parameter, λ which will then determine the whole survival curve. Referring to Equation 3.1 in which, for $\delta_i = 1$, the contribution to the likelihood is, $f(x_i; \lambda) = \lambda \exp(-\lambda x_i)$ and, for $\delta_i = 0$, the contribution is, $S(x_i; \lambda) = \exp(-\lambda x_i)$. Equation 3.1 then becomes:

$$\log L(\lambda) = k \log \lambda - \lambda \sum_{j=1}^n x_j, \quad (3.4)$$

where $k = \sum_{i=1}^n N_i(\infty)$. Differentiating this and equating with zero we find that $\hat{\lambda} = k / \sum_{j=1}^n x_j$. Differentiating a second time we obtain $I(\lambda) = k / \lambda^2$. Note, by conditioning upon the observed number of failures k , that $EI(\lambda) = I(\lambda)$, the observed information coinciding with the expected Fisher information, a property of exponential families, but which we are not generally able to recover in the presence of censoring.

An ancillary argument would nonetheless treat k as being fixed and this is what we will do as a general principle in the presence of censoring, the observed information providing the quantity of interest. Some discussion of this is given by Efron and Hinkley (1978) and Barndorff-Nielsen and Cox (1994). We can now write down an estimate of the large sample variance which, interestingly, only depends on the number of observed failures. Thus, in order to correctly estimate the average, it is necessary to take into account the total time on study for both the failures and those observations that result in censoring. On the other hand,

given this estimate of the average, the precision we will associate with this only depends on the observed number of failures. This is an important observation and will be made again in the more general stochastic process framework.

MULTIVARIATE SETTING

In the majority of applications, the parameter θ will be a vector of dimension p . The notation becomes heavier but otherwise everything is pretty much the same. The estimating equation, $U(\theta) = 0$ then corresponds to a system of p estimating equations and $I(\theta)$ is a $p \times p$ symmetric matrix in which the (q,r) th element is given by $-\partial^2 \log L(\theta)/\partial \theta_q \partial \theta_r$ where θ_q and θ_r are elements of the vector θ . Also, the system of Newton-Raphson iteration can be applied to each one of the components of θ so that we base our calculations on solving the set of equations:

$$\hat{\theta}_{j+1,q} = \hat{\theta}_{j,q} + I(\hat{\theta}_j)^{-1} U(\hat{\theta}_j), \quad q = 1, \dots, p; \quad j \geq 1, \quad (3.5)$$

where, in this case, $\hat{\theta}_1$ is a vector of starting values to the iterative cycle, again most often zero.

Example 3.1. For the Freireich data (Cox, 1972), we calculate for the 6-MP group $\hat{\lambda} = 9/359 = 0.025$. For the placebo group we obtain $\hat{\lambda} = 21/182 = 0.115$. Furthermore in the 6-MP group we have $\text{Var}(\hat{\lambda}) = 9/(359)^2 = 0.000070$ whereas for the placebo group we have $\text{Var}(\hat{\lambda}) = 21/(182)^2 = 0.0006$.

The non-parametric empirical estimate (described below) agrees well with curves based on the exponential model and this is illustrated in Figure 3.1. Infer-

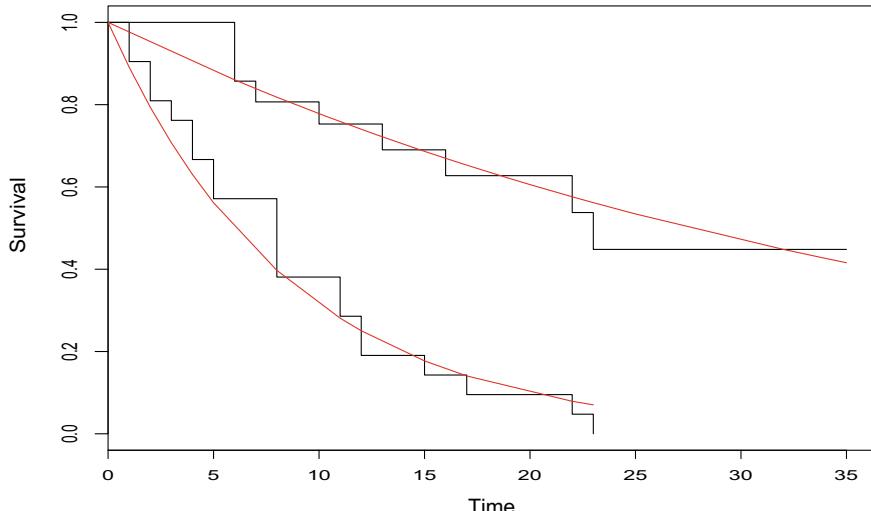


Figure 3.1: Exponential fitted curves to two-group Freireich data.

ence can be based on an appeal to the usual large sample theory. In this particular case, however, we can proceed in a direct way by recognizing that, for the case of no censoring $k = n$, the sample size and $\sum_{j=1}^n T_j$ is a sum of n independent random variables each exponential with parameter λ . We can therefore treat $n/\hat{\lambda}$ as a gamma variate with parameters (λ, n) . When there is censoring, in view of the consistency of $\hat{\lambda}$, we can take $k/\hat{\lambda}$ as a gamma variate with parameters (λ, k) , when $k < n$. This is not an exact result, since it hinges on a large sample approximation, but it may provide greater accuracy than the large sample normal approximation.

In order to be able to use standard tables we can multiply each term of the sum by 2λ since this then produces a sum of n exponential variates, each with variance 2. Such a distribution is a gamma $(2, n)$, equivalent to a chi-square distribution with $2n$ degrees of freedom Evans et al. (2001). Taking the range of values of $2k\lambda/\hat{\lambda}$ to be between $\chi_{\alpha/2}$ and $\chi_{1-\alpha/2}$ gives a $100(1-\alpha)\%$ confidence interval for λ . For the Freireich data we find 95% $CI = (0.0115, 0.0439)$. Once we have intervals for λ , we immediately have intervals with the same coverage properties for the survivorship function, this being a monotonic function of λ .

Denoting the upper and lower limits of the $100(1-\alpha)\%$ confidence interval $S_\alpha^+(t; \hat{\lambda})$ and $S_\alpha^-(t; \hat{\lambda})$ respectively, we have:

$$S_\alpha^+(t; \hat{\lambda}) = \exp \left\{ - \left(\frac{\hat{\lambda}\chi_{\alpha/2}}{2k} \right) t \right\}, \quad S_\alpha^-(t; \hat{\lambda}) = \exp \left\{ - \left(\frac{\hat{\lambda}\chi_{1-\alpha/2}}{2k} \right) t \right\}.$$

An alternative approximation, based on large sample normality, and also very simple in form, can be written as:

$$S_\alpha^+(t; \hat{\lambda}) \approx \exp \left\{ - \left(\frac{\hat{\lambda}}{\sqrt{k}} \right) (\sqrt{k} - z_{1-\alpha/2}) t \right\}, \quad (3.6)$$

where, this time, the corresponding expression for $S_\alpha^-(t; \hat{\lambda})$ obtains by replacing $z_{1-\alpha/2}$ by $z_{\alpha/2}$.

Piecewise exponential model

We can achieve considerably greater flexibility than the standard exponential model by constructing a partition of the time axis $0 = a_0 < a_1 < \dots < a_k = \infty$. Within the j th interval (a_{j-1}, a_j) , ($j = 1, \dots, k$) the hazard function is given by $\lambda(t) = \lambda_j$. Using Equations 2.12 and (3.1) and equating first derivatives to zero we find

$$\hat{\lambda}_j = k_j / \sum_{\ell: x_\ell > a_{j-1}} \{(x_\ell - a_{j-1}) I(x_\ell < a_j) + (a_j - a_{j-1}) I(x_\ell \geq a_j)\}.$$

Differentiating a second time we find $I(\lambda_1, \dots, \lambda_k)$ to be block diagonal where the (j,j) th element is given by $I_{jj} = k_j / \lambda_j^2$. In view of the orthogonality of the parameter estimates, we can construct a $100(1-\alpha)\%$ simultaneous confidence interval for $\lambda_1, \dots, \lambda_k$ by choosing a sequence of α_j such that $1 - \prod_j (1 - \alpha_j) = \alpha$.

An estimate of the survivorship function derives from $\hat{\Lambda}(t)$ in which we define $\hat{S}(t; \hat{\Lambda}) = \exp\{-\hat{\Lambda}(t)\}$ and where

$$\hat{\Lambda}(t) = \sum_{j: a_j \leq t} \hat{\lambda}_j (a_j - a_{j-1}) + \sum_{\ell} \hat{\lambda}_{\ell} (t - a_{\ell-1}) I(a_{\ell-1} \leq t < a_{\ell}). \quad (3.7)$$

Confidence intervals for this function can be based on Equation 3.3. We can view the simple exponential survival model as being at one extreme of the parametric spectrum, leaning as it does on a single parameter, the mean. It turns out that we can view the piecewise exponential model, with a division so fine that only single failures occur in any interval, as being at the other end of the parametric spectrum, i.e., a non-parametric estimate. Such an estimate corresponds to that obtained from the empirical distribution function. This is discussed below.

OTHER PARAMETRIC MODELS

The exponential and piecewise exponential models hold a special place in the survival literature for a number of reasons. The models are simple in form, simple to understand, have a clear physical property that we can interpret (lack of memory property) and the parameters can be estimated so easily that analysis based on such models clearly falls under the heading of desirable procedures as defined by Student; ones that can be calculated on the back of a railway ticket while awaiting the train. The piecewise model also allows quite considerable flexibility.

Given these facts, it is hard to justify the use of other parametric models unless motivated by some compelling physical argument. Cox (1958) used the Weibull model in analyzing the strengths of wool, but the model was not just pulled out of the air. Physical considerations and the fact that the Weibull distribution obtains as the limiting case of the minimum of a collection of random variables provide a powerful case in its favor. Another physical case might be the sum of exponential variates, each having the same parameter, since this can be seen to be a gamma variate. Generally, we may have no good reason to believe some model over most others is likely to provide a good fit for data. In these cases, given the generally good performance of empirical estimates, it may be preferable to base our analyses on these.

3.4 Empirical estimate (no censoring)

Empirical or non-parametric estimates, i.e., those making no model assumptions, can be most readily understood in the context of some finite division of the time axis. Theoretical results stem from continuous-time results, the transition from the discrete to the continuous, as a consequence of sample size n increasing without bound, presenting no conceptual difficulty. Unlike the discussion on the piecewise exponential model in which we most often anticipate having very few intervals, rarely more than three or four—the idea being to pool resources (estimating power) per interval while keeping within the bounds of serious model violation—for empirical estimates we do the opposite.

We imagine a fine division of the real line in which, given real data, the vast majority of the intervals are likely to contain no observations. Consider the time interval to be fixed, divided equally into k non-overlapping intervals $(a_{j-1}, a_j]$, $j = 1, \dots, k$, the notation “(” indicating that the interval is open on the left, and “]” closed on the right, i.e., a_{j-1} does not belong to the interval but a_j does. We have that $\Pr(T \geq a_j) = \Pr(T > a_j) = S(a_j)$ and that $\Pr(T > a_j | T > a_{j-1}) = S(a_j, a_{j-1})$. Recall from Section 2.3 that

$$S(a_j) = S(a_{j-1})S(a_j, a_{j-1}) = \prod_{\ell \leq j} S(a_\ell, a_{\ell-1}).$$

For each $t = a_\ell$, ($\ell > 0$), the empirical estimate of $S(t)$ based on a sample of size n , and denoted $S_n(t)$, is simply the observed number of observations that are greater than t . For a random sample of observations T_i , ($i = 1, \dots, n$) we use the indicator variable $I(\cdot)$ to describe whether or not the subject i survives beyond point t , i.e., for $t = a_j$,

$$S_n(a_j) = \frac{1}{n} \sum_{i=1}^n I(T_i > a_j) = \prod_{\ell=1}^j S_n(a_\ell, a_{\ell-1}) \quad (3.8)$$

in which the empirical $S_n(a_\ell, a_{\ell-1})$ is defined in an entirely analogous way to $S_n(a_\ell)$, i.e.,

$$S_n(a_\ell, a_{\ell-1}) = \frac{1}{n_{\ell-1}} \sum_{i=1}^n I(T_i > a_\ell), \quad n_\ell = \sum_{i=1}^n I(T_i \geq a_\ell). \quad (3.9)$$

It is readily seen, and instructive for understanding the Kaplan-Meier estimate of the next section, that, if no failure is observed in $(a_{\ell-1}, a_\ell]$ then $S_n(a_\ell, a_{\ell-1}) = 1$, whereas for an observed failure, $S_n(a_\ell, a_{\ell-1}) = (n_{\ell-1} - 1)/n_{\ell-1}$. The empirical distribution has been well studied and, in particular, we have:

Lemma 3.2. *For any fixed value of t , $S_n(t)$ is asymptotically normal with mean and variance given by:*

$$E\{S_n(t)\} = S(t); \quad \text{Var}\{S_n(t)\} = F(t)S(t)/n. \quad (3.10)$$

We should understand the operators E and Var to refer to expectations over a set of repetitions, the number of repetitions increasing without bound, and with n and t fixed. As an estimator of $S(t)$, the function $S_n(t)$ is then unbiased. The variance, as we might anticipate, is the variance of a binomial variate with parameters n and $S(t)$. These two moments, together with a normal approximation of DeMoivre-Laplace (see Appendix C.2), enable the calculation of approximate confidence intervals. The result, which is well known (Glivenko-Cantelli), enables us to carry out inference for $S(t)$ on the basis of $S_n(t)$ for any given point t . Very often we will be interested in the whole function $S_n(t)$, over all values of t and, in this case, it is more helpful to adopt the view of $S_n(t)$ as a stochastic process. In this regard we have

Theorem 3.1. $\sqrt{n}\{S_n(t) - S(t)\}$ is a Gaussian process with mean zero and covariance given by

$$\text{Cov}[\sqrt{n}\{S_n(s)\}, \sqrt{n}\{S_n(t)\}] = F(s)\{1 - F(t)\}. \quad (3.11)$$

An important consequence of the theorem arises when the distribution of T is uniform, for, in this case, $F(s) = s$ and all the conditions are met for the process $\sqrt{n}\{F_n(t) - t\}$ ($0 \leq t \leq 1$) to converge in distribution to the Brownian bridge.

Finally these results apply more generally than just to the uniform case for, as long as T has a continuous distribution, there exists a unique monotonic transformation from T to the uniform, such a transformation not impacting $\sqrt{n}\{F_n(t) - F(t)\}$ itself. In particular this enables us to use the result of Appendix B.2 to make inference for an arbitrary continuous cumulative distribution function, $F(t)$, whereby, for $W_n(t) = \sqrt{n}\{F_n(t) - F(t)\}$,

$$\Pr\left\{\sup_t |W_n(t)| \leq D\right\} \rightarrow 1 - 2 \sum_{k=1}^{\infty} (-1)^{k+1} \exp(-2k^2 D^2), \quad D \geq 0. \quad (3.12)$$

Most often we are interested in events occurring with small probability in which case a good approximation obtains by only taking the first term of the sum, i.e., $k = 1$. Under this approximation $|\sqrt{n}\{F_n(t) - F(t)\}|$ will be greater than about 1.4 less than 5% of the time. This is a simple and effective working rule.

3.5 Kaplan-Meier (empirical estimate with censoring)

The impact of censoring is to reduce the number of observations T_1, \dots, T_n so that some, a number strictly less than n , are actually observed. The empirical estimate of the standard distribution function is no longer available to us so that

$$\hat{F}_{\text{emp}}(t) = \frac{1}{n} \sum_{i=1}^n I(T_i \leq t), \quad 0 \leq t \leq \mathcal{T},$$

cannot be calculated. Simply taking the observed censored times as though they were failure times will clearly not work so that the estimator;

$$1 - \frac{1}{n} \sum_{i=1}^n Y_i(t) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{X_i \leq t}, \quad 0 \leq t \leq \mathcal{T}$$

will exhibit increasing bias as the amount of censoring increases. This follows since we will be estimating $P(T \leq t, C \leq t)$ and this underestimates $P(T \leq t)$. A little more work is needed although it is only really a question of treating the various needed ingredients in a sensible way for things to work. The most famous empirical estimate is that of Kaplan and Meier (1958) which we can show to be consistent under an independent censoring mechanism.

Definition 3.1. *The Kaplan-Meier estimator of F , denoted \hat{F} , is written:*

$$\hat{F}(t) = 1 - \prod_{\substack{i=1, \dots, n \\ X_i \leq t}} \left(1 - \frac{\delta_i}{\sum_{j=1}^n Y_j(X_i)} \right), \quad 0 \leq t \leq \mathcal{T}.$$

This function is piecewise continuous. At the jump points of discontinuity, for elementary theoretical reasons, it is helpful to define it as being right continuous with a left-hand limit (CADLAG, which is a French translation of that definition). When there is no censoring, for a sample of size n , these jumps all have the same size $1/n$. In the presence of censoring these jumps will become larger as we move through time. In some sense, the observations lost to censoring, redistribute a part of their probability mass, over the observations strictly greater than themselves. This idea can be made precise and we do so below. The Kaplan-Meier estimate is more general than the usual empirical estimate F_n in that, in the absence of censoring, the former will reduce to the latter. Like $F_n(t)$, the Kaplan-Meier estimator is consistent for $F(t)$ although this requires the additional condition of independence between the censoring and failure mechanisms. Also, we write $\hat{S}(t) = 1 - \hat{F}(t)$. Calculation is simple, and like the usual empirical estimate, can be done manually. To see this, consider the 7 observations in Table 3.1. The 3 observations, 7, 13, and 22 are censored times. The other observations are failure times. When $t \in [X_i, X_{i+1}[$ then $\hat{S}(t) = \hat{S}(X_i)$. The censored observations play a

X	2	4	7	9	13	16	22
δ	1	1	0	1	0	1	0

Table 3.1: Taken from Chauvel (2014) as an illustration in Kaplan-Meier calculation

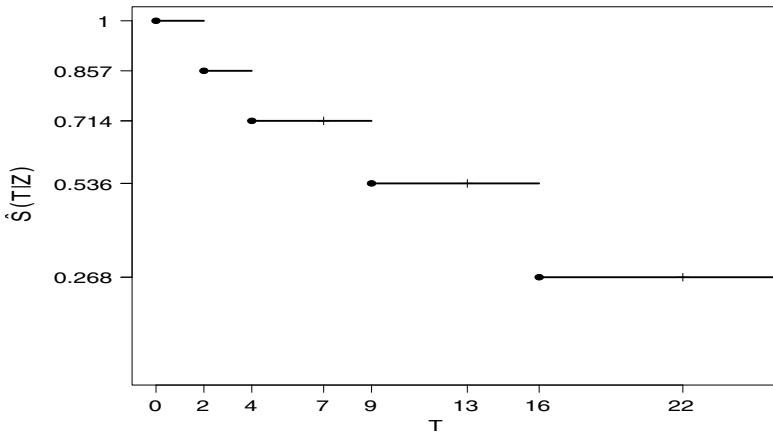


Figure 3.2: Estimated Kaplan–Meier curve

X_i	$\sum_{j=1}^n Y_j(X_i)$	$1 - \delta_i / \sum_{j=1}^n Y_j(X_i)$	$\hat{S}(X_i)$
2	7	$1-1/7$	$1-1/7 \simeq 0.857$
4	6	$1-1/6$	$(1-1/7)(1-1/6) \simeq 0.714$
7	5	1	0.714
9	4	$1-1/4$	$(1-1/7)(1-1/6)(1-1/4) \simeq 0.536$
13	3	1	0.536
16	2	$1-1/2$	$(1-1/7)(1-1/6)(1-1/4)(1-1/2) \simeq 0.268$
22	1	1	0.268

Table 3.2: Chauvel (2014) illustrates the step-by-step Kaplan-Meier calculations

role for values lower than themselves. The Kaplan-Meier curve does not change at the censored observation time and, beyond this, the role of the censored observation is indirect. It does not influence the actual calculation. Figure 3.2 shows the estimated Kaplan–Meier curve, $\hat{S}(t)$. Unlike the empirical function, $1 - F_n(t)$, the curve does not reach zero in this example and this will be the case whenever the last observation is a censored observation. The jumps in the Kaplan-Meier curve play an important role in proportional and non-proportional hazards modeling. The size of these jumps at time t , written $d\hat{S}(t)$, are readily described and we have (Table 3.2):

Proposition 3.1. For an observed event time and where $\hat{S}(t^-) = \lim_{s \rightarrow t^-} \hat{S}(s)$, we can write,

$$d\hat{S}(t) = \hat{S}(t) - \hat{S}(t^-) = \frac{\hat{S}(t^-)}{\sum_{j=1}^n Y_j(t)}, \quad (3.13)$$

In later chapters we will see how to make full use of these increments. They can be viewed as weights or as approximations to infinitesimal contributions to an integral with respect to the survival function. These are sometimes referred to as Kaplan-Meier integrals and, in the light of Helly-Bray's theorem (Appendix A), allow us to consistently estimate functions of T of interest in the presence of censoring. Details can be found in Stute (1995) where, in particular, under general conditions, we can obtain the asymptotic normality of these integrals. Another important example in which we use standard software to fit a proportional hazards model to data generated under a non-proportional hazards mechanism has been studied by Xu and O'Quigley (2000) and O'Quigley (2008). The use of the Kaplan-Meier increments enables us to consistently estimate an average regression effect $E\beta(T)$ when $\beta(t)$ is not a constant. Failing to correctly incorporate the Kaplan-Meier increments into the estimating equation—a common oversight encouraged by the availability of a lot of standard software—will lead to serious bias in the estimation of $E\beta(T)$.

APPROACH USING FINITE CENSORING SUPPORT

First, recall the model for finite censoring support, described in Section 2.5. Such a model adds no additional practical restriction but helps us to avoid much mathematical complexity, leading to a simple and useful theorem (Theorem 3.2). The theorem enables us to circumvent the difficulty arising from the fact that we are unable to obtain $\sum_{i=1}^n I(T_i \geq a_j)$ as in Equation 3.8. This is because at a_j we cannot ascertain whether or not earlier observed values of $X_i (\delta_i = 0)$, arising from the pair (T_i, C_i) in which $C_i = X_i$, are such that $T_i > a_j$. Since $X_i = \min(T_i, C_i)$ we do not observe the actual value of T_i when $X_i = C_i$. The trick is to notice in the right-hand side of (3.8) that we are able to obtain the empirical estimates, $G_n(a_\ell, a_{\ell-1})$ for $G(a_\ell, a_{\ell-1}) = \Pr(X \geq a_\ell | X > a_{\ell-1})$. But, unlike Equation 3.8

$$G_n(a_j) = \frac{1}{n} \sum_{i=1}^n I(X_i \geq a_j) \neq \prod_{\ell \leq j} G_n(a_\ell, a_{\ell-1}). \quad (3.14)$$

This is because of the non-zero masses being associated to the times at which the censorings occur. Nonetheless, the rest follows through readily, although, unlike (3.9), we now define $n_\ell = \sum_{i=1}^n I(X_i \geq a_\ell)$, noting that this definition contains (3.9) as a special case when there is no censoring.

KAPLAN-MEIER ESTIMATE BASED ON FINITE CENSORING SUPPORT

The distribution of $X = \min(T, C)$ is mostly of indirect interest but turns out to be important in view of the following theorem and corollary.

Theorem 3.2. *Under an independent censoring mechanism,*

$$G(a_\ell, a_{\ell-1}) = S(a_\ell, a_{\ell-1}). \quad (3.15)$$

Corollary 3.1. *Under an independent censoring mechanism, a consistent estimator of $S(t)$ at $t = a_j$ is given by $\hat{S}(t) = \prod_{\ell \leq j} G_n(a_\ell, a_{\ell-1})$.*

Corollary 3.2. *In the absence of censoring $\hat{S}(t) = S_n(t)$.*

The estimator of the corollary is known as the Kaplan-Meier estimator. In the light of our particular setup, notably the use of the independent censoring mechanism having finite support (Section 2.5), it can be argued that our estimate is only available at distinct values of $t = a_j$. But this is not a practical restriction since our finite division can be as fine as we wish, the restriction amounting to limiting accuracy to some given number of decimal places. One way or another it is not possible to completely avoid some mathematical fine points, our preference being to use the mechanism of Section 2.5.

In practice we talk about the Kaplan-Meier estimate at time point t , the underlying support of C and T rarely being a central concern. The one exception to this arises when the range of C is less than that for T . A single sampled T_i which is greater than the greatest value that can be taken by C_i will necessarily appear as a censored observation. In this case the observed empirical Kaplan-Meier estimate is never able to reach zero. The true distribution $F(t)$ cannot be estimated at values greater than the upper limit for the support of C . This is intuitively clear. Again, in practice, this ought not to be a real concern. If we have no information on something (in this case a certain upper region of the survival curve), it is most likely too optimistic to hope to be able to carry out useful estimation there, unless, of course, we make additional parametric assumptions, which amount to saying that information obtained over some range can tell us a lot about what is taking place elsewhere, and certainly more than just the provision of bounds on $F(t)$ for t greater than the upper limit of C .

REMARKS ON THE KAPLAN-MEIER CURVE

The Kaplan-Meier (1958) product-limit estimator provides a non-parametric estimate for the survival function $S(t)$ under the independent censorship assumption. It is a generalization of the commonly known empirical distribution function to the case of censoring since, in the absence of censoring, the two coincide. Expressing the estimate $\hat{S}(t)$ in terms of the fine division $(a_\ell, a_{\ell-1})$, $\ell = 1, \dots, N$, is conceptually useful. However, since intervals not containing events produce no change in $\hat{S}(t)$ it is, for the purposes of practical calculation, only necessary to consider

evaluation at the distinct observed failure times $t_1 < t_2 < \dots < t_k$. All divisions of the time interval $(a_\ell, a_{\ell-1})$, $\ell = 1, \dots, N$, for different N , lead to the same estimate $\hat{S}(t)$, provided that the set of observed failure points is contained within the set $\{a_\ell; \ell = 1, \dots, N\}$. A minimal division of the time axis arises by taking the set $\{a_\ell\}$ to be the same as the set of the distinct observed failure times. So, for practical purposes $a_j = t_j, j = 1, \dots, k$, and the Kaplan-Meier estimate can be defined as

$$\hat{S}(t) = \prod_{j:t_j < t} \frac{n_j - d_j}{n_j} = \prod_{j:t_j < t} 1 - \frac{d_j}{n_j}. \quad (3.16)$$

where $d_j = \sum \delta_j$, the sum being over the ties at time t_j . If there are no ties at t_j then $d_j = \delta_j$. Note that $\hat{S}(t)$ is a left-continuous step function that equals 1 at $t = 0$ and drops immediately after each failure time t_j . The estimate does not change at censoring times. When a censoring time and a failure time t_j are recorded as equal, the convention is that censoring times are adjusted an infinitesimal amount to the right so that the censoring time is considered to be infinitesimally larger than t_j . Any subjects censored at time t_j are therefore included in the risk set of size n_j , as are those that fail at t_j . This convention is sensible because a subject censored at time t_j almost certainly survives beyond t_j . Note also that when the last observation is a censoring time rather than a failure time, the KM estimate is taken as being defined only up to this last observation.

CONTINUOUS VERSION OF KAPLAN-MEIER CURVE

The Kaplan-Meier curve is so useful and so commonly employed that we reserve the most standard notation for this estimate, i.e., $\hat{S}(t)$. For the majority of applications this is all we need. However, there are some situations in which it is useful to be able to invert the function in order to estimate $S^{-1}(t)$. Now, the usual Kaplan-Meier estimate takes a constant value between adjacent failures, t_{j-1} and t_j and so, in general, we are unable to invert this function in a unique way. This is not a serious concern and any intuitive solution such as taking $\hat{S}(t)$ to be $S(t_{j-1})$ or $S(t_j)$ for all $t \in (t_{j-1}, t_j)$ would be fine. Rather than this we suggest a simple linear interpolation. We then define the continuous (and invertible since strictly decreasing) function $\bar{S}(t)$ to be given by

$$\bar{S}(t) = \hat{S}(t_{j-1}) + \left(\frac{t - t_{j-1}}{t_j - t_{j-1}} \right) \{ \hat{S}(t_j) - \hat{S}(t_{j-1}) \}; \quad t \in (t_{j-1}, t_j). \quad (3.17)$$

Note that at the distinct failure times t_j the two estimates, $\hat{S}(t)$ and $\bar{S}(t)$ coincide. An example of where we make an appeal to $\bar{S}(t)$ is illustrated in the two-

group exponential model where a transformation to exponentiality for one group is then applied to the other group.

PRECISION OF KAPLAN-MEIER ESTIMATE

Keeping in mind the results of Appendices A.2 and A.10 we can derive a Taylor series approximation to a function of random variables wherever the function of expectation converges in probability to the expectation of the function. An immediate application leads to the following theorem.

Theorem 3.3. *For each $t = a_\ell$, the estimate $\hat{S}(t)$ is asymptotically normal with mean $S(t)$ and variance:*

$$\text{Var } S(a_\ell) \approx S(a_\ell)^2 \sum_{m \leq \ell} \sum_{m \leq \ell} \frac{d_m}{n_m(n_m - d_m)}. \quad (3.18)$$

The above expression for the variance of $\hat{S}(t)$ is known as Greenwood's formula. Breslow and Crowley (1974) in a detailed large sample study of the Kaplan-Meier estimator obtained a result asymptotically equivalent to the Greenwood formula, making a slight correction to overestimation of the variation in the estimated survival probability. The formula's simplicity, however, made it the most commonly used when computing the variance of the Kaplan-Meier estimate and related quantities. We also have:

Corollary 3.3. *When there is no censoring, the approximation $\text{Var } \hat{S}(t)$ from Theorem 3.3 reduces to the usual binomial variance estimate $\hat{S}(t)\{1 - \hat{S}(t)\}/n$.*

The usual use to which we put such variance estimates is in obtaining approximate confidence intervals. Thus, using the large sample normality of $\hat{S}(t)$, adding and subtracting to this $z_{1-\alpha/2}$ (the $1 - \alpha/2$ quantile from the standard normal distribution) multiplied by the square root of the variance estimate, provides approximate $100(1 - \alpha)\%$ confidence intervals for $\hat{S}(t)$. As mentioned before the constraints on $\hat{S}(t)$, lying between 0 and 1, will impact the operating characteristics of such intervals, in particular, it may not be realistic, unless sample sizes are large, to limit attention to symmetric intervals around $\hat{S}(t)$. Borgan and Liestol (1990) investigate some potential transformations, especially the log-minus-log transformation discussed in Section 2.4, leading to

Corollary 3.4. *Let $w(\alpha) = \text{Var}^{1/2} \hat{S}(t) z_{1-\alpha/2} / \hat{S}(t) \log \hat{S}(t)$. For each $t = a_\ell$, a $100(1 - \alpha)\%$ confidence intervals for $\hat{S}(t)$ can be approximated by*

$$\{\hat{S}(t)^{\exp -w(\alpha)}, \hat{S}(t)^{\exp w(\alpha)}\} \quad (3.19)$$

The same arguments which led to Greenwood's formula also lead to approximate variance expressions for alternative transformations of the survivorship function. In particular we have

Corollary 3.5. *For each $t = a_\ell$ the estimate $\log \hat{S}(t)$ is asymptotically normal with asymptotic mean $\log S(t)$ and variance*

$$\text{Var} \log S(a_\ell) \approx \sum_{m \leq \ell} \sum_{m \leq \ell} \frac{d_m}{n_m(n_m - d_m)} \quad (3.20)$$

Corollary 3.6. *For each $t = a_\ell$ the estimate $\log \hat{S}(t)/\{1 - \hat{S}(t)\}$ is asymptotically normal with asymptotic mean $\log S(t)/\{1 - S(t)\}$ and variance*

$$\text{Var} \log \left\{ \frac{S(a_\ell)}{1 - S(a_\ell)} \right\} \approx \{1 - S(a_\ell)\}^{-2} \sum_{m \leq \ell} \sum_{m \leq \ell} \frac{d_m}{n_m(n_m - d_m)}. \quad (3.21)$$

Confidence intervals calculated using any of the above results will be of help in practice. Following some point estimate, obtained from $\hat{S}(t)$ at some given t , these intervals are useful enough to quantify the statistical precision that we wish to associate with the estimate. All of the variance estimates involve a comparable degree of complexity of calculation so that choice is to some extent a question of taste. Nonetheless, intervals based on the log-minus-log or the logit transformation will behave better for smaller samples, and guarantee that the endpoints of the intervals themselves stay within the interval (0,1). This is not so for the Greenwood formula, the main argument in its favor being that it has been around the longest and is the most well known. For moderate to large sample sizes, and for $\hat{S}(t)$ not too close to 0 or 1, all the intervals will, for practical purposes, coincide.

OBSERVED DIFFERENCES BETWEEN KAPLAN-MEIER CURVES

A common and fundamental goal in survival studies is to make a statistical decision as to whether the observed differences between two Kaplan-Meier curves can be attributed to sampling differences from the same common population or whether the observed differences are indicative of two distinct populations. We consider formal testing in later chapters and here we make a few general observations. The most common test used here is the so-called log-rank test. It can be shown to be the most powerful test against local departures that can be described as having a proportional hazards nature (Peto and Peto, 1972), in other words the log-log transformations of the two survival curves is a constant that does not change with time. If the hazard ratio is not constant over time, the log-rank test can suffer significant power losses (Leurgans, 1983, 1984), so much so that the test can fail to detect differences that are clearly indicative of real effects in many cases (Figure 3.3).

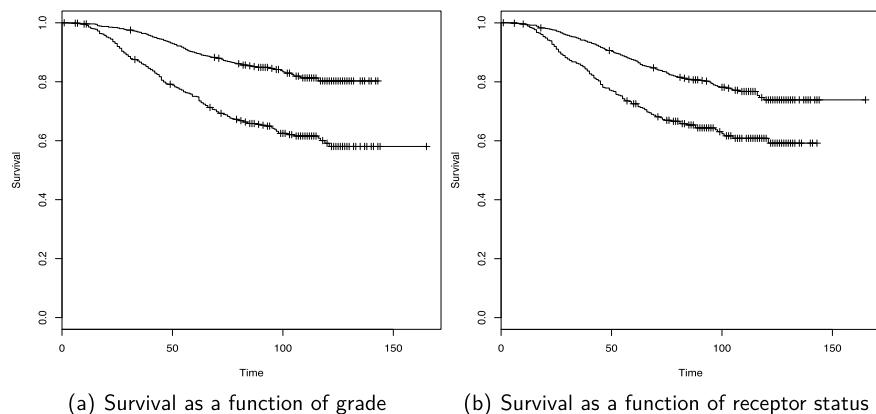


Figure 3.3: Kaplan-Meier curves obtained from the Curie breast cancer study. A proportional hazards assumption suggests itself as a possibility to model the observed differences. A good candidate test would be the log-rank test.

The right-hand figure in Figure 3.4 represents Kaplan-Meier estimators of the survival of 2 groups of patients with breast cancer followed at the Institut Curie for a period of 12 years. For the first group, the diameter of the tumor of the patients is less than 60mm and for the other group the diameter is greater than 60mm. There were a total of 339 patients in the study. The differential survival experience of the two groups is quite clear from the figure. At the same time, a rough and ready glance suggests that these differences are not likely to be well described by a proportional hazards model. During the first 60 months, the advantage seen by one group is clear and the curves move further and further apart. The effect appears to be a strong one. However, beyond this period, the slopes in the curves are much more similar and suggest a very considerable attenuation of the initial effect. Some 80% of patients with tumor size less than 60mm survive at least 75 months. This duration drops all the way down to 28 months for patients with a tumor size greater than 60mm. There would seem little doubt as to the prognostic impact of tumor size and yet the classical log-rank test fails to achieve significance in this example. Of course this could always be due to the fact that we cannot rule out with great enough conviction the possibility of having observed a large but non-significant difference. Although, it seems more likely that there is a real, and clinically important difference here, but that the lack of proportionality of risks clouds the issue and puts a spoke in the wheel of the usual log-rank test. A simple clinical explanation might be that tumor size is of considerable importance in the short term but that, after having managed to survive beyond some point, this importance is much diminished. A more likely clinical explanation is that tumor size maintains its importance but that it is measured with some degree of error. A number of patients would have

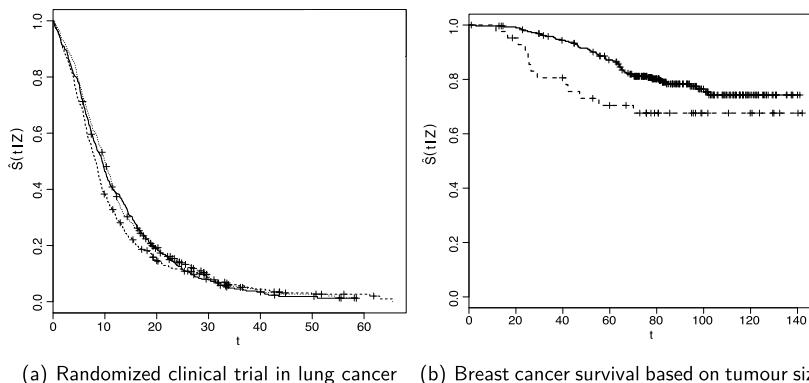


Figure 3.4: Kaplan-Meier curves obtained from a randomized clinical trial and a study in breast cancer. In both cases, a proportional hazards assumption is doubtful. The log-rank test would show poor performance in such situations.

been misclassified and, as we will see in later theoretical work, the consequences of this would be to produce an impression of a diminishing regression effect.

Figure 3.4 also illustrates two Kaplan-Meier curves taken from a randomized clinical trial in lung cancer. There are three groups and the short-term effects and long term effects appear to indicate no real treatment effect. Initially, the curves more or less coincide, as they do in the long term. However, for a significant part of the study, say between the median and the 10th percentile, there appears to be a real advantage in favor of the two active treatment groups. The log-rank test fails to detect this and more suitable tests, described in later chapters, are able to confirm the treatment differences. The clear lack of proportionality of hazards is at the root of the problem here.

Striving to obtain tests that to a greater or lesser extent can reverse the power deficit arising as a result of non-proportional hazards has stimulated the work of many authors in this field. Weighting the linear contributions to the log-rank test, leading to the so-called weighted log-rank tests, has a long history (Fleming and Harrington, 1991; Gehan, 1965; Peto and Peto, 1972). Prentice (1978) and others showed that a weighting that mirrors the actual form of the departure from proportional hazards will lead to powerful alternative tests. What is trickier, but arguably much more relevant, is to obtain tests that will have good performance in both situations, that will be close to optimal under proportional hazards, if not quite optimal, but that will retain good power in situations where the alternative is significantly remote from that of proportional hazards. We consider this in later chapters.

REDISTRIBUTION TO THE RIGHT ALGORITHM

Another way to look at the KM estimate, which turns out to be conceptually useful and of value to later developments, is to focus on the increments, or step

size, of the function at points where it changes. If we denote t_j+ the time instant immediately after t_j , then $\hat{S}(t_j) - \hat{S}(t_j+)$ is the stepsize, or jump, of the KM curve at time t_j . From Equation 3.16 we see that

$$\hat{S}(t_j+) = \hat{S}(t_j) \cdot \frac{n_j - d_j}{n_j},$$

so the stepsize is $\hat{S}(t_j) \cdot d_j/n_j$. That is to say, when the total “leftover” probability mass is $\hat{S}(t_j)$, each observed failure gets one- n_j th of it, where $n_j = \sum_{i=1}^n Y_i(t_j)$ is the number of subjects at risk at time t_j . In the absence of censoring this corresponds exactly to the way in which the empirical estimate behaves. When there is censoring, then one way of looking at a censored observation is to consider that the mass that would have been associated with it is simply reallocated to all of those observations still remaining in the risk set (hence the term “redistribution to the right.”)

KAPLAN-MEIER ESTIMATES OF MEDIAN AND MEAN SURVIVAL

Since the support of T is only on the positive real line the distributions we deal with are asymmetrical. In consequence it is more common to take as summary measures simple functions of different quantiles, most often the median or interquartile range, rather than the mean and variance. Nonetheless, the mean is of interest and in special cases, such as the exponential distribution, relates directly to the median via a constant scaling factor. Also, of course, there is no compelling reason not to work with symmetric distributions defined for say $\log T$ and then transform back to T , although this is not very commonly done. From Section 2.3 we can estimate the expected lifetime over some given interval $[0, t]$ as

$$\hat{\mu}(t) = \int_0^t \hat{S}(u) du, \quad (3.22)$$

the mean itself being then estimated by $\hat{\mu}(\infty)$. However, the theory comes a little unstuck here since, not only must we restrict the time scale to be within the range determined by the largest observation, the empirical distribution itself will not correspond to a probability distribution whenever $\hat{F}(t) = 1 - \hat{S}(t)$ fails to reach one. In practice then it makes more sense to consider mean life time $\hat{\mu}(t)$ over intervals $[0, t]$, acknowledging that t needs to be kept within the range of our observations. The following result provides the required inference for $\hat{\mu}(t)$;

Lemma 3.3. *For large samples, $\hat{\mu}(t)$ can be approximated by a normal distribution with $E\hat{\mu}(t) = \mu(t)$ and*

$$\text{Var } \hat{\mu}(t) \approx \sum_{m \leq \ell} \sum_{m \leq \ell} \frac{\{\hat{\mu}(t) - \hat{\mu}(a_m)\}^2 d_m}{n_m(n_m - d_m)}. \quad (3.23)$$

In view of the consistency of the Kaplan-Meier estimate it is, in principle, straightforward to obtain estimates for any desired quantile, or function of the quantiles, such as the median or interquartile range. However, again, we can be limited by the observations and, if $\hat{S}(t)$, for the largest observed survival time, is not less than p then we are not able to obtain point estimates of quantiles corresponding to such values, although we could obtain interval estimates. We define the p th quantile ξ_p to satisfy $F(\xi_p) = p$. The Kaplan-Meier function, or indeed the uncensored empirical distribution function itself, being a step function, is not invertible. To overcome this, we define the estimate $\hat{\xi}_p$ to be the smallest value of t such that $\hat{F}(\hat{\xi}_p) \geq p$. In order to make inferences for $\hat{\xi}_p$ we can appeal to some basic results from Appendix A.9 to obtain:

Lemma 3.4. *For large samples, $\hat{\xi}_p$ approaches a normal distribution with $E\hat{\xi}_p = \xi_p$ and*

$$\text{Var}\hat{\xi}_p \approx \sum_{m \leq \ell} \sum_{m \leq \ell} \frac{d_m(1-p^2)f^{-2}(\hat{\xi}_p)}{n_m(n_m - d_m)}. \quad (3.24)$$

It is difficult to use the above result in practice in view of the presence of the density $f(\cdot)$ in the expression. Smoothing techniques and the methods of density estimation can be used to make progress here but our recommendation would be to use a more direct, albeit more heavy, approach constructing intervals based on sequences of hypothesis tests. In principle at least the programming of these is straightforward.

3.6 Nelson-Aalen estimate of survival

An alternative approach to estimating the empirical survival distribution, adapted to accommodate censoring, and, essentially, equivalent to the Kaplan-Meier estimate arises from considerations of the basic formulae in Section 2.3. Recalling that $S(t) = \exp\{-\Lambda(t)\}$ and that $\Lambda(t) = \int_0^t \lambda(u)du$ we can consider empirical estimates of $\Lambda(t)$. The integral can be approximated by a Riemann sum so that for $t = a_j$

$$\sum_{\ell \leq j} \lambda(a_\ell)(a_\ell - a_{\ell-1}) \rightarrow \int_0^{a_j} \lambda(u)du \quad (3.25)$$

as $a_\ell - a_{\ell-1}$ goes to zero. Now, applying a local linearization, we obtain

$$\lambda(a_\ell)(a_\ell - a_{\ell-1}) \approx P(a_\ell < T < a_{\ell-1} | T > a_{\ell-1}) = 1 - S(a_\ell, a_{\ell-1}).$$

Applying Theorem 3.2 and then, first replacing $S(a_\ell, a_{\ell-1})$ by $G(a_\ell, a_{\ell-1})$, second replacing $G(a_\ell, a_{\ell-1})$ by $G_n(a_\ell, a_{\ell-1})$ i.e., d_ℓ/n_ℓ , we obtain, at $t = a_j$, $\tilde{\Lambda}(a_j) = \sum_{\ell=1}^j d_\ell/n_\ell$ as a consistent estimator for $\Lambda(t)$. The resulting estimator

$$\tilde{S}(t) = \exp\{-\tilde{\Lambda}(t)\} \quad (3.26)$$

is called the Nelson-Aalen estimate of survival. Recalling the Taylor series expansion, $\exp(x) = 1 + x + x^2/2! + \dots$, for small values of x we have $\exp(-x) = 1 - x + O(x^2)$, the error of the approximation being strictly less than $x^2/2$ since the series is convergent with alternating sign. Applying this approximation to $\tilde{S}(t)$ we recover the Kaplan-Meier estimate described above. In fact we can use this idea to obtain:

Lemma 3.5. *Under the Breslow-Crowley conditions, $|\tilde{S}(t) - \hat{S}(t)|$ converges almost surely to zero.*

In view of the lemma, large sample results for the Nelson-Aalen estimate can be deduced from those already obtained for the Kaplan-Meier estimate. This is the main reason that there is relatively little study of the Nelson-Aalen estimate in its own right. We can exploit the wealth of results for the Kaplan-Meier estimate that are already available to us. Indeed, in most practical finite sample applications, the level of agreement is also very high and the use of one estimator rather than the other is really more a question of taste than any theoretical advantage. In some ways the Nelson-Aalen estimate appears very natural in the survival setting, and it would be nice to see it used more in practice.

3.7 Model verification using empirical estimate

A natural approach to model assessment, i.e., whether or not some parametric model appears as a reasonable choice for the observed data, is to contrast the empirical estimates to those leaning on the model assumptions, the role of the data being reduced to that of providing estimates for any unknown parameters. We do not propose tackling the broad issues of goodness of fit until later but notice that, if the assumed parametric form is reasonable, then the model-based estimates and the empirical estimates ought to broadly agree. For the Weibull model, for example, we know that $S(t) = \exp\{-(\lambda t)^p\}$. Therefore, a plot, from the Kaplan-Meier, or Nelson-Aalen estimate, of $\log\{-\log \hat{S}(t)\}$ against $\log t$ should be linear with an intercept equal to $p \log \lambda$ and slope equal to p . For all the other parametric models it is usually possible to derive similar constructions. A visual impression of the adequacy of any postulated model is obtained, alongside the possibility of obtaining simple parameter estimates for the unknown parameters. Such estimates are typically less efficient than maximum likelihood estimates so, in a more thorough analysis, we may wish to use them either as a rough guide or as a first approximation in some iterative scheme. In practice it is often the case that quite different parametric assumptions, unless particularly restrictive like that for the exponential model, will produce very similar survival curves. Important differences between competing parametrizations tend

to manifest themselves mostly in the tails of the distribution where there may be few observations. As a goodness of fit tool these procedures are not usually very powerful.

EMPIRICAL ESTIMATES AND EXPONENTIAL ANALYSIS

Referring to Appendix A.6 and Theorem A.8, we have the important result that, for any continuous positive random variable T , with distribution function $F(t)$, the variate $\Lambda(T) = \int_0^T f(u)/[1-F(u)]du$ has a standard exponential distribution. As a consequence, if we consider the empirical survivorship function, $\hat{S}(t)$, then we can take the observations $-\log \hat{S}(X_i)$ as arising from a standard exponential distribution. All of the simple results that are available to us when data are generated by an exponential distribution can be used. In particular, if we wish to compare the means of two distributions, both subject to censoring, then we can transform one of them to standard exponential via its empirical survival function, then use this same transformation on the other group. The simple results for contrasting two censored exponential samples can then be applied even though, at least initially, the data arose from samples generated by some other mechanism.

3.8 Classwork and homework

1. Write down the estimating equations for a Weibull model based on maximum likelihood. Write down the estimating equations for a Weibull model based on the mean and variance.
2. Consider the following observations; 1, 3, 4, 4, 5, 6*, 6, 7, 9*, 16 where a * indicates a censored observation. Fit a Weibull model to these observations based on (i) maximum likelihood, (ii) method of moments.
3. For the data of the previous question, calculate and plot the survivorship function. Calculate an approximate 90% confidence interval for $S(4)$. Do the same for $S(7)$.
4. Describe how you might calculate a simultaneous 90% confidence interval for $S(4)$ and $S(7)$ together.
5. Compare the variance expression for $S(7)$ with that approximated by the binomial formula based on $\hat{S}(7)$ and 8 failure times.
6. Take 100 bootstrap samples, fit the Weibull model to each one separately and estimate $S(4)$ and $S(7)$. Calculate empirical variances based on the 100 sample estimates. How do these compare with those calculated on the basis of large sample theory.

7. In the previous question, rather than fit the model, we could calculate empirical estimates of $S(4)$ and $S(7)$. Describe possible difficulties with this approach.
8. For small samples generated via an exponential distribution with unknown mean, discuss the relative merits of the different possible confidence interval approximations for the survival function. Describe a study you might set up in order to make a recommendation regarding the “best” confidence interval to work with. How do you understand “best” in this context.
9. Simulate 100 uncensored observations from a standard exponential distribution. Calculate the empirical survival function $S_n(t)$. Choose two points, s and t and repeat the whole process 100 times obtaining 100 pairs of $S_{100}(s)$ and $S_{100}(t)$. Calculate the empirical covariance between $S_{100}(s)$ and $S_{100}(t)$. Use different values of s and t to suggest the validity of Theorem 3.1.
10. Explain why the result of Equation 3.12 for $W_n(t) = \sqrt{n}\{F_n(t) - F(t)\}$, when $F(t)$ is uniform continues to hold for any other continuous distribution.
11. Explain the importance of Theorem 3.2 and how it is used in order to obtain consistent estimates of survival in the presence of an independent censoring mechanism.
12. Simulate 200 uncensored observations from the log-normal distribution. Suppose that we had been led to believe that the observations had been generated from a Weibull law. Carry out graphical procedures to challenge the validity of the assumption. Repeat the exercise under the supposition that the observations had been generated via a log-logistic model.
13. For the 200 observations of the previous question, introduce an independent censoring mechanism so that approximately half of the observations are censored. Calculate the logarithm of the Kaplan-Meier and Nelson-Aalen estimates and plot one against the other. Fit a least squares line to the plot and comment on the values of the slope.
14. Use the results of Lemma 3.3 to show that $\hat{\mu}(t)$ is consistent for $\mu(t)$.
15. Show that when there is no censoring, the Greenwood estimate of the variance of the Kaplan-Meier estimate reduces to the usual variance estimate for the empirical distribution function. Conclude from this that confidence intervals based on the Greenwood estimate of variance are only valid at a single given time point, t , and would not provide bounds for the whole Kaplan-Meier curve.

16. Carry out a study on the coverage properties based on $\hat{S}(t)$, $\log \hat{S}(t)$ and $\log \hat{S}(t)/\{1 - \hat{S}(t)\}$. Describe what you anticipate to be the relative merits of the different functions.
17. Malani (1995) outlined an operationally simple approach for estimating survival in the presence of a dependent censoring mechanism. The method requires that the dependency be captured via some explanatory variable. Appealing to the redistribution to the right algorithm, each censored observation has its remaining mass redistributed. However, unlike the simple version of the algorithm, Malani proposes to only redistribute among subjects in the risk set sharing the same covariate value as the subject censored at that point. Give an intuitive explanation as to why this would work.
18. Following the idea of Malani, suppose, in the presence of dependent censoring, we obtained a Nelson-Aalen estimate of survival for each level of the covariate. Subsequently, appealing to the law of total probability, we estimate marginal survival by a linear combination of these several estimates. Comment on such an estimate and contrast it with that of Malani.
19. Recall that the uncensored Kaplan-Meier estimator, i.e., the usual empirical estimate, is unbiased. This is no longer generally so for the Kaplan-Meier estimate. Can you construct a situation in which the estimate of Equation 3.17 would exhibit less bias than the Kaplan-Meier estimate?
20. Using data from a cancer registry, show how you could make use of the piecewise exponential model to obtain conditional survival estimates for $S(T > t + s | T > s)$.

3.9 Outline of proofs

Lemma 3.1 and Theorem 3.2: There are two possibilities; the observation x_i corresponds to a failure, the observation corresponds to a censoring time. For the first possibility let dx_i be an infinitesimally small interval around this point. The probability that we can associate with this event is $\Pr(T \in dx_i, C > x_i) = \Pr(T \in dx_i) \times \Pr(C > x_i)$ i.e. $f(x_i; \theta)dx_i \times G(x_i; \theta)$. For a censored observation at time x_i ($\delta_i = 0$) we have $\Pr(C \in dx_i, T > x_i) = \Pr(C \in dx_i) \times \Pr(T > x_i)$ i.e., $g(x_i; \theta)dx_i \times S(x_i; \theta)$. We can then write the likelihood as

$$L(\theta) = \prod_{\delta_i=1} f(x_i; \theta)dx_i G(x_i; \theta) \times \prod_{\delta_i=0} g(x_i; \theta)dx_i S(x_i; \theta)$$

In most cases the further assumption that the censoring itself does not depend on θ may be reasonable. Taking logs and ignoring constants we obtain the result. For Theorem 3.2 note that

$$\begin{aligned}\Pr\{X_i \geq a_\ell | X_i > a_{\ell-1}\} &= \Pr\{T_i \geq a_\ell | X_i > a_{\ell-1}\} \times \Pr\{C_i \geq a_\ell | X_i > a_{\ell-1}\} \\ &= \Pr\{T_i \geq a_\ell | C_i > a_{\ell-1}, T_i > a_{\ell-1}\} = \Pr\{T_i \geq a_\ell | T_i > a_{\ell-1}\}\end{aligned}$$

by independence.

Theorem 3.3: We have $\text{Var}(\log S(a_\ell)) \approx \sum_{m \leq \ell} \text{Var}(\log(1 - \pi_m))$. This is $\sum_{m \leq \ell} (1 - \pi_m)^{-2} \text{Var}(\pi_m)$ which we write as $\sum_{m \leq \ell} (1 - \pi_m)^{-2} \pi_m (1 - \pi_m) / n_m$ where n_m corresponds to the number at risk. i.e. the denominator. We write $\sum_{m \leq \ell} d_m / r_m (r_m - d_m)$. Use log function and a further application of delta method to obtain:

$$\text{Var} S(a_\ell) \approx S(a_\ell)^2 \sum_{m \leq \ell} \sum_{m \leq \ell} d_m / r_m (r_m - d_m).$$

An approach not using the delta method follows Greenwood (1926). Let $t_0 < t_1 < \dots < t_k$. An estimate of the variance of the survival probability P of the form $P = p_1 \times p_2 \times \dots \times p_k$, where each p_i is the estimated probability of survival from time t_{i-1} to time t_i , $q_i = 1 - p_i$ and P is therefore the estimated probability of survival from t_0 to t_k . Assuming that the p_i 's are independent of one another, we have $\mathbf{E}(P) = \mathbf{E}(p_1) \times \mathbf{E}(p_2) \times \dots \times \mathbf{E}(p_k)$, as well as, $\mathbf{E}(P^2) = \mathbf{E}(p_1^2) \times \mathbf{E}(p_2^2) \times \dots \times \mathbf{E}(p_k^2)$, and $\mathbf{E}(p_i^2) = (\mathbf{E}p_i)^2 + \sigma_i^2$, where $\sigma_i^2 = \text{Var}(p_i)$. Then

$$\text{Var}(P) = \mathbf{E}(P^2) - \{\mathbf{E}(P)\}^2 = \{\mathbf{E}(P)\}^2 ((1 + \sigma_1^2 / (\mathbf{E}p_1)^2) (1 + \sigma_2^2 / (\mathbf{E}p_2)^2) \dots (1 + \sigma_k^2 / (\mathbf{E}p_k)^2) - 1) \approx \{\mathbf{E}(P)\}^2 (\sigma_1^2 / (\mathbf{E}p_1)^2 + \sigma_2^2 / (\mathbf{E}p_2)^2 + \dots + \sigma_k^2 / (\mathbf{E}p_k)^2).$$

Now condition on $\mathbf{E}(p_i)$ and the number of observations n_i to which p_i is applied, then $\sigma_i^2 = \mathbf{E}p_i(1 - \mathbf{E}p_i)/n_i$. Substituting p_i for $\mathbf{E}(p_i)$, we obtain estimates of the variance

$$\hat{\text{Var}}(P) = P^2 \left\{ \prod_{i=1}^k (1 + q_i/n_i p_i) - 1 \right\} \approx P^2 \sum_{i=1}^k q_i/n_i p_i.$$

Proposition 3.1: Take $l \in \{1, \dots, n\}$. If subject l fails at time t , then we have that $X_l = t$ and $\delta_l = 1$, so that denoting $\Delta \hat{S}(t) = \hat{S}(t) - \hat{S}(t^-)$. then

$$\begin{aligned}\Delta \hat{S}(t) &= \prod_{\substack{i=1, \dots, n \\ X_i \leq t}} \left(1 - \frac{\delta_i}{\sum_{j=1}^n Y_j(X_i)} \right) - \prod_{\substack{i=1, \dots, n \\ X_i \leq t^-}} \left(1 - \frac{\delta_i}{\sum_{j=1}^n Y_j(X_i)} \right) = \\ &\prod_{\substack{i=1, \dots, n \\ X_i \leq t^-}} \left(1 - \frac{\delta_i}{\sum_{j=1}^n Y_j(X_i)} \right) \left(1 - \frac{\delta_l}{\sum_{j=1}^n Y_j(X_l)} - 1 \right) = \frac{\hat{S}(t^-) \delta_l}{\sum_{j=1}^n Y_j(t)} = \frac{\hat{S}(t^-)}{\sum_{j=1}^n Y_j(t)}.\end{aligned}$$



Chapter 4

Proportional hazards models

4.1 Chapter summary

We consider several models that describe survival in the presence of observable covariates, these covariates measuring subject heterogeneity. The most general situation can be described by a model with a parameter of high, possibly unbounded, dimension. We refer to this as the general or non-proportional hazards model since dependence is expressed via a parameter, $\beta(t)$, that is not constrained or restricted. Proportional hazards models have the same form but constrain $\beta(t)$ to be a constant. We write the constant as β , sometimes β_0 , since it does not change with time. When the covariate itself is constant, the dependence structure corresponds to the Cox regression model. We describe this model, its connection to the well-known log-rank test, and its use in many applications. We recall the founding paper of Cox (Cox, 1972) and the many discussions that surrounded that paper. Some of the historical backgrounds that lay behind Cox's proposal is also recalled in order to for the new reader to quickly appreciate that, brilliant though Professor Cox's insights were, they leant on more than just his imagination. They did not emerge from a vacuum. Some discussion on how the model should be used in practice is given.

4.2 Context and motivation

The presence of subject heterogeneity, summarized by risk factors Z , known or suspected of being related to $S(t)$, is our central concern. The previous chapter dealt with the issue of marginal survival, i.e., survival ignoring any indicator of heterogeneity and which treats the data in hand as though the observations came from a single population. In Figure 4.1 there are two groups. This can be described by two distinct Kaplan-Meier curves or, possibly, two independently

calculated fitted parametric curves. If, however, the curves are related, then each estimate provides information not only about its own population curve but also about the other group's population curve. The curve estimates would not be independent. Exploiting such dependence can lead to considerable gains in our estimating power. The agreement between an approach modeling dependence and

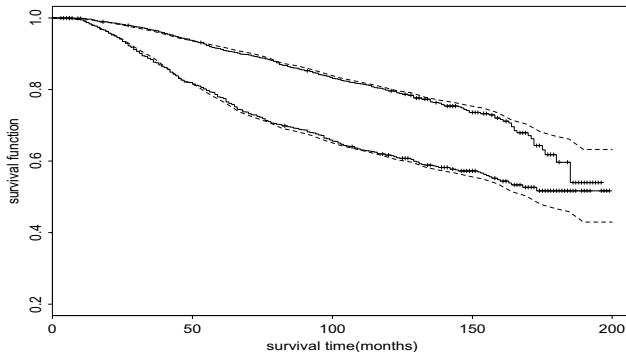


Figure 4.1: Kaplan-Meier survival curves and PH model curves for two groups defined by a binary covariate. Dashed lines represent PH estimates. The fit is adequate up to 150 months, after which the fit becomes progressively poorer.

one ignoring it can be more or less strong and, in Figure 4.1, agreement is good apart from observations beyond 150 months where a proportional hazards assumption may not hold very well. Returning to the simplest case, we can imagine a compartmental model describing the occurrence of deaths independently of group status in which all individuals are assumed to have the same hazard rates. As pointed out in the previous chapter, the main interest then is in the survival function $S(t)$ when the Z are either unobservable or being ignored. Here we study the conditional survival function given the covariates Z and we write this as $S(t|Z)$. In the more complex situations (multicompartment models, time-dependent Z) it may be difficult, or even impossible, to give an interpretation to $S(t)$ as an average over conditional distributions, but the idea of conditioning is still central although we may not take it beyond that of the probability of a change of state conditional upon the current state as well as the relevant covariate history which led to being in that state.

The goal here is to consider models with varying degrees of flexibility applied to the summary of n subjects each with an associated covariate vector Z of dimension p . The most flexible models will be able to fully describe any data at hand but, as a price for their flexibility, little reduction in dimension from the $n \times p$ data matrix we begin with. Such models will have small bias in prediction compared with large sampling errors. The most rigid models can allow for striking reductions in dimension. Their consequent impact on prediction will be associated with much smaller sampling errors. However, as a price for such gains, the biases

in prediction can be large. The models we finally work with will lie between these two extremes. Their choice then depends on an artful balance between the two conflicting characteristics. A central task, guided by the principle of parsimony, is to use as few parameters as possible to achieve whatever purpose we have in mind.

4.3 General or non-proportional hazards model

In the most straightforward cases we can express the conditional dependence of survival upon fixed covariates in terms of the hazard function. A general expression for the hazard function, given the value of the covariate Z is:

$$\lambda(t|Z) = \lambda_0(t) \exp\{\beta(t)Z\}, \quad (4.1)$$

where $\lambda(t|\cdot)$ is the conditional hazard function, $\lambda_0(t)$ the baseline hazard corresponding to $Z = 0$, and $\beta(t)$ a time-varying regression effect. Whenever Z has dimension greater than one we view $\beta(t)Z$ as an inner product in which $\beta(t)$ has the same dimension as Z so that $\beta(t)Z = \beta_1(t)Z_1 + \dots + \beta_p(t)Z_p$.

Recalling the discussion of Chapter 3, we are mostly interested in situations where observations on Z can be made in the course of any study. In Equation 4.1 Z is not allowed to depend upon time. If we also disallow the possibility of continuous covariates, which, in practice, we can approximate as accurately as we wish via high dimensional Z together with $\beta(t)$ of the same dimension, we see that model (4.1) is completely general and, as such, not really a model. It is instead a representation, or re-expression, of a very general reality, an expression that is convenient and which provides a framework to understanding many of the models described in this chapter. At the cost of losing the interpretation of a hazard function, we can immediately generalize (4.1) to

$$\lambda(t|Z) = \lambda_0(t) \exp\{\beta(t)Z(t)\}. \quad (4.2)$$

As long as we do not view $Z(t)$ as random, i.e., the whole time path of $Z(t)$ is known at $t = 0$, then a hazard function interpretation for $\lambda(t|Z)$ is maintained. Otherwise we lose the hazard function interpretation, since this requires knowledge of the whole function at the origin $t = 0$, i.e., the function is a deterministic and not a random one. In some ways this loss is of importance in that the equivalence of the hazard function, the survival function, and the density function means that we can easily move from one to another. However, when $Z(t)$ is random, we can reason in terms of intensity functions and compartmental models, a structure that enables us to deal with a wide variety of applied problems such as clinical trials using cross-over designs, studies in HIV that account for varied accumulated treatment histories and involved epidemiological investigations in which exposure history over time can be complex. The parameter $\beta(t)$ is of

infinite dimension and therefore the model would not be useful without some restrictions upon $\beta(t)$.

4.4 Proportional hazards model

Corresponding to the truth or reality under scrutiny, we can view Equation 4.2 as being an extreme point on a large scale which calibrates model complexity. The opposite extreme point on this scale might have been the simple exponential model, although we will start with a restriction that is less extreme, specifically the proportional hazards model in which $\beta(t) = \beta$ so that;

$$\lambda(t|Z) = \lambda_0(t) \exp\{\beta Z(t)\}. \quad (4.3)$$

Putting restrictions on $\beta(t)$ can be done in many ways, and the whole art of statistical modeling, not only for survival data, is in the search for useful restrictions upon the parameterization of the problem in hand. Our interpretation of the word “useful” depends very much on the given particular context. Just where different models find themselves on the infinite scale between Equation 4.3 and Equation 4.2 and how they can be ordered is a very important concept we need to master if we are to be successful at the modeling process, a process which amounts to feeling our way up this scale (relaxing constraints) or down this scale (adding constraints), guided by the various techniques at our disposal. From the outset it is important to understand that the goal is not one of establishing some unknown hidden truth. We already have this, expressed via the model described in Equation 4.1. The goal is to find a much smaller, more restrictive model, which, for practical purposes is close enough or which is good enough to address those questions that we have in mind; for example, deciding whether or not there is an effect of treatment on survival once we have accounted for known prognostic factors which may not be equally distributed across the groups we are comparing. For such purposes, no model to date has seen more use than the Cox regression model (Figure 4.2).

4.5 Cox regression model

In tackling the problem of subject heterogeneity, the Cox model has enjoyed outstanding success, a success, it could be claimed, matching that of classic multilinear regression itself. The model has given rise to considerable theoretical work and continues to provoke methodological advances. Research and development into the model and the model’s offspring have become so extensive that we cannot here hope to cover the whole field, even at the time of writing. We aim nonetheless to highlight what seems to be the essential ideas and we begin with a recollection of the seminal paper of D.R. Cox, presented at a meeting of the Royal Statistical Society in London, England, March 8, 1972.

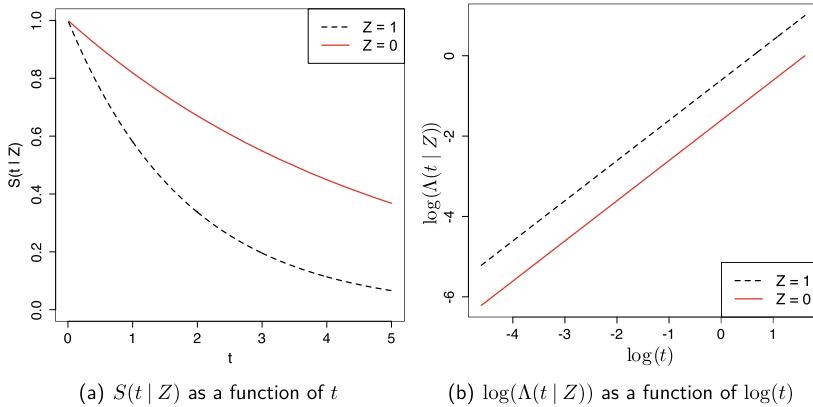


Figure 4.2: An illustration of survival curves and associated hazard functions for a proportional hazards model.

REGRESSION MODELS AND LIFE TABLES (D.R. Cox 1972)

After summarizing earlier work on the life table (Chang, 1968; Kaplan and Meier, 1958), Professor Cox introduced his, now famous, model postulating a simplified form for the relationship between the hazard function $\lambda(t)$, at time t and the value of an associated fixed covariate Z . As its name suggests, the proportional hazards model assumes that the hazard functions among subjects with different covariates are proportional to one another. The hazard function can then be written:

$$\lambda(t|Z) = \lambda_0(t) \exp\{\beta Z\}, \quad (4.4)$$

where $\lambda_0(t)$ is a fixed “baseline” hazard function, and β is a relative risk parameter to be estimated. Whenever $Z = 0$ has a concrete interpretation (which we can always obtain by re-coding) then so does the baseline hazard $\lambda_0(t)$ since, in this case, $\lambda(t|Z = 0) = \lambda_0(t)$. As mentioned just above, when Z is a vector of covariates, then the model is the same, although with the product of vectors βZ interpreted as an inner product. It is common to replace the expression βZ by $\beta'Z$ or $\beta^T Z$ where β and Z are $p \times 1$ vectors, and $a'b$, or $a^T b$ denote the inner product of vectors a and b . Usually, though, we will not distinguish notationally between the two situations since the former is just a special case of the latter. We write them both as βZ . Again we can interpret $\lambda_0(t)$ as being the hazard corresponding to the group for which the vector Z is identically zero.

The model is described as a multiplicative model, i.e., a model in which factors related to the survival time have a multiplicative effect on the hazard function. An illustration in which two binary variables are used to summarize the effects of four groups is shown in Figure 4.3. As pointed out by Cox, the function (βZ) can be replaced by any function of β and Z , the positivity of $\exp(\cdot)$ guaranteeing that, for any hazard function $\lambda_0(t)$, and any Z , we can always maintain a hazard function interpretation for $\lambda(t|Z)$. Indeed it is not necessary to restrict ourselves

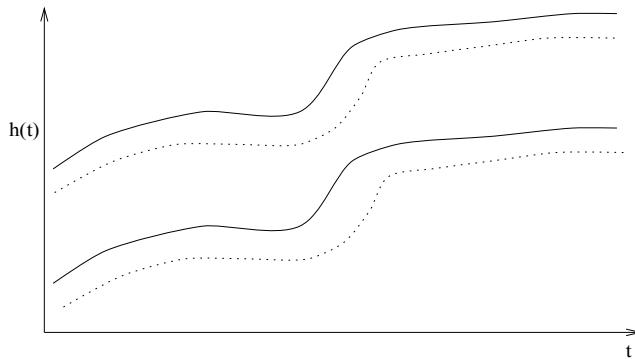


Figure 4.3: Proportional hazards with two binary covariates indicating 4 groups. Log-hazard rate written as $h(t) = \log \lambda(t)$.

to $\exp(\cdot)$, and we may wish to work with other functions $R(\cdot)$, although care is required to ensure that $R(\cdot)$ remains positive over the range of values of β and Z of interest. Figure 4.3 represents the case of two binary covariates indicating four distinct groups (in the figure we take the logarithm of $\lambda(t)$) and the important thing to observe is that the distance between any two groups on this particular scale, i.e., in terms of the log-hazards, does not change through time. In view of the relation between the hazard function and the survival function, there is an equivalent form of Equation 4.4 in terms of the survival function. Defining $S_0(t)$ to be the baseline survival function; that is, the survival function corresponding to $S(t|Z = 0)$, then, for scalar or vector Z , we have that

$$S(t|Z) = \{S_0(t)\}^{\exp(\beta Z)}. \quad (4.5)$$

When the covariate is a single binary variable indicating, for example, treatment groups, the model simply says that the survival function of one group is a power transformation of the other, thereby making an important connection to the class of Lehmann alternatives (Lehmann et al., 1953).

Cox took the view that “parametrization of the dependence on Z is required so that our conclusions about that dependence are expressed concisely”, adding that any choice “needs examination in the light of the data”. “So far as secondary features of the system are concerned ... it is sensible to make a minimum of assumptions.” This view led to focusing on inference that allowed $\lambda_0(t)$ to remain arbitrary. The resulting procedures are nonparametric with respect to t in that inference is invariant to any increasing monotonic transformation of t , but parametric in as much as concerns Z . For this reason the model is often referred to as Cox’s semi-parametric model. Let’s keep in mind, however, that it is the adopted inferential procedures that are semi-parametric rather than the model itself. Although, of course, use of the term $\lambda_0(t)$ in the model, in which $\lambda_0(t)$ is not specified, implies the use of procedures that will work for all allowable functions $\lambda_0(t)$.

Having recalled to the reader how inference could be carried out following some added assumptions on $\lambda_0(t)$, the most common assumptions being that $\lambda_0(t)$ is constant, that $\lambda_0(t)$ is a piecewise constant function, or that $\lambda_0(t)$ is equal to t^γ for some γ , Cox presented his innovative likelihood expression for inference, an expression that subsequently became known as a partial likelihood (Cox, 1975). We look more closely at these inferential questions in later chapters. First note that the quantity $\lambda_0(t)$ does not appear in the likelihood expression given by

$$L(\beta) = \prod_{i=1}^n \left\{ \frac{\exp(\beta Z_i)}{\sum_{j=1}^n Y_j(X_i) \exp(\beta Z_j)} \right\}^{\delta_i}, \quad (4.6)$$

and, in consequence, $\lambda_0(t)$ can remain arbitrary. Secondly, note that each term in the product is the conditional probability that at time X_i of an observed failure, it is precisely individual i who is selected to fail, given all the individuals at risk and given that one failure would occur. Taking the logarithm in Equation 4.6 and its derivative with respect to β , we obtain the estimating equation which, upon setting equal to zero, can generally be solved without difficulty using the Newton-Raphson method, to obtain the maximum partial likelihood estimate $\hat{\beta}$ of β . We will discuss more deeply the function $U(\beta)$ under the various approaches to inference. We can see already that it has the same form as that encountered in the standard linear regression situation where the observations are contrasted to some kind of weighted mean. The exact nature of this mean is described later. Also, even though the expression

$$U(\beta) = \sum_{i=1}^n \delta_i \left\{ Z_i - \frac{\sum_{j=1}^n Y_j(X_i) Z_j \exp(\beta Z_j)}{\sum_{j=1}^n Y_j(X_i) \exp(\beta Z_j)} \right\} \quad (4.7)$$

looks slightly involved, we might hope that the discrepancies between the Z_i and the weighted mean, clearly some kind of residual, would be uncorrelated, at least for large samples, since the Z_i themselves are uncorrelated.

All of this turns out to be so and makes it relatively easy to carry out appropriate inference. The simplest and most common approach to inference is to treat $\hat{\beta}$ as asymptotically normally distributed with mean β and large sample variance $I(\hat{\beta})^{-1}$, where $I(\beta)$, called the information in view of its connection to the likelihood, is the second derivative of $-\log L(\beta)$ with respect to β , i.e., letting

$$I(\beta) = \frac{\sum_{j=1}^n Y_j(X_i) Z_j^2 \exp(\beta Z_j)}{\sum_{j=1}^n Y_j(X_i) \exp(\beta Z_j)} - \left\{ \frac{\sum_{j=1}^n Y_j(X_i) Z_j \exp(\beta Z_j)}{\sum_{j=1}^n Y_j(X_i) \exp(\beta Z_j)} \right\}^2, \quad (4.8)$$

then $I(\beta) = \sum_{i=1}^n \delta_i I_i(\beta)$. Inferences can also be based on likelihood ratio methods. A third possibility, which is sometimes convenient, is to base tests on the score $U(\beta)$, which in large samples can be considered to be normally distributed with mean zero and variance $I(\beta)$. Multivariate extensions are completely natural, with the score being a vector and I an information matrix.

EARLY APPLICATIONS OF THE MODEL

The first success of the model was in its use for the two-sample problem, i.e., testing the null hypothesis of no difference in the underlying true survival curves for two groups. In this case Cox showed that the test statistic $U(0)/\sqrt{I(0)}$ is formally identical to a test, later known under the heading of the log-rank test, obtained by setting up at each failure point a 2×2 contingency table, group against failed/survived, and combining the many 2×2 tables. As in a standard analysis of a single such contingency table we use the marginal frequencies to obtain estimates of expected rates under the null hypothesis of no effect. Assuming, as we usually do here, no ties we can obtain a table such as described in Table 4.1 in which, at time $t = X_i$ the observed failure occurs in group A and there are $n_A(t)$ and $n_B(t)$ individuals at risk in the respective groups.

Time point $t = X_i$	Group A	Group B	Totals
Number of failures	1	0	1
Number not failing	$n_A(t) - 1$	$n_B(t)$	$n_A(t) + n_B(t) - 1$
Total at risk	$n_A(t)$	$n_B(t)$	$n_A(t) + n_B(t)$

Table 4.1: 2×2 table at failure point $t = X_i$ for group A and group B.

The observed rates and the expected rates are simply summed across the distinct failure points, each of which gives rise to its own contingency table where the margins are obtained from the available risk sets at that time. From the above, if $Z_i = 1$ when subject i is in group A and zero otherwise, then elementary calculation gives that

$$U(0) = \sum_{i=1}^n \delta_i \{Z_i - \pi(X_i)\}, \quad I(0) = \sum_{i=1}^n \delta_i \pi(X_i) \{1 - \pi(X_i)\}$$

where $\pi(t) = n_A(t)/\{n_A(t) + n_B(t)\}$. The statistic U then contrasts the observations with their expectations under the null hypothesis of no effect. This expectation is simply the probability of choosing, from the subjects at risk, a subject from group A. The variance expression is the well-known expression for a Bernoulli variable. Readers interested in a deeper insight into this test should also consult (Cochran, 1954; Mantel, 1963; Mantel and Haenszel, 1959; Peto and Peto, 1972). As pointed out by Cox, “whereas the test in the contingency table situation is, at least in principle, exact, the test here is only asymptotic ...”

This statement is not fully precise since there is still an appeal to the DeMoivre-Laplace approximation. Nonetheless, we can understand his point.

However, the real advantage of Cox's approach was that while contributing significantly toward a deeper understanding of the log-rank and related tests, it opened up the way for more involved situations; additional covariates, continuous covariates, random effects, and, perhaps surprisingly, in view of the attribute "proportional hazards", a way to tackle problems involving time-varying effects or time-dependent covariates. Cox illustrated his model via an application to the now famous Freireich data (Acute Leukemia Group B et al., 1963) describing a clinical trial in leukemia in which a new treatment was compared to a placebo. Treating the two groups independently and estimating either survivorship function using a Kaplan-Meier curve gave good agreement with the survivorship estimates derived from the Cox model. Such a result can also, of course, be anticipated by taking a $\log(-\log)$ transform of the Kaplan-Meier estimates and noting that they relate to one another via a simple shift. This shift exhibits only the weakest, if any, dependence on time itself.

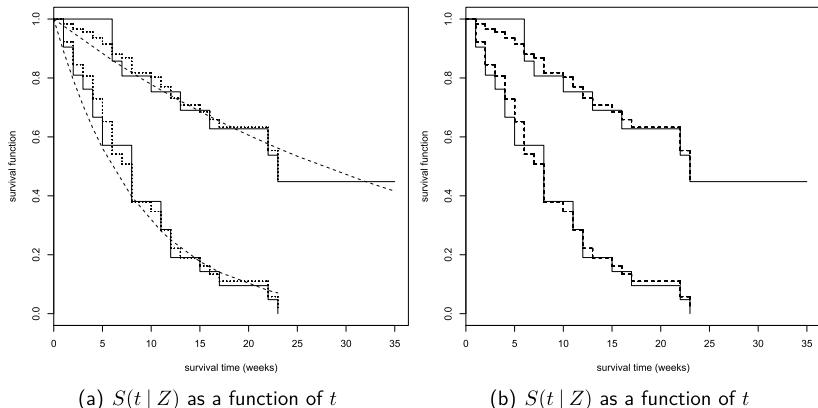


Figure 4.4: Kaplan-Meier curves and model-based curves for Freireich data. Dashed lines represent Model-based estimates; exponential model (left), Cox model (right).

Recovering the usual two-group log-rank statistic as a special case of a test based on model (4.4) is reassuring. In fact, exactly the same approach extends to the several group comparison (Breslow, 1972). More importantly, Equation 4.4 provides the framework for considering the multivariate problem from its many angles; global comparisons of course but also more involved conditional comparisons in which certain effects are controlled for while others are tested. We look at this in more detail below under the heading "Modeling multivariate problems". The partially proportional hazards model (in particular the stratified model) was to appear later to Cox's original work of 1972 and provide great

flexibility in addressing regression problems in a multivariate context. The early applications of the model provided some added theoretical structure to procedures already in use. It was not long before statisticians came to realize that the model would allow us to go well beyond those procedures and the discussion to Cox's paper already anticipated some of these later developments.

DISCUSSION OF PROFESSOR COX'S PAPER

Kalbfleisch and Prentice took issue with Cox's naming of the likelihood used for inference as a "conditional" likelihood. They pointed out that the likelihood expression is not obtainable as a quantity proportional to a probability after having conditioned on some event. Conditioning was indeed taking place in the construction of the likelihood expression but in a sequential manner, a dynamic updating whose inferential home would later be seen to lie more naturally within the context of stochastic processes, indexed by time, rather than regular likelihoods, whether marginal or conditional.

The years following this discussion gave rise to a number of papers investigating the nature of the "conditional" likelihood proposed in Cox's original paper. Given the striking success of the model, together with the suggested likelihood expression, in reproducing and taking further a wide range of statistics then in use, most researchers agreed that Cox's proposal was correct. They remained uncertain, though, as to how to justify the likelihood itself. This thinking culminated in several major contributions; those of Cox (1975), Prentice and Kalbfleisch (1979), Aalen (1978) and Andersen and Gill (1982), firmly establishing the likelihood expression of Cox.

It turned out that Cox was correct, not just on the appropriateness of his proposed likelihood expression but also in describing it as a "conditional" likelihood, this description being the source of all the debate. Although Cox's likelihood derivation may not have been conditional, in the sense of taking as observed some single statistic upon which we condition before proceeding, his likelihood is not only very much a conditional one but also it conditions in just the right way. Not in the most straightforward sense whereby all the conditioning is done in one go, but in the sense of sequentially conditioning through time. Cox's "conditional" likelihood is now called a "partial" likelihood although, as an inferential tool in its own right, i.e., as a tool for inference independent of the choice of any particular model the partial likelihood turned out not to be a particularly useful concept.

Professor Downton of the University of Birmingham and Professor Peto of the University of Oxford pointed out the connection to rank test procedures. Although the formulation of Cox allowed the user to investigate more complex structures, many existing setups, framed in terms of tests based on the ranks, could be obtained directly from the use of the Cox likelihood. The simplest example was the sign test for the median. Using permutation arguments, other tests of interest in the multivariate setting could be obtained, in particular tests analogous to the Friedman test and the Kruskal-Wallis test. Richard Peto referred

to some of his own work with Julian Peto. Their work demonstrated the asymptotic efficiency of the log-rank test and that, for the two-group problem and for Lehmann alternatives, this test was locally most powerful. Since the log-rank test coincides with a score test based on Cox's likelihood, Peto argued that Cox's method necessarily inherits the same properties.

Professor Bartholomew of the University of Kent considered a lognormal model in current use and postulated its extension to the regression situation by writing down the likelihood. Such an analysis, being fully parametric, represents an alternative approach since the structure is not nested in a proportional hazards one. Bartholomew made an insightful observation that allowing for some dependence of the explanatory variable Z on t can enable the lognormal model and a proportional hazards model to better approximate each other. This is indeed true and allows for a whole development of a class of non-proportional hazards models where Z is a function of time and within which the proportional hazards model arises as a special case.

Professors Oakes and Breslow discussed the equivalence between a saturated piecewise exponential model and the proportional hazards model. By a saturated piecewise exponential model we mean one allowing for constant hazard rates between adjacent failures. The model is data dependent in that it does not specify in advance time regions of constant hazard but will allow these to be determined by the observed failures. From an inferential standpoint, in particular making use of likelihood theory, we may expect to run into some difficulties. This is because the number of parameters of the model (number of constant hazard rates) increases at the same rate as the effective sample size (number of observed failure times). However, the approach does nonetheless work, although justification requires the use of techniques other than standard likelihood. A simple estimate of the hazard rate, the cumulative hazard rate, and the survivorship function are then available. When $\beta = 0$ the estimate of the cumulative hazard rate coincides with that of Nelson (1969).

Professor Lindley of University College London writes down the full likelihood which involves $\lambda_0(t)$ and points out that, since terms involving $\lambda_0(t)$ do not factor out we cannot justify Cox's conditional likelihood. If we take $\lambda_0(t)$ as an unknown nuisance parameter having some prior distribution, then we can integrate the full likelihood with respect to this in order to obtain a marginal likelihood (this would be different to the marginal likelihood of ranks studied later by Kalbfleisch and Prentice (1973)). Lindley argues that the impact of censoring is greater for the Cox likelihood than for this likelihood which is then to be preferred. The author of this text confesses to not fully understanding Lindley's argument and there is some slight confusion there since, either due to a typo or to a subtlety that escapes me, Lindley calls the Cox likelihood a "marginal likelihood" and what I am referring to as a marginal likelihood, an "integrated likelihood". We do, of course, integrate a full likelihood to obtain a marginal likelihood, but it seems as though Professor Lindley was making other, finer, distinctions which are best understood by those in the Bayesian school. His concern on the impact of cen-

soring is echoed by Mr. P. Glassborow of British Rail underlining the strength behind the independent censoring assumption, an assumption which would not be reasonable in many practical cases.

Professor Zelen, a pioneer in the area of regression analysis of survival data, pointed out important relationships in tests of regression effect in the proportional hazards model and tests of homogeneity of the odds ratio in the study of several contingency tables. Dr. John Gart of the National Cancer Institute also underlined parallels between contingency table analysis and Cox regression. These ideas were to be developed extensively in later papers by Ross Prentice and Norman Breslow in which the focus switched from classical survival analysis to studies in epidemiology. The connection to epidemiological applications was already alluded to in the discussion of the Cox paper by Drs. Meshalkin and Kagan of the World Health Organization. Finally, algorithms for carrying out an analysis based on the Cox model became quickly available thanks to two further important contributions to the discussion of Cox's paper. Richard Peto obtained accurate approximations to the likelihood in the presence of ties, obviating the need for computationally intensive permutation algorithms, and Susannah Howard showed how to program efficiently by exploiting the nested property of the risk sets in reversed time.

HISTORICAL BACKGROUND TO COX'S PAPER

Alternative hypotheses to a null which assumes that two probabilities are equal, such as in Equation 4.5, taking the form of a simple power transformation, have a long history in statistical modeling. Such alternatives which, in the special case where the probabilities in question are survival functions, are known as Lehmann alternatives (Lehmann et al., 1953). Lehmann alternatives are natural in that, under the restriction that the power term is positive, always achievable by reparameterizing the power term to be of an exponential form; then, whatever the actual parameter estimates, the resulting probability estimates satisfy the laws of probability. In particular, they remain in the interval (0,1).

Linear expressions for probabilities are less natural although, at least prior to the discovery of the logistic and Cox models, possibly more familiar. Feigl and Zelen (1965) postulated a linear regression for the location parameter, λ_0 , of an exponential law. In this case the location parameter and the (constant) hazard coincide so that the model could be written:

$$\lambda(t|Z) = \lambda_0 \exp\{\beta Z\}. \quad (4.9)$$

In Feigl and Zelen their model was not written exactly this way, expressed as $\lambda = \alpha + \beta Z$. However, since λ is constant, the two expressions are equivalent and highlight the link to Cox's more general formulation. Feigl and Zelen only considered the case of uncensored data. Zippin and Armitage (1966) used a modeling approach, essentially the same as that of Feigl and Zelen, although

allowing for the possibility of censoring. This was achieved by an assumption of independence between the censoring mechanism and the failure mechanism enabling an expression for the full likelihood to be obtained. Further discussion on these ideas can be found in Myers et al. (1973) and Brown (1975). The estimates of the survival function for the different groups in the Freireich study, based on a simple exponential model or a Cox model, are shown in Figure 4.4. For these data the level of agreement between the two approaches appears to be high. This early work on the exponential model certainly helped anticipate the more general development of Cox and, for many more straightforward comparisons, such as the one illustrated by the Freireich data, it is perhaps unfortunate that the exponential model has been relegated to a historical role alone and is rarely, if ever, used in current practical analysis of similar data. Estimation for the exponential model is particularly simple and can be carried out without the use of a computer. This would often make the model a good starting point before building more complex models. However, modern statistical analysis will tend to jump straight into the problem backed up by powerful programs and as a result, there is much less call for making use of methods that can be calculated using no more than a pencil and paper.

4.6 Modeling multivariate problems

The strength of the Cox model lies in its ability to describe and characterize involved multivariate situations. Crucial issues concern the adequacy of fit of the model, how to make predictions based on the model, and how strong is the model's predictive capability. These are considered in detail later. Here, in the following sections and in the chapter on inference we consider how the model can be used as a tool to formulate questions of interest to us in the multivariate setting. The simplest case is that of a single binary covariate Z taking the values zero and one. The zero might indicate a group of patients undergoing a standard therapy, whereas the group for which $Z = 1$ could be undergoing some experimental therapy. Model 4.4 then indicates the hazard rate for the standard group to be $\lambda_0(t)$ and for the experimental group to be $\lambda_0(t)\exp(\beta)$. Testing whether or not the new therapy has any effect on survival translates as testing the hypothesis $H_0 : \beta = 0$. If β is less than zero then the hazard rate for the experimental therapy is less than that for the standard therapy at all times and is such that the arithmetic difference between the respective logarithms of the hazards is of magnitude β . Suppose the problem is slightly more complex and we have two new experimental therapies. We can write:

$$\lambda(t|Z) = \lambda_0(t)\exp\{\beta_1 Z_1 + \beta_2 Z_2\}$$

and obtain Table 4.2. As we shall see the two covariate problem is very much more complex than the case of a single covariate. Not only do we need to consider the

Treatment group	Z_1	Z_2	Log of group effect
Standard therapy	0	0	0
Experimental therapy 1	1	0	β_1
Experimental therapy 2	0	1	β_2

Table 4.2: Effects for two treatment groups.

effect of each individual treatment on the hazard rate for the standard therapy but we also need to consider the effect of each treatment in the presence or absence of the other as well as the combined effect of both treatments together. The particular model form in which we express any relationships will typically imply assumptions on those relationships and an important task is to bring under scrutiny (goodness of fit) the soundness of any assumptions.

It is also worth noting that if we are to assume that a two-dimensional covariate proportional hazards model holds exactly, then, integrating over one of the covariates to obtain a one-dimensional model will not result (apart from in very particular circumstances) in a lower-dimensional proportional hazards model. The lower-dimensional model would be in a much more involved non-proportional hazards form. This observation also holds when adding a covariate to a one-dimensional proportional hazards model, a finding that compels us, in realistic modeling situations, to only ever consider the model as an approximation.

By extension the case of several covariates becomes rapidly very complicated. If, informally, we were to define complexity as *the number of things you have to worry about*, then we could, even more informally, state an important theorem.

Theorem 4.1. (Informal non-mathematical theorem). *The complexity of any regression problem grows exponentially with the number of covariates in the equation.*

Obviously such a theorem cannot hold in any precise mathematical sense without the need to add conditions and restrictions such that its simple take-home message would be lost. For instance, if each added covariate was a simple constant multiple of the previous one, then there would really be no added complexity. But, in some broad sense, the theorem does hold and to convince ourselves of this we can return to the case of two covariates. Simple combinatorial arguments show that the number of possible hypotheses of potential interest is increasing exponentially. But it is more complex than that. Suppose we test the hypothesis $H_0 : \beta_1 = \beta_2 = 0$. This translates the clinical null hypothesis: neither of the experimental therapies impacts survival against the alternative, $H_1 : \exists \beta_i \neq 0, i = 1, 2$. This is almost, yet not exactly, the same as simply regrouping the two experimental treatments together and reformulating the problem in terms of a single binary variable.

Next, we might consider testing the null hypothesis $H_0 : \beta_1 = 0$ against the alternative hypothesis $H_1 : \beta_1 \neq 0$. Such a test focuses only on the first experimental treatment, but does not, as we might at first imagine, lump together both the second experimental treatment and the standard treatment. This test makes no statement about β_2 and so this could indeed take the value zero (in which case the standard and the second experimental therapy are taken to be the same) or any other value in which case, detecting a nonzero value for β_1 translates as saying that this therapy has an effect different to the standard regardless of the effect of the second experimental therapy. Clearly this is different from lumping together the second experimental therapy with the standard and testing the two together against the first experimental therapy. In such a case, should the effect of the first experimental therapy lie somewhere between that of the standard and the second, then, plausibly, we might fail to detect a nonzero β_1 even though there exist real differences between the standard and the first therapy.

All of this discussion can be repeated, writing β_1 in the place of β_2 . Already, we can see that there are many angles from which to consider an equation such as the above. These angles, or ways of expressing the scientific question, will impact the way of setting up the statistical hypotheses. In turn, these impact our inferences.

Another example would be testing the above null hypothesis $H_0 : \beta_1 = \beta_2 = 0$ against an alternative $H_1 : 0 < \beta_1 < \beta_2$ instead of that initially considered (i.e., $H_1 : \exists \beta_i \neq 0, i = 1, 2$). The tests, and their power properties, would not typically be the same. We might consider re-coding the problem, as in Equation 4.10, so that testing $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$ corresponds to testing for an effect in either group. Given this effect we can test $H_0 : \beta_2 = 0$ against $H_1 : \beta_2 \neq 0$ which will answer the question as to whether, given that there exists a treatment effect, it is the same for both of the experimental treatments:

$$\begin{aligned}\lambda(t|Z) &= \lambda_0(t) \exp\{\beta_1 Z_1 + (\beta_1 + \beta_2) Z_2\} \\ &= \lambda_0(t) \exp\{\beta_1(Z_1 + Z_2) + \beta_2 Z_2\}. \end{aligned}\quad (4.10)$$

Note that fitting the above models needs no new procedures or software for example, since both cases come under the standard heading. In the first equation all we do is write $\alpha_1 = \beta_1$ and $\alpha_2 = \beta_1 + \beta_2$. In the second we simply redefine the covariates themselves. The equivalence expressed in the above equation is important. It implies two things. Firstly, that this previous question concerning differential treatment effects can be re-expressed in a standard way enabling us to use existing structures, and computer programs. Secondly, since the effects in our models express themselves via products of the form βZ , any re-coding of β can be artificially carried out by re-coding Z and vice versa. This turns out to be an important property and anticipates the fact that a non-proportional hazards model $\beta(t)Z$ can be re-expressed as a time-dependent proportional hazards model $\beta Z(t)$. Hence the very broad sweep of proportional hazards models.

It is easy to see how the above considerations, applied to a situation in which we have $p > 2$ covariates, become very involved. Suppose we have four ordered levels of some risk factor. We can re-code these levels using three binary covariates as in Table 4.3. For this model we can, again, write the hazard function in terms

Risk factor	Z_1	Z_2	Z_3	Log of risk factor effect
Level 1	0	0	0	0
Level 2	1	0	0	β_1
Level 3	0	1	0	β_2
Level 4	0	0	1	β_3

Table 4.3: Coding for four ordered levels of a risk factor.

of these binary coding variables, noting that, as before, there are different ways of expressing this. In standard form we write

$$\lambda(t|Z) = \lambda_0(t) \exp\{\beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3\}$$

so that the hazard rate for those exposed to the risk factor at level i , $i = 1, \dots, 4$, is given by $\lambda_0(t) \exp(\beta_i)$ where we take $\beta_0 = 0$. Our interest may be more on the incremental nature of the risk as we increase through the levels of exposure to the risk factor. The above model can be written equivalently as

$$\begin{aligned} \lambda(t|Z) &= \lambda_0(t) \exp\{\beta_1 Z_1 + (\beta_1 + \beta_2) Z_2 + (\beta_1 + \beta_2 + \beta_3) Z_3\} \\ &= \lambda_0(t) \exp\{\beta_1(Z_1 + Z_2 + Z_3) + \beta_2(Z_2 + Z_3) + \beta_3 Z_3\} \end{aligned} \quad (4.11)$$

so that our interpretation of the β_i is in terms of increase in risk. The coefficient β_1 in this formulation corresponds to an overall effect, common to all levels above the lowest. The coefficient β_2 corresponds to the amount by which the log-hazard rate for the second level differs from that at the first. Here then, a value of β_2 equal to zero does not mean that there is no effect at level 2, simply that the effect is no greater than that already quantified at level 1. The same arguments follow for levels 3 and 4.

Writing the model in these different ways is not changing the basic model. It changes the interpretation that we can give to the different coefficients. The equivalent expression shown in Equation 4.11, for example, means that we can carefully employ combinations of the covariates in order to use existing software. But we can also consider the original coding of the covariates Z . Suppose that, instead of the coding given in Table 4.3, we use the coding given in Table 4.4. This provides an equivalent description of the four levels. As we move up the levels, changing from level i to level $i+1$, the log hazard is increased by β_i .

Let's imagine a situation, taken from Table 4.4, in which $\beta_1 = \beta_2 = \beta_3$. Real situations may not give rise to strict equalities but may well provide good first approximations. The hazards at each level can now be written very simply as

Risk factor	Z_1	Z_2	Z_3	Log of risk factor effect
Level 1	0	0	0	0
Level 2	1	0	0	β_1
Level 3	1	1	0	$\beta_1 + \beta_2$
Level 4	1	1	1	$\beta_1 + \beta_2 + \beta_3$

Table 4.4: Coding for four ordered levels of a risk factor.

$\lambda_0(t) \exp(j\beta_1)$ for $j = 0, 1, 2, 3$, and this is described in Table 4.5. Taking $\beta_1 = \beta$, we are then able to write a model for this situation as $\lambda(t|Z) = \lambda_0(t) \exp(\beta Z)$, in which the covariate Z , describing group level, takes the values 0 to 3. This model has a considerable advantage over the previous one, describing the same situation of four levels, in that only a single coefficient appears in the model as opposed to three. We will use our data to estimate just a single parameter. The gain is clear.

Risk factor	Z	Log of risk factor effect
Level 1	0	0
Level 2	1	β
Level 3	2	2β
Level 4	3	3β

Table 4.5: Coding for four ordered levels of a risk factor.

The cost, however, is much less so, and is investigated more thoroughly in the chapters on prediction (explained variation, explained randomness) and goodness of fit. If the fit is good, i.e., the assumed linearity is reasonable, then we would certainly prefer the latter model to the former. If we are unsure we may prefer to make less assumptions and use the extra flexibility afforded by a model which includes three binary covariates rather than a single linear covariate. In real data analytic situations we are likely to find ourselves somewhere between the two, using the tools of fit and predictability to guide us.

Returning once more to Table 4.4 we can see that the same idea prevails for the β_i not all assuming the same values. A situation in which four ordered levels is described by three binary covariates could be recoded so that we only have a single covariate Z , together with a single coefficient β . Next, suppose that in the model, $\lambda(t|Z) = \lambda_0(t) \exp(\beta Z)$, Z not only takes the ordered values, 0, 1, 2 and 3 but also all of those in between. In a clinical study this might correspond to some prognostic indicator, such as blood pressure or blood cholesterol, recorded continuously and re-scaled to lie between 0 and 3.

Including the value of Z , as a continuous covariate, in the model amounts to making very strong assumptions. It supposes that the log hazard increases by the same amount for every given increase in Z , so that the relative risk associated

with $\Delta = z_2 - z_1$ is the same for all values of z_1 between 0 and $3 - \Delta$. Let's make things a little more involved. Suppose we have the same continuous covariate, this time let's call it Z_1 , together with a single binary covariate Z_2 indicating one of two groups. We can write

$$\lambda(t|Z_1, Z_2) = \lambda_0(t) \exp(\beta_1 Z_1 + \beta_2 Z_2).$$

Such a model supposes that a given change in exposure Z_1 results in a given change in risk, as just described, but that, furthermore, this resulting change is the same at both levels of the discrete binary covariate Z_2 . This may be so but such strong assumptions must be brought under scrutiny. Given the ready availability of software, it is not at all uncommon for data analysts to simply "throw in" all of the variables of interest, both discrete and continuous, without considering potential transformations or re-coding, turn the handle, and then try to make sense of the resulting coefficient estimates together with their standard errors. Such an exercise will rarely be fruitful. In this respect it is preferable to write one's own computer programs when possible or to use available software such as the R package, which tends to accompany the user through model development. Packages that present a "complete" one-off black-box analysis based on a single model are unlikely to provide much insight into the nature of the mechanisms generating the data at hand.

The user is advised to exercise great care when including continuous covariates in a model. We can view a continuous covariate as equivalent to an infinite-dimensional vector of indicator variables so that, in accordance with our informal theorem of complexity, the number of things we need worry about is effectively infinite. Let us not however overstate things, and it is of course useful to model continuous covariates. But be wary. Also consider the model

$$\lambda(t|Z) = \lambda_0(t) \exp(\beta_1 Z + \beta_2 Z^2).$$

If Z is binary then $Z^2 = Z$ and there is no purpose to the second term in the equation. If Z is ordinal or continuous then the effect of Z is quadratic rather than linear. And, adding yet higher-order terms enables us, at least in principle, to model other nonlinear functions. In practice, in order to carry out the analysis, we would use existing tools by simply introducing a second variable Z_2 defined by $Z_2 = Z^2$; an important observation in that the linear representation of the covariate can be relaxed with relatively little effort. For example, suppose that the log-relative risk is expressed via some smooth function $\psi(z)$ of a continuous covariate z . Writing the model

$$\lambda(t|Z) = \lambda_0(t) \exp\{\beta\psi(Z)\}$$

supposes that we know the functional form of the relative risk, at least up to the constant multiple β . Then, a power series approximation to this would allow us to

write $\psi(Z) = \sum \beta_j Z^j$ in which any constant term β_0 is absorbed into $\lambda_0(t)$. We then introduce the covariates $Z_j = Z^j$ to bring the model into its standard form.

4.7 Classwork and homework

1. One of the early points of discussion on Cox's 1972 paper was how to deal with tied data. Look up the Cox paper and write down the various different ways that Cox and the contributors to the discussion suggested that tied data be handled. Explain the advantages and disadvantages of each approach.
2. One suggestion for dealing with tied data, not in that discussion, is to simply break the ties via some random split mechanism. What are the advantages and drawbacks of such an approach?
3. As an alternative to the proportional hazards model consider the two models (i) $S(t|Z) = S_0(t) + \beta Z$, and (ii) $\text{logit}S(t|Z) = \text{logit}S_0(t) + \beta Z$. Discuss the relative advantages and drawbacks of all three models.
4. Show that the relation; $S(t|Z) = \{S_0(t)\}^{\exp(\beta Z)}$ implies the Cox model and vice versa.
5. Suppose that we have two groups and that a proportional hazards model is believed to apply. Suppose also that we know for one of the groups that the hazard rate is a linear function of time, and equal to zero at the origin. Given data from such a situation, suggest different ways in which it can be analyzed and the possible advantages and disadvantages of the various approaches.
6. Explain in what sense the components of Equation 4.7 and equation (4.8) can be viewed as an equation for the mean and an equation for the variance.
7. Using equations (4.7) and (4.8) work out the calculations explicitly for the two-group case, i.e., the case in which there are $n_1(t)$ subjects at risk from group 1 at time t and $n_2(t)$ from group 2.
8. Suppose that we have available software able to analyze a proportional hazards model with a time-dependent covariate $Z(t)$. Suppose that, for the problem in hand the covariate, Z , does not depend on time. However, the regression effect $\beta(t)$ is known to decline as an exponential function of time. How would you proceed?
9. Suppose we fit a proportional hazards model, using some standard software, to a continuous covariate Z defined on the interval (1,4). Unknown to us our model assumption is incorrect and the model applies exactly to $\log Z$ instead. What effect does this have on our parameter estimate?

10. Consider an experiment in which there are eight levels of treatment. The levels are ordered. The null hypothesis is that there is no treatment effect. The alternative is that there exists a non-null effect increasing with level until it reaches one of the levels, say level j , after which the remaining levels all have the same effect as level j . How would you test for this?
11. Write down the joint likelihood for the underlying hazard rate and the regression parameter β for the two-group case in which we assume the saturated piecewise exponential model. Use this likelihood to recover the partial likelihood estimate for β . Obtain an estimate of the survivorship function for both groups.
12. For the previous question derive an approximate large sample confidence interval for the estimate of the survivorship function for both groups in cases: (i) where the parameter β is exactly known, (ii) where the parameter is replaced by an estimate with approximate large sample variance σ^2 .
13. Carry out a large sample simulation for a model with two binary variables. Each study is balanced with a total of 100 subjects. Choose $\beta_1 = \beta_2 = 1.5$ and simulate binary Z_1 and Z_2 to be uncorrelated. Show the distribution of $\hat{\beta}_1$ in two cases: (i) where the model used includes Z_2 , (ii) where the model used includes only Z_1 . Comment on the distributions, in particular the mean value of $\hat{\beta}_1$ in either case.
14. In the previous exercise, rather than include in the model Z_2 , use Z_2 as a variable of stratification. Repeat the simulation in this case for the stratified model. Comment on your findings.
15. Consider the following regression situation. We have one-dimensional covariates Z , sampled from a density $g(z)$. Given z we have a proportional hazards model for the hazard rates. Suppose that, in addition, we are in a position to know exactly the marginal survivorship function $S(t) = \int S(t|z)g(z)dz$. How can we use this information to obtain a more precise analysis of data generated under the PH model with Z randomly sampled from $g(z)$?
16. Suppose we have two groups defined by the indicator variable $Z = \{0, 1\}$. In this example, unlike the previous in which we know the marginal survival, we know the survivorship function $S_0(t)$ for one of the groups. How can this information be incorporated into a two-group comparison in which survival for both groups is described by a proportional hazards model? Use a likelihood approach.
17. Use known results for the exponential regression model in order to construct an alternative analysis to that of the previous question based upon likelihood.
18. A simple test in the two-group case for absence of effects is to calculate the area between the two empirical survival curves. We can evaluate the null

distribution by permuting the labels corresponding to the group assignment indicator Z . Carry out this analysis for the Freireich data and obtain a p -value. How does this compare with that obtained from an analysis based on the assumption of an exponential model and that based on partial likelihood?

19. Carry out a study, i.e., the advantages, drawbacks, and potentially restrictive assumptions of the test of the previous example. How does this test compare with the score test based on the proportional hazards model?
20. Obtain a plot of the likelihood function for the Freireich data. Using simple numerical integration routines, standardize the area under the curve to be equal to one.
21. For the previous question, treat the curve as a density. Use the mean as an estimate of the unknown β . Use the upper and lower 2.5% percentiles as limits to a 95% confidence interval. Compare these results with those obtained using large sample theory.
22. Suppose we have six ordered treatment groups indicated by $Z = 1, \dots, 6$. For all values of $Z \leq \ell$ the hazards are the same. For $Z > \ell$ the hazards are again the same and either the same as those for $Z \leq \ell$ or all strictly greater than for $Z \leq \ell$. The value of ℓ is not known. How would you model and set up tests in this situation?



Chapter 5

Proportional hazards models in epidemiology

5.1 Chapter summary

The basic questions of epidemiology are reconsidered in this chapter from the standpoint of a survival model. We rework the calculations of relative risk, where the time factor is now age, and we see how our survival models can be used to control for the effects of age. Series of 2×2 tables, familiar to epidemiologists, can be structured within the regression model setting. The well-known Mantel-Haenszel test arises as a model-based score test. Logistic regression, conditional logistic regression as well as stratified regression are all considered. These various models, simple proportional hazards model, stratified models, and time-dependent models can all be exploited in order to better evaluate risk factors, how they interrelate, and how they relate to disease incidence in various situations. The use of registry data is looked at in relation to the estimation of survival in specific risk sub-groups. This motivates the topic of relative survival.

5.2 Context and motivation

By considering the time variable to correspond to age we are able to set up powerful regression models that can be used in epidemiological applications, including case-control studies, retrospective, and prospective studies. For many chronic diseases, such as cancer, age is the most important risk factor and we can control for it in several ways. Just as we do for survival time, where we sequentially condition on time to remove its influence, the same thing can be done for age. Relative risk models used in epidemiology come under these same headings. For relative risk models the time component is usually taken to be age and great generalization, e.g., period or cohort analysis is readily accomplished. Time-dependent covariates, $Z(t)$, in combination with the at-risk indicator, $Y(t)$,

can be used to describe states. Multistate models in which subjects can move in and out of different states, or into an absorbing state such as death, can then be analyzed using the same methodology.

5.3 Odds ratio, relative risk, and 2×2 tables

For arbitrary random variables X and Y with joint density $f(x,y)$, conditional densities $g(x|y)$ and $h(y|x)$, marginal densities $v(x)$ and $w(y)$, we know that

$$f(x,y) = g(x|y)w(y) = h(y|x)v(x),$$

so that, in the context of postulating a model for the pair (X,Y) , we see that there are two natural potential characterizations. Recalling the discussion from Section 2.3 note that, for survival studies, our interest in the binary pair (T,Z) , time and covariate, can be seen equivalently from the viewpoint of the conditional distribution of time given the covariate, along with the marginal distribution of the covariate, or from the viewpoint of the conditional distribution of the covariate given time, along with the marginal distribution of time. This equivalence we exploit in setting up inference where, even though the physical problem concerns time given the covariate, our analysis describes the distribution of the covariate given time.

In epidemiological studies the variable time T is typically taken to be age. Calendar time and time elapsed from some origin may also be used but, mostly, the purpose is to control for age in any comparisons we wish to make. Usually we will consider rates of incidence of some disease within small age groups or possibly, via the use of models, for a large range of values of age. Unlike the relatively artificial construction of survival analysis which exploits the equivalent ways of expressing joint distributions, in epidemiological studies our interest naturally falls on the rates of incidence for different values of Z given fixed values of age T . It is not then surprising that the estimating equations we work with turn out to be essentially the same for the two situations.

The main results of proportional hazards regression, focused on the conditional distribution of the covariate given time, rather than the other way around, apply more immediately and in a more natural way in epidemiology than in survival type studies. We return to this in the later chapters that consider inference more closely. One important distinction, although already well catered for by use of our “at risk” indicator variables, is that for epidemiological studies the subjects in different risk sets are often distinct subjects. This is unlike the situation for survival studies where the risk sets are typically nested. Even so, as we will see, the form of the equations is the same, and software which allows an analysis of survival data will also allow an analysis of certain problems in epidemiology.

For a binary outcome, indicated by $Y = 1$ or $Y = 0$, and a binary risk or exposure factor, $Z = 1$ or $Z = 0$, the relative risk is defined as the ratio of the probabilities $P(Y = 1|Z = 1)/P(Y = 1|Z = 0)$ and the, related, odds ratio ψ as

$$\psi = \frac{P(Y = 1|Z = 1)P(Y = 0|Z = 0)}{P(Y = 1|Z = 0)P(Y = 0|Z = 1)}.$$

In the above and in what follows, in order for the notation not to become too cluttered, we write $\Pr(A) = P(A)$. Under a “rare disease assumption”, i.e., when $P(Y = 0|Z = 0)$ and $P(Y = 0|Z = 1)$ are close to 1, then the odds ratio and relative risk approximate one another.

One reason for being interested in the odds ratio, as a measure of the impact of different levels of the covariate (risk factor) Z follows from the identity

$$\frac{P(Y = 1|Z = 1)P(Y = 0|Z = 0)}{P(Y = 1|Z = 0)P(Y = 0|Z = 1)} = \frac{P(Z = 1|Y = 1)P(Z = 0|Y = 0)}{P(Z = 1|Y = 0)P(Z = 0|Y = 1)}. \quad (5.1)$$

Thus, the impact of different levels of the risk factor Z can equally well be estimated by studying groups defined on the basis of this same risk factor and their corresponding incidence rates of $Y = 1$. This provides the rationale for the case-control study in which, in order to estimate ψ , we make our observations on Z over fixed groups of cases and controls (distribution of Y fixed), rather than the more natural, but practically difficult if not impossible, approach of making our observations on Y for a fixed distribution of Z . Assumptions and various subtleties are involved. The subject is vast and we will not dig too deeply into this. The points we wish to underline in this section are those that establish the link between epidemiological modeling and proportional hazards regression.

SERIES OF 2×2 TABLES

The most elementary presentation of data arising from either a prospective study (distribution of Z fixed) or a case-control study (distribution of Y fixed) is in the form of a 2×2 contingency table in which the counts of the number of observations are expressed. Estimated probabilities or proportions of interest are readily calculated.

In Table 5.1, $a_{1*} = a_{11} + a_{12}$, $a_{2*} = a_{21} + a_{22}$, $a_{*1} = a_{11} + a_{21}$, $a_{*2} = a_{12} + a_{22}$ and $a_{**} = a_{1*} + a_{2*} = a_{*1} + a_{*2}$. For prospective studies the proportions a_{11}/a_{*1} and a_{12}/a_{*2} estimate the probabilities of being a case ($Y = 1$) for both exposure groups while, for case-control studies, the proportions a_{11}/a_{1*} and a_{21}/a_{2*} estimate the probabilities of exhibiting the risk or exposure factor

	$Z = 1$	$Z = 0$	totals
$Y = 1$	a_{11}	a_{12}	a_{1*}
$Y = 0$	a_{21}	a_{22}	a_{2*}
Totals	a_{*1}	a_{*2}	a_{**}

Table 5.1: Basic 2×2 table for cases ($Y = 1$) and controls ($Y = 0$).

($Z = 1$) for both cases and controls. For both types of studies we can estimate ψ by the ratio $(a_{11}a_{22})/(a_{21}a_{12})$, which is also the numerator of the usual chi-squared test for equality of the two probabilities. If we reject the null hypothesis of the equality of the two probabilities we may wish to say something about how different they are based on the data from the table.

As explained below, in Section 5.4, quantifying the difference between two proportions is not best done via the most obvious, and simple, arithmetic difference. There is room for more than one approach, the simple arithmetic difference being perfectly acceptable when sample sizes are large enough to be able to use the De Moivre-Laplace approximation (Appendix C.2) but, more generally, the most logical in our context is to express everything in terms of the odds ratio. We can then exploit the following theorem:

Theorem 5.1. *Taking all the marginal totals as fixed, the conditional distribution of a_{11} is written*

$$P(a|a_{1*}, a_{2*}, a_{*1}, a_{*2}) = \binom{a_{1*}}{a} \binom{a_{2*}}{a_{*1}-a} \psi^a / \sum_u \binom{a_{1*}}{u} \binom{a_{2*}}{a_{*1}-u} \psi^u,$$

the sum over u being over all integers compatible with the marginal totals.

The conditionality principle appears once more, in this instance in the form of fixed margins. The appropriateness of such conditioning, as in other cases, can be open to discussion. And again, insightful conditioning has greatly simplified the inferential structure. Following conditioning of the margins, it is only necessary to study the distribution of any single entry in the 2×2 table, the other entries being then determined. This kind of approach forms the basis of the well-known Fisher's exact test. It is usual to study the distribution of a_{11} . A non-linear estimating equation can be based on $a_{11} - E(a_{11})$, expectation obtained from Theorem 5.1, and from which we can estimate ψ and associate a variance term with the estimator. The non-linearity of the estimating equation, the only approximate normality of the estimator, and the involved form of variance expressions has led to much work in the methodological epidemiology literature; improving the approximations, obtaining greater robustness and so on. However, all of this can be dealt with in the context of a proportional hazards (conditional logistic)

Table i	$Z = 1$	$Z = 0$	Totals
$Y = 1$	$a_{11}(i)$	$a_{12}(i)$	$a_{1*}(i)$
$Y = 0$	$a_{21}(i)$	$a_{22}(i)$	$a_{2*}(i)$
Totals	$a_{*1}(i)$	$a_{*2}(i)$	$a_{**}(i)$

Table 5.2: 2×2 table for i th age group of cases and controls.

regression model. Since it would seem more satisfactory to work with a single structure rather than deal with problems on a case-by-case basis, our recommendation is to work with proportional and non-proportional hazards models. Not only does a model enable us to more succinctly express the several assumptions which we may be making, it offers, more readily, well-established ways of investigating the validity of any such assumptions. In addition the framework for studying questions such as explained variation, explained randomness and partial measures of these is clear and requires no new work.

The “rare disease” assumption, allowing the odds ratio and relative risk to approximate one another, is not necessary in general. However, the assumption can be made to hold quite easily and is therefore not restrictive. To do this we construct fine strata, within which the probabilities $P(Y = 0|Z = 0)$ and $P(Y = 0|Z = 1)$ can be taken to be close to 1. For each stratum, or table, we have a 2×2 table as in Table 5.2, indexed by i . Each table provides an estimate of relative risk at that stratum level and, assuming that the relative risk itself does not depend upon this stratum, although the actual probabilities themselves composing the relative risk definition may themselves depend upon strata, then the problem is putting all these estimates of the same thing into a single expression. The most common such expression for this purpose is the Mantel-Haenszel estimate of relative risk.

MANTEL-HAENSZEL ESTIMATE OF RELATIVE RISK

The, now famous, Mantel-Haenszel estimate of relative risk was described by Mantel and Haenszel (1959) and is particularly simple to calculate. Let us first refer to the entries of observed counts in Table 5.3.

Table i	$Z = 1$	$Z = 0$	Totals
$Y = 1$	$a_{11}(i)$	$a_{12}(i)$	$a_{1*}(i)$
$Y = 0$	$a_{21}(i)$	$a_{22}(i)$	$a_{2*}(i)$
Totals	$a_{*1}(i)$	$a_{*2}(i)$	$a_{**}(i)$

Table i	$Z = 1$	$Z = 0$	Totals
$Y = 1$	$e_{11}(i)$	$e_{12}(i)$	$e_{1*}(i)$
$Y = 0$	$e_{21}(i)$	$e_{22}(i)$	$e_{2*}(i)$
Totals	$e_{*1}(i)$	$e_{*2}(i)$	$e_{**}(i)$

Table 5.3: 2×2 table for i th age group of cases and controls. Left-hand table: observed counts. Right-hand table: expected counts.

If we first define for the i th sub-table $R_i = a_{11}(i)a_{22}(i)/a_{**}(i)$ and $S_i = a_{12}(i)a_{21}(i)/a_{**}(i)$, then the Mantel-Haenszel summary relative risk estimate across the tables is given by $\hat{\psi}_{MH} = \sum_i R_i / \sum_i S_i$. Breslow (1996) makes the following useful observations concerning $\hat{\psi}_{MH}$ and $\hat{\beta}_{MH} = \hat{\psi}_{MH}$. First, $E(R_i) = \psi_i E(S_i)$ where the true odds ratio in the i th table is given by ψ_i . When all of these odds ratios coincide then $\hat{\psi}_{MH}$ is the solution to the unbiased estimating equation; $R - \psi S = 0$, where $R = \sum_i R_i$ and $S = \sum_i S_i$.

Under an assumption of binomial sampling, Breslow shows that the variances of the individual contributions to the estimating equation are such that the quantity $2a_{**}^2(i)\text{Var}(R_i - \psi S_i)$ can be equated to

$$E\{[a_{11}(i)a_{22}(i) + \psi a_{12}(i)a_{21}(i)][a_{11}(i) + a_{22}(i) + \psi(a_{12}(i) + a_{21}(i))]\},$$

from which, by a simple application of the delta method (Appendix A.10), we can obtain estimates of the variance of $\hat{\psi}_{MH}$.

5.4 Logistic regression and proportional hazards

Without any loss in generality we can express the two probabilities of interest, $P(Y = 1|Z = 1)$ and $P(Y = 1|Z = 0)$ as simple power transforms of one another. This follows, since, whatever the true values of these probabilities, there exists some positive number α such that $P(Y = 1|Z = 1) = P(Y = 1|Z = 0)^\alpha$. The parameter α is constrained to be positive in order that the probabilities themselves remain between 0 and 1. To eliminate any potential dangers that may arise, particularly in the estimation context where, even though the true value of α is positive, the estimate itself may not be, a good strategy is to re-express this parameter as $\alpha = \exp(\beta)$. We then have

$$\log -\{\log P(Y = 1|Z = 1)\} = \log -\{\log P(Y = 1|Z = 0)\} + \beta. \quad (5.2)$$

The parameter β can then be interpreted as a linear shift in the log-log transformation of the probabilities, and can take any value between $-\infty$ and ∞ , the inverse transformations being one-to-one and guaranteed to lie in the interval $(0,1)$. An alternative model to the above is

$$\text{logit } P(Y = 1|Z = 1) = \text{logit } P(Y = 1|Z = 0) + \beta. \quad (5.3)$$

where the logit transformation, again one-to-one, is defined by $\text{logit } \theta = \log\{\theta/(1-\theta)\}$. Although a natural model, the model of Equation 5.2 is not usually preferred to that of Equation 5.3, motivated in an analogous way (i.e., avoiding

constraints) but having a slight advantage from the viewpoint of interpretation. This is because the parameter β is the logarithm of the odds ratio, i.e., $\beta = \log \psi$.

In the light of the equivalence of the odds for disease given the risk factor and the odds for the risk factor given the disease, as expressed in Equation 5.1, we conclude immediately that, equivalent to the above model involving β , expressed in Equation 5.3, we have a model expressing the conditional probability of Z given Y and using the same β . This highlights an important feature of proportional hazards modeling whereby we focus attention on the conditional distribution of the covariates given an event yet, when thinking of the applied physical problem behind the analysis, we would think more naturally in terms of the conditional distribution of the event given the covariates. The essential point is that the unknown regression parameter, β , of interest to us is the same for either situation so that in place of Equation 5.3, we can write

$$\text{logit } P(Z = 1|Y = 1) = \text{logit } P(Z = 1|Y = 0) + \beta. \quad (5.4)$$

Since the groups are indicated by a binary Z , we can exploit this in order to obtain the more concise notation, now common for such models, whereby

$$\text{logit } P(Y = 1|Z) = \text{logit } P(Y = 1|Z = 0) + \beta Z. \quad (5.5)$$

As we have tried, in as much as is possible throughout this text, to restrict attention to a single explanatory variable, this is once more the case here. Extension to multiple explanatory variables, or risk factors, is immediate and, apart from the notation becoming more cumbersome, there are no other concepts to which to give thought. We write the model down, as above in Equation 5.5, and use several binary factors Z (Z now a vector) to describe the different group levels. The coefficients β (β now a vector) then allow the overall odds ratio to be modeled or, allows the modeling of partial odds ratios whereby certain risk factors are included in the model, and our interest focuses on those remaining after having taken account of those already included. The above model can also be written in the form

$$\frac{P(Y = 1|Z)}{1 - P(Y = 1|Z)} = \exp(\beta_0 + \beta Z), \quad (5.6)$$

where $\beta_0 = \text{logit } P(Y = 1|Z = 0)$. Maintaining an analogy with the usual linear model we can interpret β_0 as an intercept, simply a function of the risk for a “baseline” group defined by $Z = 0$.

Assigning the value $Z = 0$ to some group and thereby giving that group baseline status is, naturally, quite arbitrary and there is nothing special about the baseline group apart from the fact that we define it as such. We are at liberty to make other choices and, in all events, the only quantities of real interest to us are relative ones. In giving thought to the different modeling possibilities that arise when dealing with a multivariate Z , the exact same kind of considerations,

already described via several tables in the section on modeling multivariate problems will guide us (see Section 4.6 and those immediately following it). Rather than repeat or reformulate those ideas again here, the reader, interested in these aspects of epidemiological modeling, is advised to go over those earlier sections. Indeed, without a solid understanding as to why we choose to work with a particular model rather than another, and as to what the different models imply concerning the complex inter-relationships between the underlying probabilities, it is not really possible to carry out successful modeling in epidemiology.

STRATIFIED AND CONDITIONAL LOGISTIC REGRESSION

In the above model, and Z being multivariate, we may wish to include alongside the main factors under study, known risk factors, and particularly risk factors such as age, or period effects, for which we would like to control. Often age alone is the strongest factor and its effect can be such that the associated errors of estimation in quantifying its impact can drown the effect of weaker risk factors. A powerful approach in controlling for such factors, S , is to appeal to the idea of stratification. This means that analysis is carried out at each level of S and, within a level, we make the same set of assumptions concerning the principle factors under study. We write

$$\frac{P(Y = 1|Z, S)}{1 - P(Y = 1|Z, S)} = \exp(\beta_0 + \beta Z), \quad (5.7)$$

where, in the same way as before, $\beta_0 = \text{logit } P(Y = 1|Z = 0, S)$. The important aspect of a stratified model is that the levels of S only appear on the left-hand side of the equation.

We might conclude that this is the same model as the previous one but it is not quite and, in later discussions on inference, we see that it does impact the way in which inferences are made. It also impacts interpretation. In the simpler cases, in as far as β is concerned, the stratified model is exactly equivalent to a regular logistic model if we include in the regression function indicator variables, of dimension one less than the number of strata. However, when the number of strata is large, the use of the stratified model enables us to bypass estimation of the stratum-level effects. If these are not of real interest then this may be useful in that it can result in gains in estimating efficiency even though the underlying models may be equivalent. In a rough intuitive sense we are spending the available estimating power on the estimation of many less parameters, thereby increasing the precision of each one. This underlines an important point in that the question of stratification is more to do with inference than the setting up of the model itself.

This last remark is even more true when we speak of conditional logistic regression. The model will look almost the same as the unconditional one but the process of inference will be quite different. Suppose we have a large number of strata, very often in this context defined by age. A full model would be as in

Equation 5.6, including in addition to the risk factor vector Z , a vector parameter of indicator variables of dimension one less than the number of strata. Within each age group, for the sake of argument let's say age group i , we have the simple logistic model. However, rather than write down the likelihood in terms of the products $P(Y = 1|Z)$ and $P(Y = 0|Z)$ we consider a different probability upon which to construct the likelihood, namely the probability that the event of interest, the outcome or case in other words, occurred on an individual (in particular the very individual for whom the event *did* occur, given that one event occurred among the set $S\{i\}$ of the $a_{**}(i)$ cases and controls. Denoting Z_i to be the risk factor for the case, corresponding to the age group i , then this probability is simply $\exp(\beta Z_i)/\sum I[j \in S\{i\}] \exp(\beta Z_j)$. The likelihood is then the product of such terms across the number of different age groups for which a case was selected.

If we carefully define the “at-risk” indicator $Y(t)$ where t now represents age, we can write the conditional likelihood as

$$L(\beta) = \prod_{i=1}^n \left\{ \frac{\exp(\beta Z_i)}{\sum_{j=1}^n Y_j(X_i) \exp(\beta Z_j)} \right\}^{\delta_i}. \quad (5.8)$$

Here we take the at-risk indicator function to be zero unless, for the subject j , X_j has the same age, or is among the same age group as that given by X_i . In this case the at-risk indicator $Y_j(X_i)$ takes the value one. To begin with, we assume that there is only a single case per age group, that the ages are distinct between age groups, and that, for individual i , the indicator δ_i takes the value one if this individual is a case. Use of the δ_i would enable us to include in an analysis sets of controls for which there was no case. This would be of no value in the simplest case but, generalizing the ideas along exactly the same lines as for standard proportional hazards models, we could easily work with indicators $Y(t)$ taking the value one for all values less than t and becoming zero if the subject becomes incident or is removed from the study. A subject is then able to make contributions to the likelihood at different values of t , i.e., at different ages, and appears therefore in different sets of controls. Indeed, the use of the risk indicator $Y(t)$ can be generalized readily to other complex situations.

One example is to allow it to depend on two time variables, for example, an age and a cohort effect, denoting this as $Y(t,u)$. Comparisons are then made between individuals having the same age and cohort status. Another useful generalization of $Y(t)$ is where individuals go on and off risk, either because they leave the risk set for a given period or, possibly, because their status cannot be ascertained. Judicious use of the at-risk indicator Y makes it possible then to analyze many types of data that, at first glance, would seem quite intractable. This can be of particular value in longitudinal studies involving time-dependent measurements where, in order to carry out unmodified analysis we would need, at

each observed failure time, the time-dependent covariate values for all subjects at risk. These would not typically all be available. A solution based on interpolation, assuming that measurements do not behave too erratically, is often employed. Alternatively we can allow for subjects for whom, at an event time, no reliable measurement is available, to simply temporarily leave the risk set, returning later when measurements have been made.

The striking thing to note about the above conditional likelihood is that it coincides with the expression for the partial likelihood. This is no real coincidence of course and the main theorems of proportional hazards models apply equally well here. This result anticipates a very important concept, and that is the idea of sampling from the risk set. The difference between the $Y(t)$ in a classical survival study, where it is equal to one as long as the subject is under study and then drops to zero, as opposed to the $Y(t)$ in the simple epidemiological application in which it is zero most of time, taking the value one when indicating the appropriate age group, is a small one. It can be equated with having taken a small random sample from a conceptually much larger group followed since time (age) is zero. On the basis of the above conditional likelihood we obtain the estimating equation

$$U(\beta) = \sum_{i=1}^n \delta_i \left\{ Z_i - \frac{\sum_{j=1}^n Y_j(X_i) Z_j \exp(\beta Z_j)}{\sum_{j=1}^n Y_j(X_i) \exp(\beta Z_j)} \right\}, \quad (5.9)$$

which we equate to zero in order to estimate β . The equation contrasts the same quantities written down in Table 5.3 in which the expectations are taken with respect to the model. The estimating equations are then essentially the same as those given in Table 5.3 for the Mantel-Haenszel estimator. Furthermore, taking the second derivative of the expression for the log-likelihood, we have that $I(\beta) = \sum_{i=1}^n \delta_i I_i(\beta)$ where

$$I_i(\beta) = \frac{\sum_{j=1}^n Y_j(X_i) Z_j^2 \exp(\beta Z_j)}{\sum_{j=1}^n Y_j(X_i) \exp(\beta Z_j)} - \left\{ \frac{\sum_{j=1}^n Y_j(X_i) Z_j \exp(\beta Z_j)}{\sum_{j=1}^n Y_j(X_i) \exp(\beta Z_j)} \right\}^2, \quad (5.10)$$

then $I(\beta) = \sum_{i=1}^n \delta_i I_i(\beta)$. Inferences can then be carried out on the basis of these expressions. In fact, once we have established the link between the applied problem in epidemiology and its description via a proportional hazards model, we can then appeal to those model-building techniques (explained variation, explained randomness, goodness of fit, conditional survivorship function etc.) which we use for applications in time to event analysis. In this context the building of models in epidemiology is no less important, and no less delicate, than the building of models in clinical research. Although many of the regression modeling ideas came to the field of epidemiology later than they did for clinical research, several of the deeper concepts such as stratification, risk-set sampling and non-nested studies were already well known to epidemiologists. As a result some

of the most important contributions to the survival literature have come from epidemiologists. For readers looking for more of an epidemiological flavor to survival problems we would recommend taking a look at Breslow (1978), Annesi and Lellouch (1989), Rosenberg and Anderson (2010), Cologne et al. (2012), Xue et al. (2013), Moolgavkar and Lau (2018) and Fang and Wang (2020).

5.5 Survival in specific groups

The issue of competing risks is never far from our attention, no less so when we discuss problems in epidemiology. Almost without exception, outcomes other than that of the investigators' main focus, will hinder its observation and will bring into play one of the approaches, CRM-I, CRM-J or CRM-ID (Section 2.6), in order to enable our work to progress. Some obvious questions such as how would a given population fare if we were able to remove some particular cause of death turns out to be almost impossibly difficult to even correctly formulate. How probable is it that a member of some high-risk group will succumb to the disease is again a question that is very difficult to address. This may at first seem surprising. Survival methods can help and, in this and the following section dealing with genetic epidemiology, we will see how very great care is needed to avoid making quite misleading inferences. One case in point is that of the so-called breast cancer susceptibility genes, BRCA1 and BRCA2. Any additional risk to carriers, if indeed there is any additional risk, has been very greatly exaggerated.

RELATIVE SURVIVAL AND REGISTRY DATA

Under CRM-I (see Section 1.5) we can follow a large number of subjects for an interval of short duration, noting the number available to suffer the event of interest during the interval, together with the number of cases occurring. CRM-I enables us to obtain an estimate (slightly biased downwards due to censorings during the interval) of the disease rate as the ratio of the number of events to the number available at the beginning of the interval. The smaller the interval size the better but, of course, in practice we are limited by the amount of data available as well as the mechanisms by which we can make accurate recordings. Typically these intervals are of 5-years durations.

At some time and place we may have large enough observations to enable accurate estimates of disease rates. Under some assumptions differential rates can be smoothed, often using Bayesian techniques, and this is a whole field that has many overlaps with the main strands of this text. Suppose, for some population, we have large enough estimates from registry data, that we can consider the survival probabilities known, denoted as $S_0(t)$. Within this population we have all disease factors acting, and represented in accordance to their observed frequencies. We may be interested to know how some specific sub-group, say those who have been incident for disease Z fare with respect to the whole population at large. We might postulate an elementary model for this as;

$$S(t|Z) = S_0(t) \times \alpha(t) \quad (5.11)$$

where $S(t|Z)$ is the survival of the sub-group of interest and $S_0(t)$ the population survival. For this model, we refer to $\alpha(t)$ as relative survival. It is described as a measure of net survival for this group in the absence of all those other factors that influence the reference group. We often assume that the size of the group Z is not large enough to have a significant impact on $S_0(t)$. In consequence, the sub-group, $S(t|Z)$ forms a negligible component of $S_0(t)$ so that $\alpha(t)$ will be described as the ratio of survival in the sub-group with respect to the expected population survival unaffected by whatever handicap the sub-group suffers from. Equation 5.11 is very popular in this context due to its simple interpretation. As pointed out in Section 2.4 it is generally preferable to appeal to the log-minus-log transformation as the basis for a model. We would then have a non-proportional hazards model alternative to Equation 5.11 as

$$\log\{-\log S(t|Z)\} = \beta(t) + \log\{-\log S_0(t)\}. \quad (5.12)$$

Aside from taking logs it is no more difficult to estimate $\beta(t)$ than $\alpha(t)$ and, should $\beta(t)$ appear to be reasonably constant over time, then we could appeal to

$t=\text{age}$	$n(t)$	$n_C(t)$	$I(t)$	$I_H(t)$
00-05	10,000,000	0	0	0
05-10	9,895,159	2	1 in 4,947,580	16 in 4,947,580
10-15	9,887,872	26	1 in 380,303	16 in 4,947,580
15-20	9,879,578	173	1 in 57,107	16 in 57,107
20-25	9,862,951	1087	1 in 9,073	16 in 9,073
25-30	9,840,178	4795	1 in 2,052	16 in 2,052
30-35	9,810,380	14,529	1 in 675	16 in 675
35-40	9,763,506	34,102	1 in 286	16 in 286
40-45	9,681,029	69,466	1 in 139	16 in 139
45-50	9,538,425	106,633	1 in 89	16 in 89
50-55	9,324,938	136,860	1 in 68	16 in 68
55-60	9,034,317	168,931	1 in 53	16 in 53
60-65	8,634,150	189,344	1 in 46	16 in 46
65-70	8,088,535	198,438	1 in 41	16 in 41
70-75	7,365,731	189,195	1 in 39	16 in 39
00-75			1 in 8 = 12.1%	7 in 8 = 87%

Table 5.4: Based on NIH-SEER data. $n(t) = \#$ at-risk at age t , $n_C(t) = \#$ cases between t and $t+5$, $I(t) =$ overall incidence rate, $I_H(t) =$ incidence rate (high risk, BRCA1/2) on the basis of a relative risk of 16.

a simple proportional hazards model to describe the observations. An important contribution in this field is that of Perme et al. (2012).

USING MODELS CRM-I AND CRM-J TO EXPLOIT REGISTRY DATA

In epidemiology, the age-specific failure rate, $\lambda(t)$, characterizes the distribution of the random variable T , the time until the relevant event. Our illustration here focuses on breast cancer. In the simplest construct the only cause of death is breast cancer, so that the probability that the incidence of breast cancer occurs before time t is given by $F(t) = 1 - \exp\{-\Lambda(t)\}$ where $\Lambda(t)$ is the cumulative risk over time. Note that $\Lambda(t) = \int_0^t \lambda(u)du$. Under CRM-I we consider censoring random variables, C_1, \dots, C_k , describing the occurrence of competing events. If any one of them occurs before $T = t$, then we are unable to observe our outcome of interest should it take place itself any time later than t . The risk function of interest—in as much as it's the only one that can be estimated by real data under CRM-I—is then rather different to $\lambda(t)$ and can be expressed as

$$\lambda\{t|C_{\min} > t\} = \lim_{\Delta t \rightarrow 0^+} (\Delta t)^{-1} \Pr\{t < T < t + \Delta t | T > t, C_{\min} > t\}.$$

where $C_{\min} = \min(C_1, \dots, C_k)$. The conditioning event $C_{\min} > t$ is of central importance in epidemiological studies since, in practical investigations—in particular the compilation of registry data—all our observations at time t have necessarily been conditioned by the events, $T > t$ and $C_{\min} > t$. All associated probabilities are also necessarily conditional. But note that, under an independent censoring mechanism, $\lambda(t|C_{\min} > t) = \lambda(t)$. This result is crucial since, given independence of the competing causes of failure, and, by partitioning the interval $(0, t)$ into $t_0 < t_1 < \dots < t_m$ where $t_0 = 0$ and $t_m = t$ then, as a consequence of the above expression we can write:

$$\Lambda(t) = \int_0^t \lambda\{u|C_{\min} > u\} du \approx \sum_{j=1}^m \Pr\{t_{j-1} < T < t_j | T > t_{j-1}, C_{\min} > t_{j-1}\} \quad (5.13)$$

so that we can approximate $F(t)$ via $F(t) = 1 - \exp\{-\Lambda(t)\}$. When using this approximation with real data, the quality of the approximation can be investigated on a case-by-case basis. It is common, in epidemiological applications, to use gaps, $t_j - t_{j-1}$, as large as 5 years. Equation 5.13 can be directly used and interval probability estimates taken directly from registry data. Data such as the SEER data, for very large samples, indicate those subjects entering each 5-year age interval, i.e., those satisfying the conditioning restriction in Equation 5.13, and just how many observed cases could be counted during that 5-year interval. For each component to the sum we have an empirical estimate of the required probability, the numerator being simply the number of cases seen during the

interval and the denominator typically the number of individuals satisfying the criteria to be “at-risk” at the beginning of the interval. Attempts to improve the approximation may involve looking at the average number at-risk throughout the interval rather than just at the beginning or some other model to improve the numerical approximation to the integral. Working with model CRM-J, rather than CRM-I will, in many cases, and these can include breast cancer, have little impact on the estimation of the cumulative incidence rate (Satagopan and Auerbach, 2004).

Equation 5.13 is the basis for estimating probabilities of cancer incidence over given periods. Cancer registry data, such as the SEER data set, provide for the numbers exposed to risk of breast and other cancers and the number of cases observed during those 5-year intervals. These two numbers provide the numerator and the denominator to our conditional probability estimates. We can first use Table 5.4 to confirm the widely quoted figure of one woman in eight will have breast cancer in her life. If we let $H = 1/39 + 1/41 + 1/46 + \dots + 1/4,947,580$, in Equation 5.13, then we find that $1 - \exp(-H) = 0.121 \approx 1/8$, a much quoted result in not just scientific journals but in everyday popular sources. Note that this estimated probability is unaltered in the first 3 figures if we start the clock from age 20 rather than age zero. The reason for this can be seen in the table since those early rates are so small as to be negligible. On the other hand, at the top end, the summation stops for women over 75 years of age. Some authors go beyond 75 years when evaluating lifetime risk (Brose et al. (2002), for example, calculate out to age 110 years).

Our estimates are limited by the rapidly declining number of observations for the higher age groups but, theoretically, if we were able to estimate the death rate due to cancer without fixing some upper limit then it would go to one hundred percent. This is easily seen intuitively since, everyone has to die eventually, and the calculations do not allow for anyone to die of causes other than breast cancer (in technical language, other causes of death are “censored out” under CRM-I when we calculate the risk sets). When appealing to model CRM-J, instead of CRM-I, we will need to introduce into the calculation the marginal survival time to either death or breast cancer incidence (see Section 2.6) which is often approximated from life tables by registry death rates. In this case the choice of model did not have any significant impact on estimation.

5.6 Genetic epidemiology

It is not easy to find a subject more fascinating than genetics. Seyerle and Avery (2013) describe genetic epidemiology as being at the crossroads of epidemiology and genetics, the discipline’s aim being to identify the myriad of relationships between inherited risk factors and disease etiology. The field is truly a vast one and our purpose here is very limited—to consider the contribution to such an aim that can be made by modern techniques of survival analysis, proportional and non-

proportional hazards models in particular. Many of the more notable advances in genetics have not come from statistical modeling but from more direct combinatorial probabilistic calculation. Mendelian inheritance, Hardy-Weinberg equilibrium equations, and pedigree analysis all come under this heading. The genetic nature of certain diseases, their observed frequencies being well explained by given mechanisms, for example, a recessive normal allele together with a dominant abnormal allele, have added greatly to our understanding of diseases such as Huntington's. This is all the more significant as diseases like Huntington's may not become manifest until well beyond the sufferer's age of producing offspring.

Now, while we can well describe the example of Huntington's using probability models, the probabilistic mechanism of these models is fully understood. It is a combinatorial one. The situation is very different when we consider so-called "cancer susceptibility" genes. Unlike Huntington's where the dominant inherited gene has been traced to a mutation of a specific nature—an approximate tripling of the frequency of a particular nucleotide sequence—when we consider susceptibility genes we have nothing particular in mind and while several types of polymorphisms may be considered unharful others will be considered "deleterious". This number has grown over the years and the number of deleterious mutations believed to be associated with breast cancer, some on chromosome 17 and some on chromosome 13, are now estimated to be in the hundreds.

ESTABLISHING FAMILY ASSOCIATION

Inheriting a defective allele, considered either dominant or recessive, is well described by simple probabilistic models. Observations on families are necessarily very sparse and so studies will rely on collecting families and appealing to likelihood models to provide complete descriptions of the data. Even in the absence of any clear indication as to the location of an implicated gene, or genes, the distribution of diseases such as Huntington's can be accurately described. This is a very different situation to that we encounter when we set out to investigate so-called "cancer susceptibility" genes. In this latter situation the room for errors is very large; so much so that some of the more alarming claims concerning the impending risks faced by carriers of, for example, the BRCA1 and BRCA2 mutations, need to be taken with a great deal of circumspection. What is more, these errors, whether statistical (lack of precision), logical (the presence of flaws in the reasoning), or interpretative (the difficulty in translating results into something understandable) will amplify one another. Without some understanding, however limited, of these errors and their sources, the carrier is in no position to make any kind of informed decision. A decision based on little more than some simple take-home number, such as lifetime risk, would seem to be most unwise. The techniques of survival analysis can provide deeper insight. We describe this below but, first, let us revisit those approaches that will greatly overestimate the degree of family association.

Early studies on family association in breast cancer used case-control methodology. Making the usual efforts to control for age and possibly other risk factors, the following question was put to both the case and the control subject: do you have one or more family members diagnosed with breast cancer? The estimated probability of a positive response for the cases turned out to be more than twice that of the controls. When the question was changed to two or more family members, this ratio increased further to more than three. At first glance this may appear to provide solid evidence in favor of a family association with a strong relative risk. The logic is however flawed. To see this, we first define Y to be the total number of cases in the family, i.e., at least one for the case and no more than $n - 1$, where n is the number of relevant family members, for the control. We express our null hypothesis as

$$H_0 : \Pr(Y \geq 2 | Y \geq 1) = \Pr(Y \geq 1 | Y \geq 0)$$

and we estimate the relative risk, ψ , by

$$\psi = \frac{\Pr(Y \geq 2 | Y \geq 1)}{\Pr(Y \geq 1 | Y \geq 0)}.$$

We can re-express the null hypothesis—are cases no more likely than controls to have other family members with breast cancer—as; $H_0 : \psi = 1$. Now, while the hypothesis, $H_0 : \Pr(Y \geq 2 | Y \geq 1, n) = \Pr(Y \geq 1 | Y \geq 0, n)$ will correctly control the size of the test, the more obvious formulation, $H_0 : \Pr(Y \geq 2 | Y \geq 1) = \Pr(Y \geq 1 | Y \geq 0)$ will not. Indeed, under H_0 it is not generally true that $\psi = 1$. We have that;

$$\Pr(Y \geq 2 | Y \geq 1) = \sum_n \Pr(Y \geq 2 | Y \geq 1, n) \times g(n | Y \geq 1) \quad (5.14)$$

$$\Pr(Y \geq 1 | Y \geq 0) = \sum_n \Pr(Y \geq 1 | Y \geq 0, n) \times g(n | Y \geq 0) \quad (5.15)$$

Under the null hypothesis we would like for ψ , the ratio of Equation 5.14 to Equation 5.15, to be equal to one. This however is not the case. Choosing a particular distribution for $g(n)$, specifically one with support restricted to $n = \{3, 25\}$, and with $g(3)/g(25)$ for the under 50 age group twice that of the other group, we obtain the results shown in Table 5.5. In brackets are the values taken from early works that supposedly justify the conclusion of family association. Table 5.5 is based on an independence assumption and we obtain very similar results to those taken from the literature. Such results do not therefore give support to a conclusion of family association.

Of course, the particular distribution chosen for $g(n)$ is not at all plausible. This is however beside the point since it underlines what we need to know and

# 1st degree relatives	ψ (age<50)	ψ (age>50)
1 or more	2.10 (2.14)	1.76 (1.65)
2 or more	3.71 (3.84)	2.42 (2.61)

Table 5.5: Ratio of the left hand sides of Equation 5.14 to Equation 5.15 under Binomial sampling with unknown n and independence. In brackets, ψ from case control studies, the strong similarity of the numbers tending to contradict a conclusion of dependence.

that is that, even under independence, the null hypothesis will not hold. The reason for this is that knowing that $Y \geq 1$ tells us something about n , i.e., a case is more likely to be associated with another case in the family than is a control. This dependence will disappear if we are able to condition (take as fixed) n . This is not usually feasible and, in any event, the bias would not disappear due to other less obvious biases, recall bias in particular. In statistical terms, the test cannot be shown to be unbiased (note that the meaning of the term unbiased differs when we refer to a test rather than an estimator) and will fail to control for the false-positive rate.

IDENTIFYING IMPLICATED GENE OR GENES

The null hypothesis of no family association is unlikely to be fully satisfied since many factors, environmental, socioeconomic, and other less obvious ones are potentially involved and shared to some degree within families. Had early studies failed to reject a null hypothesis of no family association that would have brought the discussion, at least for the time being, to a close. Instead, the apparent discovery of strong family association was enough to conclude that genetics was playing a part, a conclusion followed by the natural question ... which gene or genes are implicated.

Moving beyond the relatively simple, albeit difficult, question of the detection of family association onto that of the identification of responsible mutations and their location is no small task. The difficulty here can only be described as herculean. The studies leading to the identification of BRCA1 made use of quite small data sets consisting of families containing varying numbers of cases. Likelihood methods (LOD scores) were calculated to see which mutations were most strongly associated with the observed outcomes, these outcomes being the numbers of cases. The statistical problem here is that of looking for a needle in a haystack, sometimes referred to as the “fishing expedition” problem. The likelihood will not identify with any reasonable degree of accuracy a specific location on the genome. For sure, some location will correspond to the largest likelihood, while others, possibly very far removed in genetic distance, will also produce likelihoods that can lie a negligibly small distance from the maximum. The point estimate is very imprecise.

Essentially, the likelihood (LOD scores) method is of little help in cases like this. For more regular problems the log-likelihood that is used to estimate an unknown parameter will furnish us with an estimating equation, the zero of which corresponds to our parameter estimate. Small perturbations in the likelihood around its maximum correspond to small perturbations in our parameter estimate. We have continuity but, usually more, we have a log-likelihood that is twice differentiable allowing us, after verifying that the order of limiting processes can be switched, to provide an assessment of the precision of our estimate. For our genome-wide search we have no such conditions and no such reassuring estimates of precision. So much so that, for the kind of limited data available, the gene location identified by a likelihood maximum has much less chances of being the quantity we are seeking than has the sum total of all other locations. If, as a statistical technique, likelihood was up to the task, it would tell us that, even if our best estimate of the location on the genome is not estimated with perfect precision, then the location must be nearby. However, no such statement can be made, even in any approximate way, telling us that a heavy reliance on likelihood is not warranted.

In order to get a better insight into the size of the challenge here, consider an experiment quite unrelated to genetics. Suppose we have a fair deck of cards and one card is randomly chosen. The outcome of interest is obtaining a black queen. The probability of this outcome is $1/26$. Suppose now that we have a second deck, this time a very flawed deck in which half of the cards are black queens. On the basis of observations our goal is to identify the flawed deck. To this end we will use the likelihood function. The likelihood that a random draw obtains a black queen is 13 times higher in the second deck than the first. If, for each deck, 10 cards were drawn with replacement and we count the number of times X , $X = 0, \dots, 10$, we observe a black queen, then, identifying (using the likelihood) the fair ($1/26$) from the flawed ($13/26$) deck, we would choose incorrectly the flawed deck less than one time in ten thousand. We are just about certain to correctly identify the faulty deck.

How would this work though when our faulty deck is surrounded by, not one, but one hundred thousand fair decks. Choosing the deck that maximizes the likelihood will very rarely lead to the correct choice. We are far more likely to incorrectly identify a fair deck than correctly identify the faulty deck. And this despite the huge difference in risk, 13:1. Note also, returning to the case of interest, that of the BRCA gene, the relative risk is not generally believed to be this large. And of course, while the card example relates to a search across one hundred thousand candidate choices, in the case of alleles defined by nucleotide variations, we are talking about millions of choices.

An approach based on the likelihood can be enhanced using techniques such as linkage analysis that will weigh more strongly the relationships mother-daughter to mother-niece. It will nonetheless do little to refine what remains a rough and ready approach involving a lot of statistical uncertainty. Although it

can be argued that the best estimate of the implicated gene is the likelihood-based estimate, we ought not to overlook the great deal of imprecision associated with this estimate. More problematic is the highly erratic nature of an estimate, such as this one, where the parameter of interest (the gene) is not a continuous function of the log-likelihood. Very small non-significant differences from the maximum of the likelihood (LOD score) can be associated with genes that are not only very far removed from the so-called BRCA genes but they may well be found on entirely different chromosomes. The situation here is quite different from that we are familiar with when the log-likelihood is a continuous differentiable function of the unknown parameter(s), enabling us to structure reliable inference based on estimating equations. All we have here is the best point estimate of the implicated gene and no measure at all regarding the amount of sampling uncertainty that ought to be associated with it.

ESTABLISHING FAMILY ASSOCIATION USING SURVIVAL ANALYSIS

The case-control methodology used to detect and quantify the degree of family association with breast cancer is too inaccurate to allow reliable estimation. Biases such as recall bias—cases are much more likely than controls to investigate the cause of death for some remote aunt, otherwise barely heard of or mentioned—and the impracticability of controlling for family size make inferences unconvincing.

Survival analysis provides a way forward. First, we need to directly acknowledge the time (age) component of the study. A case-control study mostly ignores this although some efforts are made by age matching. The key concept here is that of a time-dependent covariate indicating, for any individual, how many other family members have been incident for breast cancer. At some age, an individual may have no family members that have been incident for breast cancer, while, six months later, the same individual could signal two family members incident for breast cancer. Not only that but, if we have the ages of the individual at which the other members became incident for breast cancer, then we are in a position to structure our time-dependent covariate, $Z(t)$ so that it can assume values, $Z(t) = 0, 1, 2, \dots$, as age t ranges from 0 to some upper limit. In this way we can see that not only can a control later become a case but, the risk factor itself is time dependent.

We consider models that will accommodate such situations in the following chapter. More than one model could be of interest and, in particular, we may wish to consider making use of stratified models where the stratification variable would be indexed to the size of the family. The logarithm of the relative risk—point and interval estimates—can be obtained from inference for the regression coefficient. A test of the null hypothesis: there is no family association, is readily expressed in terms of a test on the regression coefficient, β . The maximum likelihood estimate, $\hat{\beta}$, quantifies the observed degree of family association.

QUANTIFYING INDIVIDUAL RISK

Should a healthy individual be told that she or he carries one of the genetic mutations believed to be associated with increased risk of breast cancer the question is how to best advise them. Guidance on providing any such advice is well beyond the capacity of this work. That said, if any part of this guidance is built around a statistical assessment of the risks involved, we could say to the carrier the following: focus only on short-term risks; long-term risks are hopelessly misleading. While short-term risks can be readily understood as being the ratio of those who fall victim to the disease divided by all of those who might have become such a victim, long-term risks have no such interpretation and lean wholly on complex models built on an array of untestable assumptions that are, to say the very least, somewhat unlikely. Not only that but, whereas an estimation of short-term risks will most likely produce percentages offering some reassurance, an estimation of long-term risks, unless dealing with a statistical expert, are likely to produce percentages that provoke considerable alarm. Claims such as 45% to 65% of BRCA1/2 carriers will get breast cancer (ww5.komen.org) are very misleading and have no useful scientific foundation. No less alarmists are claims that an intervention such as a double mastectomy will reduce the probability of breast cancer by a large amount. Figures of 90% or more are often quoted. Again, such claims have no statistical or scientific foundation.

O'Quigley (2017) explains why long-term risks have no fruitful meaning that could be of any help to a carrier. In that paper the example illustrates the case of 25-year-old Sophie who has been given an 87% lifetime risk of developing breast cancer. The same data and assumptions simultaneously show that Sophie has a 99.9% chance of attaining age 30 cancer free and, if she succeeds—the chances are overwhelmingly in her favor—she then has a 98% chance of reaching age 35 cancer free. Indeed, since the biggest constituent to lifetime risk comes after age 70, Sophie's chances of reaching age 65 cancer free do not differ dramatically from that of non-carriers. And these calculations make no provision for the potentially very large statistical inaccuracies in quantifying family association as well as imprecision in the location of the implicated genes. The calculations take their starting point to be the most pessimistic one—from the wide range of estimated relative risks found in the literature, close to the highest is assumed—so that, in truth, there is very much less cause for alarm than that initially provoked in Sophie upon being first informed of her lifetime risk.

Even if we were able to provide some meaningful interpretation of lifetime risk we would have to say that, as a summary of the risks facing a carrier, it is at its best, woefully inadequate. Aside from short-term risks, there are other summary statistics, over a lifetime, that the carrier can make sense of. For example, if she imagines herself together with 4 friends of similar age and health characteristics, she would readily understand that, from the 5 of them, all else being equal, her probability of not being the first one to die would be 0.8. Death from all causes is what matters and the question now becomes, given her BRCA status,

how does that probability change. The answer to that is ... not much. It can be estimated using registry data and some calculations involving order statistics (Appendix A.9). It would greatly strengthen intuitive understanding of the implications of carrier status. The idea is to come up with quantifiers of risk that can be understood on some level as opposed to lifetime risk which a carrier will not understand. More work could be done. The example of the first to die out of a small group of friends is immediately extended to not being one of the first two and how this might change given BRCA status. As well as short-term probabilities, another useful quantity from survival analysis, is the expected remaining lifetime at some point. For a 25-year-old carrier, assuming that surgery would put them in the same position as a non-carrier, we can calculate a difference in mean remaining lifetimes of 3.7 years. We might tell a carrier that they have an 87% probability of getting breast cancer without surgery. Or that, with surgery, they will be expected to gain, on average, some 3.7 years a half-century down the line. The carrier would interpret those two statements very differently. And yet they use the very same data, the same assumptions, and the same belief as to the existence of the BRCA gene as well as the unfavorable outcome associated with many of the variant polymorphisms.

5.7 Classwork and homework

1. Consider an epidemiological application in which workers may be exposed to some carcinogen during periods in which they work in some particular environment. When not working in that particular environment their risk falls back to the same as that for the reference population. Describe this situation via a proportional hazards model with time-dependent effects. How do you suggest modifying such a model if the risk from exposure rather than falling back to the reference group once exposure is removed is believed to be cumulative?
2. Write down a conditional logistic model in which we adjust for both age and cohort effects where cohorts are grouped by intervals of births from 1930-35, 1936-40, 1940-45, etc. For such a model is it possible to answer the question: was there a peak in relative risk during the nineteen sixties?
3. Consider a study of an industrial risk factor such as asbestos exposure. How might the presence of non-proportional hazards manifest itself in such a setting. Argue for some plausible situations. How might this be checked?
4. An investigation into the potential carcinogenic properties of an industrial compound has evidence to suggest that it may influence the rate of some recurrent secondary risk factor but little evidence that it would directly impact the main outcome. An example might be recurrent liver disease, a risk factor for liver cancer, but no direct influence on liver cancer itself.

Describe this situation via the use of compartment models. Complete the description via the use of regression models. What assumptions do you need to make? How can you go about testing the validity of those assumptions?

5. A biological marker is believed to be raised in the presence of tumor activity. Studies on the marginal distribution of this marker show it to have a highly skewed distribution. Also, there is a lower threshold below which values cannot be detected. What kind of model might be of value in trying to detect and quantify the role of this marker in cancer incidence.
6. Simulate the number of cases per family in the following way. First choose family size, n , based on a mixture of two Poisson laws: one with a mean equal to 6 and one with a mean equal to 20 and where $P(20) = 1 - P(8) = 0.1$. Next, use a Binomial law $\mathbb{B}(n, p)$, where $p = 0.15$, to simulate the number of observed cases. Note down this number and repeat for 200 families. Finally, produce a histogram of the observed number of cases per family; 0, 1, 2 ...
7. For the data obtained from the previous exercise, ask a classmate or colleague to look at the data and, in the absence of any further analysis, to judge whether or not there appears to be evidence of a family effect. Do they consider this effect to be absent or weak, moderate to strong, or very strong? Do not provide any explanation as to how the data were obtained.
8. If given data like that of the previous question and no indication as to how the data were obtained how might you go about testing a null hypothesis—there is no family association. What would be the nature of the statistic that would lead to a consistent test while maintaining power against the alternative hypothesis.
9. Imagine that a friend, with no knowledge of epidemiology or statistics, is told that they have a 90% lifetime probability of suffering some event. Is it possible to explain to them just what this means? If not, why not? If so, then what meaning—understandable to such a friend—could be furnished.



Chapter 6

Non-proportional hazards models

6.1 Chapter summary

The most general model, described in Chapter 4 covers a very broad spread of possibilities and, in this chapter, we consider some special cases. Proportional hazards models, partially proportional hazards models (O'Quigley and Stare, 2002), stratified models, or models with frailties or random coefficients all arise as special cases of this model (Xu and O'Quigley, 2000). One useful parameterization (O'Quigley and Pessione, 1991; O'Quigley and Prentice, 1991) can be described as a non-proportional hazards model with intercept. Changepoint models are a particular form of a non-proportional hazards model with intercept (O'Quigley and Natarajan, 2004). Any model can be viewed as a special case of the general model, lying somewhere on a conceptual scale between this general model and the most parametric extreme, which would be the simple exponential model. Models can be placed on this scale according to the extent of model constraints and, for example, a random effects model would lie somewhere between a stratified model and a proportional hazards model.

6.2 Context and motivation

It is very easy to extend the simple proportional hazards model to deal with much more general problems. Allowing regression effects, $\beta(t)$, to change with time introduces great generality to our model structure. We can then view the proportional hazards structure as a special case of a non-proportional hazards structure. Intermediary cases, where some parameters do not depend on time, and some do, is also very easily accommodated. The fact that so many different model structures can be all put under the same heading brings two major benefits. The first is that a general understanding of the overall model struc-

ture allows a better understanding of the specific structure of the special cases. The second benefit of this generality is that we only need to tackle inferential questions in the broadest setting. Applications to special cases are then almost immediate. An important result is that any non-proportional hazards model can, under a suitable transformation, be made into a proportional hazards one. The transformation depends on the unknown regression function, $\beta(t)$, and so is of no obvious practical value. However, since we have optimality results for the proportional hazards setting, we can use this result, in a theoretical framework, to investigate for instance how well any test performs when contrasted to the (generally unavailable) optimal test.

6.3 Partially proportional hazards models

In the case of a single binary variable, model (4.2) and model (4.4) represent the two extremes of the modeling options open to us. Under model (4.2) there would be no model constraint and any consequent estimation techniques would amount to dealing with each level of the variable independently. Under model (4.4) we make a strong assumption about the nature of the relative hazards, an assumption that allows us to completely share information between the two levels. There exists an important class of models lying between these extremes and, in order to describe this class, let us now imagine a more complex situation; that of three groups, A , B , and C , identified by a vector Z of binary covariates; $Z = (Z_2, Z_3)$. This is summarized in Table 6.1.

	Z_2	Z_3	Log of group effect	Relative risk wrt A
Group A	0	0	0	1
Group B	1	0	β_2	$\exp(\beta_2)$
Group C	1	1	$\beta_2 + \beta_3$	$\exp(\beta_2 + \beta_3)$

Table 6.1: Coding for three groups. Impact on relative risks.

We are mainly interested in a treatment indicator Z_1 , mindful of the fact that the groups themselves may have very different survival probabilities. Under model (4.4) we have

$$\lambda(t|Z) = \lambda_0(t) \exp\{\beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3\}. \quad (6.1)$$

Our assumptions are becoming stronger in that not only are we modeling the treatment affect via β_1 but also the group effects via β_2 and β_3 . Expressing this problem in complete generality, i.e., in terms of model (4.2), we write

$$\lambda(t|Z) = \lambda_0(t) \exp\{\beta_1(t) Z_1 + \beta_2(t) Z_2 + \beta_3(t) Z_3\}. \quad (6.2)$$

Unlike the simple case of a single binary variable where our model choices were between the two extremes of model (4.2) and model (4.4), as the situation becomes more complex, we have open to us the possibility of a large number of intermediary models. These are models that make assumptions lying between model (4.2) and model (4.4) and, following O'Quigley and Stare (2002) we call them partially proportional hazards models. A model in between (6.1) and (6.2) is

$$\lambda(t|Z) = \lambda_0(t) \exp\{\beta_1 Z_1 + \beta_2(t) Z_2 + \beta_3(t) Z_3\}. \quad (6.3)$$

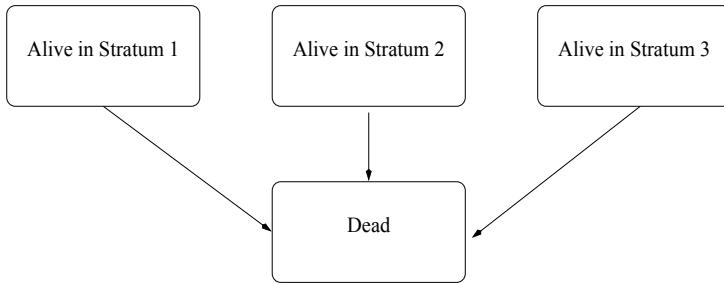


Figure 6.1: A stratified model with transitions only to death state. For all 3 transitions the log-relative risk is given by β . The base rates, however, $\lambda_{0w}, w = 1, 2, 3$ depend on the stratum w . Risk sets are stratum specific.

This model is of quite some interest in that the strongly modeled part of the equation concerns Z_1 , possibly the major focus of our study. Figure 6.1 illustrates a simple situation. The only way to leave any state is to die, the probabilities of making this transition varying from state to state and the rates of transition themselves depending on time. Below, under the heading time-dependent covariates, we consider the case where it is possible to move within states. Here it will be possible to move from a low-risk state to a high-risk state, to move from either to the death state, but to also, without having made the transition to the absorbing state, death, to move back from high-risk to low-risk.

STRATIFIED MODELS

Coming under the heading of a partially proportional hazards model is the class of models known as stratified models. In the same way these models can be considered as being situated between the two extremes of Equation 4.2 and Equation 4.3 and have been discussed by Kalbfleisch and Prentice (2002) among others. Before outlining why stratified models are simply partially proportional hazards models we recall the usual expression for the stratified model as

$$\lambda(t|Z(t), w) = \lambda_{0w}(t) \exp\{\beta Z(t)\}, \quad (6.4)$$

where w takes integer values $1, \dots, m$. If the coefficient β were allowed to depend on each stratum, indicated by w , say $\beta(w)$, then this would exactly correspond to a situation in which we consider each stratum independently, i.e., we have independent models for each stratum. This would be nothing more than m separate, independent, proportional hazards models. The estimation of $\beta(w)$ for one model has no impact on the estimation of $\beta(w)$ for another. If we take β to be common to the different strata, which is of course the whole purpose of the stratified model, then, using data, whatever we learn about one stratum tell us something about the others. They are no longer independent of one another. Stratified models are necessarily broader than (4.3), lying, in the precise sense described below, between this model and the non-proportional hazards model (4.2). To see this, consider a restricted case of model (4.2) in which we have two binary covariates $Z_1(t)$ and $Z_2(t)$. We put the restriction on the coefficient β_2 , constrained to be constant in time. The model is then

$$\lambda\{t|Z_1(t), Z_2(t)\} = \lambda_0(t) \exp\{\beta_1(t)Z_1(t) + \beta_2 Z_2(t)\}, \quad (6.5)$$

a model clearly lying, in a well-defined way, between models (4.3) and (4.2). It follows that

$$\lambda\{t|Z_1(t) = 0, Z_2(t)\} = \lambda_0(t) \exp\{\beta_2 Z_2(t)\}$$

and

$$\lambda\{t|Z_1(t) = 1, Z_2(t)\} = \lambda_0^*(t) \exp\{\beta_2 Z_2(t)\},$$

where $\lambda_0^*(t) = \lambda_0(t)e^{\beta_1(t)}$. Recoding the binary $Z_1(t)$ to take the values 1 and 2, and rewriting $\lambda_0^*(t) = \lambda_{02}(t)$, $\lambda_0(t) = \lambda_{01}(t)$ we recover the stratified PH model (6.4) for $Z_2(t)$. The argument is easily seen to be reversible and readily extended to higher dimensions so we can conclude an equivalence between the stratified model and the partially proportional hazards model in which some of the $\beta(t)$ are constrained to be constant. We can exploit this idea in the goodness of fit or the model construction context. If a PH model holds as a good approximation, then the main effect of Z_2 say, quantified by β_2 , would be similar over different stratifications of Z_1 and remain so when these stratifications are re-expressed as a PH component to a two-covariate model. Otherwise the indication is that $\beta_1(t)$ should be allowed to depend on t . The predictability of any model is studied later under the headings of explained variation and explained randomness and it is of interest to compare the predictability of a stratified model and an un-stratified one. For instance, we might ask ourselves just how strong is the predictive strength of Z_2 after having accounted for Z_1 . Since we can account for the effects of Z_1 either by stratification or by its inclusion in a single PH model we may obtain different results. Possible discrepancies tell us something about our model choice.

The relation between the hazard function and the survival function follows as a straightforward extension of (4.5). Specifically, we have

$$S(t|Z) = \sum_w \phi(w) \{S_{0w}(t)\}^{\exp(\beta Z)}, \quad (6.6)$$

where $S_{0w}(t)$ is the corresponding baseline survival function in stratum w and $\phi(w)$ is the probability of coming from that particular stratum. This is then slightly more involved than the nonstratified case in which, for two groups, the model expresses the survival function of one group as a power transformation of the other. The connection to the class of Lehmann alternatives is still there although somewhat weaker. For the stratified model, once again the quantity $\lambda_{0w}(t)$ does not appear in the expression for the partial likelihood given now by

$$L(\beta) = \prod_{i=1}^n \left\{ \frac{\exp(\beta Z_i)}{\sum_{j=1}^n Y_j \{w_i(X_i), X_i\} \exp(\beta Z_j)} \right\}^{\delta_i} \quad (6.7)$$

and, in consequence, once again, $\lambda_{0w}(t)$ can remain arbitrary. Note also that each term in the product is the conditional probability that at time X_i of an observed failure, it is precisely individual i who is selected to fail, given all the individuals at risk from stratum w and that one failure from this stratum occurs.

The notation $w_i(t)$ indicates the stratum in which the subject i is found at time t . Although we mostly consider $w_i(t)$ which do not depend on time, i.e., the stratum is fixed at the outset and thereafter remains the same, it is almost immediate to generalize this idea to time dependency and we can anticipate the later section on time-dependent covariates where the risk indicator $Y_j \{w_i(t), t\}$ is not just a function taking the value one until it drops at some point to zero, but can change between zero and one with time, as the subject moves from one stratum to another. For now the function $Y_j \{w_i(t), t\}$ will be zero unless the subject is at risk of failure from stratum w_i , i.e., the same stratum in which the subject i is to be found. Taking the logarithm in (6.7) and derivative with respect to β , we obtain the score function

$$U(\beta) = \sum_{i=1}^n \delta_i \left\{ Z_i - \frac{\sum_{j=1}^n Y_j \{w_i(X_i), X_i\} Z_j \exp(\beta Z_j)}{\sum_{j=1}^n Y_j \{w_i(X_i), X_i\} \exp(\beta Z_j)} \right\}, \quad (6.8)$$

which, upon setting equal to zero, can generally be solved without difficulty using standard numerical routines, to obtain the maximum partial likelihood estimate $\hat{\beta}$ of β . The parameter β then is assumed to be common across the different strata.

Inferences about β are made by treating $\hat{\beta}$ as asymptotically normally distributed with mean β and variance $I(\hat{\beta})^{-1}$, where, now, $I(\beta)$ is given by

$I(\beta) = \sum_{i=1}^n \delta_i I_i(\beta)$. In this case each I_i is, as before, obtained as the derivative of each component to the score statistic $U(\beta)$. For the stratified score this is

$$I_i = \frac{\sum_{j=1}^n Y_j \{w_i(X_i), X_i\} Z_j^2 \exp(\beta Z_j)}{\sum_{j=1}^n Y_j \{w_i(X_i), X_i\} \exp(\beta Z_j)} - \left\{ \frac{\sum_{j=1}^n Y_j \{w_i(X_i), X_i\} Z_j \exp(\beta Z_j)}{\sum_{j=1}^n Y_j \{w_i(X_i), X_i\} \exp(\beta Z_j)} \right\}^2.$$

The central notion of the risk set is once more clear from the above expressions and we most usefully view the score function as contrasting the observed covariates at each distinct failure time with the means of those at risk from the same stratum. A further way of looking at the score function is to see it as having put the individual contributions on a linear scale. We simply add them up within a stratum and then, across the strata, it only remains to add up the different sums. Once again, inferences can also be based on likelihood ratio methods or on the score $U(\beta)$, which in large samples can be considered to be normally distributed with mean zero and variance $I(\beta)$.

Multivariate extensions follow as before. For the stratified model the only important distinction impacting the calculation of $U(\beta)$ and $I_i(\beta)$ is that the sums are carried out over each stratum separately and then combined at the end. The indicator $Y_j \{w_i(X_i)\}$ enables this to be carried out in a simpler way as indicated by the equation. The random effects model has proved to be a valuable tool in applications and this can be seen in several detailed applications (Binder, 1992; Carlin and Hodges, 1999; Collaboration, 2009; Natarajan and O'Quigley, 2002; O'Quigley and Stare, 2002; Zhou et al., 2011; Hanson, 2012).

RANDOM EFFECTS AND FRAILTY MODELS

Also coming under the heading of partially proportional hazards model are the classes of models, which include random effects. When the effects concern a single individual such models have been given the heading frailty models (Vaupel et al., 1979) since, for an individual identified by w , we can write $\lambda_{0w}(t) = \alpha_w \lambda_0(t)$ implying a common underlying hazard $\lambda_0(t)$ adjusted to each individual by a factor, the individual's *frailty*, unrelated to the effects of any other covariates that are quantified by the regression coefficients. The individual effects are then quantified by the α_w .

Although of some conceptual interest, such models are indistinguishable from models with time-dependent regression effects and therefore, unless there is some compelling reason to believe (in the absence of frailties) that a proportional hazards model would hold, it seems more useful to consider departures from proportional hazards in terms of model (4.2). On the other hand, random effects models, as commonly described by Equation (6.9) in which the α_w identifies a potentially large number of different groups, are interesting and potentially of use. We express these as

$$\lambda(t|Z(t), w) = \alpha_w \lambda_0(t) \exp\{\beta Z(t)\}. \quad (6.9)$$

These models are also partially proportional in that some effects are allowed not to follow a proportional hazards constraint. However, unlike the stratified models described above, restrictions are imposed. The most useful view of a random effects model is to see it as a stratified model with some structure imposed upon the strata. A random effects model is usually written

$$\lambda(t|Z(t), w) = \lambda_0(t) \exp\{\beta Z(t) + w\}, \quad (6.10)$$

in which we take w as having been sampled from some distribution $G(w; \theta)$. Practically there will only be a finite number of distinct values of w , however large. For any value w we can rewrite $\lambda_0(t)e^w = \lambda_{0w}(t)$ and recover model (6.4). For the right-hand side of this equation, and as we might understand from (6.4), we suppose w to take the values 1, 2, ... The values on the left-hand side, being generated from $G(\cdot)$ would generally not be integers but this is an insignificant notational issue and not one involving concepts. Consider the equation to hold. It implies that the random effects model is a stratified model in which added structure is placed on the strata. In view of Equation 6.5 and the arguments following this equation we can view a random effects model equivalently as in Equation 4.2 where, not only are PH restrictions imposed on some of the components of $\beta(t)$, but the time dependency of the other components is subject to constraints. These latter constraints, although weaker than imposing constancy of effect, are all the stronger as the distribution of $G(w; \theta)$ is concentrated. In applications it is common to choose forms for $G(w; \theta)$ that are amenable to ready calculation (Duchateau and Janssen, 2007; Gutierrez, 2002; Hanagal, 2011; Li and Ryan, 2002; Liu et al., 2004; Wienke, 2010).

STRUCTURE OF RANDOM EFFECTS MODELS

Consider firstly the model of Equation 4.3. Suppose we have one main variable, possibly a treatment variable of interest, coded by $Z_1 = 0$ for group A and $Z_1 = 1$ for group B. The second variable, say a center variable, which may or may not have prognostic importance and for which we may wish to control for possible imbalance is denoted by Z_2 . A strong modeling approach would include both binary terms in the model so that the relationship between the hazard functions is as described in the left-hand side of Figure 6.2. If our main focus is on the effect of treatment, believed to be comparable from one center to another, even though the effects of the centers themselves are not absent, it makes sense to stratify. This means that we do not attempt to model the effects of the centers but, instead, remove any such potential effects from our analysis. This is nice in that it allows for rather greater generality than that illustrated in the left-hand side of Figure 6.2. We maintain an assumption of constant treatment effect but the center effects can be arbitrary. This is illustrated on the right-hand side of Figure

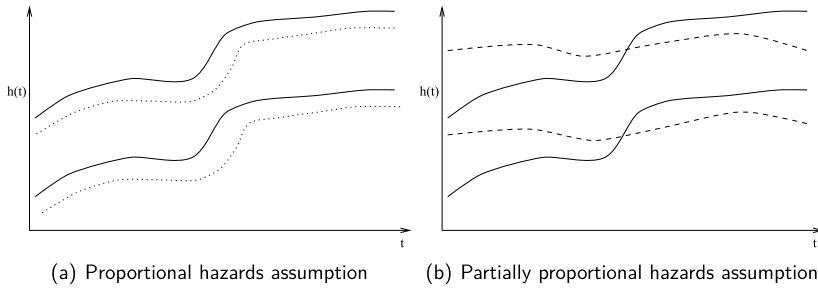


Figure 6.2: An illustration of the impact of a proportional hazards assumption and a partially proportional hazards assumption (stratified model (b)). Base rate depends on stratum while regression coefficient β is the same.

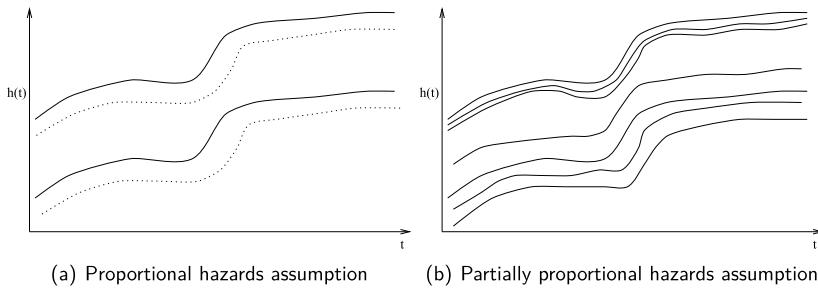


Figure 6.3: An illustration on the impact of a proportional hazards assumption and a partially proportional hazards assumption (random effects model). Stratum effect is sampled from some distribution, $G(w)$. Regression effect (β) common across strata

6.2. The illustration makes it clear that, under the assumption, a weaker one than that implied by Equation 4.3, we can estimate the treatment effect whilst ignoring center effects. A study of these figures is important to understanding what takes place when we impose a random effects model as in Equation 6.10. For many centers, Figure 6.3, rather than having two curves per center, parallel but otherwise arbitrary, we have a family of parallel curves. We no longer are able to say anything about the distance between any given centers, as we could for the model of Equation 4.3, a so-called fixed effects model, but the distribution of the distances between centers is something we aim to quantify. This is summarized by the distribution $G(w; \theta)$ and our inferences are then partly directed at θ .

RANDOM EFFECTS MODELS VERSUS STRATIFIED MODELS

The stratified model is making weaker assumptions than the random effects model. This follows since the random effects model is just a special case of a

stratified model in which some structure is imposed upon the differences between strata. The stratified model not only leaves any distribution of differences between strata unspecified, but it also makes no assumption about the form of any given stratum. Whenever the random effects model is valid, then so also is the stratified model, the converse not being the case.

It may then be argued that we are making quite a strong assumption when we impose this added structure upon the stratified model. In exchange we would hope to make non-negligible inferential gains, i.e., greater precision of our estimates of errors for the parameters of main interest, the treatment parameters.

In practice gains tend to be small for most situations and give relatively little reward for the extra effort made. Since any such gains are only obtainable under the assumption that the chosen random effects model generates the data, actual gains in practice are likely to be yet smaller and, of course, possibly negative when our additional model assumptions are incorrect. A situation where gains for the random effects model may be of importance is one where a non-negligible subset of the data include strata containing only a single subject. In such a case simple stratification would lose information on those subjects. A random effects model, assuming the approximation to be sufficiently accurate, enables us to recover such information.

EFFICIENCY OF RANDOM EFFECTS MODELS

Most of our discussion here focuses on different possible representations of the infinitely complex reality we are hoping to model. Our purpose in modeling is, ultimately, to draw simple, at least clear-cut, inferences. The question of inference no longer concerns the general but rather the specific data set we have at hand. If our main concern is on estimating risk functions then the question becomes, to what extent do we gain by including in our inferential setup the presence of random effect terms. Since our main objective is estimation and quantification of regression parameters enabling us to say something about the risk factors under study, the idea behind the inclusion of additional random effect terms is to make more precise this estimation and quantification.

As already argued above the inclusion of individual random effects (frailties) is of no practical interest and simply amounts to expressing the idea, albeit in an indirect way, of model inadequacy (O'Quigley and Stare, 2002). We therefore assume that we are dealing with groups, some of which, but not all, may only include an isolated individual. We know that a partial likelihood analysis, stratified by group, is estimating the same regression parameter. Inference is based on the stratified score statistic. We contrast the observed covariate value with its estimated expectation under the model. Different model assumptions will impact this estimated expectation and it is here that any efficiency gains can be made. For a stratified model, these estimated expectations may be with respect to relatively small risk sets. A random effects model on the other hand, via the

	100×5	250×2	25×20
Ignoring effect	0.52 (0.16)	0.51 (0.16)	0.54 (0.16)
Random effect model	1.03 (0.19)	0.99 (0.22)	1.01 (0.17)
Stratified model	1.03 (0.22)	1.02 (0.33)	1.01 (0.18)

Table 6.2: Simulations for three models under different groupings.

inclusion of a different w per group, will estimate the relevant expectations over the whole risk set and not just that relative to the group defined by the covariate value.

Comparisons for the stratified model are made with respect to the relatively few subjects of the group risk sets. This may lead us to believe that much information could be recovered were we able to make the comparison, as does the alternative random effects analysis, with respect to the whole risk set. Unfortunately this is not quite so because each contribution to the score statistic involves a difference between an observation on a covariate and its expectation under the model and the “noise” in the expectation estimate is of lower order than the covariate observations themselves. There is not all that much to be gained by improving the precision of the expectation estimate.

In other words, using the whole of the risk set or just a small sample from it will provide similar results. This idea of risk set sampling has been studied in epidemiology and it can be readily seen that the efficiency of estimates based on risk set samples of size k , rather than the whole risk set, is of the order

$$\frac{k}{k+1} \left\{ 1 + \sum_{j=1}^n \frac{1}{n(n-j+1)} \right\}. \quad (6.11)$$

This function increases very slowly to one but, with as few as four subjects on average in each risk set comparison, we have already achieved 80% efficiency. With nine subjects this figure is close to 90%. Real efficiency will be higher for two reasons: (1) the above assumes that the estimate based on the full risk set is without error, (2) in our context we are assuming that each random effect w is observed precisely.

Added to this is the fact that, since the stronger assumptions of the random effects model must necessarily depart to some degree from the truth, it is by no means clear that there is much room to make any kind of significant gains. As an aside, it is of interest to note that, since we do not gain much by considering the whole of the risk set as opposed to a small sample from it, the converse must also hold, i.e., we do not lose very much by working with small samples rather than the whole of the risk set. In certain studies, there may be great economical savings made by only using covariate information, in particular when time dependent, from a subset of the full risk set.

Table 6.2 was taken O'Quigley and Stare (2002). The table was constructed from simulated failure times where the random effects model was taken to be exactly correct. Data were generated from this model in which the gamma frailty had a mean and variance equal to one. The regression coefficient of interest was exactly equal to 1.0. Three situations were considered; 100 strata each of size 5, 250 strata each of size 2, and 25 strata each of size 20. The take-home message from the table is that, in these cases for random effects models, not much is to be gained in terms of efficiency. Any biases appear negligible and the mean of the point estimates for both random effects and stratified models, while differing notably from a crude model ignoring model inadequacy, are effectively indistinguishable. As we would expect there is a gain for the variance of estimates based on the random effects model but, even for highly stratified data (100×5), any gain is very small. Indeed for the extreme case of 250 strata, each of size 2, surely the worst situation for the stratified model, it is difficult to become enthusiastic over the comparative performance of the random effects model.

We might conclude that we only require around 80% of the comparative sample size needed for estimating relative risk based on the stratified model. But, such a conclusion, leaning entirely on the assumption that we know not only the class of distributions from which the random effects come but also the exact value of the population parameters, suggests, in practice, that the hoped for gain, in this most hopeful of cases, is more likely to be greater than the 80% indicated by our calculations. The only real situation that can be clearly disadvantageous to the stratified model is one where a non-negligible subset of the strata are seen to only contain a single observation. For such cases, and assuming a random effects model to provide an adequate fit, information from states with a single observation (which would be lost by a stratified analysis) can be recovered by a random effects analysis.

6.4 Partitioning of the time axis

Recalling the general model, i.e., the non-proportional hazards model for which there is no restriction on $\beta(t)$, note that we can re-express this so that the function $\beta(t)$ is written as a constant term, the intercept, plus some function of time multiplied by a constant coefficient. Writing this as

$$\lambda(t|Z) = \lambda_0(t) \exp\{[\beta_0 + \theta Q(t)]Z\}, \quad (6.12)$$

we can describe the term β_0 as the intercept and $Q(t)$ as reflecting the nature of the time dependency. The coefficient θ will simply scale this dependency and we may often be interested in testing the particular value, $\theta = 0$, since this value corresponds to a hypothesis of proportional hazards. Fixing the function $Q(t)$ to be of some special functional form allows us to obtain tests of proportionality against alternatives of a particular nature. Linear or quadratic decline in the

log-relative risk, changepoint, and crossing hazard situations are all then easily accommodated by this simple formulation. Tests of goodness of fit of the proportional hazards assumption can be then be constructed which may be optimal for certain kinds of departures.

Although not always needed it can sometimes be helpful to divide the time axis into r non-overlapping intervals, B_1, \dots, B_r in an ordered sequence beginning at the origin. In a data-driven situation these intervals may be chosen so as to have a comparable number of events in each interval or so as not to have too few events in any given interval. Defined on these intervals is a vector, also of dimension r , of some known or estimable functions of time, not involving the parameters of interest, β . This is denoted $Q(t) = \{Q_1(t), \dots, Q_r(t)\}$ This model is then written in the form

$$\lambda(t|Z) = \lambda_0(t) \exp\{[\beta + \theta Q(t)]Z\}, \quad (6.13)$$

where θ is a vector of dimension r . Thus, $\theta Q(t)$ (here the usual inner product) has the same dimension as β , i.e., one. In order to investigate the time dependency of particular covariates in the case of multivariate Z we would have β of dimension greater than one, in which case $Q(t)$ and θ are best expressed in matrix notation (O'Quigley and Pessione, 1989).

Here, as through most of this text, we concentrate on the univariate case since the added complexity of the multivariate notation does not bring any added light to the concepts being discussed. Also, for the majority of the cases of interest, $r = 1$ and θ becomes a simple scalar. We will often have in mind some particular form for the time-dependent regression coefficient $Q(t)$, common examples being a linear slope (Cox, 1972), an exponential slope corresponding to rapidly declining effects (Gore et al., 1984) or some function related to the marginal distribution, $F(t)$ (Breslow et al., 1984). In practice we may be able to estimate this function of $F(t)$ with the help of consistent estimates of $F(t)$ itself, in particular the Kaplan-Meier estimate. The non-proportional hazards model with intercept is of particular use in questions of goodness of fit of the proportional hazards model pitted against specific alternatives. These specific alternatives can be quantified by appropriate forms of the function $Q(t)$. We could also test a joint null hypothesis $H_0 : \beta = \theta = 0$ corresponding to no effect, against an alternative H_1 , either θ or β nonzero. This leads to a test with the ability to detect non-proportional hazards, as well as proportional hazards departures to the null hypothesis of no effect. We could also test a null hypothesis $H_0 : \theta = 0$ against $H_1 : \theta \neq 0$, leaving β itself unspecified. This would then provide a goodness of fit test of the proportional hazards assumption. We return to these issues later on when we investigate in greater detail how these models give rise to simple goodness of fit tests.

CHANGEPPOINT MODELS

A simple special case of a non-proportional hazards model with an intercept is that of a changepoint model. O'Quigley and Pessione (1991), O'Quigley (1994) and O'Quigley and Natarajan (2004) develop such models whereby we take the function $Q(t)$ to be defined by, $Q(t) = I(t \leq \gamma) - I(t > \gamma)$ with γ an unknown changepoint. This function $Q(t)$ depends upon γ but otherwise does not depend upon the unknown regression coefficients and comes under the above heading of a non-proportional hazards model with an intercept. For the purposes of a particular structure for a goodness of fit test we can choose the intercept to be equal to some fixed value, often zero (O'Quigley and Pessione, 1991). The model is then

$$\lambda(t|Z) = \lambda_0(t) \exp\{[\beta + \alpha Q(t)]Z(t)\}. \quad (6.14)$$

The parameter α is simply providing a scaling (possibly of value zero) to the time dependency as quantified by the function $Q(t)$. The chosen form of $Q(t)$, itself fixed and not a parameter, determines the way in which effects change through time; for instance, whether they decline exponentially to zero, whether they decline less rapidly or any other way in which effects might potentially change through time.

Inference for the changepoint model is not straightforward and in the later sections dealing with inference we pay particular attention to some of the difficulties raised. Note that were γ to be known, then inference would come under the usual headings with no additional work required. The changepoint model expressed by Equation 6.14 deals with the regression effect changing through time and putting the model under the heading of a non-proportional hazards model. A related, although entirely different model, is one which arises as a simplification of a proportional model with a continuous covariate and the idea is to replace the continuous covariate with a discrete classification.

The classification problem itself falls into two categories. If we are convinced of the presence of effects and simply wish to derive the most predictive classification into, say, two groups, then the methods using explained randomness or explained variation will achieve this goal. If, on the other hand, we wish to test a null hypothesis of absence of effect, and, in so doing, wish to consider all possible classifications based on a family of potential cutpoints of the continuous covariate, then, as mentioned above, special techniques of inference are required. We return to these questions in later chapters where they are readily addressed via the regression effect process.

6.5 Time-dependent covariates

In all of the above models we can make a simple change by writing the covariate Z as $Z(t)$, allowing the covariate to assume different values at different time points. Our model then becomes

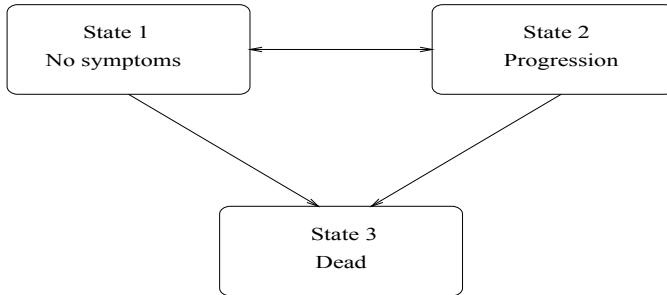


Figure 6.4: Compartment model where ability to move between states other than death state can be characterized by time-dependent indicator covariates $Z(t)$. Any paths not contradicting arrows are allowed.

$$\lambda(t|Z(t)) = \lambda_0(t) \exp\{\beta(t)Z(t)\} \quad (6.15)$$

and allows situations such as those described in Figure 6.4 to be addressed. As we change states the intensity function changes. This enables us to immediately introduce further refinement into a simple alive/dead model whereby we can suppose one or more intermediary states. A subject can move across states thereby allowing prognosis to improve or to worsen, the rates of these changes themselves depending upon other factors. The state death is described as an absorbing state and so we can move into this state but, once there, we cannot move out of it again.

Mostly we will work with the proportional hazard restriction on the above model so that

$$\lambda(t|Z(t)) = \lambda_0(t) \exp\{\beta Z(t)\}, \quad (6.16)$$

Such a simple, albeit very much more sophisticated, model than our earlier one describes a broad range of realistic situations. We will see that models with time-dependent covariates do not raise particular difficulties, either computationally or from the viewpoint of interpretation, when we deal with inference. The model simply says that the effect of the covariate remains constant, i.e., the regression coefficient remains constant, but that the covariate, or state, can itself change with time. Models with time-dependent covariates can also be used as a purely artificial construction in order to be able to express non-proportional hazards models in a proportional hazards form.

We can also imagine a slightly more involved situation than the above. Suppose that the covariate Z remains fixed, but that a second covariate, known to influence survival, also needs to be accounted for. Furthermore this second covariate is time dependent. We could, of course, simply use the above model extended to the case of two covariates. This is straightforward, apart from the fact that, as previously underlined by the complexity theorem, care is needed.

If, however, we do not wish to model the effects of this second covariate, either because it is only of indirect concern or because its effects might be hard to model, then we could appeal to a stratified model. We write:

$$\lambda(t|Z(t), w(t)) = \lambda_{0w(t)}(t) \exp\{\beta Z(t)\}, \quad (6.17)$$

where, as for the non-time-dependent case, $w(t)$ takes integer values $1, \dots, m$ indicating status. The subject can move in and out of the m strata as time proceeds. Two examples illustrate this. Consider a new treatment to reduce the incidence of breast cancer. An important time-dependent covariate would be the number of previous incidents of benign disease. In the context of inference, the above model simply means that, as far as treatment is concerned, the new treatment and the standard are only ever contrasted within patients having the same previous history. These contrasts are then summarized in final estimates and possibly tests. Any patient works her way through the various states, being unable to return to a previous state. The states themselves are not modeled. A second example might be a sociological study on the incidence of job loss and how it relates to covariates of main interest such as training, computer skills, etc. Here, a stratification variable would be the type of work or industry in which the individual finds him or herself. Unlike the previous example a subject can move between states and return to previously occupied states.

Time-dependent covariates describing states can be used in the same way for transition models in which there is more than one absorbing “death” state. Many different kinds of situations can be constructed, these situations being well described by compartment models with arrows indicating the nature of the transitions that are possible (Figure 6.5). For compartment models with time-dependent covariates there is a need for some thought when our interest focuses on the survival function. The term external covariate is used to describe any covariate $Z(t)$ such that, at $t = 0$, for all other $t > 0$, we know the value of

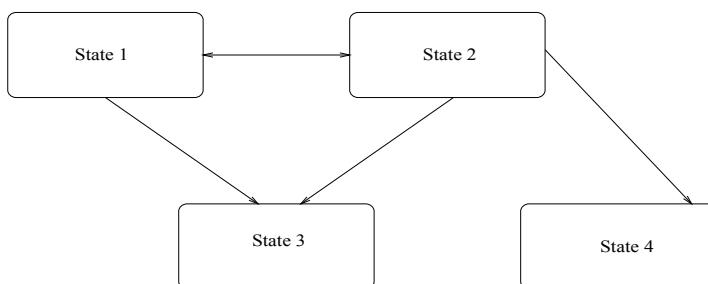


Figure 6.5: Compartment model with 2 absorbing “death” states. State 4 cannot be reached from State 3 and can only be reached from State 1 indirectly via State 2.

$Z(t)$. The paths can be described as deterministic. In the great majority of the problems that we face this is not the case and a more realistic way of describing the situation is to consider the covariate path $Z(t)$ to be random. Also open to us as a modeling possibility, when some covariate $Z_1(t)$ is of secondary interest assuming a finite number of possible states, is to use the at-risk function $Y(s, t)$. This restricts our summations to those subjects in state s as described above for stratified models.

TIME-DEPENDENT COVARIATES AND NON-PH MODELS

A non-proportional hazard model with a single constant covariate Z is written

$$\lambda(t|Z) = \lambda_0(t) \exp\{\beta(t)Z\}. \quad (6.18)$$

The multivariate extension is immediate and, in keeping with our convention of only dealing with the univariate problem whenever possible, we focus our attention on the simple product $\beta(t)Z$ at some given point in time t . If we define $\beta_0 = \beta(0) \neq 0$, we can rewrite this product as $\beta_0 Z(t)$ where $Z(t) = Z\beta(t)/\beta_0$. We could take any other time point t' and, once again, we observe that we can rewrite the product $\beta(t')Z$ as $\beta_0 Z(t')$ where $Z(t') = Z\beta(t')/\beta_0$. This equivalence is then true for any, and all, values of t . Thus, a non-proportional hazards model with a constant covariate can be re-expressed, equivalently, as a simple proportional hazards model with a time-dependent covariate.

It is almost immediate, and perhaps worth carrying out as an exercise, to show that we can reverse these steps to conclude also that any model with time-dependent covariates can be expressed in an equivalent form as a non-proportional hazards model. In conclusion, for every non-proportional hazards model there exists an equivalent proportional hazards model with time-dependent covariates. Indeed, it also clear that this argument can be extended. For, if we have the model: $\lambda(t|Z) = \lambda_0(t) \exp\{\beta(t)Z(t)\}$, then, via a re-expression of the model using $\beta_0 Z^*(t)$ where $Z^*(t) = Z(t)\beta(t)/\beta_0$, we can construct a proportional hazards model with a time-dependent regression effect from a model which began with both time-dependent regression effects as well as time changing regression coefficient. For reference purposes, it is worth putting this summary in the form of a lemma.

Lemma 6.1. *For given $\beta(t)$ and covariate $Z(t)$ there exists a constant β_0 and time-dependent covariate $Z^*(t)$ so that $\lambda(t|Z(t)) = \lambda_0(t) \exp\{\beta(t)Z(t)\} = \lambda_0(t) \exp\{\beta_0 Z^*(t)\}$.*

The important thing to note is that we have the same $\lambda_0(t)$ either side of the equation and that, whatever the value of $\lambda(t|Z(t))$, for all values of t , these values are exactly reproduced by either expression, i.e., we have equivalence.

This equivalence is a formal one and does not of itself provide any new angle on model development. The function $\beta(t)$ will not generally be known to us. This

equivalence may be exploited nonetheless in theoretical investigation. We can use it to obtain the most powerful unbiased test in chosen situations. We can also make use of the idea to adopt certain software, as well as user-written code, that already caters for time-dependent covariates. Almost no extra work is needed to use these programs should we wish to study particular types of non-proportional hazards models characterized by different $\beta(t)$.

We can end this section and round off connections to earlier sections by considering the likelihood. The sequential nature of these studies and the conditional arguments that run throughout this work means—leaving aside at least for now questions of computing—that no extra work is needed. We sequentially condition on time (this enables us in later chapters to view everything from the angle of stochastic processes) so that, at a given time $t = X_i$, all that is needed is information available at that time point X_i . The likelihood and resulting estimating equations look almost identical to what we had in the case of constant covariates. In particular, the function $\lambda_0(t)$ is not involved. The likelihood expression is given by

$$L(\beta) = \prod_{i=1}^n \left\{ \frac{\exp(\beta Z_i(X_i))}{\sum_{j=1}^n Y_j(X_i) \exp(\beta Z_j(X_i))} \right\}^{\delta_i}, \quad (6.19)$$

As before, taking the logarithm in Equation 6.19 and its derivative with respect to β , we obtain the estimating equation which, upon setting equal to zero, can generally be solved without difficulty using the Newton-Raphson method. Chapter 7 on estimating equations looks at this more closely. The form of $U(\beta)$ is as before with only the time dependency marking any distinction.

$$U(\beta) = \sum_{i=1}^n \delta_i \left\{ Z_i(X_i) - \frac{\sum_{j=1}^n Y_j(X_i) Z_j(X_i) \exp(\beta Z_j(X_i))}{\sum_{j=1}^n Y_j(X_i) \exp(\beta Z_j(X_i))} \right\} \quad (6.20)$$

The residual interpretation of the contributions to the estimating equation remains. In order for $\hat{\beta}$ to behave as in the non-time-dependent case some weak restrictions on the allowable covariate functions would be needed. We leave this to later chapters, and treating $\hat{\beta}$ to be asymptotically normally distributed with mean β and large sample variance $I(\hat{\beta})^{-1}$, where $I(\beta)$ is, once again, minus the second derivative of $\log L(\beta)$ with respect to β . The expression for $I(\beta) = \sum_{i=1}^n \delta_i I_i(\beta)$ would be analogous to that given by Equation 4.8. We might also keep in mind that Lemma 11.1 allows us to formally work with this structure for time-dependent covariates in order to cater for non-proportional hazards situations.

6.6 Linear and alternative model formulations

Throughout the history of survival analysis some researchers have preferred to frame the many questions raised within the setting of linear models. Certainly there is an advantage in having a very substantial amount of literature dealing with the linear model in other circumstances. However, this advantage is more of a theoretical than a practical one and is lost under the simple device of taking logarithms. Linear models are simply less natural in the context of rates and probabilities and, while we recall some of the main ideas here, our purpose is more to do with completeness than a belief that linear models offer something not adequately catered for by PH and NPH models. That said, we should always keep our eye on parsimony principles. In some situations an additive model may provide a more parsimonious description of the data. One example that comes to mind arises in cardiovascular studies where, for certain risk factors, a constant additive handicap for those at risk provides a simpler model than the more involved relative risk one where there are complex time dependencies that have to be catered for.

ADDITIVE MODELS

Instead of considering a model with multiplicative risks, some authors appeal to a structure more familiar to the one we know from linear regression, e.g., the additive intensity model proposed by Aalen (1989, 1980). This is written:

$$\lambda(t | Z) = \beta_0(t) + \beta^T(t)Z(t), \quad (6.21)$$

where the vector of regression coefficients, $\beta(t) = (\beta_0(t), \beta_1(t), \dots, \beta_p(t))$ depends on time. The estimation of the cumulative effects $\int_0^t \beta_i(s)ds$ for $i = 1, \dots, p$ is done by weighted least squares. McKeague and Sasieni (1994) studied the case of fixed effects for some covariates and time-dependent effects for the others creating a model analogous to the partially proportional hazards model. Lin and Ying (1995) proposed a more complex combined additive and multiplicative risk model written:

$$\lambda(t | X, Z) = Y(t)g\{\beta^T X(t)\} + \lambda_0(t)h\{\gamma^T Z(t)\},$$

where $(X(t), Z(t))$ is a vector of $p + q$ time-dependent covariables $\beta \in \mathbb{R}^p$ and $\gamma \in \mathbb{R}^q$ and where, again, $Y(t)$ is the indicator of being at risk at time t . The functions h and g are link functions. For example, we can choose $h(x) = \exp(x)$ or $1 + x$ and $g(x) = x$ or $\exp(x)$. This model class aims to include both the proportional hazard model as well as the additive model for particular values of h and g . Scheike and Zhang (2002) investigated a further combination of multiplicative and additive intensities with the model given by

$$\lambda(t | X, Z) = Y(t)\beta(t)^T X(t) \exp(\gamma^T Z(t)),$$

where $\beta(t)$ is time dependent. Aalen et al. (2008) presents an analysis of the properties of these kinds of model. Beran (1981) and McKeague and Utikal (1990) studied the model,

$$\lambda(t | Z) = Y(t)\alpha(t, Z(t)), \quad (6.22)$$

without making explicit the function α , which can be estimated by kernel methods and local smoothing.

LINEAR TRANSFORMATION MODELS

Elementary manipulation allows us to see that the proportional hazards model with fixed covariables, Z , can be written

$$\log\{-\log(S(t | Z))\} = \log\left(\int_0^t \lambda_0(s)ds\right) + \beta^T Z,$$

where, as usual, λ_0 is an arbitrary underlying risk function. Using this as a starting point we can generalize in the following way:

$$h(S(t | Z)) = g(t) + \beta^T Z, \quad (6.23)$$

where h is a link function, g an unspecified function, both of which are strictly increasing. Model (6.23) can be re-expressed as

$$g(T) = -\beta^T Z + \varepsilon, \quad (6.24)$$

where ε is a random variable with distribution function $1 - h^{-1}$. Equation (6.24) describes a class of transformation models for T that results in a linear relationship with Z (Cheng et al., 1995; Dabrowska and Doksum, 1988; Lin, 2013; Wei, 1992). Linear models have been very well studied and their properties are readily anticipated. For this reason it can seem attractive to bring these non-linear models under a linear umbrella. One particular case is the proportional odds model (Bennett, 1983a) which follows from fixing

$$h(x) = \text{logit}(x) = \log\left(\frac{x}{1-x}\right).$$

Note that, for

$$h(x) = \log\left(\frac{1-S(x | Z)}{S(x | Z)}\right) \quad \text{and} \quad g(t) = \phi \log(t),$$

we recover the log-logistic model (Bennett, 1983b). The transformation $g = \log$, produces a class of models referred to as accelerated failure time models. The survival function, given Z , is given by

$$S(t | Z) = S_0 \left(t \exp(\beta^T Z) \right), \quad 0 \leq t \leq \mathcal{T}, \quad (6.25)$$

where S_0 is an underlying survival function corresponding to the hazard λ_0 . The formula (6.25) shows that the covariates have a multiplicative effect on time, which implies that they accelerate or slow down the death rate, according to the sign of $\beta^T Z$ (Martinussen and Peng, 2013). The class of linearly transformed models is more general than the proportional hazard model but is less flexible, since it does not easily allow for the use of time-dependent covariates. Extensions to accommodate other common situations such as multiple events is also awkward. Finally, great care is needed when using mixture type models of this nature in practice. It is frequently the case (Venzon and Moolgavkar, 1988) that they fail to be invariant to simple changes in the coding. Coding the treatment 0, rather than 1, can have a significant impact on the p -value. There is no good reason to choose a particular coding, say 1 for males and 0 for females rather than the other way around, and it is more than problematic that the final result would depend on that choice. The Yang and Prentice model, described in the following section, also suffers from a lack of invariance to coding.

YANG AND PRENTICE MODEL

In the presence of a binary covariate Z describing the two-group problem, Yang and Prentice (2005) proposed modeling the survival time in both groups by

$$\lambda_T(t) = \frac{\theta_1 \theta_0}{\theta_1 + (\theta_0 - \theta_1) S_P(t)} \lambda_P(t), \quad 0 \leq t \leq \mathcal{T}, \quad (6.26)$$

where θ_1 and θ_0 are two positive parameters, S_P is the survival function in the first group, λ_P and λ_T are the hazard rate in the first and second groups, respectively. The proportional hazards model ($\theta_1 = \theta_0$) and the proportional odds model ($\theta_0 = 1$) are special cases. The model appears to be flexible allowing for the possibility of modeling risks that cross over in time. This property can be exploited as a means to obtain tests of greater generality than, say, the log-rank test. The parameters of the model are such that $\theta_1 = \lim_{t \downarrow 0} \lambda_T(t)/\lambda_P(t)$ and $\theta_0 = \lim_{t \uparrow \mathcal{T}} \lambda_T(t)/\lambda_P(t)$. Thus θ_1 represents the relative risk in the short term and θ_0 the long-term relative risk. The estimation in the model is done by maximizing an equivalent of the model-specific partial likelihood. The authors suggest the introduction of Z covariates by replacing the survival functions by $S_T(t)^{\exp(\beta^T Z)}$ and $S_P(t)^{\exp(\beta^T Z)}$.

While these ideas are very interesting and provide a lot of insight into the possible mechanisms that generate the observations, at the time of writing, the test based on this model cannot be recommended for general use. There are two reasons for this. The first, identified by Chauvel and O'Quigley (2014) is the poor control on Type 1 error offered by the test. The good power properties are in part a reflection of this. Indeed in several situations of proportional hazards the

test outperformed the log-rank test, something that is not theoretically possible. So, some kind of adjustment is needed before too much reliance be put on the calculated p -value. The second reason is rather more serious, not easily overcome, and stems from the somewhat non-linear specification of the model. Different codings will give different results. Coding the treatment variable (1,0) rather than (0,1) will lead to different answers. Since there is often no natural way to code this is very problematic. It is quite plausible, for example, that the same set of observations allows us to conclude that the survival experience of women differs significantly from that of men and that, simultaneously, the survival experience of men does not significantly differ from that of women. This is clearly not a coherent summary of the data and has to be fixed before we could have confidence in any results based on this test. A simple potential solution would be to calculate both p -values say and then take the minimum. This would solve the coherence problem but would certainly exacerbate the difficulty in exercising good control over Type 1 error. And in the situation of several levels, or several covariates, it would seem to be very difficult to anticipate the operating characteristics in any broad sense.

Nonetheless, there is something rather attractive in this formulation and it would be worth the extra effort to fix this coherence problem. The root of the difficulty would stem from the lack of symmetry in Equation 6.26 and this observation may open up a path toward a solution. Further discussion is given in Flandre and O'Quigley (2020).

6.7 Classwork and homework

1. Consider the approach of partitioning the time axis as a way to tackle non-proportional hazards. For example, we might choose a partition consisting of 3 intervals, any one of which respects the proportional hazards constraint with an interval-specific regression coefficient. Find the simplest way of writing this model down. How might you express a null hypothesis of no effect against an alternative of a non-null effect. How would this expression change if, aside from the null, the only possibility is of an effect that diminishes through time.
2. When biological or other measures are expensive to make, sampling from the risk set can enable great savings to be made. Explain why this would be the case and consider the criteria to be used to decide the size of the risk set sampled. Look up and describe the case-cohort design which has been successfully used in large epidemiological studies.
3. Show formally that a non-proportional hazards model with a constant covariate is equivalent to a proportional hazards one with a time-dependent covariate. Does any such result hold for a non-proportional hazards model with a non-constant covariate.

4. Suppose, for a binary group indicator, that the observations are generated by a linear model with a constant regression coefficient. Show that such a model is equivalent to a proportional hazards model with a time-dependent covariate. Conclude that a linear model is equivalent to a non-proportional hazards one.
5. Consider a randomized clinical trial comparing 2 treatments. In order to obtain sufficient recruitment a total of 25 medical centers participate in the trial. The prognosis varies widely among centers and so center effects need to be taken into account to avoid any real effects of the new treatment being drowned by noise. As a statistician your opinion is asked. There are 3 possibly approaches: (1) make use of a proportional hazards model that includes a term of dimension 24 to account for center effects, (2) make use of a stratified proportional hazards model to account for center effects, (3) make use of a random effects model to account for center effects, or (4) ignore center effects and carry out a simple log-rank test for possible differences in treatment outcomes. What advice would you give? And why?
6. Again consider the setting of a randomized clinical trial comparing 2 treatments. Suppose that we base our analysis on a proportional hazards model but that the mechanism generating the observations is in fact a non-proportional hazards one where the effects decline through time. Describe how inference will be impacted by this. Next, suppose that we assume a broader non-proportional hazards model in which regression effects may decline but may also remain constant. It turns out that the mechanism generating the observations is a proportional hazards one. Describe how, in this case, inference will be affected by our assumptions.
7. Show formally that, if our covariate space does not include continuous covariates, then any true situation can be modeled precisely by a non-proportional hazards model. Conclude that, for any arbitrary situation, including that of continuous covariates, we can postulate a non-proportional hazards model that is arbitrarily close to that generating the observations.



Chapter 7

Model-based estimating equations

7.1 Chapter summary

The regression effect process, described in Chapter 9, shapes our main approach to inference. At its heart are differences between observations and their model-based expectations. The flavor is very much that of linear estimating equations (Appendix D.1). Before we study this process, we consider here an approach to inference that makes a more direct appeal to estimating equations. The two chapters are closely related and complement one another. This chapter leans less heavily on stochastic processes and links in a natural and direct way to the large body of theory available for estimating equations. Focusing attention on the expectation operator, leaning upon different population models and different working assumptions, makes several important results transparent. For example, it is readily seen that the so-called partial likelihood estimator is not consistent for average effect, $E\{\beta(T)\}$, under independent censoring and non-constant $\beta(t)$. One example we show, under heavy censoring, indicates the commonly used partial likelihood estimate to converge to a value greater than 4 times its true value. Linear estimating equations provide a way to investigate statistical behavior of estimates for small samples. Several examples are considered.

7.2 Context and motivation

This chapter and Chapter 9 provide the inferential tools needed to analyze survival data. Either approach provides several techniques and results that we can exploit. Taken together we have a broad array of methods, and their properties, that, when used carefully, can help gain deep understanding of real datasets that arise in the setting of a survival study.

The earlier chapter on marginal survival is important in its own right and we lean on the results of that chapter throughout this work. We need to keep in mind the idea of marginal survival for two reasons: (1) it provides a natural backdrop to the ideas of conditional survival and (2), together with the conditional distribution of the covariate given $T = t$, we are able to consider the joint distribution of covariate and survival time T . A central concern is conditional survival, where we investigate the conditional distribution of survival given different potential covariate configurations, as well as variables such as time elapsed. More generally we are interested in survival distributions corresponding to transitions from one state to another, conditional on being in some particular state or of having mapped out some particular covariate path. The machinery that will enable us to obtain insight into these conditional distributions is that of proportional and non-proportional hazards regression.

When we consider any data at hand as having arisen from some experiment, the most common framework for characterizing the joint distribution of the covariate Z and survival T is one where the distribution of Z is fixed and known, and the conditional survivorship distribution is the subject of our inferential endeavors. Certainly, this characterization would bring the model to most closely resemble the experiment as set-up. However, in order to accommodate censoring, it is more useful to characterize the joint distribution of Z and T via the conditional distribution of Z given $T = t$ and the marginal distribution of T . This is one of the reasons why we dealt first with the marginal distribution of T . Having dealt with that we can now focus our attention on the conditional distribution of Z given T . We can construct estimating equations based on these ideas and from these build simple tests or make more general inferences.

The main theorem of proportional hazards regression, introduced by O'Quigley (2008), generalizes earlier results of Schoenfeld (1980), O'Quigley and Flandre (1994), and Xu and O'Quigley (2000). The theorem has several immediate corollaries and we can use these to write down estimating equations upon which we can then construct suitable inferential procedures for our models. The regression effect process of subsequent chapters also has a clear connection to this theorem. While a particular choice of estimating equation can result in high efficiency when model assumptions are correct or close to being correct, other equations may be less efficient but still provide estimates which can be interpreted when model assumptions are incorrect. For example, when the regression function $\beta(t)$ might vary with time, we are able to construct an estimating equation, the solution of which provides a consistent estimate of $E\{\beta(T)\}$, the average effect. The usual partial likelihood estimate fails to achieve this, and the resulting errors, depending on the level of censoring, even when independent, can be considerable. This can be seen in Table 7.1 where the resulting bias, for high censoring rates, can be more than 400%. Most currently available software makes no account of this and, given that a non-constant $\beta(t)$ will be more the rule than the exception,

we ought to be very cautious in interpreting the majority of estimates of average effect, $E\{\beta(T)\}$, unless: (i) the function $\beta(t)$ is close to being a constant or (ii) the dataset contains little censoring.

7.3 Likelihood solution for parametric models

For almost any statistical model, use of the likelihood is usually the chosen method for dealing with inference on unknown parameters. Bayesian inference, in which prior information is available, can be viewed as a broadening of the approach and, aside from the prior, it is again the likelihood function that will be used to estimate parameters and carry out tests. Maximum likelihood estimates, broadly speaking, have good properties and, for exponential families, a class to which our models either belong or are close, we can even claim some optimality. A useful property of maximum likelihood estimators of some parameter is that the maximum likelihood estimator of some monotonic function of the parameter is the same monotonic function of the maximum likelihood estimator of the parameter itself. Survival functions themselves will often come under this heading and so, once we have estimated parameters that provide the hazard rate, then we immediately have estimates of survival. Variance expressions are also obtained quite easily, either directly or by the approximation techniques of Appendix A.10. Keeping in mind that our purpose is to make inference on the unknown regression coefficients, invariant to monotonic increasing transformations on T , we might also consider lesser used likelihood approaches such as marginal likelihood and conditional likelihood. It can be seen that these kinds of approaches lead to the so-called partial likelihood. In practice we will treat the partial likelihood as though it were any regular likelihood, the justification for this being possible through several different arguments.

For fixed covariates, in the presence of parametric assumptions concerning $\lambda_0(t)$, inference can be carried out on the basis of the following theorem that simply extends that of Theorem 3.1. We suppose that the survival distribution is completely specified via some parametric model, the parameter vector being say θ . A subset of θ is a vector of regression coefficients, β , to the covariates in the model. The usual working assumption is that of a conditionally independent censoring mechanism, i.e., the pair (T, C) is independent given Z . This would mean, for instance, that within any covariate grouping, T and C are independent but that C itself can depend on the covariate. Such dependence would generally induce a marginal dependency between C and T .

Theorem 7.1. *Under a conditionally independent censoring mechanism the log-likelihood $\log L(\theta)$ can be written as $\log L(\theta) = \sum_{i=1}^n \log L_i(\theta)$ where*

$$\log L_i(\theta) = I(\delta_i = 1) \log f(x_i | z_i; \theta) + I(\delta_i = 0) \log S(x_i | z_i; \theta). \quad (7.1)$$

The maximum likelihood estimates obtain the values of θ , denoted $\hat{\theta}$, that maximize $\log L(\theta)$ over the parameter space. For $\log L(\theta)$ a differentiable function of θ , this value is then the solution to the estimating equation $U(\theta) = 0$ where $U(\theta) = \sum_i \partial \log L_i(\theta) / \partial \theta$. Next notice that, at the true value of θ , i.e., the value which supposedly generates the observations, denoted θ_0 , we have $\text{Var}(U(\theta_0)) = EU^2(\theta_0) = EI(\theta_0)$ where

$$I(\theta) = \sum_{i=1}^n I_i(\theta) = -\partial^2 \log L(\theta) / \partial \theta^2 = -\sum_{i=1}^n \partial^2 \log L_i(\theta) / \partial \theta^2.$$

As for likelihood in general, some care is needed in thinking about the meaning of these expressions and the fact that the operators $E(\cdot)$ and $\text{Var}(\cdot)$ are taken with respect to the distribution of the pairs (x_i, δ_i) but with θ_0 fixed. The score equation is $U(\hat{\theta}) = 0$ and the large sample variance is approximated by $\text{Var}(\hat{\theta}) \approx 1/I(\hat{\theta})$. Newton-Raphson iteration is set up by a simple application of the mean value theorem so that

$$\theta_{j+1} = \theta_j + I(\theta_j)^{-1} U(\theta_j), \quad j \geq 1, \tag{7.2}$$

where θ_1 is some starting value, often zero, to the iterative cycle. The iteration is brought to a halt once we achieve some desired level of precision. An interesting result that is not well known and is also, at first glance, surprising is that θ_2 is a fully efficient estimator. It does no less well than $\hat{\theta}$ and this is because, while subsequent iterations may bring us closer and closer to $\hat{\theta}$, they do not, on average, bring us any closer to θ_0 . A one step estimator such as θ_2 can save time when dealing with onerous simulations.

Note that likelihood theory would imply that we work with the expected information (called Fisher information) $E\{I(\theta)\}$ but in view of Efron and Hinkley (1978) and the practical difficulty of specifying the censoring we usually prefer to work with a quantity allowing us to consistently estimate the expected information, in particular the observed information.

Large sample inference can be based on any one of the three tests derived from the likelihood. For the score test there is no need to carry out parameter estimation or to maximize some function. Many well-established tests can be derived in this way. In exponential families, also the so-called curved exponential families (Efron et al., 1978), such tests reduce to contrasting some observed value to its expected value under the model. Good confidence intervals (Cox and Hinkley, 1979) can be constructed from “good” tests. For the exponential family class of distributions the likelihood ratio forms a uniformly most powerful test and, as such, qualifies as a “good” test in the sense of Cox and Hinkley. The other tests are asymptotically equivalent so that confidence intervals based on the above test procedures will agree as sample size increases. Also, we can use

such intervals for other quantities of interest such as the survivorship function since this function depends on these unknown parameters.

Recall from Chapter 3 that we can estimate the survival function as $S(t; \hat{\theta})$. If Θ_α provides a $100(1 - \alpha)\%$ confidence region for the vector θ , then we can obtain a $100(1 - \alpha)\%$ confidence region for $S(t; \theta)$ in the following way. For each t let

$$S_\alpha^+(t; \hat{\theta}) = \sup_{\theta \in \Theta_\alpha} S(t; \theta), \quad S_\alpha^-(t; \hat{\theta}) = \inf_{\theta \in \Theta_\alpha} S(t; \theta), \quad (7.3)$$

then $S_\alpha^+(t; \hat{\theta})$ and $S_\alpha^-(t; \hat{\theta})$ form the endpoints of the $100(1 - \alpha)\%$ confidence interval for $S(t; \theta)$. Such a quantity may not be so easy to calculate in general, simulating from Θ_α or subdividing the space being an effective way to approximate the interval. Some situations nonetheless simplify. The most straightforward is where the survival function is a monotonic function of the one-dimensional parameter θ . As an illustration, the scalar location parameter, θ , for the exponential model corresponds to the mean. We have that $S(t; \theta)$ is monotonic in θ . For such cases it is only necessary to invert any interval for θ to obtain an interval with the same coverage properties for $S(t; \theta)$. Denoting the upper limit of the $100(1 - \alpha)\%$ confidence interval for θ as θ_α^+ and the lower limit of the $100(1 - \alpha)\%$ confidence interval for θ as θ_α^- , we can then write: $S_\alpha^+(t; \hat{\theta}) = S(t; \theta_\alpha^-)$ and $S_\alpha^-(t; \hat{\theta}) = S(t; \theta_\alpha^+)$. Note that these intervals are calculated under the assumption that t is fixed. For the exponential model, since the whole distribution is defined by θ , the confidence intervals calculated pointwise at each t also provide confidence bands for the whole distribution.

LIKELIHOOD SOLUTION FOR EXPONENTIAL MODELS

As for the case of a single group, an analysis based on the exponential model is particularly simple. For this reason alone it is of interest but also (see section below on the nonparametric exponential model) the results are much more general than are often supposed. We restrict attention to the two-group case in order to enhance readability. The two groups are defined by the binary covariate Z taking the value either zero or one. The extension to higher dimensions is all but immediate. For the two-group case we will only need to concern ourselves with two parameters, λ_1 and λ_2 , for which, once we have them or consistent estimates of them, we have the whole survival experience (or estimates of this) for both groups.

Expressing the model in proportional hazards form we can write: $\lambda_1 = \lambda$ and $\lambda_2 = \lambda \exp(\beta)$. Referring to Equation 7.1 then, if individual i corresponds to group 1, his or her contribution to the likelihood is $f(x_i; \lambda) = \lambda \exp(-\lambda x_i)$ when $\delta_i = 1$, whereas for $\delta_i = 0$, the contribution is $S(x_i; \lambda) = \exp(-\lambda x_i)$. If the individual belongs to group 2 the likelihood contribution would be either $\lambda \exp(\beta) \exp(-e^\beta \lambda x_i)$ or $\exp(-e^\beta \lambda x_i)$ according to whether δ_i is equal to one

or zero. We use the variable $w_i = I(z_i = 1)$ to indicate which group the subject is from. From this and Theorem 7.1 we have:

Corollary 7.1. *For the 2-sample exponential model, the likelihood satisfies*

$$\log L(\lambda, \beta) = k \log \lambda + \beta k_2 - \lambda \left\{ \sum_{j=1}^n x_j (1 - w_j) + e^\beta \sum_{j=1}^n x_j w_j \right\},$$

where $w_i = \mathbf{1}_{z_i=1}$ and where there are k_1 distinct failures in group 1, k_2 in group 2, and $k = k_1 + k_2$.

Differentiating the log-likelihood with respect to both λ and β and equating both partial derivatives to zero we readily obtain an analytic solution to the pair of equations given by:

Corollary 7.2. *The maximum likelihood estimates $\hat{\beta}$ and $\hat{\lambda}$ for the two-group exponential model are written as*

$$\hat{\beta} = \log \frac{k_2}{\sum_{j=1}^n x_j w_j} - \log \frac{k_1}{\sum_{j=1}^n x_j (1 - w_j)}; \quad \hat{\lambda} = \frac{k_1}{\sum_{j=1}^n x_j (1 - w_j)}.$$

It follows immediately that $\hat{\lambda}_1 = \hat{\lambda}$ and that $\hat{\lambda}_2 = \hat{\lambda} \exp(\hat{\beta}) = k_2 / \sum_{j=1}^n x_j w_j$. In order to carry out tests and construct confidence intervals we construct the matrix of second derivatives of the log-likelihood, $I(\lambda, \beta)$, obtaining

$$\begin{pmatrix} -\partial^2 \log L(\lambda, \beta) / \partial \lambda^2 & -\partial^2 \log L(\lambda, \beta) / \partial \lambda \partial \beta \\ -\partial^2 \log L(\lambda, \beta) / \partial \lambda \partial \beta & -\partial^2 \log L(\lambda, \beta) / \partial \beta^2 \end{pmatrix} = \begin{pmatrix} k/\lambda^2 & e^\beta \sum_j x_j w_j \\ e^\beta \sum_j x_j w_j & \lambda e^\beta \sum_j x_j w_j \end{pmatrix}.$$

The advantage of the two parameter case is that the matrix can be explicitly inverted. We then have:

Corollary 7.3. *Let $D = \lambda^{-1} e^\beta \sum_j x_j w_j \{k - \lambda e^\beta \sum_j x_j w_j\}$. Then, for the two-group exponential model the inverse of the information matrix is given by*

$$I^{-1}(\lambda, \beta) = D^{-1} \begin{pmatrix} \lambda e^\beta \sum_j x_j w_j & -e^\beta \sum_j x_j w_j \\ -e^\beta \sum_j x_j w_j & k/\lambda^2 \end{pmatrix}.$$

The score test is given by $X_S^2 = U'(\hat{\lambda}, 0) I^{-1}(\hat{\lambda}, 0) U(\hat{\lambda}, 0)$. Following some simple calculations and recalling that $\exp(-\hat{\beta}) = \sum_{j=1}^n x_j w_j / k_2$, we have:

Corollary 7.4. *For the two-group exponential model the score test is given by*

$$X_S^2 = k^{-1} k_1 k_2 \exp(-\hat{\beta}) \{1 - \exp(\hat{\beta})\}^2.$$

At first glance the above expression, involving as it does $\hat{\beta}$, might appear to contradict our contention that the score statistic does not require estimation

of the parameter. There is no contradiction although we consider in the above expression λ to be a nuisance parameter not specified under the null hypothesis. This parameter value *does* require estimation, although still under the null. The regression parameter itself turns out to have a simple explicit form, and it is this same term that appears in the score statistic. Typically, for other models, the maximum likelihood estimate would not have an explicit analytic form. We do not need to estimate it in order to evaluate the score statistic. On the other hand, both the Wald test and the likelihood ratio test do require estimation under the alternative. The calculations in this specific case can be carried out straightforwardly and we also have a relatively simple, and again explicit solution (i.e., not requiring the finding of an iterative solution to the likelihood equation) for the likelihood ratio test. We have then the following two corollaries:

Corollary 7.5. *For the two-group exponential model the likelihood ratio test X_L^2 is given by*

$$X_L^2 = 2 \left(k_2 \log \frac{k_2}{\sum_j x_j w_j} + k_1 \log \frac{k_1}{\sum_j x_j (1 - w_j)} - k \log \frac{k}{\sum_j x_j} \right).$$

The third of the tests based on the likelihood, the Wald test, is also straightforward to calculate and we have the corresponding lemma:

Corollary 7.6. *For the two-group exponential model, the Wald test is given by*

$$X_W^2 = k^{-1} k_1 k_2 \hat{\beta}^2.$$

For large samples we anticipate the three different tests to give very similar results. For smaller samples the Wald test, although the most commonly used, is generally considered to be the least robust. In particular, a monotonic transformation of the parameter will, typically, lead to a different value of the test statistic.

For the Freireich data, the maximum likelihood estimates of the hazard rates in each group are

$$\begin{aligned} \hat{\lambda}_1 &= 9/359 = 0.025, & \text{Var}(\hat{\lambda}_1) &= 9/(359)^2 = 0.000070, \\ \hat{\lambda}_2 &= 21/182 = 0.115, & \text{Var}(\hat{\lambda}_2) &= 21/(182)^2 = 0.00063. \end{aligned}$$

We might note that the above results are those that we would have obtained had we used the exponential model separately in each of the groups. In this particular case then the model structure has not added anything or allowed us to achieve any greater precision in our analysis. The reason is simple. The exponential model only requires a single parameter. In the above model we have two groups and, allowing these to be parameterized by two parameters, the rate λ and the multiplicative factor $\exp(\beta)$, we have a saturated model. The saturated model is entirely equivalent to using two parameters, λ_1 and λ_2 , in each of the

two groups separately. More generally, for exponential models with many groups or with continuous covariates, or for other models, we will not usually obtain the same results from separate analyzes as those we obtain via the model structure. The model structure will, as long as it is not seriously misspecified, usually lead to inferential gains in terms of precision of parameter estimates.

Since $\exp(\hat{\beta}) = 21/182 \times 359/9 = 4.60$ we have that the estimate of the log-relative risk parameter $\hat{\beta}$ is 1.53. We also have that the score test $X_S^2 = 17.8$, the Wald test $X_W^2 = 14.7$, and the likelihood ratio test $X_L^2 = 16.5$. The agreement between the test statistics is good and, in all cases, the significance level is sufficiently strong to enable us to conclude in favor of clear evidence of a difference between the groups.

Had there been no censoring then $k = n$, the sample size, and $\sum_{j=1}^n t_j$ corresponds to a sum of n independent random variables each exponential with parameter λ . We could therefore treat $n/\hat{\lambda}$ as a gamma variate with parameters (λ, n) . In view of the consistency of $\hat{\lambda}$, when there is censoring, we can take $k/\hat{\lambda}$ as a gamma variate with parameters (λ, k) , when $k < n$. This is not an exact result, since it hinges on a large sample approximation, but it may provide greater accuracy than the large sample normal approximation.

Recall from Chapter 3 that we can make use of standard tables by multiplying each term of the sum by 2λ . The result of this product is a sum of n exponential variates in which each component of the sum has variance equal to 2. This corresponds to a gamma $(2, n)$ distribution which is also equivalent to a chi-square distribution with $2n$ degrees of freedom. Taking the range of values of $2k\lambda/\hat{\lambda}$ to be between $\chi_{\alpha/2}$ and $\chi_{1-\alpha/2}$ gives a $100(1-\alpha)\%$ confidence interval for λ . For the Freireich data we obtained a 95% $CI = (0.0115, 0.0439)$. On the basis of intervals for λ , we can obtain intervals for the survivorship function which is, in this particular case, a monotonic function of λ . The upper and lower limits of the $100(1-\alpha)\%$ confidence interval are denoted by $S_\alpha^+(t; \hat{\lambda})$ and $S_\alpha^-(t; \hat{\lambda})$, respectively. We write:

$$[S_\alpha^+(t; \hat{\lambda}), S_\alpha^-(t; \hat{\lambda})] = \left[\exp \left\{ - \left(\frac{\hat{\lambda}\chi_{\alpha/2}}{2k} \right) t \right\}, \exp \left\{ - \left(\frac{\hat{\lambda}\chi_{1-\alpha/2}}{2k} \right) t \right\} \right]. \quad (7.4)$$

A different approximation is described in Section 7.3 in which

$$S_\alpha^+(t; \hat{\lambda}) \approx \exp \left\{ - \left(\frac{\hat{\lambda}}{\sqrt{k}} \right) (\sqrt{k} - z_{1-\alpha/2}) t \right\}, \quad (7.5)$$

where the corresponding expression for $S_\alpha^-(t; \hat{\lambda})$ is obtained using the percentiles, $z_{1-\alpha/2}$ by $z_{\alpha/2}$ of the standard normal distribution. Agreement between these two approximations appears to be very close. It would be of interest to have a more detailed comparison between the approaches.

NONPARAMETRIC EXPONENTIAL ANALYSIS

For the one-sample case we have already seen how, referring to Section A.6 and Theorem A.8, we are able to make use of the result that, for any continuous positive random variable T , with distribution function $F(t)$, the variate $\Lambda(T) = \int_0^T f(u)/[1 - F(u)]du$ has a standard exponential distribution. For the one-sample case we can work with the empirical survival function $\hat{S}(t)$ appealing to the result that the observations $-\log \hat{S}(X_i)$ can be taken to have been sampled from a standard exponential distribution.

In the two-group and, by extension, many-group case it will be necessary to transform observations from a group other than that giving rise to the group estimate, say $\hat{S}_G(t)$. To facilitate this, as described in Chapter 3, we work with the continuous version of the Kaplan-Meier estimate, $\bar{S}(t)$. Note that a two-group proportional hazards model can be expressed as $S_2(t) = S_1^\alpha(t)$ where $\alpha = \exp(\beta)$. Taking logarithms then enables an analytic expression for β as

$$\beta = \log\{-\log S_2(t)\} - \log\{-\log S_1(t)\}. \quad (7.6)$$

For the case of three groups, defined by the pair of binary indicator variables, Z_1 and Z_2 , the model states that $S(t|Z_1, Z_2) = S_1^\alpha(t)$ where, in this more complex set-up, $\log \alpha = \beta_1 Z_1 + \beta_2 Z_2$. Here, in exactly the same way, we obtain analytic expressions for β_1 and β_2 as the arithmetic difference between the log-log transformations of the respective marginal survival curves.

For two independent groups G_1 and G_2 we can consider two separate estimators, $\hat{S}_1(t)$ and $\hat{S}_2(t)$. Since we are assuming a proportional hazards model, we will carry over this restriction to the sample-based estimates whereby $\hat{S}_2(t) = \hat{S}_1^\alpha(t)$ and where, as before, $\alpha = \exp(\beta)$. In view of the above result we have:

Lemma 7.1. *A consistent estimator of β obtains from*

$$\hat{\beta} = \frac{1}{n_m} \sum_{i=1}^{n_m} [\log\{-\log \bar{S}_2(X_i)\} - \log\{-\log \bar{S}_1(X_i)\}], \quad (7.7)$$

where $n_m = \sum_i \mathbf{1}_{X_i \leq m}$, $m_\ell = \max\{X_i | X_i \in G_\ell\}$, and $m = \min(m_1, m_2)$.

All of the simple results that are available to us when data are generated by an exponential distribution can be used. In particular, if we wish to compare the means of two distributions, both subject to censoring, then we can transform one of them to standard exponential via its empirical survival function, then use this same transformation on the other group. The simple results for contrasting two censored exponential samples can then be applied even though, at least initially, the data arose from samples generated by some other mechanism.

ESTIMATING EQUATIONS AND MODEL ADEQUACY

A goodness-of-fit test is simply a test of a particular kind of hypothesis, notably a data-driven hypothesis where certain parameters, under the original specification of the model, have been replaced by estimates. We are then able to place these kinds of test under the same heading as the other tests, whether or not they appeal to any approach using likelihood. More formally, instead of testing a hypothesis of the type $H_0 : \beta = \beta_0$, for some β_0 or set of β_0 given in advance, we test a hypothesis of the type $H_0 : \beta = \hat{\beta}$. Evaluation of adequacy that goes beyond simple tests is helpful and knowing the form that certain processes under the model will take can furnish us with powerful techniques. One good choice can be based on processes that, under the model, will approximate that of a Brownian bridge. The construction, evaluated in greater detail in Chapter 9, is based upon being able to estimate the first two moments of the covariate at each failure point X_i . For parametric models using likelihood the dependence is typically expressed via the conditional distributions of time given the covariate, i.e., $f(t|z)$, that is unlike the case based on the main theorem and the processes described earlier. However, by applying Bayes' theorem successively we can obtain necessary expectations in terms of the conditional probabilities of T given Z , i.e., via the use of $f(t|z)$. Specifically we have:

$$E[Z - \mathcal{E}_\beta(Z|t)] = \int_Z \left\{ z - \frac{\int_Z z f(t|z) dG(z)}{\int_Z f(t|z) dG(z)} \right\} \left\{ \frac{f(t|z)}{\int_Z f(t|z) dG(z)} \right\} dG(z), \quad (7.8)$$

where we use the notation \mathcal{E}_β to denote an inner expectation conditional not only on the assumed model but also on the observations k and n , in addition to the configuration of censorings. The notation \mathcal{E}_β is not of great interest here, in the context of a fully parameterized model, but is of great importance in later sections and chapters. Figure 9.3(b) shows a bridged process for the Freireich data where the fitted model is a parametric Weibull model. The very good fit is confirmed by inspection of the curve. The essential difference between parametric and nonparametric, or semi-parametric, approaches is in the calculation of the means and variances. The observations are structured in the same way and it is the model, of whatever nature, that provides the necessary first two moments for us. In the parametric set-up, since the model specifies the distribution of time given the covariate, and we need to consider the distribution of the covariate given time, we make a simple appeal to Bayes' theorem. In the two-group case with constant hazard rates and where there is no censoring the above expression can be simplified. Letting the proportion of the first group be given by π_1 , the second by π_2 such that $(\pi_1 + \pi_2 = 1)$ and writing $\psi(v) = a(v)/\{\pi_1 e^{-v} + a(v)\}$ where $a(v) = \pi_2 e^\beta \exp(-ve^\beta)$ then

$$E\{[Z - \mathcal{E}_\beta(Z|t)]^p\} = \pi_1 \int_0^t \{-\psi(v)\}^p e^{-v} dv + \pi_2 \int_0^t \{1 - \psi(v)\} e^\beta \exp(-ve^\beta) dv.$$

The most immediate departure from proportional hazards would be one where effects decline through time and, as mentioned before, perhaps even changing direction (Stablein et al., 1981). The standardized cumulative score, however the moments are calculated, would, under such departures, increase (or decrease) steadily until the increase (or decrease) dies away and the process would proceed on average horizontally or, if effects change direction, make its way back toward the origin. A test based on the maximum would have the ability to pick up this kind of behavior. Again, the visual impression given by the cumulative standardized score process can, of itself, suggest the nature of the departure from proportional hazards. This kind of test is not focused on parameters in the model other than the regression effect given by β or possibly $\beta(t)$. The goal is to consider the proportionality or lack of such proportionality and, otherwise, how well the overall model may fit is secondary (Figure 7.1).

7.4 Semi-parametric estimating equations

Remind ourselves that the data consist of the observations $(Z_i(t), Y_i(t), (t \leq X_i), X_i; i = 1 \dots n)$. The Z_i are the covariates (possibly time-dependent), the $X_i = \min(T_i, C_i)$, the observed survival which is the smallest of the censoring time and the actual survival time, and the $Y_i(t)$ are time-dependent indicators taking the value one as long as the i th subject is at risk at time t and zero otherwise. For the sake of large sample constructions we make $Y_i(t)$ to be left continuous. At some level we will be making an assumption of independence, an

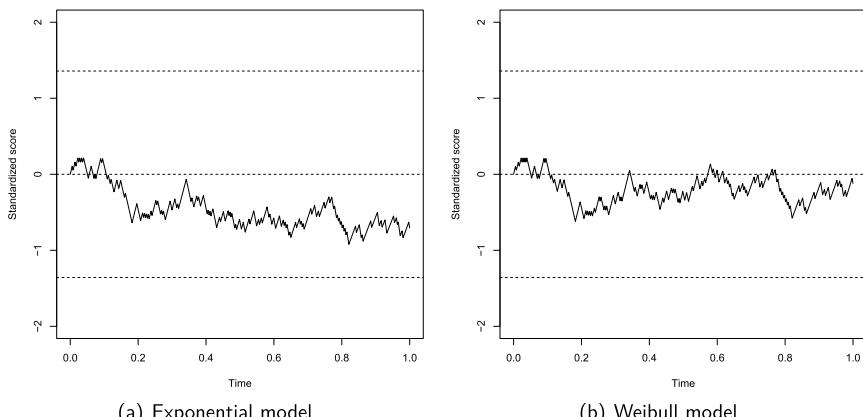


Figure 7.1: Graphical fit for two parametric models. A slight improvement in fit is seen for the Weibull model over the exponential model. Both are well within limits of significance and either assumption appears reasonable. There is a hint of weak non-significant time dependence that the Weibull model better accommodates.

assumption that can be challenged via the data themselves, but that is often left unchallenged, the physical context providing the main guide. Mostly, we think of independence as existing across the indices $i(i = 1, \dots, n)$, i.e., the triplets $\{Z_i(t), Y_i(t), X_i; i = 1, \dots, n\}$. It is helpful to our notational construction to have:

Definition 7.1. Let $Z(t)$ be a data-based step function of t , everywhere equal to zero except at the points X_i , $i = 1, \dots, n$, at which the function takes the value $Z_i(X_i)$ so that $Z(t) = \sum_{i=1}^n Z_i(t) \mathbf{1}_{X_i=t}$. We assume that $|Z_i|$ is bounded, if not the definition is readily broadened.

The reason for this definition is to unify notation. Our practical interest will be on sums of quantities such as $Z_i(X_i)$ with i ranging from 1 to n . Using the Stieltjes integral (Appendix A), we will be able to write such sums as integrals with respect to an empirical process. In view of the Helly-Bray theorem (Appendix A.2) this makes it easier to gain an intuitive grasp on the population structure behind the various statistics of interest. Both T and C are assumed to have supports on some finite interval, the first of which is denoted \mathcal{T} . The time-dependent covariate $Z(\cdot)$ is assumed to be a left-continuous stochastic process and, for notational simplicity, is taken to be of dimension one whenever possible.

We use the function $\text{Pr}(A)$ to return the probability measure associated with the event A , the reference sets for this being the largest probability space in the context. In other words, we have not restricted our outcome space by conditioning on any particular events. It is usually clear which probability space is assumed, if not we include \mathcal{F} to denote this space preceded by a colon, i.e., we write $\text{Pr}(A : \mathcal{F})$. The function $\mathbb{P}(A)$ also returns a probability measure associated with the event A and we reserve this usage for those cases where significant reduction of the original probability space has taken place. In other words, we view $\mathbb{P}(A)$ as a probability arising after conditioning, and, in general, after conditioning on a substantial part of the data structure. Again, the context is usually sufficient to know what is being conditioned on. If not this is made explicit. It is of interest, although not generally exploited, to observe that, under repeated sampling, A under $\mathbb{P}(A)$ will generally converge in distribution to that of A under $\text{Pr}(A)$. The discrete probabilities $\pi_i(\beta(t), t)$, defined in Equation 7.10, are so central to the development that they are given a notation all of their own. The expectation operator $E(\cdot)$ is typically associated with $\text{Pr}(\cdot)$ whereas the expectation operator $\mathcal{E}(\cdot | t)$ is associated with the model-based $\pi_i(\beta(t), t)$. Let $F(t) = \text{Pr}(T < t)$, $D(t) = \text{Pr}(C < t)$, and $H(t) = F(t)\{1 - D(t)\} - \int_0^t F(u)dD(u)$.

For each subject i we observe $X_i = \min(T_i, C_i)$, and $\delta_i = I(T_i \leq C_i)$ so that δ_i takes the value one if the i th subject corresponds to a failure and is zero if the subject corresponds to a censored observation. A more general situation allows a subject to be dynamically censored in that he or she can move in and out of the risk set. To do this we define the “at-risk” indicator $Y_i(t)$ where $Y_i(t) = I(X_i \geq t)$. The events on the i th individual are counted by $N_i(t) = I\{T_i \leq t, T_i \leq C_i\}$, and $\bar{N}(t) = \sum_1^n N_i(t)$ counts the number of events before t . Some other sums

of observations will frequently occur. In order to obtain an angle on empirical moments under the model, Andersen and Gill (1982) define

$$S^{(r)}(\beta, t) = n^{-1} \sum_{i=1}^n Y_i(t) e^{\beta Z_i(t)} Z_i(t)^r, \quad s^{(r)}(\beta, t) = E S^{(r)}(\beta, t),$$

for $r = 0, 1, 2$, where the expectations are taken with respect to the true distribution of $(T, C, Z(\cdot))$. Define also

$$V(\beta, t) = \frac{S^{(2)}(\beta, t)}{S^{(0)}(\beta, t)} - \frac{S^{(1)}(\beta, t)^2}{S^{(0)}(\beta, t)^2}, \quad v(\beta, t) = \frac{s^{(2)}(\beta, t)}{s^{(0)}(\beta, t)} - \frac{s^{(1)}(\beta, t)^2}{s^{(0)}(\beta, t)^2}. \quad (7.9)$$

The Andersen and Gill notation is now classic in this context. Their notation lends itself more readily to large sample theory based upon Rebollo's multivariate central limit theorem for martingales and stochastic integrals. We will keep this notation in mind for this chapter although, for subsequent chapters, we use a lighter notation since our approach to inference does not appeal to special central limit theorems (Rebollo's theorem in particular). One reason for using the Andersen and Gill notation in this chapter is to help the reader familiar with that theory to join up the dots and readily see the connections with chapters 9, 10, and 11. The required conditions for the Andersen and Gill theory to apply are slightly broader than those of our development although this advantage is more of a theoretical than a practical one. For their results, as well as ours, the censorship is restricted in such a way that, for large samples, there remains information on F in the tails. The conditional means and the conditional variances, $\mathcal{E}_{\beta(t)}(Z|t)$ $\mathcal{V}_{\beta(t)}(Z|t)$, introduced immediately below, are related to the above via $V(\beta, t) \equiv \mathcal{V}_{\beta}(Z|t)$ and $S^{(1)}(\beta, t)/S^{(0)}(\beta, t) \equiv \mathcal{E}_{\beta}(Z|t)$. In the counting process framework of Andersen and Gill (1982), we imagine n as remaining fixed and the asymptotic results obtaining as a result of asymptotic theory for n -dimensional counting processes, in which we understand the expectation operator E to be with respect to infinitely many repetitions of the process. Subsequently we allow n to increase without bound. For the quantities $\mathcal{E}_{\beta(t)}(Z^k|t)$ we take the E operator to be these same quantities when n grows without bound.

7.5 Estimating equations using moments

We most often view time as providing the set of indices to certain stochastic processes, so that, for example, we consider $Z(t)$ to be a random variable having different distributions for different t . Also, the failure time variable T can be viewed as a non-negative random variable with distribution $F(t)$ and, whenever the set of indices t to the stochastic process coincides with the support for T , then not only can we talk about the random variables $Z(t)$ for which the distribution corresponds to $\Pr(Z \leq z|T = t)$ but also marginal quantities such as

the random variable $Z(T)$ having distribution $G(z) = \Pr(Z \leq z)$. An important result concerning the conditional distribution of $Z(t)$ given $T = t$ follows. However, the true population joint distribution of (T, Z) turns out to be of little interest. Concerning T , we will view its support mostly in terms of providing indices to a stochastic process. The variable Z depends of course on rather arbitrary design features. In order to study dependency, quantified by $\beta(t)$, we focus on the conditional distribution of Z given $T = t$.

Definition 7.2. *The discrete probabilities $\pi_i(\beta(t), t)$ are given by*

$$\pi_i(\beta(t), t) = \frac{Y_i(t) \exp\{\beta(t)Z_i(t)\}}{\sum_{j=1}^n Y_j(t) \exp\{\beta(t)Z_j(t)\}}. \quad (7.10)$$

The $\pi_i(\beta(t), t)$ are easily seen to be bonafide probabilities (for all real values of $\beta(t)$) since $\pi_i \geq 0$ and $\sum_i \pi_i = 1$. Note that this continues to hold for values of $\beta(t)$ different from those generating the data, and even when the model is incorrectly specified. As a consequence, replacing β by $\hat{\beta}$ results in a probability distribution that is still valid but different from the true one. Means and variances with respect to this distribution maintain their interpretation as means and variances.

Under the proportional hazards assumption, i.e., the constraint $\beta(t) = \beta$, the product of the π 's over the observed failure times gives the partial likelihood (Cox 1972, 1975). When $\beta = 0$, $\pi_i(0, t)$ is the empirical distribution that assigns equal weight to each sample subject in the risk set. Based on the $\pi_i(\beta(t), t)$ we have:

Definition 7.3. *Conditional moments of Z with respect to $\pi_i(\beta(t), t)$ are given by*

$$\mathcal{E}_{\beta(t)}(Z^k | t) = \sum_{i=1}^n Z_i^k(t) \pi_i(\beta(t), t), \quad k = 1, 2, \dots. \quad (7.11)$$

These two definitions are all that we need in order to set about building the structures upon which inference is based. This is particularly so when we are able to assume an independent censoring mechanism, although the weaker assumption of a conditionally independent censoring mechanism (see Chapter 2) will mostly cause no conceptual difficulties; simply a slightly more burdensome notation. Another, somewhat natural, definition will also be appealed to on occasion and this concerns unconditional expectations.

Definition 7.4. *Marginal moments of Z with respect to the bivariate distribution characterized by $\pi_i(\beta(t), t)$ and $F(t)$ are given by*

$$\mathcal{E}_{\beta(t)}(Z^k) = \int \mathcal{E}_{\beta(t)}(Z^k | t) dF(t), \quad k = 1, 2, \dots. \quad (7.12)$$

Recall that for arbitrary random variables A and B , assuming expectation to be defined, we have the result of double expectation whereby $E(A) = EE(A|B)$. This is the motivation behind the above definition. Once again, these expectations are to be interpreted as population quantities in as much as $\beta(t)$ and $F(t)$ are taken to be known. They can also, of course, be viewed as sample-based quantities since n is finite and the $Y_i(t)$ are random until time point t . At the end of the study the paths of all the $Y_i(t)$ are known and we are, to use a common expression, “conditioning on the data.” The art of inference, and its understanding, stem, to a great extent, from knowing which aspects of an experiment to view as random (given that once the experiment is over there is not really anything truly random). Also which distributions are relevant and these can change so that, here for example, we should think carefully about the meaning of the expectation operators E and \mathcal{E} in its particular context. These expectations are still well defined, but with respect to different distributions; when replacing β by $\hat{\beta}$, when replacing F by F_n and \hat{F} , and when allowing n to go to infinity. The quantity ϕ of the following definition is not of any essential interest, featuring in the main theorem but disappearing afterwards.

Definition 7.5. In order to distinguish conditionally independent censoring from independent censoring we introduce $B_C(t,z) = P(C \geq t|z)$ and define $\phi(z,t)$ where

$$\phi(z^*,t) = B_C(t,z^*)^{-1} \times \int P(C \geq t|z)dG(z).$$

Note that when censoring does not depend upon z then $\phi(z,t)$ will depend upon neither z nor t and is, in fact, equal to one. Otherwise, under a conditionally independent censoring assumption, we can consistently estimate $\phi(z,t)$ and we call this $\hat{\phi}(z,t)$. This is not explored in this text.

Theorem 7.2. (O’Quigley 2003). Under model (4.2) and assuming $\beta(t)$ known, the conditional distribution function of $Z(t)$ given $T = t$ is given by

$$\mathbb{P}\{Z(t) \leq z|T = t\} = \frac{\sum_{z_i \leq z} Y_i(t) \exp\{\beta(t)z_i(t)\}\hat{\phi}(z_i,t)}{\sum_{j=1}^n Y_j(t) \exp\{\beta(t)z_j(t)\}\hat{\phi}(z_j,t)}. \quad (7.13)$$

The theorem, which we refer to as the main theorem of proportional hazards regression, has many important consequences including

Corollary 7.7. Under model (4.2) and an independent censorship, assuming $\beta(t)$ known, the conditional distribution function of $Z(t)$ given $T = t$ is given by

$$\mathbb{P}(Z(t) \leq z | T = t) = \sum_{j=1}^n \pi_j(\beta(t), t) I(Z_j(t) \leq z). \quad (7.14)$$

The observation we would like to make here is that we can fully describe a random variable indexed by t , i.e., a stochastic process. This idea underlies the development of the regression effect process described in the following chapters. All of our inferences can be based on this. In essence, we first fix t and then we fix our attention on the conditional distribution of Z given that $T = t$. This distribution brings into play the models of interest. These models are characterized by this distribution making it straightforward to construct tests, to estimate parameters, and to build confidence regions. Indeed, under the broader censoring definition of conditional independence, common in the survival context, we can still make the same basic observation. In this case we condition upon something more complex than just $T = t$. The actual random outcome that we condition upon is of less importance than the simple fact that we are able to describe sets of conditional distributions all indexed by t , i.e., a stochastic process indexed by t . Specifically

Corollary 7.8. *For a conditionally independent censoring mechanism we have*

$$\mathbb{P}(Z(t) \leq z | T = t, C > t) = \sum_{j=1}^n \pi_j(\beta(t), t) I(Z_j(t) \leq z). \quad (7.15)$$

Whether we condition on the event $T = t$ or the event $(T = t, C > t)$, we identify a random variable indexed by t . This is all we need to construct appropriate stochastic processes (functions of $Z(t)$) enabling inference. Again simple applications of Slutsky's theorem show that the result still holds, as a large sample approximation, for $\beta(t)$ replaced by any consistent estimate. In particular, when the hypothesis of proportionality of risks is correct, the result holds for the estimate $\hat{\beta}$. The following two corollaries follow immediately from those just above and form the basis for simple inference. For integer k we have:

Corollary 7.9. $\mathcal{E}_{\hat{\beta}(t)}(Z^k | t)$ provides a consistent estimate of $E_{\beta(t)}(Z^k(t) | t)$, under model (4.2). In particular $\mathcal{E}_{\hat{\beta}}(Z^k | t)$ provides a consistent estimate of $E_{\beta}(Z^k(t) | t)$, under the model expressed by Equation 4.3.

Furthermore, once again working under the model, we consider

Definition 7.6. $\mathcal{V}_{\beta(t)}(Z|t) = \mathcal{E}_{\beta(t)}(Z^2|t) - \mathcal{E}_{\beta(t)}^2(Z|t)$.

In practical data analysis the quantity $\beta(t)$ may be replaced by a value constrained by some hypothesis or an estimate. The quantity $\mathcal{V}_{\beta(t)}(Z|t)$ can be viewed as a conditional variance which may vary little with t , in a way analogously to the residual variance in linear regression which, under classic assumptions, remains constant with different levels of the independent variable. Since $\mathcal{V}_{\beta(t)}(Z|t)$ may change with t , even if not a lot, it is of interest to consider some average quantity and so we also introduce

Definition 7.7. $\sigma^2 = E\mathcal{V}_{\beta(t)}(Z) = \int \mathcal{V}_{\beta(t)}(Z|t)dF(t)$.

Interpretation requires some care. For example, although $E\mathcal{V}_{\beta}(Z|t)$ is, in some sense, a marginal quantity, it is not the marginal variance of Z since we have neglected the variance of $E_{\beta(t)}(Z(t)|t)$ with respect to the distribution of T . The easiest case to interpret is the one where we have an independent censoring mechanism (Equation 7.14). However, we do not need to be very concerned about any interpretation difficulty, arising for instance in Equation 7.15 where the censoring time appears in the expression, since, in this or the simpler case, all that matters to us is that our observations can be considered as arising from some process, indexed by t and, for this process, we are able, under the model, to consistently estimate the mean and the variance of the quantities that we observe. It is also useful to note another natural relation between $\mathcal{V}_{\beta}(Z|t)$ and $\mathcal{E}_{\beta}(Z|t)$ since

$$\mathcal{V}_{\beta}(Z|t) = \{\partial\mathcal{E}_{\theta}(Z|t)/\partial\theta.\}_{\theta=\beta} \quad (7.16)$$

This relation is readily verified for fixed β . In the case of time-dependent $\beta(t)$ then, at each given value of t , it is again clear that the same relation holds. The result constitutes one of the building blocks in the overall inferential construction and, under weak conditions, essentially no more than Z being bounded, then it also follows that

$$\int \mathcal{V}_{\beta}(Z|t) = \int \partial\mathcal{E}_{\beta}(Z|t)/\partial\beta = \partial \left\{ \int \mathcal{E}_{\beta}(Z|t) \right\} / \partial\beta.$$

Essentially all the information we need, for almost any conceivable statistical goal, arising from considerations of any of the models considered, is contained in the joint probabilities $\pi_i(\beta(t), t)$ of the fundamental definition 7.2. We are often interested, in the multivariate setting for example, in the evaluation of the effects of some factor while having controlled for others. This can be immediately accommodated. Specifically, taking Z to be of some dimension greater than one (β being of the same dimension), writing $Z^T = (Z_1^T, Z_2^T)$ and $Z_i^T = (Z_{1i}^T, Z_{2i}^T)$, and then summing over the multivariate probabilities, we have two obvious extensions to Corollaries 7.7 and 7.8.

Corollary 7.10. *Under model (4.2) and an independent censorship, assuming $\beta(t)$ known, the conditional distribution function of $Z_2(t)$ given $T = t$ is given by*

$$\mathbb{P}(Z_2(t) \leq z | T = t) = \sum_{j=1}^n \pi_j(\beta(t), t) I(Z_{2j}(t) \leq z). \quad (7.17)$$

The corollary enables component wise inference. We can consider the components of the vector Z_i individually. Also we could study some functions of the components, usually say a simple linear combination of the components such as the prognostic index. Note also that

Corollary 7.11. *For a conditionally independent censoring mechanism we have*

$$\mathbb{P}(Z_2(t) \leq z | T = t, C > t) = \sum_{j=1}^n \pi_j(\beta(t), t) I(Z_{2j}(t) \leq z), \quad (7.18)$$

where in Definition 7.2 for $\pi_j(\beta(t), t)$ we take $\beta(t)Z_j(t)$ to be an inner product, which we may prefer to write using boldface or as $\beta(t)^T Z_j(t)$ and where $Z_j(t)$ are the observed values of the vector $Z(t)$ for the j th subject. Also, by $Z_2(t) \leq z$ we mean that all of the scalar components of $Z_2(t)$ are less than or equal to the corresponding scalar components of z . As for the corollaries and definitions following Corollaries 7.7 and 7.8 they have obvious equivalents in the multivariate setting and so we can readily write down expressions for expectations, variances, and covariances as well as their corresponding estimates.

MOMENTS FOR STRATIFIED MODELS

Firstly we recall from the previous chapter that the stratified model is simply a partially proportional hazards model in which some of the components of $\beta(t)$ remain unspecified while the other components are constant terms. The definition for the stratified model was

$$\lambda(t|Z(t), s) = \lambda_{0s}(t) \exp\{\beta(t)Z(t)\},$$

where s takes integer values $1, \dots, m$. In view of the equivalence between stratified models and partially proportional hazards models described in the previous chapter, the main theorem and its corollaries apply immediately. However, in light of the special importance of stratified models, as proportional hazards models with relaxed assumptions, it will be helpful to our development to devote a few words to this case. Analogous to the above definition for $\pi_i(\beta(t), t)$, and using the, possibly time-dependent, stratum indicator $s(t)$ we now define these probabilities via

Definition 7.8. For the stratified model, having strata $s = 1, \dots, m$, the discrete probabilities $\pi_i(\beta(t), t)$ are now given by

$$\pi_i(\beta(t), t) = \frac{Y_i\{s(t), t\} \exp\{\beta(t)Z_i(t)\}}{\sum_{j=1}^n Y_j\{s(t), t\} \exp\{\beta(t)Z_j(t)\}}. \quad (7.19)$$

When there is a single stratum then this definition coincides with the earlier one and, indeed, we use the same $\pi_i(\beta(t), t)$ for both situations, since it is only used indirectly and there is no risk of confusion. Under equation (4.3), i.e., the constraint $\beta(t) = \beta$, the product of the π 's over the observed failure times gives the so-called stratified partial likelihood (Kalbfleisch and Prentice, 2002). The series of above definitions for the non-stratified model, in particular Definition 7.2, theorems, and corollaries, all carry over in an obvious way to the stratified model and we do not propose any additional notation. It is usually clear from the context although it is worth making some remarks. Firstly, we have no direct interest in the distribution of Z given t (note that this distribution depends on the distribution of Z given $T > 0$, a distribution which corresponds to our design and is quite arbitrary).

We will exploit the main theorem in order to make inferences on β and, in the stratified case, we would also condition upon the strata from which transitions can be made. In practice, we contrast the observations $Z_i(X_i)$, made at time point X_i at which an event occurs ($\delta_i = 1$) with those subjects at risk of the same event. The “at-risk” indicator, $Y(s(t), t)$, makes this very simple to express. We can use $Y(s(t), t)$ to single out appropriate groups for comparison. This formalizes a standard technique in epidemiology whereby the groups for comparison may be matched by not just age but by other variables. Such variables have then been controlled for and eliminated from the analysis. Their own specific effects can be quite general and we are not in a position to estimate them. Very complex situations, such as subjects moving in and out of risk categories, can be easily modeled by the use of these indicator variables.

MOMENTS FOR OTHER RELATIVE RISK MODELS

Instead of Equation 4.2 some authors have suggested a more general form for the hazard function whereby

$$\lambda(t|Z) = \lambda_0(t)R\{\beta(t)Z\}, \quad (7.20)$$

and where, mostly, $\beta(t)$ is not time-varying, being equal to some unknown constant. The most common choices for the function $R(r)$ are $\exp(r)$, in which case we recover the usual model, and $1+r$ which leads to the so-called additive model. Since both $\lambda(t|Z)$ and λ_0 are necessarily positive we would generally need constraints on the function $R(r)$. In practice this can be a little bothersome and is, among several other good reasons, a cause for favoring the multiplicative

risk model $\exp(r)$ over the additive risk model $1+r$. If we replace our earlier definition for $\pi_i(\beta(t), t)$ by

Definition 7.9. *The discrete probabilities $\pi_i(\beta(t), t)$ are given by*

$$\pi_i(\beta(t), t) = \frac{Y_i(t)R\{\beta(t)Z_i(t)\}}{\sum_{j=1}^n Y_j(t)R\{\beta(t)Z_j(t)\}}, \quad (7.21)$$

following which all of the above definitions, theorems, and corollaries have immediate analogues and we do not write them out explicitly. Apart from one interesting exception, which we look at more closely in the chapters dealing with inference, there are no particular considerations we need to concern ourselves over if we choose $R(r) = 1+r$ rather than $R(r) = \exp(r)$.

What is more, if we allow the regression functions, $\beta(t)$, to depend arbitrarily upon time then, given either model, the other model exists with a different function of $\beta(t)$. The only real reason for preferring one model over another would be due to parsimony; for example, we might find in some given situation that in the case of the additive model the regression function $\beta(t)$ is in fact constant unlike the multiplicative model where it may depend on time. But otherwise both functions may depend, at least to some extent, on time and then the multiplicative model ought to be preferred since it is the more natural. We say the more natural because the positivity constraint is automatically satisfied.

TRANSFORMED COVARIATE MODELS

For some transformation ψ of the covariate we can postulate a model of the form

$$\lambda(t|Z(t)) = \lambda_0(t) \exp\{\beta(t)\psi[Z(t)]\}. \quad (7.22)$$

All of the calculations proceed as above and no real new concept is involved. Such models can be considered in the case of continuous covariates, Z , which may be sufficiently asymmetric, implying very great changes of risk at the high or low values, to be unlikely to provide a satisfactory fit. Taking logarithms, or curbing the more extreme values via a defined plateau, or some other such transformation will produce models of potentially wider applicability. Note that this is a different approach to work with, say,

$$\lambda(t|Z(t)) = \lambda_0(t) \exp\{\beta(t)Z(t)\},$$

and using the main theorem, in conjunction with estimating equations described here below and basing inference upon the observations $\psi Z(X_i)$ and their expectations under this model. In this latter case we employ ψ in the estimating equation as a means to obtain greater robustness or to reduce sensitivity to large

observations. In the former case the model itself is different and would lead to different estimates of survival probabilities.

Our discussion so far has turned around the hazard function. However, it is equally straightforward to work with intensity functions and these allow for increased generality, especially when tackling complex time-dependent effects. O'Brien (1978) introduced the logit-rank test for survival data when investigating the effect of a continuous covariate on survival time. His purpose was to construct a test that was rank invariant with respect to both time and the covariate itself. O'Quigley and Prentice (1991) showed how a broad class of rank invariant procedures can be developed within the framework of proportional hazards models. The O'Brien logit-rank procedure was a special case of this class. In these cases we work with intensity rather than hazard functions. Suppose then that $\lambda_i(t)$ indicates an intensity function for the i th subject at time t . A proportional hazards model for this intensity function can be written as

$$\lambda_i(t) = Y_i(t)\lambda_0(t)\exp\{\beta Z_i(t)\},$$

where $Y_i(t)$ indicates whether or not the i^{th} subject is at risk at time t , $\lambda_0(t)$ the usual “baseline” hazard function, and $Z_i(t)$ is a constructed covariate for the i th subject at time t . Typically, $Z_i(t)$ in the estimating equation is defined as a function of measurements on the i th subject alone, but it can be defined more generally as $Z_i(t) = \psi_i(t, \mathcal{F}_t)$ for ψ some function of \mathcal{F}_t , the collective failure, censoring, and covariate information prior to time t on the entire study group. The examples in O'Quigley and Prentice (1991) included the rank of the subject's covariate at X_i and transformations on this such as the normal order statistics. This represents a departure from most regression situations because the value used in the estimating equation depends not only on what has been observed on the particular individual but also upon what has been observed on other relevant subsets of individuals.

STRUCTURED TIME EFFECTS FOR $\beta(t)$

In the review of goodness-of-fit tests for survival models, O'Quigley and Xu (2014) gave particular consideration to the non-proportional hazards model with intercept described in Chapter 4. Using this model and the usual likelihood procedures we can test for several specific departures from proportional hazards.

For the sake of simplicity of exposition we continue to limit attention to the single variable case. Extension to higher dimensions is immediate. Also we will sometimes write Z , instead of $Z(t)$, in order for the notation to not become over cluttered. It can always be replaced by $Z(t)$ without additional concerns. We write

$$\lambda(t|Z) = \lambda_0(t)\exp\{[\beta + \alpha Q(t)]Z(t)\}, \quad (7.23)$$

where $Q(t)$ is a function of time that does not depend on the parameters β and α . Under a null hypothesis that a proportional hazards model is adequate, i.e., $H_0 : \alpha = 0$, we recover a proportional hazards model. In the context of sequential group comparisons of survival data, the above model has been considered by Tsiatis (1982) and Fleming and Harrington (1984). In keeping with the usual notation we denote $\mathcal{E}_{\beta,\alpha}(Z|t)$ to be the expectation taken with respect to the probability distribution $\pi_i(\beta, \alpha, t)$, where

$$\pi_i(\beta, \alpha, t) = \frac{Y_i(t) \exp\{[\beta + \alpha Q(t)] Z_i(t)\}}{\sum_{j=1}^n Y_j(t) \exp\{[\beta + \alpha Q(t)] Z_j(t)\}}. \quad (7.24)$$

Lemma 7.2. *The components of the score vector $U(\beta, \alpha)$ can be expressed as*

$$U_\beta(\beta, \alpha) = \sum_{i=1}^n \delta_i \{Z_i(X_i) - \mathcal{E}_{\beta,\alpha}(Z|X_i)\}, \quad (7.25)$$

$$U_\alpha(\beta, \alpha) = \sum_{i=1}^n \delta_i Q(X_i) \{Z_i(X_i) - \mathcal{E}_{\beta,\alpha}(Z|X_i)\}. \quad (7.26)$$

A test would be carried out using any one of the large sample tests arising from considerations of the likelihood. If we let $\hat{\beta}$ be the maximum partial likelihood estimate of β under the null hypothesis of proportional hazards, i.e., $H_0 : \alpha = 0$, then $U_\beta(\hat{\beta}, 0) = 0$. The score test statistic arising under H_0 is $B = U_\alpha(\hat{\beta}, 0)G^{-1}U_\alpha(\hat{\beta}, 0)$, where $G = I_{22} - I_{21}I_{11}^{-1}I_{12}$ and G^{-1} is the lower right corner element of I^{-1} . The elements of the information matrix required to carry out the calculation are given below in Lemma 7.3. Under H_0 , the hypothesis of proportional hazards, B has asymptotically a χ^2 distribution with one degree of freedom.

Lemma 7.3. *Taking $k = 1, 2$ and $\ell = 1, 2$, the components of I are*

$$I(\beta, \alpha) = - \begin{pmatrix} U_{\beta\beta} & U_{\beta\alpha} \\ U_{\alpha\beta} & U_{\alpha\alpha} \end{pmatrix} = \begin{pmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{pmatrix} \text{ where}$$

$$I_{k\ell}(\beta, \alpha) = \sum_{i=1}^n \delta_i Q(X_i)^{k+\ell-2} \{\mathcal{E}_{\beta,\alpha}(Z^2|X_i) - \mathcal{E}_{\beta,\alpha}^2(Z|X_i)\}.$$

The literature on the goodness-of-fit problem for the Cox model has considered many formulations that correspond to particular choices for $Q(t)$. The first of these was given in the founding paper of Cox (1972). Cox's suggestion was equivalent to taking $Q(t) = t$. Defining $Q(t)$ as a two-dimensional vector, Staelein et al. (1981) considered $Q(t) = (t, t^2)'$, and Brown (1975), Anderson and Senthil Selvan (1982), O'Quigley and Moreau (1984), Moreau et al. (1985), and

O'Quigley and Pessione (1989) assumed $Q(t)$ to be constant on predetermined intervals of the time axis, i.e., $Q(t)$ is a step function.

Although in the latter cases there is more than one parameter associated with $Q(t)$, the computation of the test statistic is similar. Murphy (1993) studied the size and the power of the test of Moreau and colleagues and found that, although it is consistent against a wide class of alternatives to proportional hazards, the Moreau test is nonetheless an omnibus test that is used to greatest advantage when there is no specific alternative in mind.

We can choose $Q(t)$ to be an unknown function and use the available data to provide an estimate of this function. For instance, Breslow et al. (1984) chose $Q(t) = \Lambda(t)$ and replaced this unknown function by the Nelson estimator. Because the estimates at time t depend only on the history of events up to that time, the development of Cox (1975) and Andersen and Gill (1982) and thus, the usual asymptotic theory, still applies. Breslow et al. (1984) showed that their choice $Q(t) = \Lambda(t)$ has good power against the alternative of crossing hazards. Tsiatis (1982) and Harrington and Fleming (1982) used score processes based on the non-proportional hazards model with intercept to derive sequential tests. They showed that after $Q(t)$ has been replaced by its estimate, the score process at different time points converges in distribution to a multivariate normal. Harrington and Fleming focus particular interest on the G^ρ family, where $Q(t) = S(t)^\rho$.

Another special case is described in O'Quigley and Pessione (1991), O'Quigley and Natarajan (2004), and O'Quigley and Flandre (1994), where $Q(t) = I(t \leq \gamma) - I(t > \gamma)$ with γ an unknown changepoint. When γ is known the test statistic can be evaluated with no particular difficulty. For γ unknown, we would maximize over all possible values of γ and special care is required for the resulting inference to remain valid. O'Quigley and Pessione (1991) showed that tests using a changepoint model can be powerful for testing the equality of two survival distributions against the specific alternative of crossing hazards. Also, such tests suffer only moderate losses in power, when compared with their optimal counterparts, if the alternative is one of proportional hazards.

Lin (1991) and Gill and Schumacher (1987) have taken a slightly different approach to work with the function $Q(t)$ and introduce it directly into a weighted score. This can be written as

$$U_Q(\beta) = \sum_{i=1}^n \delta_i Q(X_i) \{Z_i(X_i) - E(Z|X_i; \beta)\}, \quad (7.27)$$

where the only key requirement is that $Q(\cdot)$ be a predictable process (see Section B.4) that converges in probability to a non-negative bounded function uniformly in t . Let $\hat{\beta}_Q$ be the zero of (7.27) and $\hat{\beta}$ be the partial likelihood estimate. Under the assumption that the proportional hazards model holds and that $(X_i, \delta_i, Z_i) (i = 1, \dots, n)$ are i.i.d. replicates of (X, δ, Z) , $n^{1/2}(\hat{\beta}_Q - \hat{\beta})$ is asymptotically normal with zero mean and covariance matrix that can be consistently

estimated. It then follows that a simple test can be based on the standardized difference between the two estimates. Lin (1991) showed such a test to be consistent against any model misspecification under which $\beta_Q \neq \beta$, where β_Q is the probability limit of $\hat{\beta}_Q$. In particular, it can be shown that choosing a monotone weight function for $Q(t)$ such as $\hat{F}(t)$, where $\hat{F}(\cdot)$ is the Kaplan-Meier estimate, is consistent against monotone departures (e.g., decreasing regression effect) from the proportional hazards assumption.

7.6 Incorrectly specified models

For multinormal linear regression involving p regressors we can eliminate from consideration some of these and focus our attention on models involving the remaining regressors strictly less than p . We could eliminate these by simple integration, thereby obtaining marginal distributions. Under the usual assumptions of multiple linear regression the resulting lower dimensional model remains a multinormal one. As an example, in the simple case of a two-dimensional covariate normal model, both the marginal models involving only one of the two covariates are normal models. However, for non-linear models this result would only be expected to hold under quite unusual circumstances. Generally, for non-linear models, and specifically proportional hazards models, the result will not hold so that if the model is assumed true for a covariate vector of dimension p , then, for any sub-model, of dimension less than p , the model will not hold exactly. A corollary to this is that no model of dimension greater than p could exactly follow a proportional hazards prescription if we claim that the model holds precisely for some given p covariates.

These observations led some authors to claim that “forgotten” or “overlooked” variables would inevitably lead to misleading results. Such a claim implies that *all* analyses based on proportional hazards models are misleading and since, to say the least, such a conclusion is unhelpful we offer a different perspective. This says that *all* practical models are only ever approximately correct. In other words, the model is always making a simplifying assumption, necessarily overlooking potential effects as well as including others which may impact the proportionality of those key variables of interest. Our task then focuses on interpreting our estimates when our model cannot be exactly true. In terms of analyzing real data, it makes much more sense to take as our underlying working assumption that the model is, to a greater or lesser degree, misspecified.

A model can be misspecified in one of two clear ways; the first is that the covariate form is not correctly expressed and the second is that the regression coefficient is not constant through time. An example of the first would be that the true model holds for $\log Z$ but that, not knowing this, we include Z in the model. An example of the second might have $\beta(t)$ declining through time rather than remaining constant.

It has been argued that the careful use of residual techniques can indicate which kind of model failure may be present. This is not so. Whenever a poor fit could be due to either cause it is readily seen that a misspecified covariate form can be represented correctly via a time-dependent effect. In some sense the two kinds of misspecification are unidentifiable. We can fix the model by working either with the covariate form or the regression coefficient $\beta(t)$. Of course, in certain cases, a discrete binary covariate describing two groups, for example, there can only be one cause of model failure—the time dependency of the regression coefficient. This is because the binary coding imposes no restriction of itself since all possible codings are equivalent.

AVERAGE REGRESSION EFFECT

The important issue is then the interpretation of an estimate say $\hat{\beta}$ under a proportional hazards assumption when, in reality, the data are generated under the broader non-proportional hazards model with regression coefficient function $\beta(t)$. This is not a straightforward endeavor and the great majority of the currently used procedures, including those proposed in the widely distributed R, SAS, STATA, and S-Plus packages, produce estimates which cannot be interpreted unless there is no censoring. To study this question we first define $\mu = \int \beta(t)dF(t)$, which is an average of $\beta(T)$ with respect to the distribution $F(t)$.

It is of interest to consider the approximation,

$$\mathbb{P}(Z(t) \leq z | T = t, C > t) \approx \sum_{j=1}^n \pi_j(\mu, t) I(Z_j(t) \leq z) \quad (7.28)$$

and, for the case of a model making the stronger assumption of an independent censoring mechanism as opposed to a conditionally independent censoring mechanism given the covariate, we have

$$\mathbb{P}(Z(t) \leq z | T = t) \approx \sum_{j=1}^n \pi_j(\mu, t) I(Z_j(t) \leq z). \quad (7.29)$$

For small samples it will be unrealistic to hope to obtain reliable estimates of $\beta(t)$ for all of t so that, often, we take an estimate of some summary measure, in particular μ . It is in fact possible to construct an estimating equation which provides an estimate of μ without estimating $\beta(t)$ (Xu and O'Quigley, 2000) and it is very important to stress that, unless there is no censoring, the usual estimating equation which leads to the partial likelihood estimate does not accomplish this.

Some thought needs to be given to the issues arising when our estimating equation is based on certain assumptions (in particular, a proportional hazards assumption), whereas the data themselves can be considered to have been generated by something broader (in particular, a non-proportional hazards model).

To this purpose we firstly consider a definition that will allow us to anticipate just what is being estimated when the data are generated by model (4.2) and we are working with model (4.3). This is contained in the definition for β^* just below.

Let's keep in mind the widely held belief that the partial likelihood estimate obtained when using a proportional hazards model in a situation where the data are generated by a broader model must correspond to some kind of average effect. It does correspond to something (as always) but nothing very useful and not something we can hopefully interpret as an average effect. Firstly we need

Definition 7.10. Let β^* be the constant value satisfying

$$\int_{\mathcal{T}} \mathcal{E}_{\beta^*}(Z|t)dF(t) = \int_{\mathcal{T}} \mathcal{E}_{\beta(t)}(Z|t)dF(t). \quad (7.30)$$

The definition enables us to make sense out of using estimates based on (4.3) when the data are in fact generated by (4.2). Since we can view T as being random, whenever $\beta(t)$ is not constant, we can think of having sampled from $\beta(T)$. The right-hand side of the above equation is then a double expectation and β^* , occurring in the left-hand side of the equation, is the best fitting value under the constraint that $\beta(t) = \beta$. We can show the existence and uniqueness of solutions to Equation 7.30 (Xu and O'Quigley, 2000). More importantly, β^* can be shown to have the following three properties: (i) under model (4.3) $\beta^* = \beta$; (ii) under a subclass of the broad class of models known as the Harrington-Fleming models, we have an exact result in that $\beta^* = \int_{\mathcal{T}} \beta(t)dF(t)$; and (iii) for very general situations we can write that $\beta^* \approx \int_{\mathcal{T}} \beta(t)dF(t)$, an approximation which is in fact very accurate. Estimates of β^* are discussed in Xu and O'Quigley (2000) and, in the light of the foregoing, we can take these as estimates of μ .

FAMILIES OF ESTIMATING EQUATIONS

The above setting helps us anticipate the properties of the estimators we will be using. First, recall our definition of $\mathcal{Z}(t)$ as a step function of t with discontinuities at the points X_i , $i = 1, \dots, n$, at which the function takes the value $Z_i(X_i)$. Next, consider $F_n(t)$, the empirical marginal distribution function of T . Note that $F_n(t)$ coincides with the Kaplan-Meier estimate of $F(t)$ in the absence of censoring. When there is no censoring, a sensible estimating equation (which we will see also arises as the derivative of a log likelihood, as well as the log partial likelihood) is

$$U_1(\beta) = \int \{\mathcal{Z}(t) - \mathcal{E}_{\beta}(Z|t)\}dF_n(t) = 0. \quad (7.31)$$

The above integral is simply the difference of two sums, the first the empirical mean without reference to any model and the second the average of model-based means. It makes intuitive sense as an estimating equation and the only reason

for writing the sum in the less immediate form as an integral is that it helps understand the large sample theory when $F_n(t) \xrightarrow{P} F(t)$. Each component in the above sum includes the size of the increment, $1/n$, a quantity that can then be taken outside of the summation (or integral) as a constant factor. Since the right-hand side of the equation is identically equal to zero, the incremental size $1/n$ can be canceled, enabling us to rewrite the equation as

$$U_2(\beta) = \int \{\mathcal{Z}(t) - \mathcal{E}_\beta(Z|t)\} d\bar{N}(t) = 0. \quad (7.32)$$

It is this expression the integral is taken with respect to increments $d\bar{N}(t)$, rather than with respect to $dF_n(t)$ that is the more classic representation in this context. The expression equates $U_2(\beta)$ in terms of the counting processes $N_i(t)$. These processes, unlike the empirical distribution function, are available in the presence of censoring. It is the above equation that is used to define the partial likelihood estimator, since, unless the censoring is completely absent, the quantity $U_1(\beta)$ is not defined.

Now, suppose that two observers were to undertake an experiment to estimate β . A certain percentage of observations remain unobservable to the first observer as a result of an independent censoring mechanism but are available to the second observer. The first observer uses Equation 7.32 to estimate β , whereas the second observer uses Equation 7.31. Will the two estimates agree? By “agree” we mean, under large sample theory, will they converge to the same quantity. We might hope that they would; at least if we are to be able to usefully interpret estimates obtained from Equation 7.32. Unfortunately though (especially since Equation 7.32 is so widely used), the estimates do not typically agree, the greater the degree of censoring, even when independent, the greater the disagreement. Table 7.1 below indicates just how severe the disagreement might be. However, the form of $U_1(\beta)$ remains very much of interest and, before discussing the properties of the above equations let us consider a third estimating equation which we write as

$$U_3(\beta) = \int \{\mathcal{Z}(t) - \mathcal{E}_\beta(Z|t)\} d\hat{F}(t) = \int W(t) \{\mathcal{Z}(t) - \mathcal{E}_\beta(Z|t)\} d\bar{N}(t) = 0, \quad (7.33)$$

upon defining the stochastic process $W(t) = \hat{S}(t) \{\sum_{i=1}^n Y_i(t)\}^{-1}$. For practical calculation note that $W(X_i) = \hat{F}(X_i+) - \hat{F}(X_i)$ at each observed failure time X_i , i.e., the jump in the KM curve. When there is no censoring, then clearly

$$U_1(\beta) = U_2(\beta) = U_3(\beta).$$

More generally $U_1(\beta)$ may not be available and solutions to $U_2(\beta) = 0$ and $U_3(\beta) = 0$ do not coincide or converge to the same population counterparts even under independent censoring. They would only ever converge to the same quantities under the unrealistic assumption that the data are exactly generated

by a proportional hazards model. As argued in the previous section we can assume that this never really holds in practical situations.

Many other possibilities could be used instead of $U_3(\beta)$, ones in which other consistent estimates of $F(t)$ are used in place of $\hat{F}(t)$, for example, the Nelson-Aalen estimator or, indeed, any parametric estimate for marginal survival. If we were to take the route of parametric estimates of marginal survival, we would need to be a little cautious since these estimates could also contain information on the parameter β which is our central focus. However, we could invoke a conditional argument, i.e., take the marginal survival estimate as fixed and known at its observed value or argue that the information contained is so weak that it can be ignored. Although we have not studied any of these we would anticipate the desirable properties described below to still hold. Stronger modeling assumptions are also possible (Klein et al., 1990; Moeschberger and Klein, 1985).

Note also that the left-hand side of the equation is a special case of the weighted scores under the proportional hazards model (Harrington and Fleming, 1982; Lin, 1991; Newton and Raftery, 1994). However those weighted scores were not proposed with the non-proportional hazards model in mind, and the particular choice of $W(\cdot)$ used here was not considered in those papers. Indeed other choices for the weights will lead to estimators closer to the partial likelihood itself, in the sense that under a non-proportional hazards model and in the presence of censoring, the broader class of weighted estimates will not converge to quantities that remain unaffected by an independent censoring mechanism. On the other hand, the estimating equation based on U_3 is in the same spirit as the approximate likelihood of Oakes (1986) for censored data and the M-estimate of Zhou and Li (2008) for censored linear models. Hjort (1992) also mentioned the use of the reciprocal of the Kaplan-Meier estimate of the censoring distribution as weights in parametric survival models, and these weights are the same as $W(\cdot)$ defined here. For the random effects model—a special case of this is the stratified model which, in turn, can be expressed in the form (4.2)—we can see, even when we know that (4.3) is severely misspecified, that we can still obtain estimates of meaningful quantities. The average effect resulting from the estimating equation U_3 is clearly of interest.

For the stratified model, $\mathcal{Z}(X_i)$ is contrasted with its expectation $\mathcal{E}_\beta(Z|X_i, s)$. Here, the inclusion of s is used to indicate that if Z_i belongs to stratum s then the reference risk set for $\mathcal{E}_\beta(Z|X_i, s)$ is restricted to members of this same stratum. Note that for time-dependent $s(t)$ the risk set is dynamic, subjects entering and leaving the set as they become at risk. The usual estimating equation for stratified models is again of the form $U(\beta)$ and, for the same reasons as recalled above and described more fully in Xu and O'Quigley (2000) and we might prefer to use

$$U_s(\beta) = \int \{\mathcal{Z}(t) - \mathcal{E}_\beta(Z|t, s)\} d\hat{F}(t) = 0. \quad (7.34)$$

Even weaker assumptions (not taking the marginal $F(t)$ to be common across strata) can be made and, at present, this is a topic that remains to be studied.

ZEROS OF ESTIMATING EQUATIONS

Referring back to Section 7.5 we can immediately deduce that the zeros of the estimating equations provide consistent estimates of β under the model. Below we consider zeros of the estimating equations when the model is incorrectly specified. This is important since, in practice, we can assume this to be the case. Most theoretical developments proceed under the assumptions that the model is correct. We would have that $\hat{\beta}$ where $U_2(\hat{\beta}) = 0$ is consistent for β . Also $\tilde{\beta}$ where $U_3(\tilde{\beta}) = 0$ is consistent for a parameter of interest, namely the average effect. From the mean value theorem we write

$$U_2(\hat{\beta}) = U_2(\beta_0) + (\hat{\beta} - \beta_0) \left\{ \frac{\partial U_2(\beta)}{\partial \beta} \right\}_{\beta=\xi},$$

where ξ lies strictly on the interior of the interval with endpoints β_0 and $\hat{\beta}$. Now $U_2(\hat{\beta}) = 0$ and $U'_2(\xi) = \sum_{i=1}^n \delta_i \mathcal{V}_\xi(Z|X_i)$ so that $\text{Var}(\hat{\beta}) \approx 1/\sum_{i=1}^n \delta_i \text{Var}(Z|X_i)$. This is the Cramer-Rao bound and so the estimate is a good one. Although the sums are of variables that we can take to be independent they are not identically distributed. Showing large sample normality requires verification of the Lindeburgh condition which, although awkward, is not difficult. All the necessary ingredients are then available for inference. However, as our recommended approach, we adopt a different viewpoint based on the functional central limit theorem rather than a central limit theorem for independent variables. This is outlined in some detail in Chapter 9.

PROPERTIES OF SOLUTIONS TO ESTIMATING EQUATIONS

The reason for considering estimating equations other than (7.32) is because of large sample properties. Without loss of generality, for any multivariate categorical situation, a non-proportional hazards model (Equation 4.2) can be taken to generate the observations. Suppose that for this more general situation we fit the best available model, in particular the proportional hazards model (Equation 4.3). In fact, this is what always takes place when fitting the Cox model to data. It will be helpful to have the following definition:

Definition 7.11. *The average conditional variance $A(\beta)$ is defined as*

$$A(\beta) = \int_0^\infty \{E_\beta(Z^2|t) - E_\beta^2(Z|t)\} dF(t).$$

Note that the averaging does not produce the marginal variance. For that we would need to include a further term which measures the variance of the conditional expectations. Under the conditions on the censoring of Breslow and

Crowley (1974), essentially requiring that, for each t , as n increases, the information increases at the same rate, then $nW(t)$ converges in probability to $w(t)$. Under these same conditions, recall that the probability limit as $n \rightarrow \infty$ of $\mathcal{E}_\beta(Z|t)$ under model (4.2) is $E_\beta(Z|t)$, that of $\mathcal{E}_\beta(Z^2|t)$ is $E_\beta(Z^2|t)$, and that of $\mathcal{V}_\beta(Z|t)$ is $V_\beta(Z|t)$. The population conditional expectation and variance, whether the model is correct or not, are denoted by $E(Z|t)$ and $V(Z|t)$, respectively. We have an important result due to Struthers and Kalbfleisch (1986).

Theorem 7.3. *Under model 4.2 the estimator $\hat{\beta}$, such that $U_2(\hat{\beta}) = 0$, converges in probability to the constant β_{PL} , where β_{PL} is the unique solution to the equation*

$$\int_0^\infty w^{-1}(t) \{E(Z|t) - E_\beta(Z|t)\} dF(t) = 0, \quad (7.35)$$

provided that $A(\beta_{PL})$ is strictly greater than zero.

Should the data be generated by model (4.3) then $\beta_{PL} = \beta$, but otherwise the value of β_{PL} would depend upon the censoring mechanism in view of its dependence on $w(t)$. Simulation results below on the estimation of average effect show a very strong dependence of β_{PL} on an independent censoring mechanism. Of course, under the unrealistic assumption that the data are exactly generated by the model, then, for every value of t , the above integrand is identically zero, thereby eliminating any effect of $w(t)$. In such situations the partial likelihood estimator is more efficient and we must anticipate losing efficiency should we use the estimating equation $U_3(\beta)$ rather than the estimating equation $U_2(\beta)$.

Viewing the censoring mechanism as a nuisance feature of the data we might ask the following question: was it possible to remove the censoring then to which population value do we converge? We would like an estimating equation that, in the presence of an independent censoring mechanism, produces an estimate that converges to the same quantity we would have converged to had there been no censoring. The above estimating equation (7.33) has this property. This is summarized in the following theorem of Xu and O'Quigley (2000), which is an application of Theorem 3.2 in Lin (1991).

Theorem 7.4. *Under model 4.2 the estimator $\tilde{\beta}$, such that $U_3(\tilde{\beta}) = 0$, converges in probability to the constant β^* , where β^* is the unique solution to the equation*

$$\int_0^\infty \{E(Z|t) - E_\beta(Z|t)\} dF(t) = 0, \quad (7.36)$$

provided that $A(\beta^)$ is strictly greater than zero.*

None of the ingredients in the above equation depends on an independent censoring mechanism. In consequence the solution itself, $\beta = \beta^*$, is not influenced by the censoring. Thus the value we estimate in the absence of censoring, β^* , is the same as the value we estimate when there is censoring. A visual inspection of equations (7.35) and (7.36) suffices to reveal why we argue in favor of (7.33) as a more suitable estimating equation than (7.32) in the presence of non-proportional hazard effects. Furthermore, the solution to (7.33) can be given a strong interpretation in terms of average effects. We return to this in more detail, but we can already state a compelling argument for the broader interpretability of β^* .

LARGE SAMPLE PROPERTIES OF $\tilde{\beta}$

We have that $\mathcal{E}_\beta(Z|t) = S^{(1)}(\beta, t)/S^{(0)}(\beta, t)$, and that $W(t) = \hat{S}(t)/\{nS^{(0)}(0, t)\}$. Under an independent censoring mechanism, $s^{(1)}(\beta(t), t)/s^{(0)}(\beta(t), t) = E\{Z(t)|T = t\}$, and $s^{(1)}(\beta, t)/s^{(0)}(\beta, t)$ is what we get when we impose a constant β through time in place of $\beta(t)$, both of which do not involve the censoring distribution. In addition $v(t) = v(\beta(t), t) = \text{Var}\{Z(t)|T = t\}$. We take it that $nW(t)$ converges in probability to a non-negative bounded function $w(t)$ uniformly in t . Then we have $w(t) = S(t)/s^{(0)}(0, t)$. Using the same essential approach as that of Andersen and Gill (1982) it is seen, under the model and an independent censoring mechanism, that the marginal distribution function of T can be written as

$$F(t) = \int_0^t w(s)s^{(0)}(\beta(s), s)\lambda_0(s)ds. \quad (7.37)$$

Theorem 7.5. (Xu 1996). *Under the non-proportional hazards model and an independent censorship the estimator $\tilde{\beta}$ converges in probability to the constant β^* , where β^* is the unique solution to the equation*

$$\int_0^\infty \left\{ \frac{s^{(1)}(\beta(t), t)}{s^{(0)}(\beta(t), t)} - \frac{s^{(1)}(\beta, t)}{s^{(0)}(\beta, t)} \right\} dF(t) = 0, \quad (7.38)$$

provided that $\int_0^\infty v(\beta^, t)dF(t) > 0$.*

It is clear that equation (7.38) does not involve censoring. Neither then does the solution to the equation, β^* . As a contrast the maximum partial likelihood estimator $\hat{\beta}_{PL}$ from the estimating equation $U_2 = 0$ converges to the solution of the equation

$$\int_0^\infty \left\{ \frac{s^{(1)}(\beta(t), t)}{s^{(0)}(\beta(t), t)} - \frac{s^{(1)}(\beta, t)}{s^{(0)}(\beta, t)} \right\} s^{(0)}(\beta(t), t)\lambda_0(t)dt = 0. \quad (7.39)$$

This result was obtained by Struthers and Kalbfleisch (1986). If the data be generated by the proportional hazards model, then the solutions of (7.38) and (7.39) are both equal to the true regression parameter β . In general, however, these solutions will be different, the solution to (7.39) depending on the unknown censoring mechanism through the factor $s^{(0)}(\beta(t), t)$. The simulation results of Table 7.1 serve to underline this fact in a striking way. The estimate $\hat{\beta}$ can be shown to be asymptotically normal with mean zero and variance that can be written down. The expression for the variance is nonetheless complicated and is not reproduced here since it is not used. Instead we base inference on functions of Brownian motion which can be seen to describe the limiting behavior of the regression effect process.

INTERPRETATION AS AVERAGE EFFECT

In Section 7.5 the average effect, $\mu = \int \beta(t)dF(t)$, is seen to coincide, in several special cases, with β^* , the solution to the large sample equivalent of the estimating equation $U_3(\beta)$. More generally, and this result is supported by the entries in Table 7.1, β^* and $\mu = \int \beta(t)dF(t)$ will be so close that, for practical purposes, that could be identified as being the same quantity. In other words, Equation 7.38 can be viewed as an average regression effect. In the equation $s^{(1)}(\beta(t), t)/s^{(0)}(\beta(t), t) = E\{Z(t)|T = t\}$, and $s^{(1)}(\beta^*, t)/s^{(0)}(\beta^*, t)$ results when $\beta(t)$ is restricted to be a constant; the difference between these two is zero when integrated out with respect to the marginal distribution of failure time. Suppose, for instance, that $\beta(t)$ decreases over time, then earlier on $\beta(t) > \beta^*$ and $s^{(1)}(\beta(t), t)/s^{(0)}(\beta(t), t) > s^{(1)}(\beta^*, t)/s^{(0)}(\beta^*, t)$; whereas later we would have the opposite effect whereby $\beta(t) < \beta^*$ and $s^{(1)}(\beta(t), t)/s^{(0)}(\beta(t), t) < s^{(1)}(\beta^*, t)/s^{(0)}(\beta^*, t)$. We can write: $v(\beta, t) = \partial/\partial\beta\{s^{(1)}(\beta, t)/s^{(0)}(\beta, t)\}$ and, applying a first-order Taylor series approximation to the integrand of (7.38), we have

$$\int_0^\infty v(t)\{\beta(t) - \beta^*\}dF(t) \approx 0, \quad (7.40)$$

where $v(t) = v(\beta(t), t) = \text{Var}\{Z(t)|T = t\}$. Therefore

$$\beta^* \approx \frac{\int_0^\infty v(t)\beta(t)dF(t)}{\int_0^\infty v(t)dF(t)} \quad (7.41)$$

is a weighted average of $\beta(t)$ over time. According to Equation 7.41 more weights are given to those $\beta(t)$'s where the marginal distribution of T is concentrated, which simply means that, on average, we anticipate there being more individuals subjected to those particular levels of $\beta(t)$. The approximation of Equation 7.41 also has an interesting connection with Murphy and Sen (1991), where they show that if we divide the time domain into disjoint intervals and estimate a constant β on each interval, in the limit as $n \rightarrow \infty$ and the intervals become finer at a certain rate, the resulting $\hat{\beta}(t)$ estimates $\beta(t)$ consistently. In their large sample

studies, they used a (deterministic) piecewise-constant parameter $\bar{\beta}(t)$, which is equivalent to Equation 7.41 restricted to individual intervals. They showed that $\bar{\beta}(t)$ is the best approximation to $\hat{\beta}(t)$, in the sense that the integrated squared difference $\int \{\hat{\beta}(t) - \bar{\beta}(t)\}^2 dt \rightarrow 0$ in probability as $n \rightarrow \infty$, at a faster rate than any other choice of such piecewise-constant parameters. In Equation (7.41) if $v(t)$, the conditional variance of $Z(t)$, changes relatively little with time apart from for large t , when the size of the risk sets becomes very small, we can make the approximation $v(t) \equiv c$ and it follows that

$$\beta^* \approx \int_0^\infty \beta(t)dF(t) = E\{\beta(T)\}. \quad (7.42)$$

In practice, $v(t)$ will often be approximately constant, an observation supported by our own practical experience as well as with simulated datasets. For a comparison of two groups coded as 0 and 1, the conditional variance is of the form $p(1-p)$ for some $0 < p < 1$, and this changes relatively little provided that, throughout the study, p and $1-p$ are not too close to zero. The approximate constancy of this conditional variance is used in the sample size calculation for two-group comparisons (Kim and Tsiatis, 1990). In fact, we only require the weaker condition that $\text{Cov}(v(T), \beta(T)) = 0$ to obtain Equation 7.42, a constant $v(t)$ being a special case of this. Even when this weaker condition does not hold exactly, $\int \beta(t)dF(t)$ will still be close to β^* .

β_1	β_2	t_0	% censored	β^*	$\int \beta(t)dF(t)$	$\tilde{\beta}$	$\hat{\beta}_{PL}$
1	0	0.1	0%	0.156	0.157	0.155 (0.089)	0.155 (0.089)
			17%	0.156	0.157	0.158 (0.099)	0.189 (0.099)
			34%	0.156	0.157	0.160 (0.111)	0.239 (0.111)
			50%	0.156	0.157	0.148 (0.140)	0.309 (0.130)
			67%	0.156	0.157	0.148 (0.186)	0.475 (0.161)
			76%	0.156	0.157	0.161 (0.265)	0.654 (0.188)
3	0	0.05	0%	0.721	0.750	0.716 (0.097)	0.716 (0.097)
			15%	0.721	0.750	0.720 (0.106)	0.844 (0.107)
			30%	0.721	0.750	0.725 (0.117)	1.025 (0.119)
			45%	0.721	0.750	0.716 (0.139)	1.294 (0.133)
			60%	0.721	0.750	0.716 (0.181)	1.789 (0.168)
			67%	0.721	0.750	0.739 (0.255)	2.247 (0.195)

Table 7.1: Comparison of β^* , $\int \beta(t)dF(t)$, and the estimates $\tilde{\beta}$ and $\hat{\beta}_{PL}$.

The accuracy of $\int \beta(t)dF(t)$ in approximating β^* was further studied in simulations by Xu and O'Quigley (2000). Some of those findings are shown in Table 7.1 and these are typical of the findings from a wide variety of other situations. The results are indeed striking. It is also most likely true that it is not well known just how strong is the dependence of the partial likelihood estimator

on an independent censoring mechanism when the data are generated by a non-proportional hazards model. Since, in practical data analysis, such a situation will almost always hold, we ought to be rather more circumspect about the usual estimators furnished by standard software.

In the table the data are simulated from a simple two-step time-varying regression coefficients model, with baseline hazard $\lambda_0(t) = 1$, $\beta(t) = \beta_1$ when $t < t_0$ and β_2 otherwise. The covariate Z is distributed as Uniform(0,1). At time t_0 a certain percentage of subjects at risk are censored. The value $\hat{\beta}_{PL}$ is the partial likelihood estimate when we fit a proportional hazards model to the data. Table 7.1 summarizes the results of 200 simulations with sample size of 1600. We see that $\int \beta(t)dF(t)$ is always close to β^* , for the values of β that we might see in practice. The most important observation to be made from the table is the strong dependence of $\hat{\beta}_{PL}$ on an independent censoring mechanism, the value to which it converges changing substantially as censoring increases. The censoring mechanism here was chosen to emphasize the difference between $\hat{\beta}_{PL}$ and $\tilde{\beta}$, since $\tilde{\beta}$ puts (asymptotically) the correct weights on the observations before and after t_0 . In other cases the effect of censoring may be weaker. Nonetheless, it is important to be aware of the behavior of the partial likelihood estimator under independent censoring and non-proportional hazards and the subsequent difficulties in interpreting the partial likelihood estimate in general situations.

The bracketed figures in Table 7.1 give the standard errors of the estimates from the simulations. From these we can conclude that any gains in efficiency of the partial likelihood estimate can be very quickly lost to biases due to censoring. When there is no censoring the estimators are the same. As censoring increases we see differences in the standard errors of the estimates, the partial likelihood estimate being more efficient; but we also see differences in the biases. Typically, these latter differences are at least an order of magnitude greater.

RESIDUALS FROM ESTIMATING EQUATIONS

In linear regression we postulate a linear dependence of the response variable on the explanatory variable or variables. Once unknown parameters have been fitted we can use the model to obtain the conditional mean of the response variable given the explanatory variables. For a “good” or well performing model we would like for the conditional mean of the response variable to plausibly arise from the model. Usually we will only have one observation per subject and we can study the discrepancy between this observation and that predicted by the model. These discrepancies are typically referred to as the residuals.

Trends in the residuals or any kind of systematic behavior not described by the model are indications of a lack of fit or that a more complex model is likely to provide a better description of the observations. Checking model adequacy is a fundamental part of model construction and we return to this in later chapters. The key difference to residuals in the proportional hazards setting from residuals in the classical linear setting was pointed out by Schoenfeld (1982).

In the linear setting we study the discrepancies between the observed responses and their corresponding model-based predictions. In the proportional hazards setting, rather than being the response variable, survival, it is the covariate given survival that provides the basis for suitable residuals. On a deeper level, there is no real difference between these two settings once we accept that our outcome variable is not survival (since it only determined up to its rank) and is indeed the covariate, or covariate vector, that is observed, given $T = t$. This is of course anticipated in the main theorem of this chapter.

Schoenfeld considered a vector of covariates Z that were fixed over time. Once we fix time however, the extension of the Schoenfeld residuals to time-dependent covariates is immediate and, aside from the need for great care in the calculations, presents no added difficulty. The Schoenfeld residuals are defined at each death time t by

$$r_j(t : \beta) = Z_j(t) - \mathcal{E}_\beta(Z | t), \quad (7.43)$$

where j indexes the individual failing at time t . These residuals have formed the basis for many goodness-of-fit tests as well as the basis for graphical procedures, (Grambsch and Therneau, 1994; Lin, 1991; Lin et al., 1993; O'Quigley and Pessione, 1991; Wei, 1984).

Given the basic nature of martingales as discrepancies between random variables conditional upon a history and their corresponding conditional expectation it is natural to consider such objects as a way to create a large class of residuals. All that is needed is the history, some function of this—technically referred to as a filtration, and the conditional expectations. These discrepancies have the martingale property, essentially no more than the expectations exist, and a large class of residuals becomes available to us. These are known as martingale residuals and, of course, the Schoenfeld residuals are a particular case. In our view, the Schoenfeld residuals are not just a special case of the martingale residuals but, in practice, are the only martingale residuals of real interest. We might qualify that statement by adding after some appropriate standardization. For this reason we do not dwell on any particular features of martingale residuals and we present little of the work that has been done on these. We recall the bare bones nonetheless for completeness. For subject j at time t , the martingale residuals are (Appendix B.3)

$$M_j(t) = N_j(t) - \hat{\Lambda}_j(t), \quad \text{avec} \quad \hat{\Lambda}_j(t) = \int_0^t \pi_j(\hat{\beta}, s) d\bar{N}(s).$$

The definition follows the usual Doob-Meyer decomposition of a counting process into a martingale and a compensator (Andersen and Gill, 1982). For the usual proportional hazards model with parameter β , the compensator of the process N_j at time t is $\int_0^t \lambda(s | Z_j) ds$ and is estimated by $\hat{\Lambda}_j(t)$. If we can consider the model to have generated the observations, then $M_1(t), \dots, M_n(t)$ are observations of

independent martingales (Gill, 1980), which means that they should be centered about zero and uncorrelated. These properties can be observed on a graph of the residual, $M_j(\mathcal{T})$, $j = 1, \dots, n$. Barlow and Prentice (1988) studied a class of martingale residuals by focusing on $\int \phi(t)M_j(t)dt$ where ϕ is a predictable function. For a predictable function, $\phi(t)$, conditioning on the history at time t allows us to treat $\phi(t)$ as a constant so that $\int \phi(t)M_j(t)dt$ will also be a martingale. These residuals have formed the basis of tests of fit by Kay (1977), Lin (1991), Lin et al. (1993) as well as Therneau et al. (1990). As already mentioned, we mostly focus attention on the Schoenfeld residuals which arise under the definition, $\phi = Z_j$ (Schoenfeld, 1982).

7.7 Estimating equations in small samples

The finite sample distribution of the score statistic is considered more closely. Since the other test statistics are derived from this, and the regression coefficient itself a monotonic function of the score, it is enough to restrict attention to the score statistic alone. One direct approach leads to a simple convolution expression which can be evaluated by integrated integrals. It is also possible to make improvements to the large sample normal approximation via the use of saddlepoint approximations or Cornish-Fisher expansions. For these we can use the results of Corollary 7.9. Corrections to the distribution of the score statistic can be particularly useful when the distribution of the explanatory variable is asymmetric. Corrections to the distribution of the score equation have a rather small impact in the case of the fourth moment but can be of significance in the case of the third moment. The calculations themselves are uncomplicated and simplify further in the case of an exponential distribution. Since we can transform an arbitrary marginal distribution to one of the exponential form, while preserving the ranks, we can then consider the results for the exponential case to be of broader generality. The focus of our inferential efforts, regardless of the particular technique we choose, is mostly the score statistic. For this statistic, based on the properties of the estimating equation, we can claim large sample normality.

Recall that our underlying probability model is focused on the distribution of the covariate, or covariate vector, at each time t given that the subject is still at risk at this time. From this the mean and the variance of the conditional distribution can be consistently estimated and this is typically the cornerstone for the basis of any tests or confidence interval construction. Implicitly we are summarizing these key distributions by their means and variances or, at least, our best estimates of these means and variances. The fact that it is the distributions themselves that are of key interest, and not just their first two moments, suggests that we may be able to improve the accuracy of any inference if we were to take into account higher order moments. As a consequence of the main theorem it turns out that this is particularly straightforward, at least for the third moment and relatively uncomplicated for the fourth moment. In fact, corrections based

on the fourth moment seem to have little impact and so only the third moment might be considered in practice.

We can incorporate information on these higher moments via a Cornish-Fisher expansion or via the use of a saddlepoint approximation. Potential improvements over large sample results would need to be assessed on a case-by-case basis, often via the use of simulation. Some limited simulations are given in O'Quigley (2008) and suggest that these small sample corrections can lead to more accurate inference, in particular for situations where there is strong group imbalance.

EDGEWORTH AND SADDLEPOINT APPROXIMATIONS

For the proportional hazards model it is relatively easy to obtain Edgeworth corrections (Appendix A) to the score statistic. For the particular case of the multiplicative risk function the saddlepoint approximation is also straightforward. In either case we make progress by evaluating the cumulant generating function $K(\theta)$ given by,

$$K(\theta) = \log E(e^{\theta Z}) = \log EE(e^{\theta Z}|t) = \log \int E(e^{\theta Z}|t)dF(t),$$

the trick of double expectation leading to quantities we can readily estimate in view of the results of Section 7.5. Furthermore, applying Corollary 7.9, we have that the difference between $E(e^{\theta Z}|t)$ and $\mathcal{E}_{\hat{\beta}}(e^{\theta Z}|t)$ or $\mathcal{E}_{\beta}(e^{\theta Z}|t)$ converges in probability to zero. For the multiplicative model we can then use the estimate $\hat{K}(\theta)$ where

$$\exp\{\hat{K}(\theta)\} = \int \left\{ \frac{\sum_i \sum_j Y_j(X_i) \exp[(\theta + \beta)Z_j]}{\sum_i \sum_j Y_j(X_i) \exp\{\beta Z_j\}} \right\} d\hat{F}(t),$$

leading to the results, after some elementary manipulation, that

$$\hat{K}'(0) = \int \mathcal{E}_{\beta}(Z|t)d\hat{F}(t); \quad \hat{K}''(0) = \int \mathcal{V}_{\beta}(Z|t)d\hat{F}(t).$$

The following results enable us to obtain the needed terms.

Lemma 7.4. *Letting $A(\theta) = \exp\{K(\theta)\}$ then:*

$$\left\{ \frac{\partial^p A(\theta)}{\partial \theta^p} \right\}_{\theta=0} = \int \mathcal{E}_{\beta}(Z^p(t)|t)d\hat{F}(t) \tag{7.44}$$

The first two derivatives of $A(\theta)$ are well known and widely available from any software which fits the proportional hazards model. The third and fourth derivatives are a little fastidious although, nonetheless straightforward to obtain. Pulling all of these together we have:

Lemma 7.5. *The first four derivatives of $K(\theta)$ are obtained from:*

$$\begin{aligned} A'(\theta)/A(\theta) &= K'(\theta), \\ A''(\theta)/A(\theta) &= [K'(\theta)]^2 + K''(\theta), \\ A'''(\theta)/A(\theta) &= [K'(\theta)]^3 + 3K'(\theta)K''(\theta) + K'''(\theta), \\ A^{(4)}/A(\theta) &= [K'(\theta)]^4 + 6[K'(\theta)]^2K''(\theta) + 4K'(\theta)K'''(\theta) + 3[K''(\theta)]^2 + K^{(4)}(\theta). \end{aligned}$$

We can use these results in either a saddlepoint approximation or an Edgeworth approximation allowing us to gain greater accuracy than that provided by the usual assumption of large sample normality. In the light of several comparative studies of the relative merits of the two kinds of approximation we cannot really anticipate obtaining any clear answer as to which is the best to use. It will depend on the particular case under study and, typically, there is very little to choose between the two. The above results can be used in both cases. Once the particular parameters of a study (total sample size, number of groups, group imbalance, approximate distribution of the covariates) are known, then simulations can help answer this question. Simulations show that real advantages, in particular for the additive model, can be obtained by making these adjustments. Some preliminary work can lead to further simplification, and subtracting off the mean allows us to ignore $K'(\theta)$, which is equal then to zero at $\theta = 0$. Instead of referring the test statistic to the percentage points Z_α of the normal distribution we use,

$$\begin{aligned} L_\alpha &= Z_\alpha + (Z_\alpha^2 - 1)E\{U^3(\hat{\beta}, \infty)\}/6 \\ &\quad + (Z_\alpha^3 - 3Z_\alpha) \left[E\{U^4(\hat{\beta}, \infty)\} - 3E^2\{U^2(\hat{\beta}, \infty)\} \right] /24. \end{aligned} \quad (7.45)$$

The adjustment can be used either when carrying out a test of a point hypothesis or when constructing test-based confidence intervals. In the simulations we can see that the correction, relatively straightforward to implement, leads to improved control on type I error in a number of situations.

DISTRIBUTION OF ESTIMATING EQUATION

The parameter β is estimated by equating to zero the score function, $U(\beta)$. Thus, $U(\hat{\beta}) = 0$ and the distribution of $\hat{\beta}$ can be investigated via the estimating equation, since

$$\Pr(\hat{\beta} < b) = \Pr\{U(b) < 0\} \quad (7.46)$$

We can use what we know about U to make statements about $\hat{\beta}$. This only works because $U(\beta)$ is monotonic in β and the same will continue to hold in the

multivariate setting when considering components of the vector U . Formulating the probability statement about $\hat{\beta}$ in terms of U is particularly convenient. Two simple illustrations of this are: (1) Bayesian inference and (2) the exact distribution of a sum of independent, not necessarily identically distributed, random variables. If we have prior information on β , in the form of the density $q(\beta)$, then we can write

$$\Pr(\hat{\beta} < b) = \int \Pr\{U(b) < 0\}q(b)db,$$

being an expression in terms of total probability rather than the usual Bayes' formula since, instead of a data statistic depending on the model parameter, we have a direct expression for the parameter estimate. For the small sample exact distribution of the sum we use the following lemma:

Lemma 7.6. *Let U_1, \dots, U_n be independent, not necessarily identically distributed, continuous random variables with densities $p_1(x)$ to $p_n(s)$, respectively. Let $S_n = \sum_{j=1}^n U_j$. Then the density, $q_n(s) = dQ_n(s)$, of S_n is given by*

$$dQ_n(s) = \int_{-\infty}^{\infty} dQ_{n-1}(s-u)dP_n(u)du .$$

We use the above form $dQ_n(s)$ in order to accommodate the discrete and the continuous cases in a single expression. The lemma is proved by recurrence of an elementary convolution result (see for example Kendall et al. (1987)). Following Cox (1975) and Andersen et al. (1993), Andersen (1982), and Andersen and Gill (1982) we will take the contributions to the score statistic $U(X_i)$ to be independent with different distributions given by Theorem 7.2. We can then apply the result by letting $U_i = H_i(X_i)$ where the i indices now run over the k , rather than n failure times. The distribution of U_1 is given by $G_0(X_1)$, of U_2 by $G_0(X_2)$ and we can then construct a sequence of equations based on the above expression to finally obtain the distribution $Q_k(s)$ of the sum. Any prior information can be incorporated in this expression in the same way as before.

HIGHER ORDER MOMENTS

It is possible, in a way similar to that leading to the variance of the score statistic, to obtain third and fourth moments. We can then use the estimated cumulative hazard equation to derive data-based estimates. Since, under the model, we have that $E\{U(\beta, t)\} = 0$ then

$$E\{U^3(\beta, \infty)\} = \int_0^\infty \sum_{i=1}^n E\{Y_i(s)H_i^3(s)\}\lambda_i(s)ds. \quad (7.47)$$

We can estimate $E\{U^3(\beta, \infty)\}$ consistently by replacing $\lambda_i(s)ds$ and $\lambda_i(s_1)\lambda_j(s_2)ds_1ds_2$ by $R\{\hat{\beta}Z_i(s)\}d\hat{\Lambda}_0(s)$ and $R\{\hat{\beta}Z_i(s_1)\}R\{\hat{\beta}Z_j(s_2)\}d\hat{\Lambda}_0(s_1)d\hat{\Lambda}_0(s_2)$, respectively.

$$\begin{aligned} E\{U^4(\beta, \infty)\} &= \int_0^\infty \sum_{i=1}^n E\{Y_i(s)H_i^4(s)\}\lambda_i(s)ds \\ &+ 6 \int_0^\infty \int_0^\infty \sum_{i=1}^n \sum_{j>i} E\{Y_i(s_1)H_i^2(s_1)\}E\{Y_j(s_2)H_j^2(s_2)\}\lambda_i(s_1)\lambda_j(s_2)ds_1ds_2. \end{aligned} \quad (7.48)$$

Again, we can estimate $E\{U^4(\beta, \infty)\}$ consistently by replacing $\lambda_i(s)ds$ and $\lambda_i(s_1)\lambda_j(s_2)ds_1ds_2$ by $R\{\hat{\beta}Z_i(s)\}d\hat{\Lambda}_0(s)$ and $R\{\hat{\beta}Z_i(s_1)\}R\{\hat{\beta}Z_j(s_2)\}d\hat{\Lambda}_0(s_1)d\hat{\Lambda}_0(s_2)$, respectively. Note that we can replace $E\{Y_i(s)H_i^2(s)\}$, $E\{Y_i(s)H_i^3(s)\}$, $E\{Y_i(s)H_i^4(s)\}$ by observed values at time of failure or by an average taken over the risk set.

INTEGRAL TRANSFORM OF THE BASELINE HAZARD

Note that for an arbitrary continuous distribution function, $F(t)$, the probability integral transform tells us that the variable $Y = -\log\{1 - F(T)\}$ has a standard exponential distribution. In the case of two groups, and using \hat{F} in place of F , we can transform a Kaplan-Meier curve into one approaching a standard exponential for one group and, using the same transformation, into one approaching an exponential distribution with parameter $\exp(\beta)$ in the other. For this reason a study of the moment adjustments for the special case of an exponential distribution can be of value since our interest in the baseline hazard itself is only accessory. The necessary calculations simplify.

Being a special case of the proportional hazards model, an analysis based on the case of a constant hazard enables us to compare our estimates with fixed population values. Furthermore, the estimates can still be useful when the true model is different from the exponential one. Suppose that we have a binary covariate (0,1) denoting group membership of which there are π_1 in the first group and π_2 in the second ($\pi_1 + \pi_2 = 1$), survival time is distributed exponentially with underlying hazard equal to 1 in the first group and e^β in the second and there is no censoring. Then:

$$E\left\{\left[Z - \mathcal{E}_\beta^P(Z|t)\right]^p\right\} = \pi_1 \int_0^t \{-\psi(v)\}^p e^{-v} dv + \pi_2 \int_0^t \{1 - \psi(v)\} e^\beta \exp(-ve^\beta) dv,$$

where

$$\psi(v) = \frac{\pi_2 e^\beta \exp(-ve^\beta)}{\pi_1 e^{-v} + \pi_2 e^\beta \exp(-ve^\beta)}.$$

Evaluating the above formula under $\beta = 0$, we have:

Corollary 7.12. *The second, third, and fourth moments of U are given by the following where U_s is U standardized to have unit variance:*

$$\begin{aligned} E\{U^2(0, \infty)\} &= n\pi_1\pi_2 ; \quad E\{U_s^3(0, \infty)\} = n^{-1/2}(\pi_1 - \pi_2)(\pi_1\pi_2)^{-1/2} \\ E\{U_s^4(0, \infty)\} &= n^{-1}(\pi_1^3 + \pi_2^3)(\pi_1\pi_2)^{-1} + 3n^{-1}(n-1). \end{aligned}$$

More generally, consider the case of a continuous variable Z with support I and density f . Furthermore, suppose that survival time is distributed exponentially with underlying hazard equal to $\lambda_0 R(\beta z)$ with $R(\beta z) = 1$ for $\beta = 0$, and that there is no censoring. Then, for $\beta = 0$ and $p \geq 2$, we have:

$$\int_0^\infty \sum_{i=1}^n E\{Y_i(s)H_i^p(s)\}\lambda_i(s)ds = \int_I \left[\int_0^\infty \{z - E(Z)\}^p \lambda_0 e^{-\lambda_0 t} dt \right] f(z)dz$$

and this integral can be readily evaluated so that the right-hand term becomes:

$$\int_I \{z - E(Z)\}^p \left(\int_0^\infty \lambda_0 e^{-\lambda_0 t} dt \right) f(z)dz = \int_I \{z - E(Z)\}^p f(z)dz = E\{Z - E(Z)\}^p.$$

Therefore, the required terms can easily be evaluated from the central moments of Z . Specifically, taking the subscript s to refer to the standardized variable, we obtain:

Corollary 7.13. *The second, third, and fourth moments of U are given, respectively, by:*

$$\begin{aligned} E(U^2(0, \infty)) &= nE\{Z - E(Z)\}^2, \\ E(U_s^3(0, \infty)) &= n^{-1/2}E\{Z - E(Z)\}^3 [E\{Z - E(Z)\}^2]^{-3/2}, \\ E(U_s^4(0, \infty)) &= n^{-1}E\{Z - E(Z)\}^4 [E\{Z - E(Z)\}^2]^{-2} + 3n^{-1}(n-1), \end{aligned}$$

where the subscript s refers to the standardized variable. Note that these results also hold for a discrete variable Z .

INTEGRAL TRANSFORM OF THE CONDITIONAL COVARIATE DISTRIBUTION

Consider the case of a continuous covariate Z . A more sure way, albeit a more onerous one, to correct for asymmetries in the conditional covariate distribution is to, once again, lean on the probability integral transform. The need to evaluate higher order moments then disappears since, by construction, the odd order moments will be zero and the fourth very close to its normal counterpart. Denote by $\hat{G}(z|t)$ the estimated conditional distribution of Z given that $T = t, C > t$, i.e.,

$$\hat{G}(z|t) = \hat{P}(Z(t) \leq z | T = t, C > t) = \sum_{j=1}^n \pi_j(\beta, t) I(Z_j(t) \leq z). \quad (7.49)$$

Note that the definition of $\pi_j(\beta, t)$ restricts the subjects under consideration to those in the risk set at time t . The cumulative distribution $\hat{G}(z|t)$ is restricted by both z and t . We will need to invert this function, at each point X_i corresponding to a failure. Assuming no ties in the observations (we will randomly break them if there are any) then, at each time point X_i , we order the observations Z in the risk set. We express the order statistics as $Z_{(1)} < Z_{(2)} < \dots < Z_{(n_i)}$ where there are n_i subjects in the risk set at time X_i . We define the estimator $\tilde{G}(z|X_i)$ at time $t = X_i$ and for $z \in (Z_{(m)}, Z_{(m+1)})$ by

$$\tilde{G}(z|X_i) = \hat{G}(Z_{(m)}|X_i) + \frac{z - Z_{(m)}}{Z_{(m+1)} - Z_{(m)}} \left\{ \hat{G}(Z_{(m+1)}|X_i) - \hat{G}(Z_{(m)}|X_i) \right\},$$

noting that, at the observed values $Z_{(m)}, m = 1, \dots, n_i$, the two estimators coincide so that $\tilde{G}(z|X_i) = \hat{G}(z|X_i)$ for all values of z taken in the risk set at time X_i . Otherwise, $\tilde{G}(z|X_i)$ linearly interpolates between adjacent values of the observed order statistics $Z_{(m)}, m = 1, \dots, n_i$. Also, we are assuming no ties, in which case, the function $\tilde{G}(z|X_i)$, between the values $Z_{(1)}$ and $Z_{(n_i)}$, is a strictly increasing function and can thereby be inverted. We denote the inverse function by $\tilde{G}^{-1}(\alpha), 0 < \alpha < 1$.

Our purpose is achieved by using, instead of $\tilde{G}^{-1}(\alpha)$ which would take us back to where we began, the inverse of the cumulative normal distribution $\Phi^{-1}(\alpha)$. We define the transform

$$Z_{(m)}^* = \Phi^{-1} \tilde{G}(Z_{(m)}|X_i), \quad (7.50)$$

noting that the transform is strictly increasing so that the order of the covariate observations in the risk set is respected. We are essentially transforming to normality via the observed empirical distribution of the covariate in the risk set. Under the null hypothesis that $\beta = 0$ the cumulative distribution $\tilde{G}(Z_{(m)}|X_i)$ is discrete uniform where each atom of probability has mass $1/n_i$. Thus, the $Z_{(m)}^*, m = 1, \dots, n_i$ will be close (the degree of closeness increasing with n_i) to the expectation of the m th smallest order statistic from a normal sample of size n_i (Appendix A). The statistic $U(\beta)$ is then a linear sum of zero mean and symmetric variables that will be closer to normal than that for the untransformed sequence. At the same time any information in the covariate is captured via the ranks of the covariate values among those subjects at risk and so local power to departures from the null would be model dependent. Under the null the suggested transformation achieves our purpose, the mean of $U(0)$ is zero and the distribution of $U(0)$ is symmetric. Under the alternative, however, we would effectively have changed our model by the transformation and a choice

of model which coincides with the mechanism generating the selection from the risk set would maximize power. The above choice would not necessarily be the most efficient. An expression for the statistical efficiency of using some particular covariate transformation model when another one generates the observations is given in O'Quigley and Prentice (1991).

One way to maintain exact control over type I error using $Z_{(m)}^* = \Phi^{-1}\tilde{G}(Z_{(m)}|X_i)$ is to consider, at each observed failure time, alongside $Z_{(m)}^*$, its reflection in the origin $-Z_{(m)}^*$, such values, and any more extreme in absolute value, arising with the same probability under the null hypothesis of no effect. A nonparametric test considers the distribution of the test statistic under all possible configurations of the vector of dimension equal to the number of the observed failures having entries $Z_{(m)}^*$ or $-Z_{(m)}^*$.

The number of possibilities grows exponentially so that it is possible, with even quite small samples, to achieve almost exact control over type I error. The significance level is simply the number of tests with more extreme values than those obtained by the configuration that corresponds to the observed data themselves. This approach would be very attractive apart from the drawback of the intensity of calculation. With as few as 10 observations per group, in a two-group case, the number of cases to evaluate is over one million. Finding, say, the most extreme five percent of these requires comparisons taking us into the thousands of billions. Approximations are therefore unavoidable.

Robinson (1982) developed a simple saddlepoint approximation to the densities corresponding to paired data. Relabelling the elements of this sum as y_i , $i = 1, \dots, 2^k$, and the null density as $\psi(u)$ then, letting $\sum y_j \tanh(\lambda_u y_j) = uk$, following Robinson's development and regrouping terms, we obtain

$$\psi(u) = \left\{ \frac{k^2}{2\pi \sum y_j^2 \operatorname{sech}^2(\lambda_u y_j)} \right\}^{\frac{1}{2}} \exp \left(\sum \log \cosh(\lambda_u y_j) - \lambda_u u \right), \quad (7.51)$$

For any values of u , λ_u is obtained as a solution to this second equation. In the above expressions all sums range from 1 to 2^k . We can obtain the significance level by numerical integration, in particular via the use of orthogonal polynomials (Abramowitz and Stegun, 1965). Alternatively we can work directly with an approximation to the cumulative distribution itself (Daniels, 1983, 1954, 1980, 1987). Both require numerical approximation and the results, and effort involved, are, for practical purposes, the same.

ILLUSTRATION OF SMALL SAMPLE CORRECTIONS

Rashid et al. (1982) investigated the prognostic influence of the levels of five different acute phase reactant proteins on survival in gastric cancer. The currently recognized most important prognostic factor in the disease is stage, determined at the time of operation. A staging classification simpler than the usual TNM one was developed in the paper of Rashid and proved itself to be very predictive of survival. On the basis of this system, it was possible to strongly discriminate among different prognostic groups. However, since the staging information was only available at the time of surgical intervention, it could not be used to identify the patients with a poorer prognosis for whom intervention should possibly be avoided. The rationale for the use of the acute phase reactant protein is that any such information can be used pre-operatively. A statistical difficulty, however, arises due to the highly non-normal, in particular highly skewed, distributional behavior exhibited by these protein measurements. It has been noted by Kalbfleisch and Prentice (1973) that outliers can have an unbounded effect on estimates and test statistics.

A first approach might be to dichotomize, whereby all the high values are grouped together, and indeed this is the usual way in which such data are handled, the notion of “normal” and “raised” levels being common in the medical literature. Such an approach, however, sacrifices information and, given that power may be lacking due to modest sample sizes, this may not be the best approach. A second approach (O’Brien, 1978; O’Brien and Fleming, 1987; O’Quigley and Prentice, 1991) transforms the explanatory variables to some familiar scale (uniform or normal order statistics for example). A third approach leaves the covariate scale as observed and makes higher order corrections to the score statistic to compensate for the induced lack of normality. This third approach has an advantage over the second in that the (arbitrary) choice of scaling, necessarily impacting the result, is avoided. The second approach has an advantage over the third in that inference is rank invariant, not only with respect to the time variable, but also with respect to the explanatory variable. The data are taken from Rashid et al. (1982) where the focus was on the continuously measured variable C-reactive protein and its impact on prognosis. In the original study, in addition to the well-known prognostic indicators such as stage and tumor histology, there was interest in the degree to which the pre-operative biological measurements might on their own indicate prognostic effects. In such studies it is common to define some kind of cut-off for such measures below which the patients are considered to be within the normal range, and beyond which the tumor is suspected of being particularly aggressive. The reason to consider the original measurements rather than a new variable defined on the basis of a cut-off is that there may be a gradual worsening of prognosis rather than any sudden phenomenon in effect.

A two-sided test based on the score statistic produces the value $p = 0.034$, in reasonably close agreement with the Wald test ($p = 0.042$), although sufficiently removed from the likelihood ratio test ($p = 0.066$) to suggest that the large

sample approximations may be slightly suspect. Carrying out a third moment correction to the score statistic, the value $p = 0.060$ was obtained and increases to $p = 0.063$ when we apply a fourth moment correction. This is in closer agreement with the likelihood ratio test and indicates that the uncorrected score statistic may be slightly underconservative. The values $p = 0.06$ or 0.07 appear then to more accurately reflect the percentile under the null. Whether making a test-based decision or using a test as a means to construct confidence intervals, inference will be more accurate when the p-value is more accurately obtained. It could be argued of course that, in this particular case, it would have been more straightforward to just calculate the likelihood ratio test which seems to be more accurate. However, in other cases, in which there is lack of agreement between the likelihood ratio test and the score test, we have no way of knowing which is the more reliable. Indeed all of the corrections outlined here could be equally well applied to the likelihood ratio test (via a Taylor expansion) instead of the score statistic.

ASSESSING ACCURACY

The most useful tool in assessing which of the several approaches is likely to deliver the best rewards is that of simulation. It is difficult otherwise because, even when we can show that taking into account higher moments will reduce the order of error in an estimate, the exact value of these moments is not typically known. The further error involved in replacing them by estimates involving error can often lead us back to an overall order of error no less than we had in the first place. In some cases we can carry out exact calculation. Even here though caution is needed since if we need to evaluate integrals numerically, although there is no statistical error involved, there is a risk of approximation error. Among the three available tests based on the likelihood; the score test, likelihood ratio, and the Wald test, the score test is arguably the most satisfactory. Although all three are asymptotically equivalent, the Wald test's sensitivity to parameterization has raised questions as to its value in general situations. For the remaining two, the score test (log-rank test) has the advantage of not requiring estimation under the alternative hypothesis and has nice interpretability in terms of simple comparisons between observed and expected quantities. Indeed it is this test, the log-rank test in the case of a discrete covariate, that is by far the most used. The higher moments are also evaluated very easily, again not requiring estimation under the alternative hypothesis, and therefore it is possible to improve the accuracy of inference based on the score test at little cost. Only tests of the hypothesis $H_0 : \beta = 0$ have been discussed. More generally, we may wish to consider testing $H_0 : \beta = \beta_0$, $\beta_0 \neq 0$, such a formulation enabling us to construct confidence intervals about non-null values of β . The same arguments apply to this case also and, by extension, will lead to intervals with more accurate coverage properties.

7.8 Classwork and homework

ESTIMATION AND TESTS

1. Show that, under an independent censoring mechanism, $\hat{H}(t)$, as defined in Section 7.4, provides a consistent estimate of $H(t)$.
2. Show that the variance expression $V(\beta, t)$ using the Andersen and Gill notation (see Section 7.4) is the same as $\mathcal{V}_\beta(Z|t)$ using the notation of Section 7.5. Explain why $\text{Var}(Z|t)$ is consistently estimated by $\mathcal{V}_{\hat{\beta}}(Z|t)$ but that $\text{Var}(Z|t)$ is not generally equal to $v(\beta, t)$.
3. For the general model, suppose that $\beta(t)$ is linear so that $\beta(t) = \alpha_0 + \beta t$. Show that $\mathcal{E}_{\beta(t)}(Z^k|t)$ does not depend upon α_0 .
4. Sketch an outline of a proof that $\text{Var}(Z|t)$ is consistently estimated by $\mathcal{V}_{\hat{\beta}}(Z|t)$ and that $E\text{Var}(Z|t)$ is consistently estimated by $E\mathcal{V}_{\hat{\beta}}(Z|t)$.
5. As for the previous question, indicate why $\int \mathcal{V}_{\hat{\beta}}(Z|t)d\hat{F}(t)$ would be consistent for $E\text{Var}(Z|t)$.
6. Show that $\mathcal{V}_\beta(Z|t) = \partial \mathcal{E}_\beta(Z|t) / \partial \beta$ and identify the conditions for the relationship $\int \mathcal{V}_\beta(Z|t) = \int \partial \mathcal{E}_\beta(Z|t) / \partial \beta = \partial \left\{ \int \mathcal{E}_\beta(Z|t) \right\} / \partial \beta$ to hold.
7. Consider some parametric non-proportional hazards model (see Chapter 2), in which the conditional density of T given $Z = z$ is expressed as $f(t|z)$. Suppose the marginal distribution of Z is $G(z)$. Write down estimating equations for the unknown parameters based on the observations Z_i at the failure times X_i .
8. Use some dataset to fit the proportional hazards model. Estimate the parameter β on the basis of estimating equations for the observations Z_i^2 rather than Z_i . Derive another estimate based on estimating equations for $\sqrt{Z_i}$. Compare the estimates.
9. Write down a set of estimating equations based on the observations Z_i^p , $p > 0, i = 1, \dots, n$. Index the estimate $\hat{\beta}$ by p , i.e., $\hat{\beta}(p)$. For a given dataset, plot $\hat{\beta}(p)$ as a function of p .
10. Use analytical or heuristic arguments to describe the expected behavior of $\hat{\beta}(p)$ as a function of p under (1) data generated under a proportional hazards model, (2) data generated under a non-proportional hazards model where the effect declines monotonically with time.
11. Consider a proportional hazards model in which we also know that the marginal survival is governed by a distribution $F(t; \theta)$ where θ is not known. Suppose that it is relatively straightforward to estimate θ , by maximum

likelihood or by some graphical technique. Following this we base an estimating equation for the unknown regression coefficient, β , on $U(\beta|\hat{\theta}) = \int \{\mathcal{Z}(t) - \mathcal{E}_\beta(Z|t)\} dF(t; \hat{\theta})$. Comment on this approach and on the properties you anticipate it conferring on the estimate $\hat{\beta}$.

12. Use the approach of the preceding question on some dataset by (1) approximating the marginal distribution by an exponential distribution, (2) approximating the marginal distribution by a log-normal distribution.
13. Using again the approach of the previous two questions show that if the proportional hazards models are correctly specified then the estimate $\hat{\beta}$ based on $F(t; \theta)$ is consistent whether or not the marginal model $F(t; \theta)$ is correctly specified.
14. Supposing that the function $\beta(t)$ is linear so that $\beta(t) = \alpha_0 + \beta t$. Show how to estimate the function $\beta(t)$ in this simple case. Note that we can use this model to base a test of the proportional hazards assumption via a hypothesis test that $H_0 : \beta = 0, \alpha_0 \neq 0$ (Cox 1972).
15. Investigate the assertion that it is not anticipated for $v(t)$, the conditional variance of $Z(t)$, to change much with time. Use the model-based estimates of $v(t)$ and different datasets to study this question informally.
16. In epidemiological studies of breast cancer it has been observed that the tumor grade is not well modeled on the basis of a proportional hazards assumption. A model allowing a monotonic decline in the regression coefficient $\beta(t)$ provides a better fit to observed data. On the basis of observations some epidemiologists have argued that the disease is more aggressive (higher grade) in younger women. Can you think of other explanations for this observed phenomenon?
17. Try different weights in the weighted log-rank test and apply these to a dataset. Suppose we decide to use the weight that leads to the most significant result. Would such an approach maintain control over Type I error under the null hypothesis of no association? Suggest at least two ways in which we might improve control over the Type I error rate.
18. Use a two-sample dataset such as the Freireich data and carry out a one-sided test at the 5% level. How does the p-value change if we make the Edgeworth correction given in Equation 7.45.
19. Repeat the above question but this time using a saddlepoint approximation.
20. Using bootstrap resampling, calculate a 95% confidence interval for the estimated regression coefficient for the above data by the percentile method on $\hat{\beta}$. Compare this to a 95% confidence interval obtained by inverting the

monotone function $U(b)$ and by determining values of b for which the estimated $\Pr\{U(b) < 0\} \leq 0.025$ and $\Pr\{U(b) > 0\} \leq 0.025$.

21. Consider the following two priors on β : (1) $\Pr(\beta < -1) = 0.1$; $\Pr(\beta > 2) = 0.1$; $\Pr(-1 < \beta < 2) = 0.8$, (2) $\Pr(\beta < 0) = 0$; $\Pr(\beta > 1) = 0.2$; $\Pr(0 < \beta < 1) = 0.8$. Using these priors repeat the above confidence interval calculations and comment on the impact of the priors.
22. In clinical trials and many epidemiological investigations we are often in a position to know ranges of implausible values of the regression coefficient. Should we incorporate this knowledge into our inferential calculations, and if so, how?
23. Either use an existing dataset or generate censored data with a single continuous covariate. Evaluate the empirical distribution of the covariate at each failure time in the risk set. Use several transformations of this distribution, e.g., to approximately normal, exponential , or uniform, and take as a test statistic the maximum across all considered transformations. How would you ensure correct control of type I error for this test? What are the advantages and drawbacks of this test?

MODELS AND LIKELIHOOD

24. Describe why the assumption of marginal independence is a stronger assumption than that of conditional independence. Describe situations in which each of these assumptions appears to be reasonable. How is the likelihood function impacted by the assumptions?
25. Suppose that the censoring mechanism is not independent of the survival mechanism, in particular suppose that

$$\log \Pr(T > x + u | C = u, T > u) = 2 \log \Pr(T > x + u | T > u).$$

Write down the likelihood for a parametric model for which the censoring mechanism is governed by this equation. Next, suppose that we can take the above equation to represent the general form for the censoring model but that, instead of the constant value 2, it depends on an unknown parameter, i.e., the number 2 is replaced by α . What kind of data would enable us to estimate the parameter α ?

26. As a class project, simulate data with a dependent censoring mechanism as above with an unknown parameter α . Investigate the distribution of $\hat{\alpha}$ via 1000 simulations.
27. Fit a Weibull proportional hazards model to data including at least two binary regressors; Z_1 and Z_2 . Calculate a 90% confidence intervals for the

probability that the most unfavorable prognosis among the 4 groups has a survival greater than the marginal median. Calculate a 90% confidence interval that a subject chosen randomly from either the most unfavorable, or the second most unfavorable, group has a survival greater than the marginal median.

28. As a generalization of the previous exercise, consider a parametric model with covariate vector Z of dimension p . The p -dimensional model is considered to provide a good approximation to the underlying mechanism generating the observations. On the basis of the fitted p -dimensional model, write down an expression for a confidence interval for the probability of surviving longer than t for a subject in which Z_1 , the first component of Z , is equal to z_1 .
29. For the two-sample exponential model, write down the likelihood and confirm the maximum likelihood estimates given in Section 7.3. Calculate the score test, the likelihood ratio test, and Wald's test, and compare these with the expressions given in Section 7.3. For a dataset with a single binary covariate calculate and compare the three test statistics.
30. On the basis of data, estimate the unknown regression coefficient, β , as the expected value of the conditional likelihood (see Appendix D.4). Do this for both an exponential based likelihood and the partial likelihood. Next, consider the distribution of $\log \beta$ in this context and take an estimate as $\exp E(\log \beta)$. Do you anticipate these two estimators to agree? Note that the corresponding maximum likelihood estimators do agree exactly. Comment.

7.9 Outline of proofs

Theorem 7.2 Again, since $\lambda(t|Z(t) = z) = \lambda_0(t) \exp\{\beta(t)z\}$, then $f(t|Z(t) = z) = \lambda_0(t) \exp\{\beta(t)z\} S(t|Z(t) = z)$, where $S(t|Z(t) = z)$ is the conditional survival function. By Bayes' rule, we can express the conditional density of $Z(t)$ given $T = t$ as

$$f_t(z|T = t) = \frac{f(t|z)g(z)}{\int f(t|z)g(z)dz} = \frac{\lambda_0(t)e^{\beta(t)z}S(t|z)g(z)}{\int \lambda_0(t)e^{\beta(t)z}S(t|z)g(z)dz} = \frac{e^{\beta(t)z}h(z|T \geq t)}{\int e^{\beta(t)z}h(z|T \geq t)}$$

where $h(z|T \geq t)$ is the conditional density of $Z(t)$ given $T \geq t$. From elementary probability calculus, for sets \mathcal{A} and \mathcal{B} we have:

$$P(\mathcal{A}) = P(\mathcal{A}|\mathcal{B})P(\mathcal{B})/P(\mathcal{B}|\mathcal{A})$$

so that we can write:

$$h(z|T \geq t) = f(z|T \geq t, C \geq t)P(C \geq t)/P(C \geq t|z, T \geq t)$$

and, by our conditional independence assumption, the denominator simplifies so that

$$P(C \geq t|z, T \geq t) = P(C \geq t|z).$$

Now, by the law of total probability, we have $P(C \geq t) = \int P(C \geq t|z)g(z)dz$, where the integral is over the domain of definition of Z . Next we replace $P(Z \leq z|T \geq t, C \geq t)$ by the consistent estimate $\sum_{\{Z_j(t) \leq z\}} Y_j(t)/\sum_1^n Y_j(t)$, which is simply the empirical distribution in the risk set, which leads to

$$\hat{P}\{Z(t) \leq z|T = t\} = \frac{\sum_{z_i \leq z} Y_i(t) \exp\{\beta(t)z_i(t)\}\hat{\phi}(z_i, t)}{\sum_{j=1}^n Y_j(t) \exp\{\beta(t)z_j(t)\}\hat{\phi}(z_j, t)}.$$

An application of Slutsky's theorem enables us to claim the result continues to hold whenever $\beta(t)$ is replaced by any consistent estimator $\hat{\beta}(t)$, in particular the partial likelihood estimator when we assume the more restricted model to hold.



Chapter 8

Survival given covariate information

8.1 Chapter summary

We begin by considering the probability that one subject with particular covariates will have a greater survival time than another subject with different covariates, i.e., $\Pr(T_i > T_j | Z_i, Z_j)$. Note that this also provides a Kendall τ -type measure of predictive strength (Gönen and Heller, 2005) and, although not explored in this work, provides a potential alternative to the R^2 that we recommend. Confidence intervals are simple to construct and maintain the same coverage properties as those for β . Using the main results of Chapter 7 we obtain a simple expression for survival probability given a particular covariate configuration, i.e., $S(t|Z \in H)$ where H is some given covariate subspace. When the subspace is the full covariate space then this function coincides with $S(t)$ and the estimate coincides with the Kaplan-Meier estimate. Simple adjustments to cater for the classical difficulty of Kaplan-Meier estimates not necessarily reaching zero are provided. Several different situations are highlighted including survival under informative censoring.

8.2 Context and motivation

We would like to know how marginal survival is impacted by a knowledge of covariate information. In the simplest case, given such information, we might ask how much more likely is it that one individual outlasts another. One of the main purposes of survival analysis is to obtain an estimate of the survivorship function given certain covariate patterns. Although inference for the proportional hazards model ignores specification of the baseline hazard rate, thereby leaving the baseline survivorship function as well as conditional survivorship functions undetermined, it is common to carry out further estimation on these quantities.

The provision of such information may help guide decision making in an applied context.

While it is usually technically difficult to estimate densities and hazards (some kind of smoothing typically being required), it is easier to estimate cumulative hazards and distribution (survivorship) functions. These have already been smoothed, in some sense, via the summing inherently taking place.

Breslow (1972), Breslow and Crowley (1974), using an equivalence between the proportional hazards model and a piecewise exponential regression model, with as many parameters as there are failure times, derived a simple expression for conditional survival given covariate information. An expression for the variance of the Breslow estimate was derived by O'Quigley (1986). Appealing to Bayes' rule, Xu and O'Quigley (2000) obtained the simple expression

$$S(t|Z \in H) = \int_t^{\infty} P(Z \in H|u)dF(u) / \int_0^{\infty} P(Z \in H|u)dF(u),$$

from which an estimate of conditional survival, making a direct appeal to the main theorem, follows (O'Quigley et al., 2005; Xu and Adak, 2002; Xu and O'Quigley, 2000). In most practical applications the two estimators behave similarly. We prefer the second in view of its closer association with basic inference. It is more readily generalized to deal with non-proportional hazards models, in particular the stratified model and models that include random effects. Another predictive quantity of interest is $\Pr(T_i > T_j | Z_i, Z_j)$ which gives the probability for an individual to outlast another given their respective covariate information. Finally, survival, given a non-independent competing risk or non-independent censoring, is considered (Flandre and O'Quigley, 1995; O'Quigley and Flandre, 2006).

The question we would like to answer is expressed straightforwardly as the probability of survival time being greater than t given that Z belongs to some subset H , i.e., $\Pr(T > t | Z \in H)$. Via Bayes' rule, this probability can be immediately expressed in terms of the conditional distribution of Z at $T = t$, together with the marginal distribution of T . An estimate of this conditional distribution is available as a consequence of the main theorem of Chapter 7. The unknown marginal distribution of T can be replaced by the Kaplan-Meier estimate, or some other distribution if we wish to investigate effects in different contexts. We mostly limit attention to the case where the covariate Z is assumed to be time-invariant. The situation becomes more complicated in the presence of time-dependent covariates because of certain restrictions, for example, $Z(\cdot)$ should be an external covariate in order for $S(t|z)$ to be interpretable (Kalbfleisch and Prentice, 2002; Keiding, 1999; Keiding and Gill, 1990; Lin, 1994; Lin and Ying, 1994).

8.3 Probability that T_i is greater than T_j

If two individuals are independently sampled from the same distribution then, by simple symmetry arguments, it is clear that the probability of the first having a longer survival time than the second is just 0.5. If, instead of sampling from the same distribution, each individual is sampled from a distribution determined by the value of their covariate information, then, the stronger the impact of this covariate information, the further away from 0.5 will this probability be. When the covariates do not depend on time then this probability is very easily evaluated using:

Theorem 8.1. *For subjects i and j , having covariate values Z_i and Z_j then, under the proportional hazards model, we can write*

$$\Pr(T_i > T_j | Z_i, Z_j) = \frac{\exp(\beta Z_j)}{\exp(\beta Z_i) + \exp(\beta Z_j)}.$$

An important observation to make is that the expression does not involve $\Lambda_0(t)$. If we define $\psi(a, b : \beta)$ to be $\exp(\beta b)/\{\exp(\beta b) + \exp(\beta a)\}$ we then have:

Corollary 8.1. *A consistent estimate of $\Pr(T_i > T_j | Z_i, Z_j)$, under the proportional hazards model is given by $\psi(Z_i, Z_j : \hat{\beta})$ and $\text{Var log}\{\psi/(1 - \psi)\} \approx (Z_j - Z_i)^2 \text{Var}(\hat{\beta})$.*

The approximation in the corollary arises from an immediate application of the mean value theorem (Appendix A). In the theorem and corollary it is assumed that Z_i and Z_j are scalars and that the model involves only a one-dimensional covariate. Extension to the multivariate case is again immediate, and instead of $\hat{\beta}Z_i$ in $\psi(Z_i, Z_j : \hat{\beta})$ being a scalar it can be replaced by the usual inner product (prognostic index). Suppose that the dimension of β and Z is p and that we use the notation Z_{jr} to indicate, for subject j , the r th component of Z_j . Applying the delta method (Appendix A.10),

$$\text{Var log}\{\psi/(1 - \psi)\} \approx \sum_{r=1}^p \sum_{s=1}^p (Z_{jr} - Z_{ir})(Z_{js} - Z_{is}) \text{Cov}(\hat{\beta}_r, \hat{\beta}_s).$$

Use of ψ would seem to be a particularly simple and transparent way in which to summarize the impact, or predictive strength, of regression effects. We know that significance levels alone, directly dependent as they are on sample sizes (more precisely the amount of uncensored observations), are not indicative of predictive strength. For this we need to make use of explained variation or explained randomness measures. Whereas these measures are being averaged over some covariate distribution it is also helpful to have specific measures given two particular covariate configurations. If a patient is told that they have an unfavorable prognosis in the light of studies on their prognostic variables it can be helpful to add to that some idea on just how unfavorable it is. If $\psi(Z_i, Z_j : \hat{\beta})$ remains close

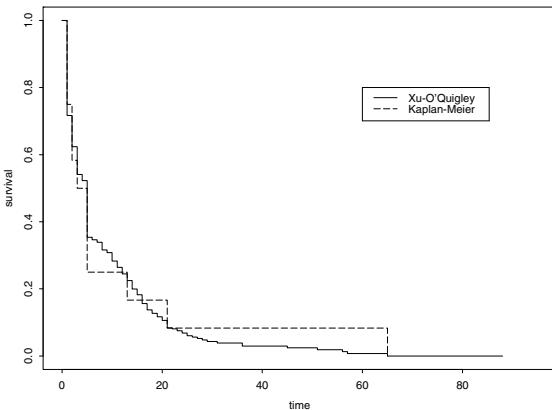


Figure 8.1: Kaplan-Meier survival plot for gastric cancer data and model-based plot based on selection of the covariate CEA to lie in interval $(0,100)$.

to 0.5, then, regardless of significance which may be of importance in considering the public health effects for large groups, an individual might reasonably feel that he or she is not particularly disadvantaged by such an unfavorable prognosis. All of the above is very straightforward when dealing with the evaluation of covariate “points” as given by Z_i and Z_j .

More generally, as discussed in Section 8.4, we may wish to consider some range within the covariate space in which case we specify $\Pr(T_i > T_j | Z_i \in H_i, Z_j \in H_j)$. To do this in practice we would need to integrate or sum over the range of points of Z_i and Z_j , contained in H_i and H_j , respectively, with respect to the densities of Z_i and Z_j within these sets.

The connection between the population explained variation Ω^2 and $\psi(Z_i, Z_j : \beta)$ would be worthy of further investigation. Xu (1996) showed that, for fixed values of the covariates, and, regarding both Ω^2 and $\psi(Z_i, Z_j : \beta)$ as functions of strength of effect as measured by β , then these two quantities are monotonic increasing functions of one another. There is a one-to-one correspondence between them so that, clearly, in some sense, they are measuring the same phenomenon but on different scales.

Another potential application of $\psi(Z_i, Z_j : \beta)$ is in relative survival where we take Z_j to be a value across some reference group or population and Z_i to be a value for some group under study, for instance, a group having recently been treated for some particular chronic disease. The negative effects of belonging to that group, as opposed to the reference group, are then directly quantified by ψ . Further developments to these ideas immediately suggest themselves and, foremost, making use of time itself. Rather than limiting our attention to $\Pr(T_i > T_j | Z_i, Z_j)$ we can explicitly introduce into this quantity the amount of time elapsed. We would not only condition on the values of Z_i and Z_j but also that both T_i and T_j are greater than t , bring our attention then to $\Pr(T_i > T_j | Z_i, Z_j, T_i > t, T_j > t)$. The resulting expressions would be more involved than

in Theorem 8.1 but could be worked out. Applications to cure studies follow since we could conceive of situations in which this expression diminishes with t , becoming, at some point, sufficiently close to the value 0.5 to claim that the exposed group no longer carries a disadvantage when compared to the reference group.

8.4 Conditional survival given $Z \in H$

Under the proportional hazards model, the conditional survival probability $S(t|z) = \Pr(T > t|Z = z)$ can be estimated using the development of Breslow (1972), Breslow and Crowley (1974) whereby

$$\hat{S}(t|z) = \exp\left(-\hat{\Lambda}_0(t)e^{\hat{\beta}z}\right); \quad \hat{\Lambda}_0(t) = \sum_{X_i \leq t} \frac{\delta_i}{\sum_{j=1}^n Y_j(X_i)e^{\hat{\beta}Z_j}}. \quad (8.1)$$

An expression for the large sample variance of $Y = \log - \log S(t|z)$ was obtained by O'Quigley (1986). Symmetric intervals for Y can then be transformed into more plausible ones, at least having better coverage properties according to the arguments of O'Quigley (1986), by simply applying the exponential function twice. The Breslow estimate concerns a single point z . It is a natural question to ask what is the survival probability given that the covariates belong to some subset H . The set H may denote for example an age group, or a certain range of continuous measurement, or a combination of those.

In general we assume H to be a subset of the p -dimensional Euclidean space. A natural approach may be to take the above formula, which is applied to a point, and average a set of curves over all points belonging to the set H of interest. For this we would need some distribution for the z across the set H . Keiding (1995) has a discussion on expected survival curves over a historical, or background, population, where the main approaches are to take an average of the individual survival curves obtained from the above equation. See also Sasieni (2003). Following that, one might use the equation to estimate $S(t|z)$ for all z in H , then average over an estimated distribution of Z . Xu and O'Quigley (2000) adopted a different starting point in trying to estimate directly the survival probabilities given that $Z \in H$. Apart from being direct, this approach is the more natural in view of the main theorem of Section 7.5. What is more, the method can also have application to situations in which the regression effect varies with time. In the following, for notational simplicity, we will assume $p = 1$. Extensions to $p > 1$ are immediate. As for almost all of the quantities we have considered it turns out to be most useful to work with the conditional distribution of Z given $T = t$ rather than the other way around. Everything is fully specified by the joint distribution of (T, Z) and we keep in mind that this can be expressed either as the conditional distribution of T given Z , together with the marginal distribution of Z or as the conditional distribution of Z given T , together with the marginal distribution of T .

Using Bayes' formula to rewrite the conditional distribution of T given information on Z , we have:

$$S(t|Z \in H) = \frac{\int_t^\infty P(Z \in H|u)dF(u)}{\int_0^\infty P(Z \in H|u)dF(u)}. \quad (8.2)$$

This is a very simple and elegant expression and we can see from it how conditioning on the covariates modifies the underlying survival distribution. If H were to be the whole domain of definition of Z , in which case Z is contained in H with probability one, then the left-hand side of the equation simply reduces to the marginal distribution of T . This is nice and, below, we will see that we have something entirely analogous when dealing with sample-based estimates whereby, if we are to consider the whole of the covariate space, then we simply recover the usual empirical estimate. In particular this is just the Kaplan-Meier estimate when the increments of the right-hand side of the equation are those of the Kaplan-Meier function. The main theorem of Section 7.5 implies that $P(Z \in H|t)$ can be consistently estimated from

Lemma 8.1. $\hat{P}(Z \in H|t)$ is consistent for the probability $P(Z \in H|t)$ where

$$\hat{P}(Z \in H|t) = \sum_{\{j: Z_j \in H\}} \pi_j(\hat{\beta}, t) = \frac{\sum_H Y_j(t) \exp\{\hat{\beta} Z_j\}}{\sum Y_j(t) \exp\{\hat{\beta} Z_j\}}. \quad (8.3)$$

This striking, and simple, result is the main ingredient needed to obtain survival function estimates conditional on particular covariate configurations. The rest essentially the step increments in the Kaplan-Meier curve are readily available. Unfortunately, a problem that is always present when dealing with censored data remains and that is the possibility that the estimated survival function does not decrease all the way to zero. This will happen when the largest observation is not a failure. To look at this more closely, let $\hat{F}(\cdot) = 1 - \hat{S}(\cdot)$ be the left-continuous Kaplan-Meier (KM) estimator of $F(\cdot)$. Let $0 = t_0 < t_1 < \dots < t_k$ be the distinct failure times, and let $W(t_i) = d\hat{F}(t_i)$ be the stepsize of \hat{F} at t_i . If the last observation is a failure, then,

$$\hat{S}(t|Z \in H) = \frac{\int_t^\infty \hat{P}(Z \in H|u)d\hat{F}(u)}{\int_0^\infty \hat{P}(Z \in H|u)d\hat{F}(u)} = \frac{\sum_{t_i > t} \hat{P}(Z \in H|t_i)W(t_i)}{\sum_{i=1}^k \hat{P}(Z \in H|t_i)W(t_i)}. \quad (8.4)$$

When the last observation is not a failure and $\sum_1^k W(t_i) < 1$, an application of the law of total probability indicates that the quantity B_1 where $B_1 = \hat{P}(Z \in H|T > t_k)\hat{S}(t_k)$ should be added to both the numerator and the denominator

in (8.4) This is due to the fact that the estimated survival distribution is not summing to one. Alternatively, we could simply reduce the support of the time frame to be less than or equal to the greatest observed failure. In addition, using the empirical estimate over all the subjects that are censored after the last observed failure, we have:

$$\hat{P}(Z \in H | T > t_k) = \frac{\sum_H Y_j(t_k+)}{\sum Y_j(t_k+)}, \quad (8.5)$$

where t_k+ denotes the moment right after time t_k . Therefore we can write:

$$\hat{S}(t|Z \in H) = \frac{\sum_{t_i > t} \hat{P}(Z \in H | t_i) W(t_i) + \hat{P}(Z \in H | T > t_k) \{1 - \sum_1^k W(t_i)\}}{\sum_1^k \hat{P}(Z \in H | t_i) W(t_i) + \hat{P}(Z \in H | T > t_k) \{1 - \sum_1^k W(t_i)\}}.$$

The above estimate of the conditional survival function is readily calculated, since each term derives from standard procedures of survival analysis to fit the Cox model. An attractive aspect of the approach is that when H includes all the possible values of z , the estimator simply becomes the Kaplan-Meier estimator of the marginal survival function. The estimate of the conditional survival probability $P(T > t + u | T > u, Z \in H)$ can also be nicely written in the form of a simple ratio where the numerator is given by $\sum_{t_i > t+u} C_i + B_1$ and the denominator by $\sum_{t_i > t} C_i + B_1$ and where $C_i = \hat{P}(Z \in H | t_i) W(t_i)$. For the gastric cancer data it is interesting to contrast the survival estimate based on these calculations for the quantity $\Pr\{T > t | Z_1 \in (0, 100)\}$, where Z_1 is the tumor marker CEA and the simple Kaplan-Meier estimator based on the subset of the data defined by tumor marker CEA less than 100. This is shown in Figure 8.1 and there is good agreement between the model-based estimator of Xu-O'Quigley and the Kaplan-Meier estimate. This, although not used here as a goodness-of-fit test in its own right, can be taken to indicate that the model appears reasonable over the specified range of the covariate.

VARIANCE APPROXIMATIONS FOR $\hat{S}(t|Z \in H)$

In this section we assume that the proportional hazards model holds with true $\beta = \beta_0$. To obtain the asymptotic variance of (8.3) at each t , we use the approach of Link (1984) and O'Quigley (1986) which is tractable and can be computed using standard packages for fitting the Cox regression model. Our experience is that this provides good estimates when compared with methods such as the bootstrap. There are two sources of variation in $\hat{S}(t|Z \in H)$; one caused by the estimate of conditional probability of survival with given β , the other by the uncertainty in $\hat{\beta}$. Using a first-order Taylor series expansion we have:

$$\hat{S}(t|Z \in H) = \hat{S}(t|Z \in H)|_{\beta_0} + (\hat{\beta} - \beta_0) \frac{\partial \hat{S}(t|Z \in H)}{\partial \beta} \Big|_{\beta=\dot{\beta}}, \quad (8.6)$$

where $\dot{\beta}$ lies on the line segment between β_0 and $\hat{\beta}$. We then need to bring together some results.

Lemma 8.2. *The quantity $\{\partial\hat{S}(t|Z \in H)/\partial\beta\}|_{\beta=\hat{\beta}}$ is asymptotically constant.*

Lemma 8.3. *The quantity $\hat{\beta} - \beta_0$ is asymptotically uncorrelated with $\hat{S}(t|Z \in H)|_{\beta_0}$.*

Corollary 8.2. *The variance of $\hat{S}(t|Z \in H)$ is approximated by*

$$\text{Var}\{\hat{S}(t|Z \in H)\} \approx \text{Var}\{\hat{S}(t|Z \in H)|_{\beta_0}\} + \left\{ \frac{\partial\hat{S}(t|Z \in H)}{\partial\beta} \Big|_{\beta=\hat{\beta}} \right\}^2 \text{Var}(\hat{\beta}). \quad (8.7)$$

The first term in (8.7) gives the variation due to the estimation of the conditional survival, the second term the variation caused by $\hat{\beta}$. Details are given at the end of the chapter. In addition, we have:

Theorem 8.2. *Under the proportional hazards model $\hat{S}(t|Z \in H)$ is asymptotically normal.*

As a consequence one can use the above estimated variance to construct confidence intervals for $S(t|Z \in H)$ at each t .

Theorem 8.3. $\sqrt{n}U(\beta_0)$ is asymptotically equivalent to $n^{-1/2} \sum_1^n \omega_i(\beta_0)$, where

$$\omega_i(\beta) = \int_0^1 \left\{ Z_i - \frac{s^{(1)}(\beta, t)}{s^{(0)}(\beta, t)} \right\} dN_i(t) - \int_0^1 Y_i(t) e^{\beta Z_i} \left\{ Z_i - \frac{s^{(1)}(\beta, t)}{s^{(0)}(\beta, t)} \right\} \lambda_0(t) dt$$

and $N_i(t) = I\{T_i \leq t, T_i \leq C_i\}$.

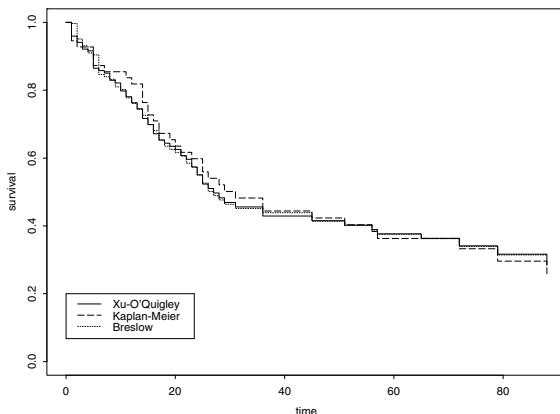


Figure 8.2: Survival probabilities for myeloma data based on prognostic index of lower 0.33 percentile using Kaplan-Meier, Breslow, and Xu-O'Quigley estimators.

So all that remains to show the asymptotic normality of $\hat{S}(t^*|Z \in H)$ is to show that the numerator of $\hat{S}(t^*|Z \in H)|_{\beta_0}$ is also asymptotically equivalent to $1/n$ times a sum of n i.i.d. random variables like the above, since the denominator of it we know is consistent for $P(Z \in H)$. To avoid becoming too cluttered we drop the subscript of β_0 in $\hat{S}(t^*|Z \in H)|_{\beta_0}$. The numerator of $\hat{S}(t^*|Z \in H)$ is $\int_{t^*}^{\infty} \hat{P}(Z \in H|t) d\hat{F}(t)$. Note that $\sqrt{n}\{\int_{t^*}^{\infty} \hat{P}(Z \in H|t) d\hat{F}(t) - P(Z \in H, T > t^*)\} = \sqrt{n}\int_{t^*}^{\infty} P(Z \in H|t) d\{\hat{F}(t) - F(t)\} + \sqrt{n}\int_{t^*}^{\infty} \{\hat{P}(Z \in H|t) - P(Z \in H|t)\} d\{\hat{F}(t) - F(t)\} + \sqrt{n}\int_{t^*}^{\infty} \{\hat{P}(Z \in H|t) - P(Z \in H|t)\} dF(t)$. Now $\sqrt{n}\{\hat{F}(t) - F(t)\}$ converges in distribution to a zero-mean Gaussian process. Therefore, the second term in the above expression is $o_p(1)$. The last term is $A_1 + o_p(1)$ where

$$\begin{aligned} A_1 &= \sqrt{n} \int_{t^*}^1 \left\{ \frac{S^{(H)}(\beta_0, t)}{s^{(0)}(\beta_0, t)} - \frac{s^{(H)}(\beta_0, t)S^{(0)}(\beta_0, t)}{s^{(0)}(\beta_0, t)^2} \right\} dF(t) \\ &= n^{-1/2} \sum_{i=1}^n \int_{t^*}^1 \frac{Y_i(t)e^{\beta_0 Z_i}}{s^{(0)}(\beta_0, t)} \left\{ Q_i - \frac{s^{(H)}(\beta_0, t)}{s^{(0)}(\beta_0, t)} \right\} dF(t). \end{aligned}$$

As for the first term, we can use Theorem II.5 of Xu (1996), which derives from a result of Stute (1995). With $\phi(t) = 1_{[t^*, 1]}(t)P(Z \in H|t)$ in the theorem, the first term is equal to $n^{-1/2} \sum_{i=1}^n \nu_i + \sqrt{n}R_n$, where $|R_n| = o_p(n^{-1/2})$ and ν 's are i.i.d. with mean zero, each being a function of X_i and δ_i .

RELATIVE MERITS OF COMPETING ESTIMATORS

A thorough study of the relative merits of the different estimators has yet to be carried out. For any such study the first thing to consider would be the chosen yardstick with which to evaluate any estimate. For example, should the whole curve be considered or only some part of it? Should a "distance" measure be an average discrepancy, the greatest discrepancy over some range or some weighted discrepancy? It is even very possible that some estimator would outperform another with respect to one distance measure and perform less well than the competitor with respect to another distance measure. In addition to this it is also quite possible for some estimator to maintain an advantage for certain population situations but to lose this advantage in other situations. In the light of these remarks it may then appear difficult to obtain a simple unequivocal finding in favor of one or another estimator. Nonetheless it would be nice to know more and further work here would be of help. In the meantime it helps to provide us with some insight by considering various situations which have arisen when looking at real datasets.

PROGNOSTIC BIOMARKERS IN GASTRIC CANCER

Rashid et al. (1982) studied a group of gastric cancer patients. The goal of the study was to determine the prognostic impact of certain acute phase reac-

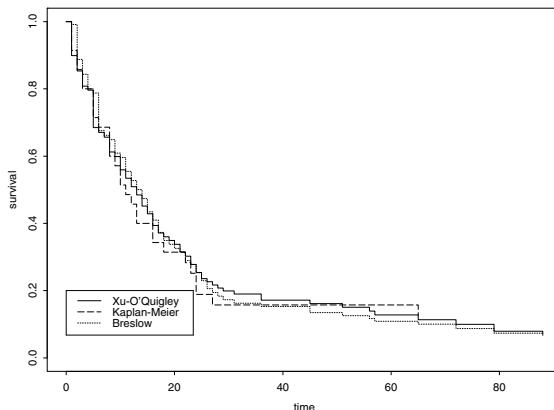


Figure 8.3: Survival probabilities for myeloma data based on prognostic index between 0.33 and 0.66 percentiles using Kaplan-Meier, Breslow, and Xu-O'Quigley estimators.

tant proteins measured pre-operatively. This biological information could then be used in conjunction with clinical information obtained at the time of surgical intervention to investigate the relative prognostic impact of the different factors. There were 104 patients and five covariates: stage (degree of invasion), ACT protein (α_1 -anti chymotrypsin), CEA (carcino embryonic antigen), CRP (C-reactive protein), and AGP (alpha glyco protein).

Although it is known that stage has strong predictive capability for survival, it can only be determined after surgery. Of interest was the prediction of a patient's survival based on pre-operative measurements alone, an assessment of which might be used to help guide clinical decision making. Values of certain covariates such as CEA are very skew and have a wide range from below 1 to over 900. After a log transformation of CEA, a proportional hazards model with the four pre-operative covariates included was not rejected by the data. In fitting such a model, CRP and AGP were found to be insignificant at 0.05 level in the presence of ACT and logCEA. Therefore only the latter two were retained for subsequent analysis. The regression coefficients for ACT and logCEA were calculated to be 1.817 and 0.212, with standard errors 0.41 and 0.07, respectively. This gives a range of 0.92–4.48 for the estimated prognostic index $\beta'z$ (Andersen et al., 1983; Altman and Andersen, 1986).

If we divide the patients into three groups, with low, median, and high risks, according to the prognostic index <2, 2–3, and >3, we can predict the survival probabilities in each risk group. It was then possible to estimate the survival curves for these three groups, and these were calculated by Xu and O'Quigley (2000). The curves can be compared with the empirical Kaplan-Meier curves which make no appeal to the model. Agreement is strong. We also chose to define the set H by all those patients having values of CEA less than 50. This

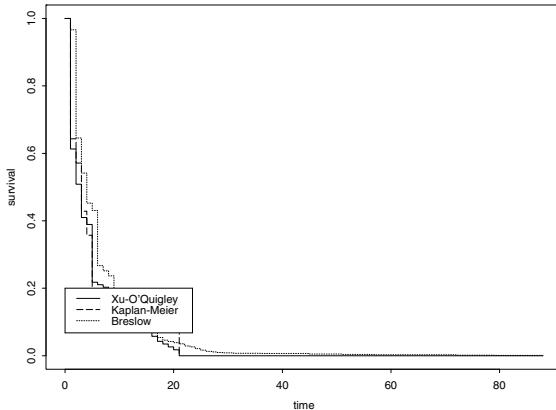


Figure 8.4: Survival probabilities for myeloma data based on prognostic index for values greater than the 0.33 percentile using 3 different estimators.

was not very far from the median value and provided enough observations for a good empirical Kaplan-Meier estimate. The empirical estimate and the model-based estimates show good agreement. For the three prognostic groups, defined on the basis of a division of the prognostic index from the multivariate model, the empirical Kaplan-Meier estimates, the Breslow estimates, and the Xu and O'Quigley estimates are shown in Figures 8.2, 8.3, and 8.4. Again agreement is strong among the three estimators.

8.5 Other relative risk forms

The estimator described above is not limited to proportional hazards models. The formula itself came from a simple application of Bayes' rule, and the marginal distribution of T can always be estimated by the Kaplan-Meier estimator or some other estimator if we wish to use other assumptions. Taking a general form of relative risk $r(t; z)$, so that $\lambda(t|z) = \lambda_0(t)r(t; z)$. Assume also that $r(t; z)$ can be estimated by $\hat{r}(t; z)$, for example, that it has a known functional form and a finite or infinite dimensional parameter that can be consistently estimated. Special cases of $r(t; z)$ are $\exp(\beta z)$, $1 + \beta z$, and $\exp\{\beta(t)z\}$. Since the main theorem of Section 7.5 extends readily to other relative risk models it is straightforward to derive analogous results to those above. We are still able to estimate $S(t|Z \in H)$, with $\hat{P}(Z \in H|t) = \sum_{\{j: Z_j \in H\}} \pi_j(t)$. This is an important extension of the estimator since we may wish to directly work with some form of a non-proportional hazards model.

STRATIFIED AND RANDOM EFFECTS MODELS

In studies where there is stratification, in particular, when there are many strata, the approach can be a relatively simple one to estimate conditional survivorship

as it does not require the estimation of the baseline hazards. Suppose that V is the stratification variable, and we are interested in the survival given $Z \in H$. Note $P(Z \in H|t) = \sum_v P(Z \in H|t, V = v)P(V = v)$. We can estimate $P(Z \in H|t, V = v)$ by $\sum_{\{j: Z_j \in H\}} \pi_j^v(\hat{\beta}, t)$, where $\pi_j^v(\beta, t)$ is the conditional probability defined within strata v , and estimate $P(V = v)$ by the empirical distribution of V . Similarly to the stratified case, (8.3) can also be used to estimate survival under random effects models arising from clustered, such as genetic or familial data. The frailty models under such settings can be written as

$$\lambda_{ij}(t) = \lambda_0(t)\omega_i \exp\{\beta z_{ij}\}, \quad (8.8)$$

where λ_{ij} is the hazard function of the j th individual in the i th cluster. This is the same as a stratified model, except that we do not observe the values of the “stratification variable” ω ; but such values are not needed in the calculation described above for stratified models. So the procedure described above can be used to estimate $S(t|Z \in H)$. In both cases considered here, we need reasonable stratum or cluster sizes in order to get a good estimate of $P(Z \in H|t, v)$ or $P(Z \in H|t, \omega)$.

CONDITIONAL INDEPENDENT CENSORSHIP

It is also possible to generalize (8.3) to the cases where C and T are independent given Z . First let us assume that the covariate Z is discrete with finitely many categories. Then, the KM estimate can be replaced by a weighted Kaplan-Meier (WKM) estimate (Murray and Tsiatis, 1999, 2001, 1996), which still consistently estimates $F(\cdot)$ under the conditional independent censorship. The WKM estimate calculates the subgroup KM estimates within each category of the covariate values, and then weights these subgroup estimates by the empirical distribution of Z . For the conditional distribution of Z given $T = t$, from the proof of the main theorem (Section 7.5),

$$f(z|T = t) = \frac{e^{\beta z} S(t|z)g(z)}{\int e^{\beta z} S(t|z)g(z)dz}. \quad (8.9)$$

If we estimate $S(t|z)$ by the subgroup KM estimate within the category of value z , and the marginal $g(z)$ by the empirical probabilities, we are still able to consistently estimate the conditional distribution of Z given $T = t$ and thus $S(t|Z \in H)$. For other types of covariate distribution such as continuous covariates, we need to incorporate the covariates into categories, and Murray and Tsiatis (1996) suggested guidelines which could be useful in practice.

SPARSE DATA IN HIGH DIMENSIONS

When data are sparse in high dimensions, i.e., multiple covariates, the $\pi_j(\hat{\beta}, t)$'s used to estimate the conditional distribution of Z given $T = t$ may encounter some difficulties, because they are like empirical distributions. In fact, they are obtained through the empirical distribution of Z given $T \geq t$. In this case, as seen in the gastric cancer example, we consider H as the Cartesian product of the range for that component with $(-\infty, \infty)$ for the remaining components. Of course the actual range of the covariate vector is reflected in the data itself. However, sometimes we might specify H as a relatively small “block” in a p -dimensional space. For example, for the gastric cancer data, we might ask what is the survival probabilities given that CEA is greater than 10 but less than 15. In this type of situation there might be so few observations in H that quite early on the risk sets may no longer contain any subjects with covariates in H , allowing for $\{\pi_j(\beta, t)\}_j$ to be a sufficiently reliable estimate of the conditional distribution of Z given $T = t$.

In many practical cases one is likely to ask for the conditional survival probabilities given a range for a single component of the covariate vector, while allowing the other components to vary freely. We can proceed as follows. Denote the prognostic index $\eta = \beta' z$. Under the model two individuals should have the same survival probabilities as long as $\eta_1 = \eta_2$. That is, conditioning on $Z \in H$ is equivalent to conditioning on $\eta \in \beta' H$. Therefore, $S(t|Z \in H) = S(t|\eta \in \beta' H)$, where $\beta' H = \{\beta' z | z \in H\}$ and should contain potentially more observations than H does. This way the p -dimensional vector of covariates is reduced to the one-dimensional prognostic index, and the same number of observations in the risk set is now used to estimate the one-dimensional conditional distribution of η . Thus $\hat{S}(t|Z \in H)$ becomes:

$$\hat{S}(t|\eta \in \beta' H) = \frac{\sum_{t_i > t} \hat{P}(\eta \in \beta' H | t_i) W(t_i) + \hat{P}(\eta \in \beta' H | T > t_k) \hat{S}(t_k)}{\sum_{i=1}^k \hat{P}(\eta \in \beta' H | t_i) W(t_i) + \hat{P}(\eta \in \beta' H | T > t_k) \hat{S}(t_k)},$$

where $\hat{P}(\eta \in \beta' H | t) = \sum_{\{j: \eta_j \in \hat{\beta}' H\}} \pi_j(\hat{\beta}, t)$, and

$$\hat{P}(\eta \in \beta' H | T > t_k) = \sum_{\{j: \eta_j \in \hat{\beta}' H\}} Y_j(t_k+) / \sum_{j: \eta_j \in \hat{\beta}' H} Y_j(t_k+).$$

As before, since one can consistently estimate β , the above expression still provides a consistent estimate of $S(t|Z \in H)$. Note that when z is a single covariate, $z \in H$ is exactly the same as $\eta \in \beta H$ (unless $\beta = 0$ in which case the covariates have no predictive capability), so the above is consistent with the one-dimensional case developed earlier. While we regard the above as one possible approach under high dimensions when there are not “enough” observations falling into the ranges

of covariates of interest, the variance estimation and the asymptotic properties seem to be more complicated as the estimate of β enters both η and the set $\beta'H$. When there is a need to use the estimate based on $\beta'H$, resampling methods such as bootstrap could be employed for inferential purposes. There are many potential areas for research here in order to further develop these techniques. Comparative studies would be of value. These would not be straightforward since comparing competing estimated curves requires considering the whole of the curve. For instance, at some points in time one estimator may outperform another whereas, at a later point, it may be the converse that holds. Some measure of overall distance such as the maximum or average discrepancy could be considered.

8.6 Informative censoring

Events that occur through time, alongside the main outcome of interest, may often provide prognostic information on the outcome itself. These can be viewed as time-dependent covariates and, as before, in light of the main theorem (Section 7.5), it is still straightforward to use such information in the expression of the survivorship function. Since, in essence, we sum, or integrate, future information, it can be necessary to postulate paths that the covariate process might take. Emphasis remains on the conditional distribution of survival given current and evolving covariate information. This differs slightly from an approach, more even handed with respect to the covariate process alongside the survival endpoints, that makes an appeal to joint modeling. Even so, the end goal is mostly the same, that of characterizing the survival experience given covariate information.

Paths that remain constant are the easiest to interpret and, in certain cases, the simple fact of having a value tells us that the subject is still at risk for the event of interest Kalbfleisch and Prentice (2002). The use of the main theorem (Section 7.5) in this context makes things particularly simple since the relevant probabilities express themselves in terms of the conditional distribution of the covariate at given time points. We can make an immediate generalization of Equation 8.2 if we also wish to condition on the fact that $T > s$. We have:

$$S(t+s|Z \in H, T > s) = \frac{\int_{t+s}^{\infty} P(Z \in H|u)dF(u|u > s)}{\int_s^{\infty} P(Z \in H|u)dF(u|u > s)},$$

and, in exactly the same way as before, and assuming that the last observation is a failure, we replace this expression in practice by its empirical equivalent

$$\hat{S}(t+s|Z \in H, T > s) = \frac{\int_s^{\infty} \hat{P}(Z \in H|u)d\hat{F}(u|u > s)}{\int_s^{\infty} \hat{P}(Z \in H|u)d\hat{F}(u|u > s)} = \frac{\sum_{t_i > t+s} \hat{P}(Z \in H|t_i)W(t_i)}{\sum_{t_i > s} \hat{P}(Z \in H|t_i)W(t_i)}.$$

When the last observation is not a failure and $\sum_1^k W(t_i) < 1$ we can make a further adjustment for this in the same way as before.

One reason for favoring the Xu-O'Quigley estimate of survival over the Breslow estimate is the immediate extension to time-dependent covariates and to time-dependent covariate effects. Keeping in mind the discussion of Kalbfleisch and Prentice (2002), concerning internal and external time-dependent covariates, whether or not these are determined in advance or can be considered as an individual process generated through time, we can, at least formally, leaving aside interpretation questions, apply the above formulae. The interpretation questions are solved by sequentially conditioning on time as we progress along the time axis and, thereby, the further straightforward extension which seeks to quantify the probability of the event $T > t$, conditioned by $T > s$ where $s < t$, is particularly useful.

SURVIVAL ESTIMATION USING SURROGATE ENDPOINTS

For certain chronic diseases, a notable example being HIV, researchers have considered the need for clinical evaluation that might yield quicker results. Surrogate endpoints, viewed as time-dependent covariates, can help in addressing this issue and have received attention in the medical statistical literature (Ellenberg and Hamilton, 1989; Hillis and Seigel, 1989; Wittes et al., 1989). Herson (1989) wrote that “a surrogate endpoint is one that an investigator deems as correlated with an endpoint of interest but that can perhaps be measured at lower expense or at an earlier time than the endpoint of interest.” Prentice (1989) defined a surrogate endpoint as “a response variable for which a test of the null hypothesis of no relationship to the treatment groups under comparison is also a valid test of the corresponding null hypothesis based on the true endpoint.”

We can view a surrogate endpoint to be a time-dependent response variable of prognostic value obtained during follow-up, which indicates an objective progression of disease. In the survival context a surrogate variable is most often a discrete endpoint indicating, in some way, disease progression. Flandre and O'Quigley (1995) proposed a two-stage procedure for survival studies when a surrogate or intermediate time-dependent response variable is available for some patients. In the presence of a time-dependent intermediary event we can write:

$$S(t) = \int_0^\infty S(t|c)g(c)dc = \int_0^t S(t|c < t)g(c)dc + \int_t^\infty S(t|c \geq t)g(c)dc.$$

If the effect of the occurrence of the intermediary event is to change the hazard function of death $\lambda(t)$ from $\lambda_1(t)$ to $\lambda_2(t)$, that is: $\lambda(t) = \lambda_1(t)$ if $t \leq C$ and is equal to $\lambda_2(t)$ otherwise then $\lambda_1(t) = \lambda_2(t)$ when the intermediary response variable has no influence on survival. When $\lambda_2(t) < \lambda_1(t)$ or $\lambda_2(t) > \lambda_1(t)$ then the intermediary, or surrogate, response variable carries relatively favorable or unfavorable predictions of survival. Thus the quantity $\pi(t) = \lambda_2(t)/\lambda_1(t)$ is a

measure of the effect of the surrogate response on survival. When $f_1(t)$ and $f_2(t)$ are the density functions, $S_1(t)$ and $S_2(t)$ the survivorship functions corresponding to the hazard functions $\lambda_1(t)$ and $\lambda_2(t)$, respectively, then the marginal survival function is

$$S(t) = \int_0^t \exp - \left[\int_0^c \lambda_1(u) du + \int_c^t \lambda_2(u) du \right] dG(c) + \exp \left[- \int_0^t \lambda_1(u) du \right] G(t).$$

In the first stage of a two-stage design, all patients are followed to the endpoint of primary concern and for some subset of the patients there will be the surrogate information collected at an intermediary point during the follow-up. The purpose of the first stage is to estimate the relationship between the occurrence of the surrogate response variable and the remaining survival time. This information can then be used in the second stage, at which time, for patients who reach the surrogate endpoint, follow-up is terminated. Such patients could be considered as censored under a particular dependent censorship model, the censoring being, in general “informative.” The Kaplan-Meier estimator will not generally be consistent if the survival time and an informative censoring time are dependent but treated as though they were independent. Flandre and O’Quigley (1995) proposed a nonparametric estimator of the survival function for data collected in a two-stage procedure. A nonparametric permutation test for comparing the survival distributions of two treatments using the two-stage procedure is also readily derived.

The idea behind a two-stage design in the context of a time-dependent surrogate endpoint is to reduce the overall duration of the study. This potential reduction occurs at the second stage, where follow-up is terminated on patients for whom the surrogate variable has been observed. The first stage is used to quantify the strength of the relationship between occurrence of the surrogate variable and subsequent survival. It is this information, obtained from the first stage analysis, that will enable us to make inferences on survival on the basis of not only observed failures, but also observed occurrences of the surrogate. In the context of clinical trials, as pointed out by Prentice (1989), the surrogate variable must attempt to “capture” any relationship between the treatment and the true endpoint. We may wish to formally test the validity of a surrogate variable before proceeding to the second stage using a standard likelihood ratio test.

In the first stage N_1 patients are enrolled and followed to the endpoint of primary concern (e.g., death) or to censorship, as in a classical study, and information concerning the surrogate variable is recorded. Survival time is then either the true survival time or the informative censoring time. The information available for some patients will consist of both time until the surrogate variable and survival time, while for others (i.e., those who die or are censored without the occurrence of the surrogate variable) it consists only of survival time. In the second stage a new set of patients (N_2) is enrolled in the study. For those patients,

the follow-up is completed when the surrogate variable has been observed. Thus, the information collected consists only of one time, either the time until the surrogate variable is reached or the survival time. In some cases the two stages may correspond to separate and distinct studies; the second stage being the clinical trial of immediate interest while the first stage would be an earlier trial carried out under similar conditions.

When parametric models are assumed, then the likelihood function can be obtained directly and provides the basis for inference. The main idea follows that of Lagakos et al. (1978), Lagakos (1976, 1977) who introduced a stochastic model that utilizes the information on a time-dependent event (auxiliary variable) that may be related to survival time. By taking $\lambda_0(t) = \lambda_2$, where $\lambda_0(\cdot)$ is the hazard function for the occurrence of the surrogate response, $\lambda_1(t) = \lambda_1$, and $\lambda_2(t) = \lambda_3$, the survival function has the marginal distribution function given by Lagakos. The Lagakos model itself is a special case of the bivariate model of Freund (1961), applicable to the lifetimes of certain two-component systems where a failure of the first or second component alters the failure rate of the second or first component from β to β' or (α to α'). By taking $\alpha = \lambda_1(t)$, $\alpha' = \lambda_2(t)$, and $\beta = \beta' = \lambda_0(t)$, the Freund model can be viewed as a special case of the model described above.

Slud and Rubinstein (1983) make simple nonparametric assumptions on the joint density of (T, C) and consider the function $\rho(t)$ defined by

$$\rho(t) = \lim_{\delta \rightarrow 0} \frac{\Pr(t < T < t + \delta | T > t, C < t)}{\Pr(t < T < t + \delta | T > t, C \geq t)}.$$

It is possible to make use of this function in order to derive a nonparametric estimation of $S(t)$ for use with the two-stage procedure with surrogate endpoints and which accounts for the dependent censoring. The authors present a class of nonparametric assumptions on the conditional distribution of T given C which leads to a consistent generalization of the Kaplan-Meier survival curve estimator. Individual values of $\rho(t) > 1$ mean that, after censoring, the hazard risk of death for the patient is increased, and thus if we make an assumption of independence of T and C , the Kaplan-Meier estimator will tend to overestimate survival. On the other hand, if $\rho(t) < 1$, the independence assumption will lead to underestimation of the survival curve. Recalling the usual notation whereby $\delta_i = 1$ if $T_i \leq C_i$ and $X_i = \min(T_i, C_i)$, then X_i defines the observed portion of survival time. The Slud and Rubinstein estimator is given by

$$\hat{S}_\rho(t) = N^{-1} \left\{ n(t) + \sum_{k=0}^{d(t)-1} W_k \prod_{i=k+1}^{d(t)} \frac{n_i - 1}{n_i + \rho_i - 1} \right\}, \quad (8.10)$$

where $n(t) = \sum I(X_i > t)$, $n_j = \sum_i I(X_i \geq X_j)$, $d(t) = \sum I(\delta_i = 1, X_i \leq t)$, $\rho_i = \rho(t_i)$ and W_j is the number of patients censored between two consecutively

ordered failure times X_j and X_{j+1} . When $\rho_i = 1$ it follows that $\hat{S}_\rho(t)$ reduces to the usual Kaplan-Meier estimator. This model is a special case of the nonparametric assumption presented by Slud and Rubinstein. The focus here is not on the dependence of T and C but on the dependence of T and C_s where C_s is a dependent censoring indicator, in particular a surrogate endpoint. The function of interest is

$$\rho_s(t) = \lim_{\delta \rightarrow 0} \frac{\Pr(t < T < t + \delta | T > t, C_s < t)}{\Pr(t < T < t + \delta | T > t, C_s \geq t)}.$$

This function is equivalent to the function $\pi(t)$ and can be estimated from data from the first stage. Suppose that the conditional hazard, $\lambda(t|z)$, of death at t given $Z = z$ has the form $h_0(t)\exp(\beta z(t))$ where $z_i(t)$ takes the value 0 if $t_i \leq c_i$ and value 1 if $t_i > c_i$ then $\rho_s(t) = \rho_s = \exp(\beta)$. Thus, an estimate of ρ_s is given by $\exp(\hat{\beta})$. The estimate of β using data from the first stage quantifies the increase in the risk of death occurring after the surrogate variable has been observed. The first stage is viewed as a training set of data to learn about the relationship between the potential surrogate endpoint and the survival time.

The estimator is constructed from the entire sample N ($N = N_1 + N_2$). In the sample of size N , the ordered failure times X_i for which $\delta_i = 1$ are $X_1 \leq \dots \leq X_d$, where d is the number of deaths of patients enrolled either in the first stage or in the second stage. Using the notation that $\epsilon_i = 1$ if $X_i > c_i$ and 0 otherwise, the random variable V defines either the observed survival time or the time to the surrogate variable. For patients in the second stage, let us denote the number of X_i with $\epsilon_i = 1$ and $\delta_i = 0$ between X_j and X_{j+1} by W_j . In this way, W_j denotes the number of individuals from the second stage having a surrogate response between two consecutive failure times X_j and X_{j+1} . Let W'_j denote either the number of individuals censored in the first stage or the number of individuals censored without a surrogate response in the second stage between X_j and X_{j+1} . Clearly, W_j is the number of patients censored with “informative censoring” between two consecutively ordered failure times X_j and X_{j+1} while W'_j is the number of patients censored with “non-informative censoring” between two consecutively ordered failure times X_j and X_{j+1} . Finally n_j is the number of i with $X_i \geq X_j$. The product-limit estimator then becomes

$$\hat{S}(t; \rho_s) = N^{-1} \left\{ n(t) + \sum_{k=0}^{d(t)-1} W_k \prod_{i=k+1}^{d(t)} \frac{n_i - 1}{n_i + \rho_s - 1} + \sum_{k=0}^{d(t)-1} W'_k \prod_{i=k+1}^{d(t)} \frac{n_i - 1}{n_i} \right\}.$$

Notice that when $\rho_s = 1$ (i.e., the occurrence of the surrogate variables has no influence of survival) then $\hat{S}(t; \rho_s)$ is simply the Kaplan-Meier estimator. Considering the two-sample case, for example, a controlled clinical trial where group membership is indicated by a single binary variable Z , then the survival

function, corresponding to $z = 0$, is denoted by $S_0(t)$ and the survival function corresponding to $z = 1$ is denoted by $S_1(t)$. A nonparametric permutation test for testing the null hypothesis $H_0 : S_0 = S_1$ versus $H_1 : S_0 \neq S_1$ can be easily constructed (Flandre and O'Quigley, 1995). The test accounts for the presence of dependent censoring as described and, since we would not want to assume that the effect of the surrogate is the same in both groups, then ρ_{s0} need not be equal to ρ_{s1} . A test can be based on the statistic Y defined by

$$Y(w; \psi) = \int_0^\infty w(t)\psi(\hat{S}_0(t; \hat{\rho}_{s0}), \hat{S}_1(t; \hat{\rho}_{s1}))d\hat{S}_0(t; \hat{\rho}_{s0}), \quad (8.11)$$

where $\psi(a, b)$ is some distance function (metric) between a and b at the point s and $w(s)$ is some positive weighting function, often taken to be 1 although a variance-stabilizing definition such as $w(t)^2 = \hat{S}_0(t; \hat{\rho}_{s0})(1 - \hat{S}_0(t; \hat{\rho}_{s0}))$ can be useful in certain applications. The choice $\psi(a, b) = a - b$ leads to a test with good power against alternatives of stochastic ordering, whereas the choice $\psi(a, b) = |a - b|$ or $\psi(a, b) = (a - b)^2$, for instance, would provide better power against crossing hazard alternatives. Large sample theory is simplified by assuming that the non-informative censoring distributions are identical in both groups and this may require more critical examination in practical examples. Given data we can observe some value $Y = y_0$ from which the significance level can be calculated by randomly permuting the $(0, 1)$ labels corresponding to treatment assignment. For the i th permutation ($i = 1, \dots, n_p$) the test statistic can be calculated, resulting in the value say y_i . Out of the n_p permutations suppose that there are n^+ values of y_i greater than or equal to y_0 and, therefore, $n_p - n^+$ values of y_i less than y_0 . The significance level for a two-sided test is then given by $2 \min(n^+, n_p - n^+)/n_p$. In practice we sample from the set of all permutations so that n_p does not correspond to the total number of possible permutations but, rather, the number actually used, of which some may even be repeated. This is the same idea that is used in bootstrap resampling.

8.7 Classwork and homework

1. For subjects i and j with covariate values Z_i and Z_j , write down the probability that the ratio of the survival time for subject i to the survival time for subject j is greater than $2/3$.
2. Calculate an estimate of the above probability for the Freireich data in which $Z_i = 1$ and $Z_j = 0$. Derive an expression for a 95% confidence interval for this quantity and use it to derive a confidence interval for use with the Freireich data.
3. Referring to the paper of Kent and O'Quigley (1988), consider the approximation suggested in that paper for the coefficient of randomness. Use this

to motivate the quantity $\Pr(T_i > T_j | Z_i, Z_j)$ as a measure of dependence analogous to explained variation.

4. Consider some large dataset in which there exist two prognostic groups. Divide the time scale into m non-overlapping intervals, $a_0 = 0 < a_1 < \dots < a_m = \infty$. Calculate $\Pr(T_i > T_j | Z_i, Z_j, T_j > a_k)$ for all values of k less than m and use this information to make inferences about the impact of group effects through time.
5. Write down the likelihood for the piecewise exponential model in which, between adjacent failure times, the hazard can be any positive value (Breslow, 1972). Find an expression for the cumulative hazard function and use this to obtain an expression for the survivorship function. Although such an estimator can be shown to be consistent (Breslow and Crowley, 1974), explain why the usual large sample likelihood theory would fail to apply.
6. Consider again the two-group case. Suppose we are told that the survival time of a given subject is less than t_0 . We also know that the groups are initially balanced. Derive an expression for the probability that the subject in question belongs to group 1. For the Freireich data, given that a subject has a survival time less than 15 weeks, estimate the probability that the subject belongs to the group receiving placebo.
7. Derive an estimator analogous to Equation 8.4 but based on the Nelson-Aalen estimator for marginal survival. By adding to the marginal cumulative hazard some arbitrary increasing unbounded function of time, obtain a simpler expression for the conditional survival estimate.
8. Use the delta method (Section A.10) to obtain Equation 8.7.
9. Use the results of Andersen and Gill (1982) to conclude that $\hat{\beta} - \beta_0$ is asymptotically uncorrelated with $\hat{S}(t|Z \in H)|_{\beta_0}$.
10. On the basis of a large dataset construct some simple prognostic indices using the most important risk factors. Divide into 3 groups the data based on the prognostic index and calculate the different survival estimates for each subgroup. Comment on the different features of these estimators as observed in this example. How would you investigate more closely the relative benefits of the different estimators?
11. Show that when T_i and T_j have the same covariate values, then $\Pr(T_i > T_j) = 0.5$.
12. For the Freireich data calculate the probability that a randomly chosen subject from the treated group lives longer than a randomly chosen subject from the control group.

13. Generate a two-variate proportional hazards model in which Z_1 is binary and Z_2 is uniform on the interval $(0, 1)$. The regression coefficients β_1 and β_2 take values (i) 0.5, 0.5; (ii) 0.5, 1.0; and (iii) 1, 2, respectively. For all three situations, and taking $\lambda_0 = 1$, calculate and compare the six survival curves for $Z_1 = 0, 1$ and $Z_2 = 0.25, 0.50$, and 0.75.

8.8 Outline of proofs

Theorem 8.2 We know how to estimate the asymptotic variance of $\hat{\beta}$ under the model. So all that remains for the second term on the right-hand side of (8.7) is to calculate the partial derivative of $\hat{S}(t|Z \in H)$ with respect to β . For this we have:

$$\frac{\partial}{\partial \beta} \hat{S}(t|Z \in H) = \frac{(\sum_{t_i > t} D_i)(\sum_{t_i \leq t} C_i) - (\sum_{t_i \leq t} D_i)(\sum_{t_i > t} C_i + B_1)}{(\sum_{i=1}^k C_i + B_1)^2}, \quad (8.12)$$

where C_i and B_1 are the same as above and

$$D_i = \frac{\partial}{\partial \beta} \hat{P}(Z \in H|t_i) W(t_i),$$

with

$$\frac{\partial}{\partial \beta} \hat{P}(Z \in H|t) = \hat{P}(Z \in H|t) \{E_\beta(Z|t; \pi^H) - E_\beta(Z|t; \pi)\},$$

where

$$E_\beta(Z|t; \pi) = \sum Y_j(t) Z_j \exp\{\hat{\beta} Z_j\} / \sum Y_j(t) \exp\{\hat{\beta} Z_j\},$$

and

$$E_\beta(Z|t; \pi^H) = \sum_H Y_j(t) Z_j \exp\{\hat{\beta} Z_j\} / \sum_H Y_j(t) \exp\{\hat{\beta} Z_j\}.$$

The first term on the right-hand side of (8.7) can be estimated using Greenwood's formula

$$\hat{\text{Var}}\{\hat{S}(t|Z \in H)|_{\beta_0}\} \approx \hat{S}(t|Z \in H)|_{\beta_0}^2 \left\{ \prod_{t_i \leq t} \left(1 + \frac{\hat{q}_i}{n_i \hat{p}_i}\right) - 1 \right\}, \quad (8.13)$$

where

$$\hat{p}_i = \hat{P}(T > t_i | T > t_{i-1}, Z \in H) = \frac{\sum_{j=i+1}^k C_j + B_1}{\sum_i^k C_j + B_1},$$

$\hat{q}_i = 1 - \hat{p}_i$ and $n_i = \sum Y_j(t_i)$. Then each \hat{p}_i is a binomial probability based on a sample of size n_i and $\hat{S}(t|Z \in H) = \prod_{t_i \leq t} \hat{p}_i$. The \hat{p}_i 's may be treated as conditionally independent given the n_i 's, with β_0 fixed. Thus, Greenwood's formula applies. All the quantities involved in (8.12) and (8.13) are those routinely calculated in a Cox model analysis.

Theorem 8.3 Xu (1996) derived the asymptotic normality of $\hat{S}(t|Z \in H)$ under the proportional hazards model at fixed points $t = t^*$. Let's take $Q_i = 1$ if $Z_i \in H$, and 0 otherwise and follow, fairly closely, the notational set-up of Andersen and Gill (1982) and Gill from which

$$\begin{aligned} S^{(r)}(t) &= n^{-1} \sum_{i=1}^n Y_i(t) e^{\beta(t)' Z_i(t)} Z_i(t)^r, & s^{(r)}(t) &= E S^{(r)}(t), \\ S^{(r)}(\beta, t) &= n^{-1} \sum_{i=1}^n Y_i(t) e^{\beta' Z_i(t)} Z_i(t)^r, & s^{(r)}(\beta, t) &= E S^{(r)}(\beta, t), \\ S^{(H)}(\beta, t) &= n^{-1} \sum_1^n Q_i Y_i(t) e^{\beta Z_i}, & s^{(H)}(\beta, t) &= E S^{(H)}(\beta, t), \\ S^{(H1)}(\beta, t) &= n^{-1} \sum_1^n Q_i Y_i(t) Z_i e^{\beta Z_i}, & s^{(H1)}(\beta, t) &= E S^{(H1)}(\beta, t), \end{aligned}$$

for $r = 0, 1, 2, .$. Next rewrite

$$\hat{P}(Z \in H|t) = S^{(H)}(\hat{\beta}, t)/S^{(0)}(\hat{\beta}, t), \quad E_\beta(Z|t; \pi^H) = S^{(H1)}(\hat{\beta}, t)/S^{(H)}(\hat{\beta}, t).$$

Using the main theorem of Section 7.5 we have $s^{(H)}(\beta_0, t)/s^{(0)}(\beta_0, t) = P(Z \in H|t)$. Under the usual regularity and continuity conditions (Xu, 1996) it can be shown that $\{\partial \hat{S}(t^*|Z \in H)/\partial \beta\}|_{\beta=\check{\beta}}$ is asymptotically constant. Now $\hat{\beta} - \beta_0 = I^{-1}(\check{\beta})U(\beta_0)$ where $\check{\beta}$ is on the line segment between $\hat{\beta}$ and β_0 , $U(\beta) = \partial \log L(\beta)/\partial \beta$ and $I(\beta) = -\partial U(\beta)/\partial \beta$. Combining these we have:

$$\sqrt{n}\hat{S}(t^*|Z \in H) = \sqrt{n}\hat{S}(t^*|Z \in H)|_{\beta_0} + I^{-1}(\check{\beta})\sqrt{n}U(\beta_0)\frac{\partial \hat{S}(t^*|Z \in H)}{\partial \beta}|_{\beta=\check{\beta}}.$$

Andersen and Gill (1982) show that $I(\check{\beta})$ converges in probability to a well-defined population parameter. In the following theorem, Lin and Wei (1989) showed that $U(\beta_0)$ is asymptotically equivalent to $1/n$ times a sum of i.i.d. random variables:

Theorem 8.4. (Lin and Wei, 1989) $\sqrt{n}U(\beta_0)$ is asymptotically equivalent to $n^{-1/2} \times \sum_1^n \omega_i(\beta_0)$, where $N_i(t) = I\{T_i \leq t, T_i \leq C_i\}$ and

$$\omega_i(\beta) = \int_0^1 \left\{ Z_i - \frac{s^{(1)}(\beta, t)}{s^{(0)}(\beta, t)} \right\} dN_i(t) - \int_0^1 Y_i(t) e^{\beta Z_i} \left\{ Z_i - \frac{s^{(1)}(\beta, t)}{s^{(0)}(\beta, t)} \right\} \lambda_0(t) dt.$$

It only then remains to show the asymptotic normality of $\hat{S}(t^*|Z \in H)$. We see that the numerator of $\hat{S}(t^*|Z \in H)|_{\beta_0}$ is asymptotically equivalent to $1/n$ times a sum of n i.i.d. random variables like the above, since the denominator of it we know is consistent for $P(Z \in H)$. We drop the subscript of β_0 in $\hat{S}(t^*|Z \in H)|_{\beta_0}$. The numerator of $\hat{S}(t^*|Z \in H)$ is $\int_{t^*}^\infty \hat{P}(Z \in H|t)d\hat{F}(t)$. Note that

$$\begin{aligned} &\sqrt{n} \left\{ \int_{t^*}^\infty \hat{P}(Z \in H|t)d\hat{F}(t) - P(Z \in H, T > t^*) \right\} \\ &= \sqrt{n} \int_{t^*}^\infty P(Z \in H|t)d\{\hat{F}(t) - F(t)\} \end{aligned}$$

$$\begin{aligned}
& + \sqrt{n} \int_{t^*}^{\infty} \{ \hat{P}(Z \in H|t) - P(Z \in H|t) \} d\{\hat{F}(t) - F(t)\} \\
& + \sqrt{n} \int_{t^*}^{\infty} \{ \hat{P}(Z \in H|t) - P(Z \in H|t) \} dF(t).
\end{aligned}$$

Now $\sqrt{n}\{\hat{F}(t) - F(t)\}$ converges in distribution to a zero-mean Gaussian process. Therefore the second term on the right-hand side of the preceding equation is $o_p(1)$. The last term is $A_1 + o_p(1)$ (see also Lemma II.4 of Xu (1996), where

$$\begin{aligned}
A_1 &= \sqrt{n} \int_{t^*}^1 \left\{ \frac{S^{(H)}(\beta_0, t)}{s^{(0)}(\beta_0, t)} - \frac{s^{(H)}(\beta_0, t)S^{(0)}(\beta_0, t)}{s^{(0)}(\beta_0, t)^2} \right\} dF(t) \\
&= n^{-1/2} \sum_{i=1}^n \int_{t^*}^1 \frac{Y_i(t)e^{\beta_0 Z_i}}{s^{(0)}(\beta_0, t)} \left\{ Q_i - \frac{s^{(H)}(\beta_0, t)}{s^{(0)}(\beta_0, t)} \right\} dF(t).
\end{aligned}$$

As for the first term on the right-hand side of the equation preceding the one immediately above, we use Theorem II.5 of Xu (1996). With $\phi(t) = 1_{[t^*, 1]}(t)P(Z \in H|t)$ in the theorem, the first term in this equation is equal to

$$n^{-1/2} \sum_{i=1}^n \nu_i + \sqrt{n}R_n,$$

where $|R_n| = o_p(n^{-1/2})$ and ν 's are i.i.d. with mean zero, each being a function of X_i and δ_i . Thus the proof is complete.



Chapter 9

Regression effect process

9.1 Chapter summary

In this chapter we describe the regression effect process. This can be established in different ways and provides all of the essential information that we need in order to gain an impression of departures from some null structure, the most common null structure corresponding to an absence of regression effect. Departures in specific directions enable us to make inferences on model assumptions and can suggest, of themselves, richer more plausible models. The regression effect process, in its basic form, is much like a scatterplot for linear regression telling us, before any formal statistical analysis, whether the dependent variable really does seem to depend on the explanatory variable as well as the nature, linear or more complex, of that relationship. Our setting is semi-parametric and the information on the time variable is summarized by its rank within the time observations. We make use of a particular time transformation and see that a great body of known theory becomes available to us immediately. An important objective is to glean what we can from a graphical presentation of the regression effect process. The two chapters following this one—building test statistics for particular and general situations, and robust, effective model-building—lean heavily on the results of this chapter.

9.2 Context and motivation

O'Quigley (2003) developed a process based on the regression effect of a proportional or a non-proportional hazards model. In that paper it was shown how an appropriate transformation of the time scale allowed us to visualize this process in relation to standard processes such as Brownian motion and the Brownian bridge, two processes that arise as limiting results for the regression effect process

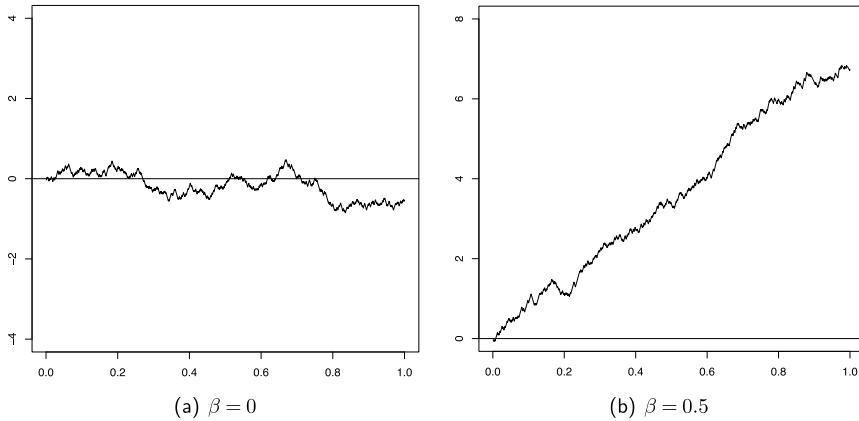


Figure 9.1: The regression effect process as a function of time for simulated data under the proportional hazards model for different values of β .

under certain conditions. O'Quigley (2003) and O'Quigley (2008) indicate how we can move back to a regression effect process on the original scale if needed. While conceptually the multivariate situation is not really a greater challenge than the univariate case, the number of possibilities increases greatly. For instance, two binary covariates give rise to several processes; two marginal processes, two conditional processes, and a process based on the combined effect as expressed through the prognostic index. In addition we can construct processes for stratified models. All of these processes have known limiting distributions when all model assumptions are met. However, under departures from working assumptions, we are faced with the problem of deciding which processes provide the most insight into the collection of regression effects, taken as a whole or looked at individually (Figure 9.1).

In the presentation of this chapter we follow the outline given by O'Quigley (2003) and O'Quigley (2008). The key results that enable us to appeal to the form of Brownian motion and the Brownian bridge are proven in O'Quigley (2003), O'Quigley (2008), Chauvel and O'Quigley (2014, 2017), and Chauvel (2014). The processes of interest to us are based on cumulative quantities having the flavor of weighted residuals, i.e., the discrepancy between observations and some mean effect determined by the model. The model in turn can be seen as a mechanism that provides a prediction and this prediction can then be indexed by one or more unknown parameters. Different large sample results depend on what values we choose to use for these parameters. They may be values fixed under an assumption given by a null hypothesis or they may be some or all of the values provided by estimates from our data. In this latter case the process will tell us something about goodness of fit (O'Quigley, 2003). In the former case the process will tell us about different regression functions that can be considered to be compatible with the observations.

9.3 Elements of the regression effect process

In the most basic case of a single binary covariate, the process developed below in Section 9.4 differs from that of Wei (1984) in three very simple ways: the sequential standardization of the cumulative process, the use of a transformed time scale, and the direct interpolation that leads to the needed continuity. These three features allow for a great deal of development, both on a theoretical and practical level.

Taking the regression parameter to be β , then at time t , the process of Wei is given by

$$U(\beta, t) = \sum_{i=1}^n \int_0^t \{Z_i(s) - \mathcal{E}_\beta(Z | s)\} dN_i(s), \quad 0 \leq t \leq \mathcal{T}. \quad (9.1)$$

The process corresponds to a sum which can be broken down into a series of sequential increments, each increment being the discrepancy between the observation and the expected value of this observation, conditional upon time s and the history up to that point. It is assumed that the model generates the observations. Following the last usable observation (by usable we mean that the conditional variance of the covariate in the risk set is strictly greater than zero), the resulting quantity is the same as the score obtained by the first derivative of the log likelihood before time transformation. For this reason the process is also referred to as the score process. In view of the immediate generalization of the score process and the use we will make of this process we prefer to refer to it as the regression effect process.

The point of view via empirical processes has been adopted by other authors. Arjas (1988) developed a process that is similar to that considered by Wei. Barlow and Prentice (1988) developed a class of residuals by conditioning in such a way—not unlike our own approach—that the residuals can be treated as martingales. The term martingale residuals is used. In essence any weight that we can calculate at time t , using only information available immediately prior to that time, can be used without compromising the martingale property. Such quantities can be calculated for each individual and subsequently summed. The result is a weighted process that contains the usual score process as a particular case (Lin et al., 1993; Therneau and Grambsch, 2000). The constructions of all of these processes take place within the context of the proportional hazards model. Extending the ideas to non-proportional hazards does not appear to be very easy and we do not investigate it outside of our own construction based on sequential standardization and a transformation of the time scale. Wei (1984) was interested in model fit and showed how a global, rather than a sequential standardization, could still result in a large sample result based on a Brownian bridge.

Sequential standardization and time transformation lead to great simplification, to immediate extensions to the multivariate case, and to covariate dependent censoring. The simplicity of the resulting graphics is also compelling and, finally, the simple theorems enable us to use these results in practice with confidence. Although more challenging in our view, other authors have had success in extending the results of Wei (1984). Haara (1987) was able to find analogous results in the case of non-binary covariables while Therneau et al. (1990) considered a global standardization leading to a process that would look like a Brownian bridge under the condition that the covariables were uncorrelated. This would not typically be a realistic assumption. Given the very limited number of situations, under global standardization, in which we can appeal to large sample results looking like Brownian motion or the Brownian bridge, Lin et al. (1993) adopted a different approach based on martingale simulation. This is more general and while the resulting process has no form that can be anticipated we can appeal to computer power to generate a large number of simulated processes under the appropriate hypothesis. Tracing out all of these processes, over the points of evaluation, provides us with envelopes that are indexed by their percent coverage properties. If the observed process lies well within the envelope, we may take that as evidence that any violation of the model assumptions is not too great. “Well within” is not easily defined so that, again, it seems preferable to have procedures based on more analytic results. Our own preference is for sequential rather than global standardization and time transformation since, as we see more fully below, this avoids having to deal with a lot of messy looking figures. More importantly, it provides an immediate impression as to directions in which to look if we have doubts about the underlying assumption of proportionality of hazards. Lin et al. (1996) extended this envelope approach for situations of covariate dependent censoring.

In order to give ourselves a foretaste as to where this is all pointing, consider a simple example taken from the breast cancer study at the Curie Institute in Paris. One covariable of interest was age at diagnosis categorized into two groups, the hypothesis being that the younger group may harbor more aggressive tumors. The results are illustrated in Figure 9.2. The leftmost figure shows the Kaplan-Meier curves for the two groups with some suggestion that there may be a difference. The middle figure gives the regression effect process in which the arrival point of the process is close to the value -2.0 , indicating a result that would be considered to be borderline significant at 5%. Of equal importance is the form of the regression effect process which, as we will show in this chapter, will look like Brownian motion with linear drift under a non-null effect of a proportional hazards nature. In Appendix B we see how this transforms to a Brownian bridge, the 95% limits of the maximum which are shown for reference in the rightmost figure. The bridge process lies well within those limits and shows no obvious sign of departure from a Brownian bridge assumption allowing us to conclude that there is no compelling reason to not accept a working assumption of proportional

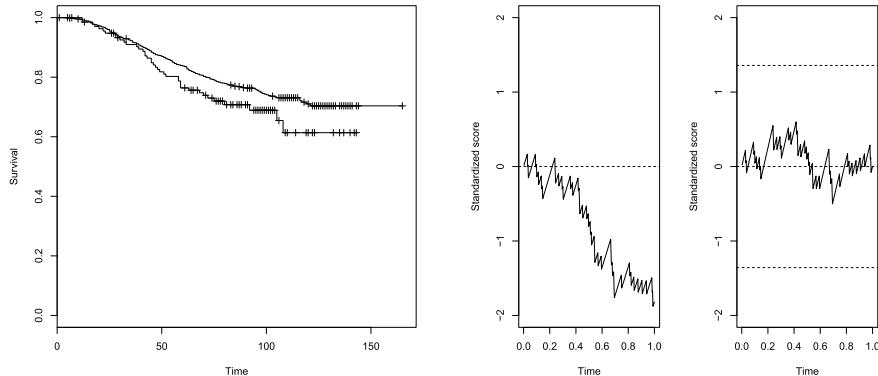


Figure 9.2: Kaplan-Meier curves for the two age groups from the Curie Institute breast cancer study. The regression effect process $U^*(t)$ approximates that of a linear drift while the transformation $U^*(t) - tU^*(1)$ of the rightmost figure appears well approximated by a Brownian bridge. This indicates a satisfactory fit.

hazards. All of this can be quantified and framed within more formal inferential structures. The key point here is that before any formal inference takes place, the eyeball factor—a quick glance at the Kaplan-Meier curves, the regression effect process, and its transform—tells us a lot.

The empirical process we construct in the following sections allows us to perform hypothesis tests on the value of the parameter and study the adequacy of the non-proportional hazards model without having to estimate the regression coefficient. In addition, increments of the process are standardized at each time point, rather than adopting an overall standardization for all increments as in Wei (1984). This makes it possible to take into account multiple and correlated covariates because, unlike the non-standardized score process (9.1), the correlation between covariates is not assumed to be constant over time. We plot the process according to ranked times of failure, in the manner of Arjas (1988), to obtain a simple and explicit asymptotic distribution. Before constructing the process, we need to change the time scale. This transformation is crucial for obtaining asymptotic results for the process without having to use the theorem of Rebolledo (1980), nor the inequality of Lenglart (1977), which are classical tools in survival analysis. In the following section, we present the time transformation used. In Section 9.4, we define the univariate regression effect process and study its asymptotic behavior. In Section 9.5, we look at the multivariate case. Section 9.8 gathers together the proofs of the theorems and lemmas that are presented.

TRANSFORMING THE TIME SCALE

In the proportional hazards model, a strictly monotonically increasing transformation in time does not affect inference on the regression coefficient β . Indeed, the

actual times of death and censored values are not used for inference; only their order of occurrence counts. Among all possible monotonic increasing transformations we choose one of particular value. The order of observed times is preserved and, by applying the inverse transformation, we can go back to the initial time scale. Depending on how we carry out inference, there can be a minor, although somewhat sticky, technical difficulty, in that the number of distinct, and usable, failure times, depends on the data and is not known in advance. Typically we treat this number as though it were known and, indeed, our advice is to do just that in practice. This amounts to conditioning on the number of distinct and usable failure times. We use the word “usable” to indicate that information on regression effects is contained in the observations at these time points. In the case of 2 groups for example, once one group has been extinguished, then, all failure times greater than the greatest for this group carry no information on regression effects.

Definition 9.1. *With this in mind we define the effective sample size as*

$$k_n = \sum_{i=1}^n \delta_i \times \Delta(X_i) \text{ where } \Delta(t) = \mathbf{1}_{\mathcal{V}_{\beta(t)}(Z|t) > 0}$$

If the variance is zero at a time of death, this means that the set of individuals at risk is composed of a homogeneous population sharing the same value of the covariate. Clearly the outcome provides no information on differential rates as quantified by the regression parameter β . This translates mathematically as zero contribution of these times of death to inference, since the value of the covariate of the individual who dies is equal to its expectation.

We want to avoid such variances from a technical point of view, because later we will want to normalize by the square root of the variance. For example, in the case of a comparison of two treatment groups, a null conditional variance corresponds to the situation in which one group is now empty but individuals from the other group are still present in the study. Intuitively, we understand that these times of death do not contribute to estimation of the regression coefficient β since no information is available to compare the two groups. Note that the nullity of a conditional variance at a given time implies nullity at later time points. Thus, according to the definition of k_n , the conditional variances $\mathcal{V}_{\beta(t)}(Z|t)$ calculated at the first k_n times of death t are strictly greater than zero. For a continuous covariate when the last time point observed is a death, we have the equality (algebraic if conditioning, almost sure otherwise) that $k_n = \sum_{i=1}^n \delta_i - 1$. In effect, the conditional variance $\mathcal{V}_{\beta(t)}(Z|t)$ is zero at the last observed time point t , and almost surely is the only time at which the set of at-risk individuals shares the same value of the covariate. If, for a continuous covariate, the last observed time point is censored, then we have that $k_n = \sum_{i=1}^n \delta_i$.

Definition 9.2. Set $\phi_n(0) = 0$ and consider the following transformation ϕ_n of the observed times $X_i, i = 1, \dots, n$:

$$\phi_n(X_i) = \frac{1}{k_n} \left[\bar{N}(X_i) + (1 - \delta_i) \frac{\#\{j : j = 1, \dots, n, X_j < X_i, \bar{N}(X_j) = \bar{N}(X_i)\}}{\#\{j : j = 1, \dots, n, \bar{N}(X_j) = \bar{N}(X_i)\}} \right]$$

We define the inverse ϕ_n^{-1} of ϕ_n on $[0, 1]$ by

$$\phi_n^{-1}(t) = \inf \{X_i, \phi_n(X_i) \geq t, i = 1, \dots, n\}, \quad 0 \leq t \leq 1.$$

Recall that the counting process $\{\bar{N}(t)\}_{t \in [0, T]}$ has a jump of size 1 at each time of death. Thus, on the new time scale representing the image of the observed times $\{X_1, \dots, X_n\}$ under ϕ_n , the values in the set $\{1/k_n, 2/k_n, \dots, 1\}$ correspond to times of death, and the i th time of death t_i is such that $t_i = i/k_n, i = 1, \dots, k_n$ where $t_0 = 0$. The set $\{1/k_n, 2/k_n, \dots, 1\}$ is included in but is not necessarily equal to the transformed set of times of death.

INFERENCE BASED ON TRANSFORMED AND ORIGINAL TIME SCALE

Inference for the regression coefficient is unaffected by monotonic increasing transformations on the time scale and, in particular, the transformation of Definition 9.2. The nature of effects will be very visible on the transformed time scale and less so on the original scale. Nonetheless we can use the transform, and its inverse, to move back and forward between scales. The transformation, $\phi_n(X_i)$, is not independent of the censoring and so some caution is needed in making statements that we wish to hold regardless of the censoring. A regression effect process with linear drift implies and is implied by proportional hazards. Constancy of effect on one scale will then imply constancy of effect on the other. We can say more. Suppose we have a changepoint model. The ratio of the early regression coefficient to the later regression coefficient will be the same on either scale.

Working with the first k_n of the transformed failures, we will therefore restrict ourselves to values less than or equal to 1. As the importance of the transformation applied to the observed times is to preserve order, various transformations—treating censoring times differently—could be used. We choose to uniformly distribute censoring times between adjacent times of death. The time τ_0 on this new scale corresponds to the $(100 \times \tau_0)$ th percentile for deaths in the sample. For example, at time $\tau_0 = 0.5$, half of the deaths have been observed. At time $\tau_0 = 0.2$, 20% of the deaths have occurred. The inverse of ϕ_n can be obtained easily and with it we can interpret the results on the initial time scale.

CONDITIONAL MEANS AND VARIANCES

Our next task is that of sequential standardization. We now work with the standardized time scale in $[0, 1]$. With this time scale, for $t \in [0, 1]$, we define the hazard index $Y_i^*(t)$ and individual counting process $N_i^*(t)$ for individuals $i = 1, \dots, n$ by

$$Y_i^*(t) = \mathbf{1}_{\phi_n(X_i) \geq t}, \quad N_i^*(t) = \mathbf{1}_{\phi_n(X_i) \leq t, \delta_i = 1}.$$

Each time of death t_i is a $(\mathcal{F}_t^*)_{t \in [0,1]}$ -stopping time where, for $t \in [0, 1]$, the σ -algebra \mathcal{F}_t^* is defined by

$$\mathcal{F}_t^* = \sigma \left\{ N_i^*(u), Y_i^*(u^+), Z_i; i = 1, \dots, n; u = 0, 1, \dots, \lfloor tk_n \rfloor \right\},$$

where $\lfloor \cdot \rfloor$ is the floor function and $Y_i^*(t^+) = \lim_{s \rightarrow t^+} Y_i^*(s)$. Notice that if $0 \leq s < t \leq 1$, $\mathcal{F}_s^* \subset \mathcal{F}_t^*$.

Remark. Denote $a \wedge b = \min(a, b)$, for $a, b \in \mathbb{R}$. If the covariates are time-dependent, for $t \in [0, 1]$, the σ -algebra \mathcal{F}_t^* is defined by

$$\mathcal{F}_t^* = \sigma \left\{ N_i^*(u), Y_i^*(u^+), Z_i \left(\phi_n^{-1} \left(\frac{u}{k_n} \right)^+ \wedge X_i \right); i = 1, \dots, n; u = 0, 1, \dots, \lfloor tk_n \rfloor \right\}$$

since, as we recall, the covariate $Z_i(\cdot)$ is not observed after time X_i , $i = 1, \dots, n$. To simplify notation in the following, we will write $Z_i(\phi_n^{-1}(t))$ in the place of $Z_i(\phi_n^{-1}(t)^+ \wedge X_i)$, $t \in [0, 1]$. We define the counting process associated with transformed times, with jumps of size 1 at times of death in the new scale, by

$$\bar{N}^*(t) = \sum_{i=1}^n \mathbf{1}_{\phi_n(X_i) \leq t, \delta_i = 1}, \quad 0 \leq t \leq 1.$$

This counting process satisfies the following result:

Proposition 9.1. (Chauvel 2014). $\forall t \in [0, 1]$, $\bar{N}^*(t) = \lfloor k_n t \rfloor$.

In other words, $t_0 = 0$ and $\bar{N}^*(t)$ has jumps at times $t_j = j/k_n$ for $j = 1, \dots, k_n$. This leads to the following lemma, which will be of use in proving Theorem 9.2:

Lemma 9.1. (Chauvel and O'Quigley 2014). *Let $t_0 = 0$, $n \in \mathbb{N}$ and let $\{A_n(t), t \in \{t_1, \dots, t_{k_n}\}\}$ be a process with values in \mathbb{R} . Suppose that*

$$\sup_{i=1, \dots, k_n} |A_n(t_i) - a(t_i)| \xrightarrow[n \rightarrow \infty]{P} 0,$$

where a is a function defined and bounded on $[0, 1]$. Then,

$$\sup_{s \in [0,1]} \left| \frac{1}{k_n} \int_0^s A_n(t) d\bar{N}^*(t) - \int_0^s a(t) dt \right| \xrightarrow[n \rightarrow \infty]{P} 0.$$

The proofs are given at the end of the chapter. To ease notation, let us define a process \mathcal{Z} which is 0 everywhere except at times of death. At each such time, the process takes the value of the covariate of the individual who fails at that time.

Definition 9.3. We denote $\mathcal{Z} = \{\mathcal{Z}(t), t \in [0,1]\}$ a process such that

$$\mathcal{Z}(t) = \sum_{i=1}^n Z_i(X_i) \mathbf{1}_{\phi_n(X_i) = t, \delta_i = 1}, \quad t \in [0,1]. \quad (9.2)$$

The family of probabilities $\{\pi_i(\beta(t), t), i = 1, \dots, n\}$, with $t \in [0, T]$ can be extended to a non-proportional hazards model on the transformed scale as follows.

Definition 9.4. For $i = 1, \dots, n$ and $t \in [0, 1]$, the probability that individual i dies at time t under the non-proportional hazards model with parameter $\beta(t)$, conditional on the set of individuals at risk at time of death t , is defined by

$$\pi_i(\beta(t), t) = \frac{Y_i^*(t) \exp(\beta(t)^T Z_i(\phi_n^{-1}(t))) \mathbf{1}_{k_n t \in \mathbb{N}}}{\sum_{j=1}^n Y_j^*(t) \exp(\beta(t)^T Z_j(\phi_n^{-1}(t)))}.$$

Remark. The only values at which these quantities will be calculated are the times of death t_1, \dots, t_{k_n} . We nevertheless define $\mathcal{Z}(t)$ and $\{\pi_i(\beta(t), t), i = 1, \dots, n\}$ for all $t \in [0, 1]$ in order to be able to write their respective sums as integrals with respect to the counting process \bar{N}^* . The values 0 taken by them at times other than that of death are chosen arbitrarily as long as they remain bounded by adjacent failures. Note also that at $t_0 = 0$, we have $\mathcal{Z}(0) = 0$. Expectations and variances with respect to the family of probabilities $\{\pi_i(\beta(t), t), i = 1, \dots, n\}$ can now be defined.

Definition 9.5. Let $t \in [0, 1]$. The expectation and variance of Z with respect to the family of probabilities $\{\pi_i(\beta(t), t), i = 1, \dots, n\}$ are respectively a vector in \mathbb{R}^p and a matrix in $\mathcal{M}_{p \times p}(\mathbb{R})$ such that

$$\mathcal{E}_{\beta(t)}(Z | t) = \sum_{i=1}^n Z_i(\phi_n^{-1}(t)) \pi_i(\beta(t), t), \quad (9.3)$$

and

$$\mathcal{V}_{\beta(t)}(Z | t) = \sum_{i=1}^n Z_i(\phi_n^{-1}(t))^{\otimes 2} \pi_i(\beta(t), t) - \mathcal{E}_{\beta(t)}(Z | t)^{\otimes 2}. \quad (9.4)$$

It can be shown that the Jacobian matrix of $\mathcal{E}_{\beta(t)}(Z | t)$ is $\frac{\partial}{\partial \beta} \mathcal{E}_{\beta(t)}(Z | t) = \mathcal{V}_{\beta(t)}(Z | t)$, for $t \in [0, 1]$.

Proposition 9.2. Let $t \in [0, 1]$. Under the non-proportional hazards model with parameter $\beta(t)$, the expectation $E_{\beta(t)}(\mathcal{Z}(t) | \mathcal{F}_{t-}^*)$ and variance $V_{\beta(t)}(\mathcal{Z}(t) | \mathcal{F}_{t-}^*)$ of the random variable $\mathcal{Z}(t)$ given the σ -algebra \mathcal{F}_{t-}^* are

$$E_{\beta(t)}(\mathcal{Z}(t) | \mathcal{F}_{t-}^*) = \mathcal{E}_{\beta(t)}(Z | t), \quad V_{\beta(t)}(\mathcal{Z}(t) | \mathcal{F}_{t-}^*) = \mathcal{V}_{\beta(t)}(Z | t).$$

In conclusion, for $j = 1, \dots, k_n$ and time of death t_j , by conditioning on the set of at-risk individuals, the expectation of $\mathcal{Z}(t_j)$ is $\mathcal{E}_{\beta(t_j)}(Z | t_j)$ and its variance-covariance matrix is $\mathcal{V}_{\beta(t_j)}(Z | t_j)$. Knowledge of these moments allows us to define the regression effect process, which we will study separately for the univariate ($p = 1$) and multivariate ($p > 1$) cases. It is worth adding that the univariate process has an interest well beyond the case of simple models involving only a single covariate. For multivariate models there are several univariate processes that will reveal much of what is taking place. These include the process for the prognostic index, processes for individual covariates after having adjusted via the model for the effects of some or all of the other covariates, processes for stratified models, and indeed any process that can be derived from the model.

9.4 Univariate regression effect process

In this section, we will focus more closely on the univariate framework ($p = 1$), i.e., when there is only one covariate $Z(t)$ in the model. In some sense higher dimensions represent an immediate generalization of this but, for higher dimensions, in view of the complexity theorem (Theorem 4.1) there are many more things to consider, both conceptually and operationally. In order to gain the firmest grip possible on this we look closely at the univariate case which contains all of the key ideas. It is always a good principle in statistics to condition on as many things as we can. We can go too far and condition on aspects of the data that are not uninformative with regard to our unknown regression parameter. This would certainly lead to error so care is always called for. In some ways conditioning is an art, knowing just how far to go. One thing that can guide us is to what extent, after such conditioning, we can recover procedures that have been established under quite different principles. Here, we will see for instance that a lot of conditioning will result in the well-known log-rank test which, at the very least, gives us some confidence.

Sequential conditioning through time, i.e., conditioning on the times of death and the sets of individuals at risk, as first proposed by Cox (1972), leads to a very natural framework for inference. An individual i who dies at t_j is viewed as having been selected randomly (under the model) from the set

$\{i : Y_i^*(t_j) = 1; i = 1, \dots, n\}$ of individuals at risk of death at time t_j with probability $\pi_i(\beta(t_j), t_j)$. Under this conditioning, at the time of death t_j , the information necessary to calculate the expectation and variance of $\mathcal{Z}(t_j)$ with respect to $\{\pi_i(\beta(t_j), t_j); i = 1, \dots, n\}$ is available for $j = 1, \dots, k_n$. The conditional variances $\mathcal{V}_{\beta(t)}(Z | t)$ calculated for the first k_n times of death t are strictly greater than zero. Thus,

$$\forall t \in \{t_1, \dots, t_{k_n}\}, \mathcal{V}_{\beta(t)}(Z | t) > 0 \quad \text{a.s.}$$

The variables $\mathcal{Z}(t_j)$ may be standardized. By sequentially summing all of these standardized variables, we obtain a standardized score which is the basis of our regression effect process. A further conditioning argument whereby the number of discrete failure times at which the variance $\mathcal{V}_{\beta(t)}(Z | t)$ is strictly greater than zero leads, once more, to great simplification. For completeness the unconditioned case is worked through at the end of the chapter. As argued above, the most useful approach is a conditional one in which we take k_n to be equal to its observed value.

Definition 9.6. Let $j = 0, 1, \dots, k_n$. The regression effect process evaluated for the parameter $\beta(t_j)$ at time t_j is defined by $U_n^*(\cdot, \cdot, 0) = 0$ and

$$U_n^*(\alpha(t_j), \beta(t_j), t_j) = \frac{1}{\sqrt{k_n}} \int_0^{t_j} \mathcal{V}_{\alpha(s)}(Z | s)^{-1/2} \{ \mathcal{Z}(s) - \mathcal{E}_{\beta(s)}(Z | s) \} d\bar{N}^*(s).$$

where, by using two parameter functions, $\alpha(t)$ and $\beta(t)$, we allow ourselves increased flexibility. The $\alpha(t)$ concerns the variance and allowing this to not be tied to $\beta(t)$ may be of help in some cases where we would like estimates of the variance to be valid both under the null and the alternative. This would be similar to what takes place in an analysis of variance where the residual variance is estimated in such a way as to remain consistent both under the null and the alternative. Our study on this, in this particular context, is very limited, and this could be something of an open problem. For the remainder of this text we will suppose that $\alpha(t) = \beta(t)$ and we will consequently write $U_n^*(\beta(t_j), t_j)$ as only having two arguments.

Remark. For computational purposes, and for $j = 1, \dots, k_n$, the variable $U_n^*(\beta(t_j), t_j)$ can also be written as the sum:

$$U_n^*(\beta(t_j), t_j) = \frac{1}{\sqrt{k_n}} \sum_{i=1}^j \mathcal{V}_{\beta(t_i)}(Z | t_i)^{-1/2} \{ \mathcal{Z}(t_i) - \mathcal{E}_{\beta(t_i)}(Z | t_i) \}, \quad (9.5)$$

noting that all of the ingredients in the above expression are obtained routinely from all of the currently available software packages.

Remark. Let w be a deterministic or random $(\mathcal{F}_t^*)_{t \in [0,1]}$ -predictable function from $[0, 1]$ to \mathbb{R} . According to Proposition 9.2, under the non-proportional hazards model with parameter $\beta(t)$, conditional on the set of individuals at risk at time of death t_j , the expectation of $w(t_j)\mathcal{Z}(t_j)$ is $w(t_j)\mathcal{E}_{\beta(t_j)}(Z | t_j)$ and its variance $w(t_j)^2\mathcal{V}_{\beta(t_j)}(Z | t_j)$ with respect to the family of probabilities $\{\pi_i(\beta(t_j), t_j); i = 1, \dots, n\}$. By summing the standardized weighted variables, we obtain the same definition of the process as in Definition 9.6. We will come back to this remark in Section 11.4 for the weighted log-rank test.

We can now define the process U_n^* on $[0, 1]$ by linearly interpolating the $k_n + 1$ random variables $\{U_n^*(\beta(t_j), t_j), j = 0, 1, \dots, k_n\}$.

Definition 9.7. The standardized score process evaluated at $\beta(t)$ is defined by $\{U_n^*(\beta(t), t), t \in [0, 1]\}$, where for $j = 0, \dots, k_n$ and $t \in [t_j, t_{j+1}]$,

$$U_n^*(\beta(t), t) = U_n^*(\beta(t_j), t_j) + (tk_n - j) \{U_n^*(\beta(t_{j+1}), t_{j+1}) - U_n^*(\beta(t_j), t_j)\}.$$

By definition, the process $U_n^*(\beta(\cdot), \cdot)$ is continuous on the interval $[0, 1]$. The process depends on two parameters: the time t and regression coefficient $\beta(t)$. For the sake of clarity, we recall that the temporal function $\beta : t \rightarrow \beta(t)$ is denoted $\beta(t)$, which will make it easier to distinguish the proportional hazards model with parameter β from the non-proportional hazards model with parameter $\beta(t)$. Under the non-proportional hazards model with $\beta(t)$, the increments of the process U_n^* are centered with variance 1. Moreover, the increments are uncorrelated, as shown in the following proposition.

Proposition 9.3. (Cox 1975). Let $D_n^*(\beta(t_j), t_j) = U_n^*(\beta(t_j), t_j) - U_n^*(\beta(t_{j-1}), t_{j-1})$. Under the non-proportional hazards model with parameter $\beta(t)$, the random variables $D_n^*(\beta(t_j), t_j)$ for $j = 1, 2, \dots, k_n$ are uncorrelated.

The uncorrelated property of the increments is also used in the calculation of log-rank statistic. All of these properties together make it possible to obtain our essential convergence results for the process $U_n^*(\beta(t), t)$.

WORKING ASSUMPTIONS FOR LARGE SAMPLE THEORY

The space $(D[0, 1], \mathbb{R})$ comes equipped with uniform convergence. For $0 \leq t \leq 1$ and for $r = 0, 1, 2$, we can define:

$$S^{(r)}\{\beta(t), t\} = \frac{1}{n} \sum_{i=1}^n Y_i^*(t) Z_i (\phi_n^{-1}(t))^r \exp(\beta(t) Z_i (\phi_n^{-1}(t))).$$

Note that for $t \in \{t_1, t_2, \dots, t_{k_n}\}$,

$$\mathcal{E}_{\beta(t)}(Z | t) = \left. \frac{\partial}{\partial b} \log \left\{ S^{(0)}(b, t) \right\} \right|_{b=\beta(t)} = \frac{S^{(1)}(\beta(t), t)}{S^{(0)}(\beta(t), t)},$$

and

$$\mathcal{V}_{\beta(t)}(Z | t) = \frac{\partial^2}{\partial b^2} \log \left\{ S^{(0)}(b, t) \right\} \Big|_{b=\beta(t)} = \frac{S^{(2)}(\beta(t), t)}{S^{(0)}(\beta(t), t)} - \left(\frac{S^{(1)}(\beta(t), t)}{S^{(0)}(\beta(t), t)} \right)^2.$$

Suppose now the following conditions:

A1 (Asymptotic stability). There exists some $\delta_1 > 0$, a neighborhood $\mathbb{B} = \left\{ \gamma, \sup_{t \in [0, 1]} |\gamma(t) - \beta(t)| < \delta_1 \right\}$ of β of radius δ_1 containing the zero function, and functions $s^{(r)}$ defined on $\mathbb{B} \times [0, 1]$ for $r = 0, 1, 2$, such that

$$\sqrt{n} \sup_{t \in [0, 1], \gamma \in \mathbb{B}} \left| S^{(r)}(\gamma(t), t) - s^{(r)}(\gamma(t), t) \right| \xrightarrow[n \rightarrow \infty]{P} 0. \quad (9.6)$$

A2 (Asymptotic regularity). The deterministic functions $s^{(r)}$, $r = 0, 1, 2$, defined in A1 are uniformly continuous for $t \in [0, 1]$ and bounded in $\mathbb{B} \times [0, 1]$. Furthermore, for $r = 0, 1, 2$, and $t \in [0, 1]$, $s^{(r)}(\cdot, t)$ is a continuous function in \mathbb{B} . The function $s^{(0)}$ is bounded below by a strictly positive constant.

By analogy with the empirical quantities, for $t \in [0, 1]$ and $\gamma \in \mathbb{B}$, denote

$$e(\gamma(t), t) = \frac{s^{(1)}(\gamma(t), t)}{s^{(0)}(\gamma(t), t)}, \quad v(\gamma(t), t) = \frac{s^{(2)}(\gamma(t), t)}{s^{(0)}(\gamma(t), t)} - e(\gamma(t), t)^2. \quad (9.7)$$

A3 (Homoscedasticity). For any $t \in [0, 1]$ and $\gamma \in \mathbb{B}$, we have $\frac{\partial}{\partial t} v(\gamma(t), t) = 0$.

A4 (Uniformly bounded covariates). There exists $L \in \mathbb{R}^{*+}$ such that

$$\sup_{i=1, \dots, n} \sup_{u \in [0, T]} |Z_i(u)| \leq L.$$

The first two conditions are standard in survival analysis. They were introduced by Andersen and Gill (1982) in order to use theory from counting processes and martingales such as the inequality of Lenglart (1977) and the theorem of Rebollo (1980). The variance $\mathcal{V}_{\beta(t)}(Z | t)$ is, by definition, an estimator of the variance of Z given $T = t$ under the non-proportional hazards model with parameter $\beta(t)$. Thus, the homoscedasticity condition A3 means that the asymptotic variance is independent of time for parameters close to the true regression coefficient $\beta(t)$.

This condition is often implicitly encountered in the use of the proportional hazards model. This is the case, for example, in the estimation of the variance of the parameter β , or in the expression for the log-rank statistic, where the contribution to the overall variance is the same at each time of death. Indeed, the overall variance is an unweighted sum of the conditional variances. The

temporal stability of the variance has been pointed out by several authors, notably Grambsch and Therneau (1994), Xu (1996), and Xu and O'Quigley (2000). Next, we need two lemmas. Proofs are given at the end of the chapter.

Lemma 9.2. *Under conditions A1 and A3, for all $t \in [0, 1]$ and $\gamma \in \mathbb{B}$, $v(\gamma(t), t) > 0$.*

Lemma 9.3. *Under hypotheses, A1, A2, and A3, and if $k_n = k_n(\beta_0)$, for all $\gamma_1 \in \mathbb{B}$, there exist constants $C(\gamma_1)$ et $C(\beta_0) \in \mathbb{R}^{*+}$ such that*

$$\sqrt{n} \sup_{t \in \{t_1, t_2, \dots, t_{k_n}\}} |\mathcal{V}_{\gamma_1(t)}(Z | t) - C(\gamma_1)| \xrightarrow[n \rightarrow \infty]{P} 0, \quad (9.8)$$

$$\sqrt{n} \sup_{t \in \{t_1, t_2, \dots, t_{k_n}\}} \left| \frac{\mathcal{V}_{\gamma_1(t)}(Z | t)}{\mathcal{V}_{\beta_0}(Z | t)} - \frac{C(\gamma_1)}{C(\beta_0)} \right| \xrightarrow[n \rightarrow \infty]{P} 0, \quad (9.9)$$

$$\sqrt{n} \sup_{t \in \{t_1, t_2, \dots, t_{k_n}\}} \left| \frac{\mathcal{V}_{\gamma_1(t)}(Z | t)}{\mathcal{V}_{\beta_0}(Z | t)^{1/2}} - \frac{C(\gamma_1)}{C(\beta_0)^{1/2}} \right| \xrightarrow[n \rightarrow \infty]{P} 0. \quad (9.10)$$

Theorem 9.1. *Under the non-proportional hazards model with parameter $\beta(t)$,*

$$U_n^*(\beta(\cdot), \cdot) \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} \mathcal{W}(t),$$

where $\mathcal{W}(t)$ is a standard Brownian motion.

The proof, given at the end of the chapter, is built on the use of the functional central limit theorem for martingale differences introduced by Helland (1982). This theorem can be seen as an extension of the functional central limit theorem of Donsker (1951) to standardized though not necessarily independent nor identically distributed variables. Figure 9.3(a) shows the result of simulating a standardized score process $U_n^*(0, t)$ as a function of time t , when the true parameter is constant. The immediate visual impression would lead us to see Brownian motion—standard or with linear drift—as a good approximation. We will study how to calculate this approximation on a given sample in Chapters 10 and 11.

Theorem 9.2. (Chauvel and O'Quigley 2014). *Let β_0 be a constant function on \mathbb{B} . Under the non-proportional hazards model with parameter $\beta(t)$, consider*

$$A_n(t) = \frac{1}{k_n} \int_0^t \{\beta(s) - \beta_0\} \frac{\mathcal{V}_{\tilde{\beta}(s)}(Z | s)}{\mathcal{V}_{\beta_0}(Z | s)^{1/2}} d\bar{N}^*(s), \quad 0 \leq t \leq 1, \quad (9.11)$$

where $\tilde{\beta}$ is in the ball with center β and radius $\sup_{t \in [0, 1]} |\beta(t) - \beta_0|$. Then, there exist two constants $C_1(\beta, \beta_0)$ and $C_2 \in \mathbb{R}^{*+}$ such that $C_1(\beta, \beta_0) = 1$ and

$$U_n^*(\beta_0, \cdot) - \sqrt{k_n} A_n \xrightarrow[n \rightarrow \infty]{\mathcal{L}} C_1(\beta, \beta_0) \mathcal{W}(t). \quad (9.12)$$

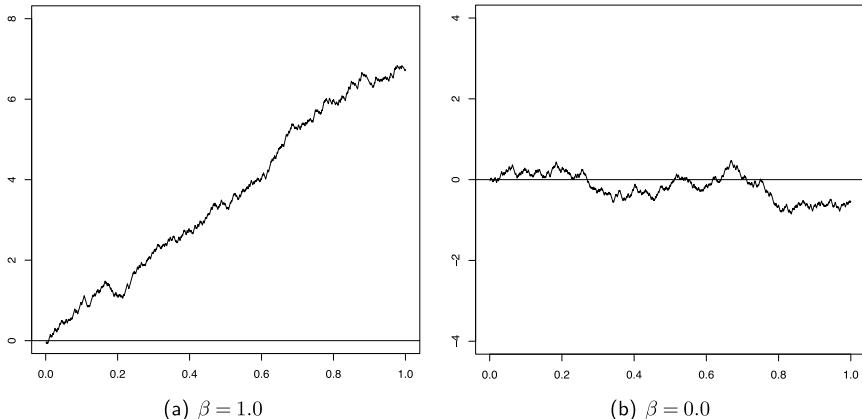


Figure 9.3: The process $U_n^*(0, \cdot)$ as a function of time for simulated data under the proportional hazards model. For left-hand figure, $\beta = 1.0$. For right-hand figure, the absence of effects manifests itself as approximate standard Brownian motion.

Furthermore,

$$\sup_{0 \leq t \leq 1} \left| A_n(t) - C_2 \int_0^t \{\beta(s) - \beta_0\} ds \right| \xrightarrow[n \rightarrow \infty]{P} 0. \quad (9.13)$$

Remark. Theorem 9.1 is a special case of Theorem 9.2 when $\beta(t) = \beta_0$. In effect, in Theorem 9.2, when $\beta(t) = \beta_0$, the constant $C_1(\beta, \beta)$ is equal to 1, the drift is zero, and the limit process of $U_n^*(\beta_0, \cdot)$ is a standard Brownian motion. The proof of the theorem can be found in Section 9.8. The drift of the standardized score process at time t tends to infinity when $k_n \rightarrow \infty$, according to equation (9.12). In practice, we work with a large enough fixed k_n and $\beta_0 = 0$. The theorem indicates that the drift of the process $U_n^*(0, \cdot)$ is proportional to $\beta(s)$. We will distinguish between two cases in which the regression coefficient $\beta(t)$ is non-zero. In the first, consider the proportional hazards model, that is, $\beta(t)$ is constant over time. From Theorem 9.2, the process $U_n^*(0, \cdot)$ can be approximated by a Brownian motion with a linear drift. This is illustrated in Figure 9.3(a) with a process that has been simulated under the proportional hazards model with non-zero β . The drift can be approximated by a straight line, indicating that the regression coefficient is constant in time. The coefficient β , which has the same sign as the slope of the line according to Theorem 9.2, is positive.

The second case corresponds to the situation where there is a non-constant regression coefficient over time. Several scenarios are possible. For example, the effect could be constant for an initial period and then quickly decline beyond

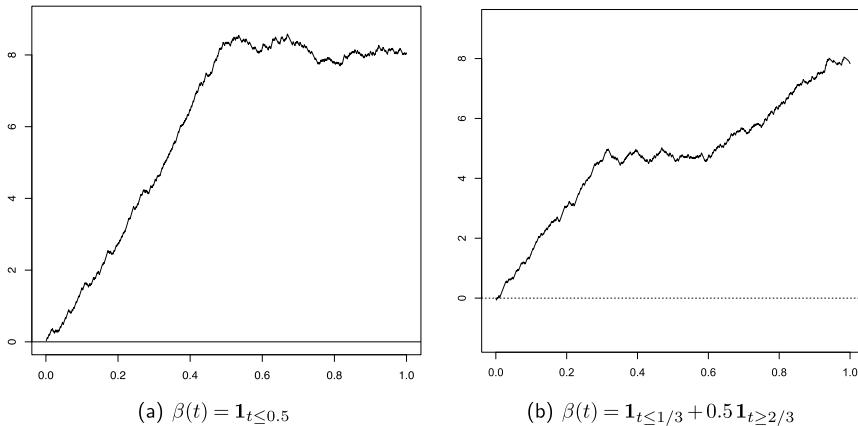


Figure 9.4: The process $U_n^*(0, \cdot)$ as a function of time for simulated data under the non-proportional hazards model with different values of $\beta(t)$. Temporal effects are clearly reflected in the process on the transformed time scale.

some point τ , piecewise constant over time, increase or decrease continuously, and so on. In all these situations, the cumulative effect $\int \beta(s)ds$ will be seen in the drift of the process (Theorem 9.2) and will, of itself, describe to us the nature of the true effects. For example, Figure 9.4(a) shows a simulated process with a piecewise-constant effect $\beta(t) = \mathbf{1}_{t \leq 0.5}$. Before $t = 0.5$, we see a positive linear trend corresponding to $\beta = 1$, and for $t > 0.5$, the coefficient becomes zero and the drift of the process $U_n^*(0, \cdot)$ is parallel to the time axis. Figure 9.4(b) shows the standardized score process for simulated data under failure time model with regression function, $\beta(t) = \mathbf{1}_{t \leq 1/3} + 0.5 \mathbf{1}_{t \geq 2/3}$. The drift of the process can be separated into three linear sections reflecting the effect's strength, each of which can be approximated by a straight line. The slope of the first line appears to be twice that of the last, while the slope of the second line is zero.

In summary, whether the effect corresponds to proportional or non-proportional hazards, all information about the regression coefficient $\beta(t)$ can be found in the process $U_n^*(0, \cdot)$. We can then use this process to test the value of the regression coefficient and evaluate the adequacy of the proportional hazards model. A simple glance at the regression effect process is often all that is needed to let us know that there exists an effect and the nature of that effect. More precise inference can then follow. Before delving into these aspects in Chapters 10 and 11, we consider more closely the standardized score process for the multivariate case. As mentioned earlier the univariate process, even in the multivariate setting, will enable us to answer most of the relevant questions, such as the behavior of many variables combined in a prognostic index, or, for example, the impact of some variable after having controlled for the effects of other variables, whether via the model or via stratification.

9.5 Regression effect processes for several covariates

Here we extend the regression effect process of the previous section to the case of several covariates grouped in a vector $Z(t)$ of dimension $p > 1$. Suppose that we have the non-proportional hazards model with regression function $\beta(t)$ and a vector of covariates $Z(t)$ which are functions from $[0, 1]$ to \mathbb{R}^p . Following Chauvel (2014), let $t \in [0, 1]$ and

$$\begin{aligned} S^{(0)}(\beta(t), t) &= n^{-1} \sum_{i=1}^n Y_i^*(t) \exp\left(\beta(t)^T Z_i(\phi_n^{-1}(t))\right) \in \mathbb{R}, \\ S^{(1)}(\beta(t), t) &= n^{-1} \sum_{i=1}^n Y_i^*(t) Z_i(\phi_n^{-1}(t)) \exp\left(\beta(t)^T Z_i(\phi_n^{-1}(t))\right) \in \mathbb{R}^p, \\ S^{(2)}(\beta(t), t) &= n^{-1} \sum_{i=1}^n Y_i^*(t) Z_i(\phi_n^{-1}(t))^{\otimes 2} \exp\left(\beta(t)^T Z_i(\phi_n^{-1}(t))\right) \in \mathcal{M}_{p \times p}(\mathbb{R}). \end{aligned}$$

For $t \in \{t_1, \dots, t_{k_n}\}$, the conditional expectation $\mathcal{E}_{\beta(t)}(Z | t)$ and the conditional variance-covariance matrix $\mathcal{V}_{\beta(t)}(Z | t)$ defined in (9.3) and (9.4) can be written as functions of $S^{(r)}(\beta(t), t)$ ($r = 0, 1, 2$):

$$\mathcal{E}_{\beta(t)}(Z | t) = \frac{S^{(1)}(\beta(t), t)}{S^{(0)}(\beta(t), t)}, \quad \mathcal{V}_{\beta(t)}(Z | t) = \frac{S^{(2)}(\beta(t), t)}{S^{(0)}(\beta(t), t)} - \mathcal{E}_{\beta(t)}(Z | t)^{\otimes 2}.$$

which is the same as $\partial \mathcal{E}_{\beta(t)}(Z | t) / \partial \beta$. As the matrix $\mathcal{V}_{\beta(t)}(Z | t)$ is symmetric and positive definite, there exists an orthogonal matrix $P_{\beta(t)}(t)$ and a diagonal matrix $D_{\beta(t)}(t)$ such that

$$\mathcal{V}_{\beta(t)}(Z | t) = P_{\beta(t)}(t) D_{\beta(t)}(t) P_{\beta(t)}(t)^T.$$

This leads us to define the symmetric matrix $\mathcal{V}_{\beta(t)}(Z | t)^{-1/2}$ as

$$\mathcal{V}_{\beta(t)}(Z | t)^{-1/2} = P_{\beta(t)}(t) D_{\beta(t)}(t)^{-1/2} P_{\beta(t)}(t)^T.$$

By the definition of $\widehat{k}_n = \widehat{k}_n(\beta)$ in the multivariate case, and since $k_n \leq \widehat{k}_n$ almost surely (Equation (9.23)), the conditional variance-covariance matrices $\mathcal{V}_{\beta(t)}(Z | t)$ calculated for the first k_n times of death t are positive definite, and

$$\forall t \in \{t_1, \dots, t_{k_n}\}, \quad \|\mathcal{V}_{\beta(t)}(Z | t)^{-1}\| \leq C_{\mathcal{V}}^{-1} \quad \text{a.s.}$$

Definition 9.8. Let $j = 0, 1, \dots, k_n$. The multivariate standardized score calculated at time t_j for parameter $\beta(t_j)$ is defined by

$$U_n^*(\beta(t_j), t_j) = \frac{1}{\sqrt{k_n}} \int_0^{t_j} \mathcal{V}_{\beta(s)}(Z | s)^{-1/2} \{Z(s) - \mathcal{E}_{\beta(s)}(Z | s)\} d\bar{N}^*(s),$$

where

$$\int_0^{t_j} a(s) d\bar{N}^*(s) = \left(\int_0^{t_j} a_1(s) d\bar{N}^*(s), \dots, \int_0^{t_j} a_p(s) d\bar{N}^*(s) \right),$$

for $a = (a_1, \dots, a_p)$, where the a_i are functions from $[0, 1]$ to \mathbb{R} .

Remark. As in the univariate case, the random variable $U_n^*(\beta(t_j), t_j)$ can be written as a sum:

$$U_n^*(\beta(t_j), t_j) = \frac{1}{\sqrt{k_n}} \sum_{i=1}^j \mathcal{V}_{\beta(t_i)}(Z | t_i)^{-1/2} \{ \mathcal{Z}(t_i) - \mathcal{E}_{\beta(t_i)}(Z | t_i) \}.$$

The standardized score is a random function defined at times t_1, \dots, t_{k_n} taking values in \mathbb{R}^p , whose definition can be extended to $[0, 1]$ by linear interpolation.

Definition 9.9. The multivariate standardized score process $\{U_n^*(\beta(t), t), t \in [0, 1]\}$ evaluated at $\beta(t)$ is such that, for $j = 0, 1, \dots, k_n$ and $t \in [t_j, t_{j+1}]$,

$$U_n^*(\beta(t), t) = U_n^*(\beta(t_j), t_j) + (tk_n - j) \{ U_n^*(\beta(t_{j+1}), t_{j+1}) - U_n^*(\beta(t_j), t_j) \}.$$

Remark. We can readily construct several univariate processes on the basis of the multivariate model. Suppose for example that we have 2 binary covariates Z_1 and Z_2 . If we stratify on the basis of Z_1 then we would have a single regression effect processes for Z_2 that can be studied, as in the simple case. The same is true if we stratify on the basis of Z_2 and study Z_1 . If, instead of stratifying, we fit both covariates into the model, then, we can consider processes for each variable separately. We could also consider conditional processes, for example where we look at Z_2 and fix Z_1 , leading to two conditional processes. And, of course, the prognostic index would define a joint process where both variables are included in a linear combination. Definition 9.9 is interesting because it allows us to take into account the degree of association between the covariates themselves as well as how they relate, individually and collectively, to survival outcome.

WORKING ASSUMPTIONS IN THE MULTIVARIATE CASE

Recall that the norms of vectors in \mathbb{R}^p and matrices in $\mathcal{M}_{p \times p}(\mathbb{R})$ are their sup norms, as defined in Appendix A. The function space $(D[0, 1], \mathbb{R}^p)$ is equipped with the Skorokhod product topology. To establish the limit behavior of the process, we refer to Chauvel and O'Quigley (2014, 2017), and suppose the following conditions to hold:

B1 (Asymptotic stability). There exists $\delta_2 > 0$, a bounded neighborhood \mathbb{B}' of β of radius δ_2 containing the zero function: $\mathbb{B}' = \left\{ \gamma, \sup_{t \in [0,1]} \|\gamma(t) - \beta(t)\| < \delta_2 \right\}$, and functions $s^{(r)}$ defined on $\mathbb{B}' \times [0,1]$ for $r = 0, 1$, and 2, respectively, taking values in \mathbb{R} , \mathbb{R}^p , and the matrix space $\mathcal{M}_{p \times p}(\mathbb{R})$, such that

$$\sqrt{n} \sup_{t \in [0,1], \gamma \in \mathbb{B}'} \|S^{(r)}(\gamma(t), t) - s^{(r)}(\gamma(t), t)\| \xrightarrow[n \rightarrow \infty]{P} 0.$$

B2 (Asymptotic regularity). The deterministic functions $s^{(r)}$, $r = 0, 1, 2$, defined in **B1** are uniformly continuous for $t \in [0,1]$ and bounded in $\mathbb{B}' \times [0,1]$. Furthermore, for $r = 0, 1, 2$ and $t \in [0,1]$, $s^{(r)}(\cdot, t)$ is a continuous function in \mathbb{B}' . The function $s^{(0)}$ is bounded below by a strictly positive constant.

B3 (Homoscedasticity). There exists a symmetric and positive definite matrix $\Sigma \in \mathcal{M}_{p \times p}(\mathbb{R})$ and a sequence of positive constants $(M_n)_n$ such that $\lim_{n \rightarrow \infty} M_n = 0$ which satisfy for all $\gamma \in \mathbb{B}'$, $t \in [0,1]$ and $i, j = 1, \dots, p$,

$$\left\| \frac{\partial^2}{\partial \beta_i \partial \beta_j} \mathcal{E}_\beta(Z | t) \Big|_{\beta=\gamma(t)} \right\| \leq M_n \quad a.s., \quad (9.14)$$

$$\sqrt{n} \sup_{t \in \{t_1, t_2, \dots, t_{k_n}\}, \gamma \in \mathbb{B}'} \left\| \mathcal{V}_{\gamma(t)}(Z | t)^{1/2} - \Sigma^{1/2} \right\| \xrightarrow[n \rightarrow \infty]{P} 0. \quad (9.15)$$

B4 (Uniformly bounded covariates). There exists some $L' \in \mathbb{R}^{*+}$ such that

$$\sup_{i=1, \dots, n} \sup_{u \in [0, T]} \sup_{j=1, \dots, p} \left| Z_i^{(j)}(u) \right| \leq L'.$$

These conditions are extensions of hypotheses **A1**, **A2**, **A3**, and **A4** from Section 9.4 to the multiple covariates case. We make the additional assumption that the conditional variance-covariance matrices do not depend on the parameter when it is close to the true regression parameter. This is a technical assumption required due to the use of the multidimensional Taylor-Lagrange inequality in the proofs. In the univariate case, we do not need this hypothesis thanks to the availability of a Taylor-Lagrange equality.

Proposition 9.4. (Chauvel 2014). *Under hypotheses **B1**, **B3**, and **B4**, we have the following:*

$$\sqrt{n} \sup_{t \in \{t_1, t_2, \dots, t_{k_n}\}, \gamma \in \mathbb{B}'} \left\| \mathcal{V}_{\gamma(t)}(Z | t) - \Sigma \right\| \xrightarrow[n \rightarrow \infty]{P} 0. \quad (9.16)$$

The proposition is proved at the end of the chapter. The following theorem gives the limit behavior of the univariate regression effect process. Let $a = (a_1, \dots, a_p)$ be a function from \mathbb{R}^p to $[0, 1]$. Denote

$$\int_0^t a(s)ds = \left(\int_0^t a_1(s)ds, \dots, \int_0^t a_p(s)ds \right), \quad 0 \leq t \leq 1.$$

Suppose $\gamma \in \mathbb{B}'$. Let $U_n^*(\gamma, \cdot)$ be the multivariate regression effect process calculated at γ . To construct this process, we consider $k_n = k_n(\gamma)$, which can be estimated by $\widehat{k}_n(\gamma)$. Recall that $k_n(\gamma) \leq \widehat{k}_n(\gamma)$ almost surely (Equation 9.23) and by the definition of $\widehat{k}_n(\gamma)$ in the multivariate case, we have:

$$\forall t \in \{t_1, \dots, t_{k_n}\}, \quad \|\mathcal{V}_{\gamma(t)}(Z | t)^{-1}\| \leq C_{\mathcal{V}}^{-1} \quad \text{a.s.} \quad (9.17)$$

Theorem 9.3. (Chauvel and O'Quigley 2014). *Let β_0 be a constant function on \mathbb{B}' . Under the non-proportional hazards model with regression coefficient $\beta(t)$, if the multivariate regression effect process is calculated at parameter β_0 , we have the convergence in distribution result:*

$$U_n^*(\beta_0, \cdot) - \sqrt{k_n} B_n \xrightarrow[n \rightarrow \infty]{\mathcal{L}} \mathcal{W}_p(t), \quad (9.18)$$

where $\mathcal{W}_p(t)$ is a standard p -dimensional Brownian motion, and for each $t \in [0, 1]$,

$$B_n(t) = \frac{1}{k_n} \sum_{i=1}^{\lfloor tk_n \rfloor} \mathcal{V}_{\beta_0}(Z | t_i)^{-1/2} \{ \mathcal{E}_{\beta(t_i)}(Z | t_i) - \mathcal{E}_{\beta_0}(Z | t_i) \}.$$

Furthermore,

$$\sup_{t \in [0, 1]} \left\| B_n(t) - \Sigma^{1/2} \int_0^t \{ \beta(s) - \beta_0 \} ds \right\| \xrightarrow[n \rightarrow \infty]{P} 0. \quad (9.19)$$

We recall that a p -dimensional Brownian motion is a vector of p independent Brownian motions, each in \mathbb{R} .

Remark. The constant $C_1(\beta, \beta_0)$ in Theorem 9.2 in which the convergence of the univariate process U_n^* is established has no multivariate equivalent. This is a consequence of Equation 9.14 in hypothesis B3, which is used in the proof in a multidimensional Taylor-Lagrange inequality, but is not needed in the univariate case. The constant C_2 in Theorem 9.2 is equivalent to the matrix $\Sigma^{1/2}$. In practice, for a large enough fixed value of n , the multivariate standardized score process $U_n^*(0, \cdot)$ can thus be approximated by a Brownian motion whose drift depends on the true parameter $\beta(t)$ of the model. Each

component of the drift $\sqrt{k_n} \Sigma^{1/2} \int_0^t \beta(s) ds$ represents a linear combination of the cumulative effects $\int_0^t \beta_1(s) ds, \dots, \int_0^t \beta_p(s) ds$, except when the variance-covariance matrix Σ is diagonal. Under the theorem's hypotheses, a direct consequence of Equation 9.16 is

$$\hat{\Sigma} := \frac{1}{k_n} \sum_{i=1}^{k_n} \mathcal{V}_{\beta_0}(Z | t_i) \xrightarrow[n \rightarrow \infty]{P} \Sigma, \quad (9.20)$$

which means that $\hat{\Sigma}$ is an estimator that converges to Σ . The following corollary allows us to decorrelate the components of the process U_n^* so that the drift of each component represents the corresponding effect (in $\beta_1(t), \dots, \beta_p(t)$) separately.

Corollary 9.1. (Chauvel 2014). *Under the conditions of Theorem 9.3,*

$$\hat{\Sigma}^{-1/2} U_n^*(\beta_0, \cdot) - \sqrt{k_n} \hat{\Sigma}^{-1/2} B_n \xrightarrow[n \rightarrow \infty]{\mathcal{L}} \Sigma^{-1/2} \mathcal{W}_p,$$

and

$$\sup_{t \in [0, 1]} \left\| \hat{\Sigma}^{-1/2} B_n(t) - \int_0^t \{\beta(s) - \beta_0\} ds \right\| \xrightarrow[n \rightarrow \infty]{P} 0.$$

Thus, when there are several covariates, the drifts of the process $\hat{\Sigma}^{-1/2} U_n^*(\beta_0, \cdot)$ help us to make inference on the regression coefficients. Each drift represents a cumulative coefficient and inference is done separately for each drift, as in the univariate case.

9.6 Iterated logarithm for effective sample size

This section is focused on a technical detail first studied by Chauvel (2014). The issue is thoroughly resolved if we are prepared to condition on the observed effective sample size. That would be our recommendation. If we do not wish to make use of such a simplifying argument, and we wish to appeal to results that are available in the framework of the martingale central limit theorem, then we need to pay some small amount of attention to a technical detail concerning effective sample size. The martingale central limit theorem requires us to make use of the concept of predictability. Our useable number of failure points, k_n , is, technically speaking, not predictable since, until the study is finished or until such a time as the variance becomes zero, we do not know what its value is. Our solution to this is simple, just take the value as fixed and known at its observed value. This amounts to conditioning on a future value, just like obtaining the Brownian bridge (not a martingale) from Brownian motion (a martingale) by conditioning on its final value. But, this very simple solution, while fine in our view, does take the inference away from under the umbrella of the martingale central limit theorem. A general solution is not hard to find but requires a little

attention to some finicky details. We consider this more closely here. Let us suppose that there exists $\alpha_0 \in]0, 1]$ such that

$$\frac{\widehat{k}_n(\beta)}{n} \xrightarrow[n \rightarrow \infty]{\text{a.s.}} \alpha_0.$$

Furthermore, suppose that there exists a sequence $(a_n)_n$ converging to 1 and $\tilde{N} \in \mathbb{N}^*$ such that for each $n \geq \tilde{N}$,

$$a_n \geq 1, \quad \frac{\widehat{k}_n(\beta)}{n} a_n \geq \alpha_0 \quad \text{a.s.} \quad (9.21)$$

This would imply in particular that $\widehat{k}_n(\beta)$ tends to infinity when n tends to infinity, i.e., the number of deaths increases without bound as the sample size increases. If Z is a discrete variable, this means that for a sample of infinite size, no group will run out of individuals. If Z is continuous, $\widehat{k}_n(\beta) = \sum_{i=1}^n \delta_i - 1$ or $\widehat{k}_n(\beta) = \sum_{i=1}^n \delta_i$ almost surely. If so, $n^{-1}\widehat{k}_n(\beta)$ is an estimator that converges to $P(T \leq C)$ and $\alpha_0 = P(T \leq C)$. The law of the iterated logarithm then allows us to construct a sequence (a_n) satisfying Equation 9.21.

Remark. Suppose we have one continuous covariate with censoring at the last observed time point, implying that $\widehat{k}_n(\beta) = \sum_{i=1}^n \delta_i = \sum_{i=1}^n \mathbf{1}_{X_i \leq C_i}$. The variables $\delta_1, \dots, \delta_n$ are independent and identically distributed with finite second order moments. Their expectation is $\alpha_0 = P(T \leq C)$ and variance $\alpha_0(1 - \alpha_0)$. Using the law of the iterated logarithm, we have:

$$\liminf_{n \rightarrow \infty} \frac{\sqrt{n} \left(n^{-1} \widehat{k}_n(\beta) - \alpha_0 \right)}{\sqrt{2\alpha_0(1 - \alpha_0) \log \log(n)}} = -1 \quad \text{a.s.}$$

Thus,

$$\forall \varepsilon > 0, \exists N_\varepsilon \in \mathbb{N}^*, \forall n \geq N_\varepsilon, \quad \frac{\sqrt{n} \left(n^{-1} \widehat{k}_n(\beta) - \alpha_0 \right)}{\sqrt{2\alpha_0(1 - \alpha_0) \log \log(n)}} \geq -1 - \varepsilon \quad \text{a.s.}$$

Setting $\varepsilon = 1/2$, $\exists N_{1/2} \in \mathbb{N}^*$, $\forall n \geq N_{1/2}$,

$$\frac{\sqrt{n} \left(n^{-1} \widehat{k}_n(\beta) - \alpha_0 \right)}{\sqrt{2\alpha_0(1 - \alpha_0) \log \log(n)}} \geq -\frac{3}{2} \quad \text{a.s.}, \quad \sqrt{\frac{\log \log(n)}{n}} \leq \frac{1}{3} \sqrt{\frac{2\alpha_0}{1 - \alpha_0}}. \quad (9.22)$$

Suppose that $n \geq N_{1/2}$. Then, with the help of Equation 9.22,

$$\frac{\widehat{k}_n(\beta)}{n} \geq \alpha_0 - \frac{3}{\sqrt{2n}} \sqrt{\alpha_0(1 - \alpha_0) \log \log(n)},$$

and by setting

$$a_n = \left(1 - \frac{3}{\sqrt{2\alpha_0 n}} \sqrt{(1-\alpha_0) \log \log(n)} \right)^{-1},$$

we find $n^{-1} \widehat{k}_n(\beta) a_n \geq \alpha_0$. From Equation 9.22, $a_n \geq 1$ and we therefore have $\lim_{n \rightarrow \infty} a_n = 1$. We have thus constructed a sequence (a_n) which satisfies Equation 9.21.

The random variable $\widehat{k}_n(\beta)$ is known only at the end of data collection. It cannot be used during the experiment to construct a predictable process (in the sense of an increasing family of σ -algebras as described below). To overcome this we define the deterministic quantity $k_n(\beta) = n a_n^{-1} \alpha_0$, in order to obtain theoretical results. We have the almost sure convergence $\frac{\widehat{k}_n(\beta)}{k_n(\beta)} \xrightarrow{n \rightarrow \infty} 1$, and Equation 9.21 implies that

$$\forall n \geq \tilde{N}, \quad \widehat{k}_n(\beta) \geq k_n(\beta) \quad \text{a.s.} \quad (9.23)$$

9.7 Classwork and homework

1. Simulate 30 i.i.d. uniform variates, X_i , $i = 1, \dots, 30$. Calculate $S_i = \sum_{j=1}^i X_j$. Join, using straight lines, the points S_i and S_{i+1} . We call this a process. Make a graphical plot of the process based on S_i versus i , for $i = 1$ through $i = 30$. Evaluate theoretically, $E(S_{10})$, $E(S_{20})$, and $E(S_{30})$. Next evaluate $\text{Var}(S_{10})$, $\text{Var}(S_{20})$, and $\text{Var}(S_{30})$. Graph the theoretical quantities against their observed counterparts.
2. Repeat the experiment in the above question for $X_i^C = X_i - 0.5$.
3. Repeat the experiment of Question 1, for $X_i^C = (X_i - 0.5)/\sqrt{2.5}$. What can we say here? Where does the number 2.5 come from? If you had to make a guess at $\text{Var}(S_{19.5})$ what value would you choose?
4. Let $\mathcal{W}(t)$ be Brownian motion on the interval $(0, 1)$. Consider the time points t where $t = 1/30, 2/30, \dots, 29/30, 1$. Write down $E\mathcal{W}(t)$ and $\text{Var}\mathcal{W}(t)$ for these points. Compare these values with those calculated above.
5. Calculate $S_i^0 = S_i - i \times S_{30}/30$, based upon X_i^C . Make some general observations on this process.
6. Repeat each set of simulations; $X_i^C = (X_i - 0.5)/\sqrt{2.5}$, 1000 times. For each of these repetitions, note down the empirically observed averages of the means and variances at each value of i for $i = 1, \dots, 30$. Rescale the interval for i to be the interval $(0, 1)$. Plot the average mean and the average variance against $i/30$. What conclusions can be made?

7. Let X_i^C take the value $X_i/\sqrt{2.5}$ for $i=1,\dots,10$; the value $(X_i - 0.25)/\sqrt{2.5}$, for $i=11,\dots, 20$; and the value $(X_i - 0.50)/\sqrt{2.5}$ otherwise. Plot these series of values against i . What conclusions can be drawn? Calculate $S_i^0 = S_i - i \times S_{30}/30$. What can be said about this process?
8. The condition of homoscedasticity is helpful in obtaining several of the large sample approximations. In practice the condition is typically respected to a strong degree. Can you construct a situation in which the assumption may not be a plausible one? What would be the implications of this for inference?

9.8 Outline of proofs

We provide here some detail to the proofs needed to support the main findings of this chapter. Broad outlines for these proofs have been given in Chauvel and O'Quigley (2014, 2017) and further details can be found in Chauvel (2014).

Proposition 9.1 The counting process \bar{N} on the initial time scale takes values in \mathbb{N} . Thus, for all $u \in [0, \mathcal{T}]$,

$$\{x \in \mathbb{R} : \bar{N}(u) \leq x\} = \{x \in \mathbb{R} : \bar{N}(u) \leq \lfloor x \rfloor\}.$$

The inverse process is such that $\bar{N}^{-1}(j)$ is the j th ordered time of death, $j = 1, \dots, k_n$. Note that $\bar{N}(\bar{N}^{-1}(j)) = j$. Thus, by the definition of \bar{N}^* , we have for $0 \leq t \leq 1$,

$$\begin{aligned} \bar{N}^*(t) &= \sum_{i=1}^n \mathbf{1}_{\{\bar{N}(X_i) \leq \lfloor k_n t \rfloor, \delta_i = 1\}} = \sum_{i=1}^n \mathbf{1}_{\{X_i \leq \bar{N}^{-1}(\lfloor k_n t \rfloor), \delta_i = 1\}} \\ &= \bar{N}(\bar{N}^{-1}(\lfloor k_n t \rfloor)) = \lfloor k_n t \rfloor \end{aligned}$$

Lemma 9.1 Suppose that $\varepsilon > 0$. Note that

$$\begin{aligned} &P \left\{ \sup_{0 \leq s \leq 1} \left| \frac{1}{k_n} \int_0^s A_n(t) d\bar{N}^*(t) - \int_0^s a(t) dt \right| > \varepsilon \right\} \\ &\leq P \left\{ \sup_{0 \leq s \leq 1} \frac{1}{k_n} \left| \int_0^s \{A_n(t) - a(t)\} d\bar{N}^*(t) \right| > \frac{\varepsilon}{2} \right\} \\ &\quad + P \left\{ \sup_{0 \leq s \leq 1} \left| \frac{1}{k_n} \int_0^s a(t) d\bar{N}^*(t) - \int_0^s a(t) dt \right| > \frac{\varepsilon}{2} \right\} \end{aligned} \tag{9.24}$$

Suppose also that $s \in [0, 1]$. For the first term we have:

$$\begin{aligned} \frac{1}{k_n} \left| \int_0^s \{A_n(t) - a(t)\} d\bar{N}^*(t) \right| &= \frac{1}{k_n} \left| \sum_{i=1}^{\lfloor k_n s \rfloor} (A_n(t_i) - a(t_i)) \right| \\ &\leq \frac{\lfloor k_n s \rfloor}{k_n} \sup_{i=1, \dots, \lfloor k_n s \rfloor} |A_n(t_i) - a(t_i)| \leq \sup_{i=1, \dots, k_n} |A_n(t_i) - a(t_i)|. \end{aligned}$$

Therefore,

$$\sup_{0 \leq s \leq 1} \frac{1}{k_n} \left| \int_0^s \{A_n(t) - a(t)\} d\bar{N}^*(t) \right| \leq \sup_{i=1, \dots, k_n} |A_n(t_i) - a(t_i)|,$$

and $\sup_{i=1, \dots, k_n} |A_n(t_i) - a(t_i)| \xrightarrow[n \rightarrow \infty]{P} 0$. As a result,

$$\lim_{n \rightarrow \infty} P \left\{ \sup_{0 \leq s \leq 1} \frac{1}{k_n} \left| \int_0^s \{A_n(t) - a(t)\} d\bar{N}^*(t) \right| > \frac{\varepsilon}{2} \right\} = 0.$$

For the second term on the right-hand side of Equation 9.24 we have:

$$\sup_{0 \leq s \leq 1} \left| \frac{1}{k_n} \int_0^s a(t) d\bar{N}^*(t) - \int_0^s a(t) dt \right| = \sup_{0 \leq s \leq 1} \left| \frac{1}{k_n} \sum_{i=1}^{\lfloor k_n s \rfloor} a(t_i) - \int_0^s a(t) dt \right|.$$

Note that a_n is piecewise constant on the interval $[0, 1]$ such that $a_n(t) = a(t_i)$ for $t \in [t_i, t_{i+1}[$, $i = 0, \dots, k_n$. We then have:

$$\frac{1}{k_n} \sum_{i=1}^{\lfloor k_n s \rfloor} a(t_i) = \int_0^s a_n(t) dt - \left(s - \frac{\lfloor k_n s \rfloor}{k_n} \right) a\left(\frac{\lfloor k_n s \rfloor}{k_n}\right).$$

As a result we have:

$$\begin{aligned} &\sup_{0 \leq s \leq 1} \left| \frac{1}{k_n} \int_0^s a(t) d\bar{N}^*(t) - \int_0^s a(t) dt \right| \\ &\leq \sup_{0 \leq s \leq 1} \left| \int_0^s a_n(t) dt - \int_0^s a(t) dt \right| + \sup_{0 \leq s \leq 1} \left| \left(s - \frac{\lfloor k_n s \rfloor}{k_n} \right) a\left(\frac{\lfloor k_n s \rfloor}{k_n}\right) \right| \\ &\leq \sup_{0 \leq s \leq 1} \left| \int_0^s a_n(t) dt - \int_0^s a(t) dt \right| + \frac{1}{k_n} \sup_{0 \leq s \leq 1} \left| a\left(\frac{\lfloor k_n s \rfloor}{k_n}\right) \right|, \end{aligned}$$

Note that $\lim_{n \rightarrow \infty} \sup_{s \in [0, 1]} |a_n(s) - a(s)| = 0$ by the uniform continuity of the function a . This implies uniform convergence

$$\lim_{n \rightarrow \infty} \sup_{s \in [0, 1]} \left| \int_0^s a_n(t) dt - \int_0^s a(t) dt \right| = 0.$$

Furthermore, a is bounded so that $\sup_{0 \leq s \leq 1} \left| a\left(\frac{\lfloor k_n s \rfloor}{k_n}\right) \right| < \infty$, which shows the result since

$$\lim_{n \rightarrow \infty} \sup_{0 \leq s \leq 1} \left| \frac{1}{k_n} \int_0^s a(t) d\bar{N}^*(t) - \int_0^s a(t) dt \right| = 0, \quad (9.25)$$

Proposition 9.2 Let $t \in [0, 1]$. The expectation of $\mathcal{Z}(t)$ (defined in (9.2)) conditional on the σ -algebra $\mathcal{F}_{t^-}^*$ is

$$\begin{aligned} E_{\beta(t)}(\mathcal{Z}(t) | \mathcal{F}_{t^-}^*) &= E_{\beta(t)} \left(\sum_{i=1}^n Z_i(X_i) \mathbf{1}_{\phi_n(X_i) = t, \delta_i = 1} | \mathcal{F}_{t^-}^* \right) \\ &= E_{\beta(t)} \left(\sum_{i=1}^n Z_i(\phi_n^{-1}(t)) \mathbf{1}_{\phi_n(X_i) = t, \delta_i = 1} | \mathcal{F}_{t^-}^* \right) \\ &= \sum_{i=1}^n Z_i(\phi_n^{-1}(t)) E_{\beta(t)} \left(\mathbf{1}_{\phi_n(X_i) = t, \delta_i = 1} | \mathcal{F}_{t^-}^* \right) \\ &= \sum_{i=1}^n Z_i(\phi_n^{-1}(t)) P_{\beta(t)}(\phi_n(X_i) = t, \delta_i = 1 | \mathcal{F}_{t^-}^*). \end{aligned}$$

Indeed, Z and ϕ_n^{-1} are left continuous, so $Z_i(\phi_n^{-1}(t)) = Z_i(\phi_n^{-1}(t^-))$ is $\mathcal{F}_{t^-}^*$ measurable. By definition, the probability that individual i has time of death t under the model with parameter $\beta(t)$, conditional on the set of at-risk individuals at time t , is $P_{\beta(t)}(\phi_n(X_i) = t, \delta_i = 1 | \mathcal{F}_{t^-}^*) = \pi_i(\beta(t), t)$. Thus,

$$E_{\beta(t)}(\mathcal{Z}(t) | \mathcal{F}_{t^-}^*) = \mathcal{E}_{\beta(t)}(Z | t).$$

The same reasoning for the second-order moment gives the variance result.

Proposition 9.3 Let $i, j = 1, 2, \dots, k_n$, $i < j$. By definition, the random variable

$$\mathcal{V}_{\beta(t_i)}(Z | t_i)^{-1/2} (\mathcal{Z}(t_i) - \mathcal{E}_{\beta(t_i)}(Z | t_i))$$

is $\mathcal{F}_{t_i}^*$ measurable. As the σ -algebras satisfy $\mathcal{F}_{t_i}^* \subset \mathcal{F}_{t_j^-}^*$, this random variable is also $\mathcal{F}_{t_j^-}^*$ measurable. Furthermore, from Proposition 9.2,

$$\begin{aligned} &E \left(\mathcal{V}_{\beta(t_j)}(Z | t_j)^{-1/2} (\mathcal{Z}(t_j) - \mathcal{E}_{\beta(t_j)}(Z | t_j)) \middle| \mathcal{F}_{t_j^-}^* \right) \\ &= \mathcal{V}_{\beta(t_j)}(Z | t_j)^{-1/2} E \left(\mathcal{Z}(t_j) - \mathcal{E}_{\beta(t_j)}(Z | t_j) \middle| \mathcal{F}_{t_j^-}^* \right) = 0. \end{aligned}$$

Thus,

$$\begin{aligned} & E \left(\mathcal{V}_{\beta(t_i)}(Z | t_i)^{-1/2} \{ \mathcal{Z}(t_i) - \mathcal{E}_{\beta(t_i)}(Z | t_i) \} \mathcal{V}_j^Z \{ \mathcal{Z}(t_j) - \mathcal{E}_{\beta(t_j)}(Z | t_j) \} \right) \\ &= E \left[E \left(\mathcal{V}_{\beta(t_i)}(Z | t_i)^{-1/2} \{ \mathcal{Z}(t_i) - \mathcal{E}_{\beta(t_i)}(Z | t_i) \} \mathcal{V}_j^Z \{ \mathcal{Z}(t_j) - \mathcal{E}_{\beta(t_j)}(Z | t_j) \} \middle| \mathcal{F}_{t_j^-}^* \right) \right] \\ &= E \left[\mathcal{V}_{\beta(t_i)}(Z | t_i)^{-1/2} \{ \mathcal{Z}(t_i) - \mathcal{E}_{\beta(t_i)}(Z | t_i) \} E \left(\mathcal{V}_j^Z (\mathcal{Z}(t_j) - \mathcal{E}_{\beta(t_j)}(Z | t_j)) \middle| \mathcal{F}_{t_j^-}^* \right) \right] = 0. \end{aligned}$$

where we have written \mathcal{V}_j^Z in place of $\mathcal{V}_{\beta(t_j)}(Z | t_j)^{-1/2}$

Lemma 9.2 Appealing to the law of large numbers, condition A1 implies that for $r = 0, 1, 2$, we have:

$$s^{(r)}(\gamma(0), 0) = E \left(S^{(r)}(\gamma(0), 0) \right) = E \{ Y(0) Z(0)^r \exp(\gamma(0) Z(0)) \},$$

which is $E \{ Z(0)^r \exp(\gamma(0) Z(0)) \}$ since $Y(0) = \mathbf{1}_{X \geq 0} = 1$. As a result,

$$\begin{aligned} v(\gamma(0), 0) &= \frac{E \{ Z(0)^2 \exp(\gamma(0) Z(0)) \}}{E \{ \exp(\gamma(0) Z(0)) \}} - \left(\frac{E \{ Z(0) \exp(\gamma(0) Z(0)) \}}{E \{ \exp(\gamma(0) Z(0)) \}} \right)^2 \\ &= E \{ Z(0)^2 \Phi_{\gamma(0)}(Z(0)) \} - (E \{ Z(0) \Phi_{\gamma(0)}(Z(0)) \})^2, \end{aligned}$$

where

$$\Phi_{\gamma(0)}(z) = \frac{\exp(\gamma(0)z)}{E(\exp(\gamma(0)Z(0)))}, \quad z \in \mathbb{R}.$$

Furthermore, denoting $P_{Z(0)}$ to be the distribution function of the random variable $Z(0)$ and X having distribution function P_X such that $dP_X(x) = \Phi_{\gamma(0)}(x)dP_{Z(0)}(x)$, $x \in \mathbb{R}$, we have:

$$v(\gamma(0), 0) = E(X^2) - E(X)^2 = V(X).$$

The function $\Phi_{\gamma(0)}$ being invertible on \mathbb{R} , X is distributed according to a Dirac law if and only if $Z(0)$ follows a Dirac law. We know that $V(Z(0)) > 0$ which is equivalent to saying that $Z(0)$ does not follow a Dirac law. Thus, X does not follow a Dirac law and $v(\gamma(0), 0) = V(X) > 0$. The hypothesis A3 concerning homoscedasticity indicates that, for all $t \in [0, 1]$, $v(\gamma(t), t) = v(\gamma(0), 0)$, which leads to the conclusion.

Lemma 9.3 Assume that conditions A1 and A2 hold and that $\gamma_1 \in \mathbb{B}$. Recall the definition of $v(\gamma_1(\cdot), \cdot)$. We write \sup_k^n to denote $\sqrt{n} \sup_{t \in \{t_1, t_2, \dots, t_{k_n}\}}$ so that we have:

$$\begin{aligned}
& \sup_k^n |\mathcal{V}_{\gamma_1(t)}(Z | t) - v(\gamma_1(t), t)| \\
&= \sup_k^n \left| \frac{S^{(2)}(\gamma_1(t), t)}{S^{(0)}(\gamma_1(t), t)} - \left(\frac{S^{(1)}(\gamma_1(t), t)}{S^{(0)}(\gamma_1(t), t)} \right)^2 - \left[\frac{s^{(2)}(\gamma_1(t), t)}{s^{(0)}(\gamma_1(t), t)} - \left(\frac{s^{(1)}(\gamma_1(t), t)}{s^{(0)}(\gamma_1(t), t)} \right)^2 \right] \right| \\
&\leq \sqrt{n} \sup_{t \in [0, 1]} \left| \frac{S^{(2)}(\gamma_1(t), t)}{S^{(0)}(\gamma_1(t), t)} - \left(\frac{S^{(1)}(\gamma_1(t), t)}{S^{(0)}(\gamma_1(t), t)} \right)^2 - \left[\frac{s^{(2)}(\gamma_1(t), t)}{s^{(0)}(\gamma_1(t), t)} - \left(\frac{s^{(1)}(\gamma_1(t), t)}{s^{(0)}(\gamma_1(t), t)} \right)^2 \right] \right| \\
&\leq V_n + W_n,
\end{aligned}$$

where

$$V_n = \sqrt{n} \sup_{t \in [0, 1]} \left| \frac{S^{(2)}(\gamma_1(t), t)}{S^{(0)}(\gamma_1(t), t)} - \frac{s^{(2)}(\gamma_1(t), t)}{s^{(0)}(\gamma_1(t), t)} \right|,$$

and

$$W_n = \sqrt{n} \sup_{t \in [0, 1]} \left| \left(\frac{S^{(1)}(\gamma_1(t), t)}{S^{(0)}(\gamma_1(t), t)} \right)^2 - \left(\frac{s^{(1)}(\gamma_1(t), t)}{s^{(0)}(\gamma_1(t), t)} \right)^2 \right|.$$

We can study the large sample behavior of these two terms separately. Denote m_0 , M_1 , and M_2 to be strictly positive constants such that $s^{(0)}(\gamma(t), t) \geq m_0$ et $|s^{(i)}(\gamma(t), t)| \leq M_i$, ($i = 1, 2$) for all $t \in [0, 1]$ and $\gamma \in \mathbb{B}$. Their existence follows from condition A2. We have:

$$\begin{aligned}
V_n &\leq \sqrt{n} \sup_{t \in [0, 1]} \left| \frac{1}{S^{(0)}(\gamma_1(t), t)} (S^{(2)}(\gamma_1(t), t) - s^{(2)}(\gamma_1(t), t)) \right| \\
&\quad + \sqrt{n} \sup_{t \in [0, 1]} \left| s^{(2)}(\gamma_1(t), t) \left(\frac{1}{S^{(0)}(\gamma_1(t), t)} - \frac{1}{s^{(0)}(\gamma_1(t), t)} \right) \right| \\
&\leq \sqrt{n} \sup_{t \in [0, 1]} \frac{1}{S^{(0)}(\gamma_1(t), t)} \sup_{t \in [0, 1]} \left| S^{(2)}(\gamma_1(t), t) - s^{(2)}(\gamma_1(t), t) \right| \\
&\quad + M_2 \sqrt{n} \sup_{t \in [0, 1]} \left| \frac{1}{S^{(0)}(\gamma_1(t), t)} - \frac{1}{s^{(0)}(\gamma_1(t), t)} \right|. \tag{9.26}
\end{aligned}$$

Furthermore,

$$\begin{aligned}
& \sqrt{n} \sup_{t \in [0, 1], \gamma \in \mathbb{B}} \left| \frac{1}{S^{(0)}(\gamma(t), t)} - \frac{1}{s^{(0)}(\gamma(t), t)} \right| \\
&= \sqrt{n} \sup_{t \in [0, 1], \gamma \in \mathbb{B}} \frac{\left| s^{(0)}(\gamma(t), t) - S^{(0)}(\gamma(t), t) \right|}{S^{(0)}(\gamma(t), t) s^{(0)}(\gamma(t), t)} \\
&\leq m_0^{-1} \sqrt{n} \sup_{t \in [0, 1], \gamma \in \mathbb{B}} \left| s^{(0)}(\gamma(t), t) - S^{(0)}(\gamma(t), t) \right| \sup_{t \in [0, 1], \gamma \in \mathbb{B}} \frac{1}{S^{(0)}(\gamma(t), t)}.
\end{aligned}$$

From Equation 9.6, $\sqrt{n} \sup_{t \in [0,1], \gamma \in \mathbb{B}} |s^{(0)}(\gamma(t), t) - S^{(0)}(\gamma(t), t)| = o_P(1)$, and since $s^{(0)}$ is bounded below uniformly by m_0 , we have $\sup_{t \in [0,1], \gamma \in \mathbb{B}} S^{(0)}(\gamma(t), t)^{-1} = O_P(1)$, which indicates that this term is bounded after a certain rank with high probability. As a result,

$$\sqrt{n} \sup_{t \in [0,1], \gamma \in \mathbb{B}} \left| \frac{1}{S^{(0)}(\gamma(t), t)} - \frac{1}{s^{(0)}(\gamma(t), t)} \right| \xrightarrow[n \rightarrow \infty]{P} 0. \quad (9.27)$$

In addition, under condition A1,

$$\sqrt{n} \sup_{t \in [0,1], \gamma \in \mathbb{B}} |S^{(2)}(\gamma(t), t) - s^{(2)}(\gamma(t), t)| \xrightarrow[n \rightarrow \infty]{P} 0,$$

and since $\sup_{t \in [0,1], \gamma \in \mathbb{B}} S^{(0)}(\gamma(t), t)^{-1} = O_P(1)$, we have convergence for

$$\sqrt{n} \sup_{t \in [0,1]} \frac{1}{S^{(0)}(\gamma_1(t), t)} \sup_{t \in [0,1]} |S^{(2)}(\gamma_1(t), t) - s^{(2)}(\gamma_1(t), t)| \xrightarrow[n \rightarrow \infty]{P} 0.$$

Equation 9.26 allows us to conclude that V_n converges in probability to 0 when $n \rightarrow \infty$. Similar arguments show that we can bound the term W_n by

$$\begin{aligned} W_n &\leq \sqrt{n} \sup_{t \in [0,1]} \left| \frac{1}{S^{(0)}(\gamma_1(t), t)^2} (S^{(1)}(\gamma_1(t), t)^2 - s^{(1)}(\gamma_1(t), t)^2) \right| \\ &\quad + \sqrt{n} \sup_{t \in [0,1]} \left| s^{(1)}(\gamma_1(t), t)^2 \left(\frac{1}{S^{(0)}(\gamma_1(t), t)^2} - \frac{1}{s^{(0)}(\gamma_1(t), t)^2} \right) \right| \\ &\leq \sqrt{n} \sup_{t \in [0,1]} \frac{1}{S^{(0)}(\gamma_1(t), t)^2} \sup_{t \in [0,1]} |S^{(1)}(\gamma_1(t), t)^2 - s^{(1)}(\gamma_1(t), t)^2| \\ &\quad + M_1^2 \sqrt{n} \sup_{t \in [0,1]} \left| \frac{1}{S^{(0)}(\gamma_1(t), t)^2} - \frac{1}{s^{(0)}(\gamma_1(t), t)^2} \right|. \end{aligned} \quad (9.28)$$

We have:

$$\begin{aligned} &\sqrt{n} \sup_{t \in [0,1], \gamma \in \mathbb{B}} \left| \frac{1}{S^{(0)}(\gamma(t), t)^2} - \frac{1}{s^{(0)}(\gamma(t), t)^2} \right| \\ &= \sqrt{n} \sup_{t \in [0,1], \gamma \in \mathbb{B}} \left| \left(\frac{1}{S^{(0)}(\gamma(t), t)} - \frac{1}{s^{(0)}(\gamma(t), t)} \right) \left(\frac{1}{S^{(0)}(\gamma(t), t)} + \frac{1}{s^{(0)}(\gamma(t), t)} \right) \right| \\ &\leq \sqrt{n} \sup_{t \in [0,1], \gamma \in \mathbb{B}} \left| \frac{1}{S^{(0)}(\gamma(t), t)} - \frac{1}{s^{(0)}(\gamma(t), t)} \right| \left(\sup_{t \in [0,1], \gamma \in \mathbb{B}} \frac{1}{S^{(0)}(\gamma(t), t)} + m_0^{-1} \right). \end{aligned}$$

Equation 9.27 and the fact that $\sup_{t \in [0,1], \gamma \in \mathbb{B}} S^{(0)}(\gamma(t), t)^{-1} = O_P(1)$ imply that

$$\sqrt{n} \sup_{t \in [0,1], \gamma \in \mathbb{B}} \left| \frac{1}{S^{(0)}(\gamma(t), t)^2} - \frac{1}{s^{(0)}(\gamma(t), t)^2} \right| \xrightarrow[n \rightarrow \infty]{P} 0.$$

Similar arguments can be used to show that

$$\sqrt{n} \sup_{t \in [0,1], \gamma \in \mathbb{B}} \left| S^{(1)}(\gamma(t), t)^2 - s^{(1)}(\gamma(t), t)^2 \right| \xrightarrow[n \rightarrow \infty]{P} 0.$$

Since $\sup_{t \in [0,1], \gamma \in \mathbb{B}} S^{(0)}(\gamma(t), t)^{-2} = O_P(1)$, we have the convergence of

$$\sqrt{n} \sup_{t \in [0,1]} \frac{1}{S^{(0)}(\gamma_1(t), t)^2} \sup_{t \in [0,1]} \left| S^{(1)}(\gamma_1(t), t)^2 - s^{(1)}(\gamma_1(t), t)^2 \right| \xrightarrow[n \rightarrow \infty]{P} 0.$$

As a result, following Equation 9.28, W_n converges in probability to 0 when $n \rightarrow \infty$. In conclusion,

$$\sqrt{n} \sup_{t \in \{t_1, \dots, t_{k_n}\}} |\mathcal{V}_{\gamma_1(t)}(Z | t) - v(\gamma_1(t), t)| \xrightarrow[n \rightarrow \infty]{P} 0.$$

Under condition A3, there exists a positive constant $C(\gamma_1)$ such that $v(\gamma_1(t), t) = C(\gamma_1)$ for all $t \in [0, 1]$. Lemma 9.2 indicates the strict positivity of $C(\gamma_1)$, and Equation 9.8 is thus demonstrated.

Equations 9.9 and 9.10. Let $\gamma_1 \in \mathbb{B}$. Following Equation 9.8, under conditions A1, A2, and A3, there exist constants $C(\gamma_1)$, $C(\beta_0) > 0$ such that

$$\sqrt{n} \sup_{t \in \{t_1, t_2, \dots, t_{k_n}\}} |\mathcal{V}_{\gamma_1(t)}(Z | t) - C(\gamma_1)| \xrightarrow[n \rightarrow \infty]{P} 0, \quad (9.29)$$

$$\sqrt{n} \sup_{t \in \{t_1, t_2, \dots, t_{k_n}\}} |\mathcal{V}_{\beta_0}(Z | t) - C(\beta_0)| \xrightarrow[n \rightarrow \infty]{P} 0. \quad (9.30)$$

Since $k_n = k_n(\beta_0)$, Equation 9.9 is confirmed:

$$\forall t \in \{t_1, \dots, t_{k_n}\}, \mathcal{V}_{\beta_0}(Z | t) > C_V \text{ w.p. 1}$$

We then have:

$$\begin{aligned} \sup_k^n \left| \frac{\mathcal{V}_{\gamma_1(t)}(Z | t)}{\mathcal{V}_{\beta_0}(Z | t)} - \frac{C(\gamma_1)}{C(\beta_0)} \right| &\leq \sup_k^n \left| \frac{1}{\mathcal{V}_{\beta_0}(Z | t)} (\mathcal{V}_{\gamma_1(t)}(Z | t) - C(\gamma_1)) \right| \\ &+ C(\gamma_1) \sup_k^n \left| \frac{1}{\mathcal{V}_{\beta_0}(Z | t)} - \frac{1}{C(\beta_0)} \right| \\ &\leq C_V^{-1} \sup_k^n |\mathcal{V}_{\gamma_1(t)}(Z | t) - C(\gamma_1)| + C(\gamma_1) \sup_k^n \left| \frac{1}{\mathcal{V}_{\beta_0}(Z | t)} - \frac{1}{C(\beta_0)} \right|. \end{aligned}$$

Moreover,

$$\begin{aligned} \sup_{t \in \{t_1, t_2, \dots, t_{k_n}\}} \left| \frac{1}{\mathcal{V}_{\beta_0}(Z | t)} - \frac{1}{C(\beta_0)} \right| &= \sup_k^n \frac{|\mathcal{V}_{\beta_0}(Z | t) - C(\beta_0)|}{C(\beta_0)\mathcal{V}_{\beta_0}(Z | t)} \\ &\leq C(\beta_0)^{-1} C_V^{-1} \sup_k^n |\mathcal{V}_{\beta_0}(Z | t) - C(\beta_0)|. \end{aligned}$$

The result 9.30 then implies:

$$\sqrt{n} \sup_{t \in \{t_1, t_2, \dots, t_{k_n}\}} \left| \frac{1}{\mathcal{V}_{\beta_0}(Z | t)} - \frac{1}{C(\beta_0)} \right| \xrightarrow[n \rightarrow \infty]{P} 0.$$

This convergence together with that of Equation 9.29 lead to the conclusion that

$$\sqrt{n} \sup_{t \in \{t_1, t_2, \dots, t_{k_n}\}} \left| \frac{\mathcal{V}_{\gamma_1(t)}(Z | t)}{\mathcal{V}_{\beta_0}(Z | t)} - \frac{C(\gamma_1)}{C(\beta_0)} \right| \xrightarrow[n \rightarrow \infty]{P} 0.$$

The same arguments can be used to prove Equation 9.10.

Theorem 9.1 This is a special case of the more general proof given for Theorem 9.2.

Theorem 9.2 Let $t \in [0, 1]$. Under hypotheses A1, A2, A3, and A4, we suppose that $\beta(t)$ is the true value for the model and that the standarized score, U_n^* , is evaluated at the value $\beta_0 \in \mathbb{B}$. Recall the equalities of Proposition 9.2:

$$\mathcal{E}_{\beta(t)}(Z | t) = E_{\beta(t)}(\mathcal{Z}(t) | \mathcal{F}_{t^-}^*), \quad \mathcal{V}_{\beta_0}(Z | t) = V_{\beta_0}(\mathcal{Z}(t) | \mathcal{F}_{t^-}^*).$$

Define the process $\{U_n^{**}(\beta_0, t), t \in [0, 1]\}$ by

$$U_n^{**}(\beta_0, t) = \int_0^t \frac{\mathcal{Z}(s) - E_{\beta_0}(\mathcal{Z}(s) | \mathcal{F}_{s^-}^*)}{(k_n V_{\beta_0}(\mathcal{Z}(s) | \mathcal{F}_{s^-}^*))^{1/2}} d\bar{N}^*(s), \quad 0 \leq t \leq 1. \quad (9.31)$$

This process is piecewise constant with a jump at each time of event. The outline of the proof is the following. Firstly, we will decompose $U_n^{**}(\beta_0, \cdot)$ into 2 processes belonging to $(D[0, 1], \mathbb{R})$. We will show that the first process converges in distribution to a Brownian motion by appealing to a theorem for the convergence of differences of martingales. The convergence of the second process can be obtained by making use of Lemma 9.1. The large sample behavior of the process $U_n^*(\beta_0, \cdot)$ belonging to $(C[0, 1], \mathbb{R})$, which is the linear interpolation of the variables $\{U_n^{**}(\beta_0, t_i), i = 1, \dots, k_n\}$, will be obtained by showing that the difference between U_n^* and U_n^{**} tends in probability to 0 when $n \rightarrow \infty$. A Taylor-Lagrange expansion of the conditional expectation $E_{\beta(t)}\{\mathcal{Z}(t) | \mathcal{F}_{t^-}^*\}$ at the point $\beta(t) = \beta_0$ gives:

$$E_{\beta_0} (\mathcal{Z}(t) \mid \mathcal{F}_{t^-}^*) = E_{\beta(t)} (\mathcal{Z}(t) \mid \mathcal{F}_{t^-}^*) + (\beta_0 - \beta(t)) \frac{\partial}{\partial \beta} E_\beta (\mathcal{Z}(t) \mid \mathcal{F}_{t^-}^*) \Big|_{\beta=\tilde{\beta}(t)}, \quad (9.32)$$

where $\tilde{\beta}$ belongs to the ball centered at β having radius $\sup_{t \in [0,1]} |\beta(t) - \beta_0|$. Recall that

$$\frac{\partial}{\partial \beta} E_\beta (\mathcal{Z}(t) \mid \mathcal{F}_{t^-}^*) = V_\beta (\mathcal{Z}(t) \mid \mathcal{F}_{t^-}^*).$$

Therefore $U_n^{**}(\beta_0, t) = X_n(t) + \sqrt{k_n} A_n(t)$, where

$$\begin{aligned} A_n(t) &= \frac{1}{k_n} \int_0^t \{\beta(s) - \beta_0\} \frac{V_{\tilde{\beta}(s)} (\mathcal{Z}(s) \mid \mathcal{F}_{s^-}^*)}{V_{\beta_0} (\mathcal{Z}(s) \mid \mathcal{F}_{s^-}^*)^{1/2}} d\bar{N}^*(s), \\ X_n(t) &= \frac{1}{k_n^{1/2}} \int_0^t \frac{\mathcal{Z}(s) - E_{\beta(s)} (\mathcal{Z}(s) \mid \mathcal{F}_{s^-}^*)}{V_{\beta_0} (\mathcal{Z}(s) \mid \mathcal{F}_{s^-}^*)^{1/2}} d\bar{N}^*(s). \end{aligned}$$

We need to consider the limiting behavior of A_n . We have:

$$\begin{aligned} &\sup_{t \in [0,1]} \left| A_n(t) - C_2 \int_0^t \{\beta(s) - \beta_0\} ds \right| \\ &\leq \sup_{t \in [0,1]} |\beta(t) - \beta_0| \sup_{t \in [0,1]} \left| \frac{1}{k_n} \int_0^t \frac{V_{\tilde{\beta}(s)} (\mathcal{Z}(s) \mid \mathcal{F}_{s^-}^*)}{V_{\beta_0} (\mathcal{Z}(s) \mid \mathcal{F}_{s^-}^*)^{1/2}} d\bar{N}^*(s) - C_2 t \right| \\ &\leq \delta_1 \sup_{t \in [0,1]} \left| \frac{1}{k_n} \int_0^t \frac{V_{\tilde{\beta}(s)} (\mathcal{Z}(s) \mid \mathcal{F}_{s^-}^*)}{V_{\beta_0} (\mathcal{Z}(s) \mid \mathcal{F}_{s^-}^*)^{1/2}} d\bar{N}^*(s) - C_2 t \right|, \end{aligned} \quad (9.33)$$

since $\beta_0 \in \mathbb{B}$ with radius δ_1 (condition A1). Under conditions A1, A2, and A3, Equation 9.10 of Lemma 9.3 indicates the existence of a constant $C_2 > 0$ such that

$$\sup_{s \in \{t_1, t_2, \dots, t_{k_n}\}} \left| \frac{V_{\tilde{\beta}(s)} (\mathcal{Z}(s) \mid \mathcal{F}_{s^-}^*)}{V_{\beta_0} (\mathcal{Z}(s) \mid \mathcal{F}_{s^-}^*)^{1/2}} - C_2 \right| \xrightarrow[n \rightarrow \infty]{P} 0.$$

Lemma 9.1 then implies:

$$\sup_{t \in [0,1]} \left| \frac{1}{k_n} \int_0^t \frac{V_{\tilde{\beta}(s)} (\mathcal{Z}(s) \mid \mathcal{F}_{s^-}^*)}{V_{\beta_0} (\mathcal{Z}(s) \mid \mathcal{F}_{s^-}^*)^{1/2}} d\bar{N}^*(s) - C_2 t \right| \xrightarrow[n \rightarrow \infty]{P} 0.$$

Equation (9.33) now indicates that the convergence of Equation 9.13 is shown. We see that X_n converges in distribution to Brownian motion by making use of the functional central limit theorem for the differences of martingales developed by Helland (1982) and described in Theorem C.4. Let $j = 1, \dots, k_n$. We write ξ_{j,k_n} as the j th increment of X_n :

$$\xi_{j,k_n} = \frac{\mathcal{Z}(t_j) - E_{\beta(t_j)} \left(\mathcal{Z}(t_j) \mid \mathcal{F}_{t_j^-}^* \right)}{\left[k_n V_{\beta_0} \left(\mathcal{Z}(t_j) \mid \mathcal{F}_{t_j^-}^* \right) \right]^{1/2}}.$$

The increment ξ_{j,k_n} is $\mathcal{F}_{t_j}^*$ -measurable and $E_{\beta(t_j)} \left(\xi_{j,k_n} \mid \mathcal{F}_{t_j^-}^* \right) = 0$. The process \bar{N}^* contains a jump at each time of an event $t_j = j/k_n$ (Proposition 9.1), we then have:

$$X_n(t) = \int_0^{\lfloor tk_n \rfloor / k_n} \frac{\mathcal{Z}(s) - E_{\beta(s)} \left(\mathcal{Z}(s) \mid \mathcal{F}_{s^-}^* \right)}{\{k_n V_{\beta_0}(\mathcal{Z}(s) \mid \mathcal{F}_{s^-}^*)\}^{1/2}} d\bar{N}^*(s) = \sum_{j=1}^{\lfloor tk_n \rfloor} \xi_{j,k_n},$$

where the function $t \rightarrow \lfloor tk_n \rfloor$ is positive, increasing with integer values and right continuous. To check the assumption of Theorem C.4 we need to consider the expectation, the variance, and the Lyapunov condition: We begin with the expectation.

a. (Expectation). The inclusion of the σ -algebras $\mathcal{F}_{t_{j-1}}^* \subset \mathcal{F}_{t_j^-}^*$ leads to

$$E_{\beta(t_{j-1})} \left(\xi_{j,k_n} \mid \mathcal{F}_{t_{j-1}}^* \right) = E_{\beta(t_{j-1})} \left(E_{\beta(t_j)} \left(\xi_{j,k_n} \mid \mathcal{F}_{t_j^-}^* \right) \mid \mathcal{F}_{t_{j-1}}^* \right) = 0. \quad (9.34)$$

b. (Variance). Under the hypotheses A1, A2, and A3, following Lemma 9.3, there exists a constant $C_1(\beta, \beta_0) > 0$ such that

$$\sup_{j=1, \dots, k_n} \left| \frac{V_{\beta(t_j)} \left(\mathcal{Z}(t_j) \mid \mathcal{F}_{t_j^-}^* \right)}{V_{\beta_0} \left(\mathcal{Z}(t_j) \mid \mathcal{F}_{t_j^-}^* \right)} - C_1(\beta, \beta_0)^2 \right| \xrightarrow[n \rightarrow \infty]{P} 0.$$

In particular, $C_1(\beta, \beta_0) = 1$ whenever $\beta(t) = \beta_0$. Let $s \in \{t_1, t_2, \dots, t_{k_n}\}$. From Equation 9.9,

$$V_{\beta_0} \left(\mathcal{Z}(s) \mid \mathcal{F}_{s^-}^* \right) > C_V \quad \text{a.s.}$$

We also have, according to the hypothesis A4, the upper bound,

$$\begin{aligned} |\mathcal{Z}(s)| &= \left| \sum_{i=1}^n Z_i(X_i) \mathbf{1}_{\phi_n(X_i) = s, \delta_i = 1} \right| \\ &\leq \sup_{i=1, \dots, n} |Z_i(X_i)| \sum_{i=1}^n \mathbf{1}_{\phi_n(X_i) = s, \delta_i = 1} \leq L, \end{aligned} \quad (9.35)$$

since $s \in \{t_1, t_2, \dots, t_{k_n}\}$ donc $\sum_{i=1}^n \mathbf{1}_{\phi_n}(X_i) = s, \delta_i = 1 = 1$. As a result,

$$V_{\beta(s)}(\mathcal{Z}(s) \mid \mathcal{F}_{s^-}^*) \leq E_{\beta(s)}(\mathcal{Z}(s)^2 \mid \mathcal{F}_{s^-}^*) \leq L^2.$$

In conclusion, we have, almost surely, the upper bound,

$$\sup_{j=1, \dots, k_n} \frac{V_{\beta(t_j)}(\mathcal{Z}(t_j) \mid \mathcal{F}_{t_j^-}^*)}{V_{\beta_0}(\mathcal{Z}(t_j) \mid \mathcal{F}_{t_j^-}^*)} \leq \frac{L^2}{C_V},$$

and an application of the dominated convergence theorem implies that

$$\sup_{j=1, \dots, k_n} \left| \frac{V_{\beta(t_j)}(\mathcal{Z}(t_j) \mid \mathcal{F}_{t_j^-}^*)}{V_{\beta_0}(\mathcal{Z}(t_j) \mid \mathcal{F}_{t_j^-}^*)} - C_1(\beta, \beta_0)^2 \right| \xrightarrow[n \rightarrow \infty]{L^1} 0. \quad (9.36)$$

Next, we have that

$$\sum_{j=1}^{\lfloor tk_n \rfloor} E_{\beta(t_{j-1})}(\xi_{j, k_n}^2 \mid \mathcal{F}_{t_{j-1}}^*) \xrightarrow[n \rightarrow \infty]{L^1} C_1(\beta, \beta_0)^2 t.$$

and, to keep things uncluttered, we first write $W_0^{\mathcal{F}}(t_j) = \frac{V_{\beta(t_j)}(\mathcal{Z}(t_j) \mid \mathcal{F}_{t_j^-}^*)}{V_{\beta_0}(\mathcal{Z}(t_j) \mid \mathcal{F}_{t_j^-}^*)}$, then,

$$\begin{aligned} & \left| \sum_{j=1}^{\lfloor tk_n \rfloor} E_{\beta(t_{j-1})}(\xi_{j, k_n}^2 \mid \mathcal{F}_{t_{j-1}}^*) - C_1(\beta, \beta_0)^2 t \right| \\ &= \left| \sum_{j=1}^{\lfloor tk_n \rfloor} E_{\beta(t_{j-1})} \left(E_{\beta(t_j)} \left(\xi_{j, k_n}^2 \mid \mathcal{F}_{t_j^-}^* \right) \mid \mathcal{F}_{t_{j-1}}^* \right) - C_1(\beta, \beta_0)^2 t \right| \\ &= \left| \frac{1}{k_n} \sum_{j=1}^{\lfloor tk_n \rfloor} E_{\beta(t_{j-1})} \left(W_0^{\mathcal{F}}(t_j) \mid \mathcal{F}_{t_{j-1}}^* \right) - C_1(\beta, \beta_0)t \right| \\ &\leq \frac{1}{k_n} \sum_{j=1}^{\lfloor tk_n \rfloor} E_{\beta(t_{j-1})} \left(\left| W_0^{\mathcal{F}}(t_j) - C_1(\beta, \beta_0)^2 \right| \mid \mathcal{F}_{t_{j-1}}^* \right) + C_1(\beta, \beta_0)^2 \left| \frac{\lfloor tk_n \rfloor}{k_n} - t \right|. \end{aligned}$$

As a result of the above we now have:

$$\begin{aligned}
& E \left(\left| \sum_{j=1}^{\lfloor tk_n \rfloor} E_{\beta(t_{j-1})} \left(\xi_{j,k_n}^2 \mid \mathcal{F}_{t_{j-1}}^* \right) - C_1(\beta, \beta_0)^2 t \right| \right) \\
& \leq \frac{1}{k_n} \sum_{j=1}^{\lfloor tk_n \rfloor} E \left(\left| W_0^{\mathcal{F}}(t_j) - C_1(\beta, \beta_0)^2 \right| \right) + C_1(\beta, \beta_0)^2 \left| \frac{\lfloor tk_n \rfloor}{k_n} - t \right| \\
& \leq \frac{\lfloor tk_n \rfloor}{k_n} \sup_{j=1, \dots, \lfloor tk_n \rfloor} E \left(\left| W_0^{\mathcal{F}}(t_j) - C_1(\beta, \beta_0)^2 \right| \right) + C_1(\beta, \beta_0)^2 \left| \frac{\lfloor tk_n \rfloor}{k_n} - t \right| \\
& \leq \frac{\lfloor tk_n \rfloor}{k_n} E \left(\sup_{j=1, \dots, \lfloor tk_n \rfloor} \left| W_0^{\mathcal{F}}(t_j) - C_1(\beta, \beta_0)^2 \right| \right) + C_1(\beta, \beta_0)^2 \left| \frac{\lfloor tk_n \rfloor}{k_n} - t \right| \\
& \leq \frac{\lfloor tk_n \rfloor}{k_n} E \left(\sup_{j=1, \dots, k_n} \left| W_0^{\mathcal{F}}(t_j) - C_1(\beta, \beta_0)^2 \right| \right) + C_1(\beta, \beta_0)^2 \left| \frac{\lfloor tk_n \rfloor}{k_n} - t \right|.
\end{aligned}$$

Equation 9.36 then leads to the conclusion that $\sum_{j=1}^{\lfloor tk_n \rfloor} E_{\beta} \left(\xi_{j,k_n}^2 \mid \mathcal{F}_{t_{j-1}}^* \right)$ converges in mean and, as a result, in probability to $C_1(\beta, \beta_0)^2 t$ when $n \rightarrow \infty$.

- c. We have the convergence in mean of $\sum_{j=1}^{\lfloor tk_n \rfloor} E_{\beta(t_{j-1})} \left(\xi_{j,k_n}^3 \mid \mathcal{F}_{t_{j-1}}^* \right)$ to 0 when $n \rightarrow \infty$. Furthermore,

$$\begin{aligned}
E \left(\left| \sum_{j=1}^{\lfloor tk_n \rfloor} E_{\beta(t_{j-1})} \left(\xi_{j,k_n}^3 \mid \mathcal{F}_{t_{j-1}}^* \right) \right| \right) & \leq \sum_{j=1}^{\lfloor tk_n \rfloor} E \left(E_{\beta(t_{j-1})} \left(|\xi_{j,k_n}^3| \mid \mathcal{F}_{t_{j-1}}^* \right) \right) \\
& = \sum_{j=1}^{\lfloor tk_n \rfloor} E \left(|\xi_{j,k_n}^3| \right), \tag{9.37}
\end{aligned}$$

and keeping in mind that for $j = 1, \dots, k_n$, then

$$\xi_{j,k_n} = \frac{\mathcal{Z}(t_j) - E_{\beta(t_j)} \left(\mathcal{Z}(t_j) \mid \mathcal{F}_{t_j^-}^* \right)}{\left\{ k_n V_{\beta_0} \left(\mathcal{Z}(t_j) \mid \mathcal{F}_{t_j^-}^* \right) \right\}^{1/2}}.$$

Equation 9.35 implies that

$$E \left(|\mathcal{Z}(t) - E_{\beta(t)} (\mathcal{Z}(t) \mid \mathcal{F}_{t^-}^*)|^3 \right) \leq 8L^3.$$

Also, on the basis of Equation 9.9, $t \in \{t_1, t_2, \dots, t_{k_n}\}$, $V_{\beta_0}(\mathcal{Z}(t)|\mathcal{F}_{t^-}^*)$ is strictly greater than C_V . As a result, Equation 9.37 becomes:

$$\begin{aligned} & E \left(\left| \sum_{j=1}^{\lfloor tk_n \rfloor} E_{\beta(t_{j-1})} \left(\xi_{j,k_n}^3 | \mathcal{F}_{t_{j-1}}^* \right) \right| \right) \\ & \leq \sum_{j=1}^{k_n} \frac{1}{(k_n C_V)^{3/2}} E \left(\left| \mathcal{Z}(t_j) - E_{\beta(t_j)} \left(\mathcal{Z}(t_j) | \mathcal{F}_{t_j^-}^* \right) \right|^3 \right) \leq \frac{8L^3}{k_n^{1/2} C_V^{3/2}}, \end{aligned} \quad (9.38)$$

and $\sum_{j=1}^{\lfloor tk_n \rfloor} E_{\beta(t_{j-1})} \left(\xi_{j,k_n}^3 | \mathcal{F}_{t_{j-1}}^* \right)$ converges in mean to 0 when $n \rightarrow \infty$ and, in consequence, in probability. The Lyapunov condition is verified.

We can now apply Theorem C.4 and conclude that

$$X_n \xrightarrow[n \rightarrow \infty]{\mathcal{L}} C_1(\beta, \beta_0) \mathcal{W},$$

where the constant $C_1(\beta, \beta_0)^2$ is the ratio of two asymptotic conditional variances v evaluated at β and β_0 . When $\beta = \beta_0$, we have $C_1(\beta, \beta_0) = 1$. We have the relation

$$\begin{aligned} U_n^*(\beta_0, \cdot) - \sqrt{k_n} A_n &= U_n^*(\beta_0, \cdot) - U_n^{**}(\beta_0, \cdot) + U_n^{**}(\beta_0, \cdot) - \sqrt{k_n} A_n \\ &= U_n^*(\beta_0, \cdot) - U_n^{**}(\beta_0, \cdot) + X_n. \end{aligned} \quad (9.39)$$

We have then established the limiting behavior of X_n when $n \rightarrow \infty$. Next we look at the convergence in probability of $U_n^*(\beta_0, \cdot) - U_n^{**}(\beta_0, \cdot)$ to zero when $n \rightarrow \infty$ with respect to the uniform norm. Let $\varepsilon > 0$,

$$\begin{aligned} P(\|U_n^*(\beta_0, \cdot) - U_n^{**}(\beta_0, \cdot)\| \geq \varepsilon) &= P \left(\sup_{t \in [0, 1]} |U_n^*(\beta_0, t) - U_n^{**}(\beta_0, t)| \geq \varepsilon \right) \\ &= P \left(\sup_{i=1, \dots, k_n} |U_n^*(\beta_0, t_i) - U_n^{**}(\beta_0, t_{i-1})| \geq \varepsilon \right) \\ &= P \left(\frac{1}{\sqrt{k_n}} \sup_{i=1, \dots, k_n} \left| \frac{\mathcal{Z}(t_j) - E_{\beta_0}(\mathcal{Z}(t_j) | \mathcal{F}_{t_j^-}^*)}{V_{\beta_0}(\mathcal{Z}(t_j) | \mathcal{F}_{t_j^-}^*)^{1/2}} \right| \geq \varepsilon \right) \leq P \left(\frac{2L}{\sqrt{k_n} C_V} \geq \varepsilon \right), \end{aligned}$$

Therefore, $U_n^*(\beta_0, \cdot) - U_n^{**}(\beta_0, \cdot) \xrightarrow[n \rightarrow \infty]{P} 0$. Slutsky's theorem applied to the decomposition of Equation 9.39 leads us to conclude that

$$U_n^*(\beta_0, \cdot) - k_n^{1/2} A_n \xrightarrow[n \rightarrow \infty]{\mathcal{L}} C_1(\beta, \beta_0) \mathcal{W},$$

in $(C[0,1], \mathbb{R})$ with respect to the topology of uniform convergence.

Proposition 9.4 Recall that $\|Ax\| \leq p\|A\|\|x\|$ et $\|AB\| \leq p\|A\|\|B\|$ for $A, B \in \mathcal{M}_{p \times p}(\mathbb{R})$ et $x \in \mathbb{R}^p$. Note that, if A is diagonal and its elements are positive, we have:

$$\left\| A^{1/2} \right\| = \max_{l=1,\dots,p} (A^{1/2})_{ll} \leq \left(\max_{l=1,\dots,p} (A)_{ll} \right)^{1/2} = \|A\|^{1/2}.$$

Following the condition B3, Σ is a symmetric matrix having real coefficients so that $\Sigma^{1/2}$ exists and $\left\| \Sigma^{1/2} \right\| < \infty$. We have:

$$\begin{aligned} & \sqrt{n} \sup_{t \in \{t_1, t_2, \dots, t_{k_n}\}, \gamma \in \mathbb{B}'} \left\| \mathcal{V}_{\gamma(t)}(Z | t) - \Sigma \right\| \\ &= \sqrt{n} \sup_{t \in \{t_1, t_2, \dots, t_{k_n}\}, \gamma \in \mathbb{B}'} \left\| \mathcal{V}_{\gamma(t)}(Z | t) - \mathcal{V}_{\gamma(t)}(Z | t)^{1/2} \Sigma^{1/2} + \mathcal{V}_{\gamma(t)}(Z | t)^{1/2} \Sigma^{1/2} - \Sigma \right\| \\ &\leq p \sqrt{n} \sup_{t \in \{t_1, t_2, \dots, t_{k_n}\}, \gamma \in \mathbb{B}'} \left\| \mathcal{V}_{\gamma(t)}(Z | t)^{1/2} - \Sigma^{1/2} \right\| \\ &\quad \times \left(\sup_{t \in \{t_1, t_2, \dots, t_{k_n}\}, \gamma \in \mathbb{B}'} \left\| \mathcal{V}_{\gamma(t)}(Z | t)^{1/2} \right\| + \left\| \Sigma^{1/2} \right\| \right). \end{aligned}$$

Now, for all $t \in \{t_1, \dots, t_{k_n}\}$ and $\gamma \in \mathbb{B}'$, according to the condition B4,

$$\left\| \mathcal{V}_{\gamma(t)}(Z | t) \right\| \leq \left\| \mathcal{E}_\gamma(Z^{\otimes 2} | t) \right\| + \left\| \mathcal{E}_\gamma(Z | t)^{\otimes 2} \right\| \leq 2L'^2. \quad (9.40)$$

Therefore, in choosing an orthogonal transformation matrix $P_{\gamma(t)}(t)$ of norm 1 in the diagonalisation of the symmetric positive definite matrix $\mathcal{V}_{\gamma(t)}(Z | t) = P_{\gamma(t)}(t) D_{\gamma(t)}(t) P_{\gamma(t)}(t)^T$ with $D_{\gamma(t)}(t)$ diagonal, we have:

$$\begin{aligned} & \left\| \mathcal{V}_{\gamma(t)}(Z | t)^{1/2} \right\| \leq p^2 \left\| P_{\gamma(t)}(t) \right\|^2 \left\| D_{\gamma(t)}(t)^{1/2} \right\| \leq p^2 \left\| D_{\gamma(t)}(t) \right\|^{1/2} \\ &= p^2 \left\| P_{\gamma(t)}(t)^T P_{\gamma(t)}(t) D_{\gamma(t)}(t) P_{\gamma(t)}(t)^T P_{\gamma(t)}(t) \right\|^{1/2} \\ &\leq p^2 \left(p^2 \left\| P_{\gamma(t)}(t) D_{\gamma(t)}(t) P_{\gamma(t)}(t)^T \right\| \left\| P_{\gamma(t)}(t) \right\|^2 \right)^{1/2} = p^3 \left\| \mathcal{V}_{\gamma(t)}(Z | t) \right\|^{1/2} \\ &\leq \sqrt{2}p^3 L'. \end{aligned}$$

from which, $\sup_{t \in \{t_1, t_2, \dots, t_{k_n}\}, \gamma \in \mathbb{B}'} \left\| \mathcal{V}_{\gamma(t)}(Z | t)^{1/2} \right\| \leq \sqrt{2}p^3 L'$, and

$$\begin{aligned} & \sqrt{n} \sup_{t \in \{t_1, t_2, \dots, t_{k_n}\}, \gamma \in \mathbb{B}'} \left\| \mathcal{V}_{\gamma(t)}(Z | t) - \Sigma \right\| \\ &\leq p \left(\sqrt{2}p^3 L' + \left\| \Sigma^{1/2} \right\| \right) \sqrt{n} \sup_{t \in \{t_1, t_2, \dots, t_{k_n}\}, \gamma \in \mathbb{B}'} \left\| \mathcal{V}_{\gamma(t)}(Z | t)^{1/2} - \Sigma^{1/2} \right\|. \quad (9.41) \end{aligned}$$

The result then follows from Equation 9.15.

Theorem 9.3 The proof follows the same outline as in the univariate setting. Recall that the space of cadlag functions of $(D[0,1], \mathbb{R}^p)$ is equipped with the product topology of Skorokhod, described in Definition A.5. Furthermore, the norm of the vector in \mathbb{R}^p or a matrix of $\mathcal{M}_{p \times p}(\mathbb{R})$ is the sup norm. The process being evaluated at $\beta_0 \in \mathbb{B}'$, we have $k_n = k_n(\beta_0)$ and the upper bound of Equation 9.17 becomes:

$$\forall t \in \{t_1, \dots, t_{k_n}\}, \|V_{\beta_0(t)}(Z | t)^{-1}\| \leq C_{\mathcal{V}}^{-1} \quad \text{a.s.} \quad (9.42)$$

We take U_n^{**} to be a cadlag process having a jump at each event time $t_i = i/k_n, i = 1, \dots, k_n$, such that

$$U_n^{**}(\beta_0, t) = \frac{1}{\sqrt{k_n}} \sum_{i=1}^{\lfloor tk_n \rfloor} V_{\beta_0}(\mathcal{Z}(t_i) | \mathcal{F}_{t_i^-}^*)^{-1/2} \{ \mathcal{Z}(t_i) - E_{\beta_0}(\mathcal{Z}(t_i) | \mathcal{F}_{t_i^-}^*) \}, \quad 0 \leq t \leq 1.$$

Let $\gamma \in \mathbb{B}'$. Recall the equalities of Proposition 9.2:

$$\mathcal{E}_{\gamma(t_i)}(Z | t_i) = E_{\gamma(t_i)}(\mathcal{Z}(t_i) | \mathcal{F}_{t_i^-}^*), \quad V_{\gamma(t_i)}(Z | t_i) = V_{\gamma(t_i)}(\mathcal{Z}(t_i) | \mathcal{F}_{t_i^-}^*).$$

Note that $U_n^{**}(\beta_0, t) = X_n(t) + \sqrt{k_n} B_n(t)$ for $t \in [0, 1]$,

$$\begin{aligned} X_n(t) &= \frac{1}{\sqrt{k_n}} \sum_{i=1}^{\lfloor tk_n \rfloor} \mathcal{V}_{\beta_0}(Z | t_i)^{-1/2} \{ \mathcal{Z}(t_i) - \mathcal{E}_{\beta(t_i)}(Z | t_i) \} \\ B_n(t) &= \frac{1}{k_n} \sum_{i=1}^{\lfloor tk_n \rfloor} \mathcal{V}_{\beta_0}(Z | t_i)^{-1/2} \{ \mathcal{E}_{\beta(t_i)}(Z | t_i) - \mathcal{E}_{\beta_0}(Z | t_i) \}. \end{aligned}$$

We have:

$$\begin{aligned} &\left\| U_n^*(\beta_0, \cdot) - \sqrt{k_n} B_n - \mathcal{W}_p \right\| \\ &\leq \|U_n^*(\beta_0, \cdot) - U_n^{**}(\beta_0, \cdot)\| + \left\| U_n^{**}(\beta_0, \cdot) - \sqrt{k_n} B_n - \mathcal{W}_p \right\| \\ &= \|U_n^*(\beta_0, \cdot) - U_n^{**}(\beta_0, \cdot)\| + \|X_n - \mathcal{W}_p\|. \end{aligned} \quad (9.43)$$

Let us consider the limiting behavior of the two terms on the right of the inequality. Considering just the first term and letting $\varepsilon > 0$, we have:

$$\begin{aligned}
& P \left(\sup_{t \in [0,1]} \|U_n^*(\beta_0, t) - U_n^{**}(\beta_0, t)\| \geq \varepsilon \right) \\
&= P \left(\sup_{i=1, \dots, k_n} \left\| U_n^{**} \left(\beta_0, \frac{i}{k_n} \right) - U_n^{**} \left(\beta_0, \frac{i-1}{k_n} \right) \right\| \geq \varepsilon \right) \\
&= P \left(\frac{1}{\sqrt{k_n}} \sup_{i=1, \dots, k_n} \left\| \mathcal{V}_{\beta_0}(Z | t_i)^{-1/2} \{ \mathcal{Z}(t_i) - \mathcal{E}_{\beta_0}(Z | t_i) \} \right\| \geq \varepsilon \right) \\
&\leq P \left(\frac{p}{\sqrt{k_n}} \sup_{i=1, \dots, k_n} \left\| \mathcal{V}_{\beta_0}(Z | t_i)^{-1/2} \right\| \left\| \mathcal{Z}(t_i) - \mathcal{E}_{\beta_0}(Z | t_i) \right\| \geq \varepsilon \right). \quad (9.44)
\end{aligned}$$

We also have, for $i = 1, \dots, k_n$,

$$\begin{aligned}
\left\| V_{\beta_0} \left(\mathcal{Z}(t_i) \middle| \mathcal{F}_{t_i^-}^* \right)^{-1/2} \right\| &= \left\| V_{\beta_0} \left(\mathcal{Z}(t_i) \middle| \mathcal{F}_{t_i^-}^* \right)^{-1} V_{\beta_0} \left(\mathcal{Z}(t_i) \middle| \mathcal{F}_{t_i^-}^* \right)^{1/2} \right\| \\
&\leq p \left\| V_{\beta_0} \left(\mathcal{Z}(t_i) \middle| \mathcal{F}_{t_i^-}^* \right)^{-1} \right\| \left\| V_{\beta_0} \left(\mathcal{Z}(t_i) \middle| \mathcal{F}_{t_i^-}^* \right)^{1/2} \right\| \\
&\leq C_{\mathcal{V}}^{-1} \sqrt{2} p^4 L',
\end{aligned} \quad (9.45)$$

where the last inequality is obtained from Equations 9.41 and 9.42. Also we have:

$$\left\| \mathcal{Z}(t_i) - E_{\beta_0} \left(\mathcal{Z}(t_i) \middle| \mathcal{F}_{t_i^-}^* \right) \right\| \leq \|\mathcal{Z}(t_i)\| + \left\| E_{\beta_0} \left(\mathcal{Z}(t_i) \middle| \mathcal{F}_{t_i^-}^* \right) \right\| \leq 2L'.$$

The inequality 9.44 implies that

$$P \left(\sup_{t \in [0,1]} \|U_n^*(\beta_0, t) - U_n^{**}(\beta_0, t)\| \geq \varepsilon \right) \leq P \left(\frac{2\sqrt{2}}{\sqrt{k_n}} p^5 C_{\mathcal{V}}^{-1} (L')^2 \geq \varepsilon \right), \quad (9.46)$$

and $\lim_{n \rightarrow \infty} P \left(\sup_{t \in [0,1]} \|U_n^*(\beta_0, t) - U_n^{**}(\beta_0, t)\| \geq \varepsilon \right) = 0$. This implies the convergence in probability of $U_n^*(\beta_0, \cdot) - U_n^{**}(\beta_0, \cdot)$ to 0 when $n \rightarrow \infty$ in the space $(D[0,1], \mathbb{R}^p)$ equipped with the Skorohod product topology as a consequence of the results of Section A.5. We now consider the second term. The convergence in distribution of X_n to \mathcal{W}_p is obtained by an application of the multivariate functional central limit theorem C.5 described in Helland (1982). Before applying this theorem we need to check some working hypotheses. Let $i \in \{1, 2, \dots, k_n\}$ and $t \in [0, 1]$. Note that

$$\xi_{i,k_n} = (\xi_{i,k_n}^1, \dots, \xi_{i,k_n}^p) = \frac{1}{\sqrt{k_n}} V_{\beta_0} \left(\mathcal{Z}(t_i) \middle| \mathcal{F}_{t_i^-}^* \right)^{-1/2} \{ \mathcal{Z}(t_i) - E_{\beta(t_i)} \left(\mathcal{Z}(t_i) \middle| \mathcal{F}_{t_i^-}^* \right) \}$$

The i th increment in \mathbb{R}^p of the process X_n . We have $X_n(t) = \sum_{i=1}^{\lfloor tk_n \rfloor} \xi_{i,k_n}$. Note that ξ_{i,k_n} is $\mathcal{F}_{t_i}^*$ -measurable and that $E_{\beta(t_i)} \left(\xi_{i,k_n} \mid \mathcal{F}_{t_i^-}^* \right) = 0$. Let $l, m \in \{1, \dots, p\}$, $l \neq m$. We have e_l the l th vector of the canonical base of \mathbb{R}^p : all of the elements are zero with the exception of the l th which is 1. Also, $e_l^T e_m = 0$, $e_l^T e_l = 1$ and $\|e_l\| = \max_{i=1, \dots, p} (e_l)_i = 1$. For $i = 1, \dots, k_n$, we have:

$$\xi_{i,k_n}^l = e_l^T \xi_{i,k_n} = \xi_{i,k_n}^T e_l \in \mathbb{R}. \quad (9.47)$$

a.' (Table of martingale differences.) The inclusion of the σ -algebras $\mathcal{F}_{t_{i-1}}^* \subset \mathcal{F}_{t_i^-}^*$ and the centering of the increments implies that

$$E_{\beta(t_{i-1})} \left(\xi_{i,k_n}^l \mid \mathcal{F}_{t_{i-1}}^* \right) = E_{\beta(t_{i-1})} \left(E_{\beta(t_i)} \left(\xi_{i,k_n}^l \mid \mathcal{F}_{t_i^-}^* \right) \mid \mathcal{F}_{t_{i-1}}^* \right) = 0.$$

b.' (Non-correlation.) Using Equation 9.47, we have:

$$\begin{aligned} E_{\beta(t_i)} \left(\xi_{i,k_n}^l \xi_{i,k_n}^m \mid \mathcal{F}_{t_i^-}^* \right) &= e_l^T E_{\beta(t_i)} \left(\xi_{i,k_n} \xi_{i,k_n}^T \mid \mathcal{F}_{t_i^-}^* \right) e_m \\ &= \frac{1}{k_n} e_l^T V_{\beta_0} \left(\mathcal{Z}(t_i) \mid \mathcal{F}_{t_i^-}^* \right)^{-1/2} V_{\beta(t_i)} \left(\mathcal{Z}(t_i) \mid \mathcal{F}_{t_i^-}^* \right) V_{\beta_0} \left(\mathcal{Z}(t_i) \mid \mathcal{F}_{t_i^-}^* \right)^{-1/2} e_m \\ &= \frac{1}{k_n} e_l^T V(t_i^-, \beta_0, \beta(t_i)) e_m, \end{aligned}$$

where

$$V(t_i^-, \beta_0, \beta(t_i)) = V_{\beta_0} \left(\mathcal{Z}(t_i) \mid \mathcal{F}_{t_i^-}^* \right)^{-1/2} V_{\beta(t_i)} \left(\mathcal{Z}(t_i) \mid \mathcal{F}_{t_i^-}^* \right) V_{\beta_0} \left(\mathcal{Z}(t_i) \mid \mathcal{F}_{t_i^-}^* \right)^{-1/2}.$$

Note that I_p is the identity matrix of $\mathcal{M}_{p \times p}(\mathbb{R})$. Also, $e_l^T I_p e_m = 0$. We have:

$$\begin{aligned} &E \left(\left| \sum_{i=1}^{\lfloor tk_n \rfloor} E_{\beta(t_{i-1})} \left(\xi_{i,k_n}^l \xi_{i,k_n}^m \mid \mathcal{F}_{t_{i-1}}^* \right) \right| \right) \\ &= E \left(\left| \frac{1}{k_n} \sum_{i=1}^{\lfloor tk_n \rfloor} E_{\beta(t_{i-1})} \left(e_l^T V(t_i^-, \beta_0, \beta(t_i)) e_m \mid \mathcal{F}_{t_{i-1}}^* \right) \right| \right) \\ &\leq \frac{1}{k_n} \sum_{i=1}^{\lfloor tk_n \rfloor} E \left(\left| e_l^T V(t_i^-, \beta_0, \beta(t_i)) e_m \right| \right) \\ &= \frac{1}{k_n} \sum_{i=1}^{\lfloor tk_n \rfloor} E \left(\left| e_l^T (V(t_i^-, \beta_0, \beta(t_i)) - I_p) e_m \right| \right) \end{aligned}$$

$$\begin{aligned}
&\leq \frac{\lfloor tk_n \rfloor}{k_n} \sup_{i=1, \dots, \lfloor tk_n \rfloor} E \left(\left| e_l^T (V(t_i^-, \beta_0, \beta(t_i)) - I_p) e_m \right| \right) \\
&\leq \frac{\lfloor tk_n \rfloor}{k_n} E \left(\sup_{i=1, \dots, \lfloor tk_n \rfloor} \left| e_l^T (V(t_i^-, \beta_0, \beta(t_i)) - I_p) e_m \right| \right) \\
&\leq p^2 \frac{\lfloor tk_n \rfloor}{k_n} \|e_l^T\| \|e_m\| E \left(\sup_{i=1, \dots, \lfloor tk_n \rfloor} \|(V(t_i^-, \beta_0, \beta(t_i)) - I_p)\| \right) \\
&= p^2 \frac{\lfloor tk_n \rfloor}{k_n} E \left(\sup_{i=1, \dots, k_n} \|(V(t_i^-, \beta_0, \beta(t_i)) - I_p)\| \right) \\
&\leq p^2 \frac{\lfloor tk_n \rfloor}{k_n} E \left(\sup_{i=1, \dots, k_n} \|(V(t_i^-, \beta_0, \beta(t_i)) - I_p)\| \right). \tag{9.48}
\end{aligned}$$

We need to show that

$$\sup_{i=1, \dots, k_n} \|(V(t_i^-, \beta_0, \beta(t_i)) - I_p)\| \xrightarrow[n \rightarrow \infty]{L^1} 0.$$

We have:

$$\begin{aligned}
&\sup_{i=1, \dots, k_n} \|(V(t_i^-, \beta_0, \beta(t_i)) - I_p)\| \\
&= \sup_{i=1, \dots, k_n} \left\| V_{\beta_0} \left(\mathcal{Z}(t_i) \middle| \mathcal{F}_{t_i^-}^* \right)^{-1/2} \left(V_{\beta(t_i)} \left(\mathcal{Z}(t_i) \middle| \mathcal{F}_{t_i^-}^* \right) \right. \right. \\
&\quad \left. \left. - V_{\beta_0} \left(\mathcal{Z}(t_i) \middle| \mathcal{F}_{t_i^-}^* \right) \right) V_{\beta_0} \left(\mathcal{Z}(t_i) \middle| \mathcal{F}_{t_i^-}^* \right)^{-1/2} \right\| \\
&\leq p^2 \left(\sup_{i=1, \dots, k_n} \left\| V_{\beta_0} \left(\mathcal{Z}(t_i) \middle| \mathcal{F}_{t_i^-}^* \right)^{-1/2} \right\|^2 \right)^2 \sup_{i=1, \dots, k_n} \left\| V_{\beta(t_i)} \left(\mathcal{Z}(t_i) \middle| \mathcal{F}_{t_i^-}^* \right) \right. \\
&\quad \left. - V_{\beta_0} \left(\mathcal{Z}(t_i) \middle| \mathcal{F}_{t_i^-}^* \right) \right\|.
\end{aligned}$$

Equation 9.45 implies that

$$\begin{aligned}
&\sup_{i=1, \dots, k_n} \|(V(t_i^-, \beta_0, \beta(t_i)) - I_p)\| \\
&\leq 2p^{10} C_{\mathcal{V}}^{-2} (L')^2 \sup_{i=1, \dots, k_n} \left\| V_{\beta(t_i)} \left(\mathcal{Z}(t_i) \middle| \mathcal{F}_{t_i^-}^* \right) - V_{\beta_0} \left(\mathcal{Z}(t_i) \middle| \mathcal{F}_{t_i^-}^* \right) \right\| \\
&\leq 2p^{10} C_{\mathcal{V}}^{-2} (L')^2 \left(\sup_{i=1, \dots, k_n} \left\| V_{\beta(t_i)} \left(\mathcal{Z}(t_i) \middle| \mathcal{F}_{t_i^-}^* \right) - \Sigma \right\| + \right. \\
&\quad \left. \sup_{i=1, \dots, k_n} \left\| \Sigma - V_{\beta_0} \left(\mathcal{Z}(t_i) \middle| \mathcal{F}_{t_i^-}^* \right) \right\| \right).
\end{aligned}$$

Finally, the convergence of Equation 9.16 implies the convergence in probability,

$$\sup_{i=1,\dots,k_n} \|V(t_i^-, \beta_0, \beta(t_i)) - I_p\| \xrightarrow[n \rightarrow \infty]{P} 0.$$

Equations 9.40 and 9.45 indicate that $\sup_{i=1,\dots,k_n} \|V(t_i^-, \beta_0, \beta(t_i)) - I_p\|$ is bounded by a finite constant. The convergence is then again a convergence in mean.

$$\sup_{i=1,\dots,k_n} \|V(t_i^-, \beta_0, \beta(t_i)) - I_p\| \xrightarrow[n \rightarrow \infty]{L^1} 0. \quad (9.49)$$

Equation 9.48 leads us to conclude that

$$\sum_{i=1}^{\lfloor k_n t \rfloor} E_{\beta(t_{i-1})}(\xi_{i,k_n}^l \xi_{i,k_n}^m | \mathcal{F}_{t_{i-1}}^*) \xrightarrow[n \rightarrow \infty]{L^1} 0,$$

and the convergence is one in probability.

c.' (Variance.) Similar arguments to a.' and b.' can be used, implying the equality,

$$\begin{aligned} & \sum_{i=1}^{\lfloor tk_n \rfloor} E_{\beta(t_{i-1})} \left((\xi_{i,k_n}^l)^2 \middle| \mathcal{F}_{t_{i-1}}^* \right) - t \\ &= \frac{1}{k_n} \sum_{i=1}^{\lfloor tk_n \rfloor} E_{\beta(t_{i-1})} \left(e_l^T \{V(t_i^-, \beta_0, \beta(t_i)) - I_p\} e_l \middle| \mathcal{F}_{t_{i-1}}^* \right) + \left(\frac{\lfloor k_n t \rfloor}{k_n} - t \right). \end{aligned}$$

This leads to

$$\begin{aligned} & E \left(\left| \sum_{i=1}^{\lfloor tk_n \rfloor} E_{\beta(t_{i-1})} \left((\xi_{i,k_n}^l)^2 \middle| \mathcal{F}_{t_{i-1}}^* \right) - t \right| \right) \\ & \leq \frac{1}{k_n} \sum_{i=1}^{\lfloor tk_n \rfloor} E \left(\left| e_l^T \{V(t_i^-, \beta_0, \beta(t_i)) - I_p\} e_l \right| \right) + \left| \frac{\lfloor k_n t \rfloor}{k_n} - t \right| \\ & \leq \frac{p^2}{k_n} \sum_{i=1}^{\lfloor tk_n \rfloor} E \left(\|V(t_i^-, \beta_0, \beta(t_i)) - I_p\| \right) + \left| \frac{\lfloor k_n t \rfloor}{k_n} - t \right| \\ & \leq p^2 \frac{\lfloor k_n t \rfloor}{k_n} \sup_{i=1,\dots,k_n} E \left(\|V(t_i^-, \beta_0, \beta(t_i)) - I_p\| \right) + \left| \frac{\lfloor k_n t \rfloor}{k_n} - t \right| \\ & \leq p^2 \frac{\lfloor k_n t \rfloor}{k_n} E \left(\sup_{i=1,\dots,k_n} \|V(t_i^-, \beta_0, \beta(t_i)) - I_p\| \right) + \left| \frac{\lfloor k_n t \rfloor}{k_n} - t \right|. \end{aligned}$$

Equation 9.49 then leads to the conclusion that

$\sum_{i=1}^{\lfloor tk_n \rfloor} E_{\beta(t_{i-1})} \left((\xi_{i,k_n}^l)^2 \middle| \mathcal{F}_{t_{i-1}}^* \right) \xrightarrow[n \rightarrow \infty]{L^1} t$, indicating convergence in probability

d' (Lyapunov condition.) We have $|\xi_{i,k_n}^l| \leq \max_{l=1,\dots,p} |\xi_{i,k_n}^l| = \|\xi_{i,k_n}\|$. Therefore,

$$E\left(\sum_{i=1}^{\lfloor k_n t \rfloor} E_{\beta(t_{i-1})}\left(|\xi_{i,k_n}^l|^3 \mid \mathcal{F}_{t_{i-1}}^*\right)\right) \leq \sum_{i=1}^{\lfloor k_n t \rfloor} E\left(|\xi_{i,k_n}^l|^3\right) \leq \sum_{i=1}^{k_n} E\left(\|\xi_{i,k_n}\|^3\right).$$

Using Equation 9.45 and condition B4, we have the bound,

$$\begin{aligned} & E\left(\sum_{i=1}^{\lfloor k_n t \rfloor} E_{\beta(t_{i-1})}\left(|\xi_{i,k_n}^l|^3 \mid \mathcal{F}_{t_{i-1}}^*\right)\right) \\ & \leq \frac{p^3}{k_n^{3/2}} \sum_{i=1}^{k_n} E\left(\left\|V_{\beta_0}(\mathcal{Z}(t_i) \mid \mathcal{F}_{t_{i-1}}^*)^{-1/2}\right\|^3 \|\mathcal{Z}(t_i) - E_{\beta(t_i)}(\mathcal{Z}(t_i) \mid \mathcal{F}_{t_{i-1}}^*)\|^3\right) \\ & \leq \frac{p^3}{k_n^{1/2}} (C_{\mathcal{V}}^{-1} \sqrt{2} p^4 L')^3 (2L')^3. \end{aligned}$$

In consequence, $\sum_{i=1}^{\lfloor k_n t \rfloor} E_{\beta(t_{i-1})}\left(|\xi_{i,k_n}^l|^3 \mid \mathcal{F}_{t_{i-1}}^*\right)$ converge to 0 in mean and therefore in probability when $n \rightarrow \infty$.

In conclusion, all of the hypotheses of Theorem C.5 are met, which allows us to conclude that X_n converges to \mathcal{W}_p when $n \rightarrow \infty$ in the space $(D[0,1], \mathbb{R}^p)$ equipped with the product topology of Skorohod. It then follows, as a result of Equation 9.43 and the convergence result of the first term on the right of Equation 9.46, Equation 9.18 is proved. To close the demonstration of the theorem it remains to prove Equation 9.19. Let $t \in [0,1]$. The multivariate Taylor-Lagrange inequality implies that for $s \in \{t_1, \dots, t_{k_n}\}$,

$$\|\mathcal{E}_{\beta(s)}(Z \mid s) - \mathcal{E}_{\beta_0}(Z \mid s) - \mathcal{V}_{\beta_0}(Z \mid s)(\beta(s) - \beta_0)\| \leq \frac{1}{2} M_n \|\beta(s) - \beta_0\|^2. \quad (9.50)$$

As a result,

$$\begin{aligned} & \left\| B_n(t) - \Sigma^{1/2} \int_0^t \{\beta(s) - \beta_0\} ds \right\| \\ & = \left\| \frac{1}{k_n} \sum_{i=1}^{\lfloor tk_n \rfloor} \mathcal{V}_{\beta_0}(Z \mid t_i)^{-1/2} \{ \mathcal{E}_{\beta(t_i)}(Z \mid t_i) - \mathcal{E}_{\beta_0}(Z \mid t_i) \} - \Sigma^{1/2} \int_0^t \{\beta(s) - \beta_0\} ds \right\| \\ & \leq \frac{1}{k_n} \sum_{i=1}^{\lfloor tk_n \rfloor} \left\| \mathcal{V}_{\beta_0}(Z \mid t_i)^{-1/2} \{ \mathcal{E}_{\beta(t_i)}(Z \mid t_i) - \mathcal{E}_{\beta_0}(Z \mid t_i) - \mathcal{V}_{\beta_0}(Z \mid t_i) \{\beta(t_i) - \beta_0\} \} \right\| \\ & \quad + \frac{1}{k_n} \sum_{i=1}^{\lfloor tk_n \rfloor} \left\| (\mathcal{V}_{\beta_0}(Z \mid t_i)^{1/2} - \Sigma^{1/2}) \{\beta(t_i) - \beta_0\} \right\| \end{aligned}$$

$$\begin{aligned}
& + \left\| \Sigma^{1/2} \left(\frac{1}{k_n} \sum_{i=1}^{\lfloor tk_n \rfloor} \beta(t_i) - \int_0^t \beta(s) ds + \beta_0 \left(\frac{\lfloor tk_n \rfloor}{k_n} - t \right) \right) \right\| \\
& \leq \frac{p M_n}{2 k_n} \sum_{i=1}^{\lfloor tk_n \rfloor} \left\| \mathcal{V}_{\beta_0}(Z | t_i)^{-1/2} \right\| \|\beta(t_i) - \beta_0\|^2 + \frac{p}{k_n} \sum_{i=1}^{\lfloor tk_n \rfloor} \left\| \mathcal{V}_{\beta_0}(Z | t_i)^{1/2} - \Sigma^{1/2} \right\| \|\beta(t_i) - \beta_0\| \\
& \quad + p \left\| \Sigma^{1/2} \right\| \left\| \frac{1}{k_n} \sum_{i=1}^{\lfloor tk_n \rfloor} \beta(t_i) - \int_0^t \beta(s) ds \right\| + \left| \frac{\lfloor tk_n \rfloor}{k_n} - t \right| \left\| \Sigma^{1/2} \beta_0 \right\| \\
& \leq \frac{\lfloor tk_n \rfloor}{k_n} \frac{p M_n}{2} \sup_{i=1, \dots, \lfloor tk_n \rfloor} \|\beta(t_i) - \beta_0\|^2 \sup_{i=1, \dots, \lfloor tk_n \rfloor} \left\| \mathcal{V}_{\beta_0}(Z | t_i)^{-1/2} \right\| \\
& \quad + p \frac{\lfloor tk_n \rfloor}{k_n} \sup_{i=1, \dots, \lfloor tk_n \rfloor} \|\beta(t_i) - \beta_0\| \sup_{i=1, \dots, \lfloor tk_n \rfloor} \left\| \mathcal{V}_{\beta_0}(Z | t_i)^{1/2} - \Sigma^{1/2} \right\| \\
& \quad + p \left\| \Sigma^{1/2} \right\| \sup_{l=1, \dots, p} \left| \frac{1}{k_n} \sum_{i=1}^{\lfloor tk_n \rfloor} \beta_l(t_i) - \int_0^t \beta_l(s) ds \right| + p \left| \frac{\lfloor tk_n \rfloor}{k_n} - t \right| \left\| \Sigma^{1/2} \right\| \|\beta_0\|.
\end{aligned}$$

Since β_0 and $0 \in \mathbb{B}'$ donc $\|\beta_0\| \leq 2\delta_2$ et $\sup_{i=1, \dots, \lfloor tk_n \rfloor} \|\beta(t_i) - \beta_0\| \leq \sup_{t \in [0, 1]} \|\beta(t) - \beta_0\| \leq \delta_2$. Now, Equation 9.45 implies that

$$\sup_{i=1, \dots, k_n} \left\| V_{\beta_0} \left(\mathcal{Z}(t_i) \middle| \mathcal{F}_{t_i^-}^* \right)^{-1/2} \right\| \leq \sqrt{2} C_{\mathcal{V}}^{-1} p^4 L'.$$

Therefore,

$$\begin{aligned}
& \left\| B_n(t) - \Sigma^{1/2} \int_0^t \{\beta(s) - \beta_0\} ds \right\| \\
& \leq M_n \frac{\lfloor tk_n \rfloor}{k_n} \frac{p^5}{\sqrt{2}} \delta_2^2 C_{\mathcal{V}}^{-1} L' + p \frac{\lfloor tk_n \rfloor}{k_n} \delta_2 \sup_{i=1, \dots, k_n} \left\| \mathcal{V}_{\beta_0}(Z | t_i)^{1/2} - \Sigma^{1/2} \right\| \\
& \quad + p \left\| \Sigma^{1/2} \right\| \sup_{l=1, \dots, p} \left| \frac{1}{k_n} \sum_{i=1}^{\lfloor tk_n \rfloor} \beta_l(t_i) - \int_0^t \beta_l(s) ds \right| + 2\delta_2 p \left| \frac{\lfloor tk_n \rfloor}{k_n} - t \right| \left\| \Sigma^{1/2} \right\|.
\end{aligned}$$

Bringing in the sup, we have:

$$\begin{aligned}
& \sup_{t \in [0, 1]} \left\| B_n(t) - \Sigma^{1/2} \int_0^t \{\beta(s) - \beta_0\} ds \right\| \\
& \leq M_n \frac{p^5}{\sqrt{2}} \delta_2^2 C_{\mathcal{V}}^{-1} L' + p \delta_2 \sup_{i=1, \dots, k_n} \left\| \mathcal{V}_{\beta_0}(Z | t_i)^{1/2} - \Sigma^{1/2} \right\| \\
& \quad + p \left\| \Sigma^{1/2} \right\| \sup_{t \in [0, 1]} \sup_{l=1, \dots, p} \left| \frac{1}{k_n} \sum_{i=1}^{\lfloor tk_n \rfloor} \beta_l(t_i) - \int_0^t \beta_l(s) ds \right| + \frac{2}{k_n} \delta_2 p \left\| \Sigma^{1/2} \right\|. \tag{9.51}
\end{aligned}$$

According to B3, $\lim_{n \rightarrow \infty} M_n = 0$, and

$$\sup_{t \in \{t_1, t_2, \dots, t_{k_n}\}, \gamma \in \mathbb{B}'} \left\| \mathcal{V}_{\gamma(t)}(Z | t)^{1/2} - \Sigma^{1/2} \right\| \xrightarrow[n \rightarrow \infty]{P} 0.$$

Let $l = 1, \dots, p$. The function β_l is bounded, allowing us to apply Equation 9.25 of the proof of Lemma 9.1 so that

$$\lim_{n \rightarrow \infty} \sup_{t \in [0, 1]} \left| \frac{1}{k_n} \sum_{i=1}^{\lfloor tk_n \rfloor} \beta_l(t_i) - \int_0^t \beta_l(s) ds \right| = 0.$$

It follows that

$$\sup_{t \in [0, 1]} \sup_{l=1, \dots, p} \left| \frac{1}{k_n} \sum_{i=1}^{\lfloor tk_n \rfloor} \beta_l(t_i) - \int_0^t \beta_l(s) ds \right| \leq \sum_{l=1}^p \sup_{t \in [0, 1]} \left| \frac{1}{k_n} \sum_{i=1}^{\lfloor tk_n \rfloor} \beta_l(t_i) - \int_0^t \beta_l(s) ds \right|,$$

and since p is finite,

$$\lim_{n \rightarrow \infty} \sup_{t \in [0, 1]} \sup_{l=1, \dots, p} \left| \frac{1}{k_n} \sum_{i=1}^{\lfloor tk_n \rfloor} \beta_l(t_i) - \int_0^t \beta_l(s) ds \right| = 0.$$

According to Equation 9.51, the convergence of Equation 9.19 is then established.



Chapter 10

Model construction guided by regression effect process

10.1 Chapter summary

The basic ideas of linear regression run through this chapter. It is structured as follows: In Section 10.3, we present a literature review of graphical methods and goodness of fit tests for proportional hazards models. In Section 10.6, we consider the R^2 coefficient for non-proportional hazards models, which is the measure of predictive ability used from here on. Although other suggestions for R^2 have been discussed extensively in the literature O'Quigley (2008), we limit our attention to this particular one in view of its many desirable properties, most importantly a key property, not known to exist for rival measures, that the population equivalent for R^2 , Ω^2 , is maximized for the model closest to that generating the observations. Section 10.4 presents a graphical method and a goodness of fit test for proportional hazards models, based on the regression effect process. We also present a method for constructing non-proportional hazards models from this process and the R^2 coefficient. Next, simulations are presented in Section 10.9 comparing the performance of the goodness of fit test we have developed with some standard tests. The method for constructing non-proportional hazards models will then be illustrated on simulated data. Applications on real data are presented in Section 10.10. We terminate the chapter with a discussion.

10.2 Context and motivation

In linear regression it is clear that bringing a non-linear model, via transformation, under a linear framework will improve predictability. Predictability is also going to improve if, via regression on covariates, we reduce the residual variance. We see in this chapter that these two distinct, although not orthogonal, perspectives

can be investigated to improve predictive strength. These two perspectives, fit and the quantification of the strength of regression effect, are not only related to one another but each relates directly to the regression effect process of the previous chapter. These loose intuitive ideas are made firm via some important theorems and corollaries. These results can then be brought together to guide model building. Several detailed examples highlight their practical importance.

The statistical properties of the regression effect process established in Chapter 9 can be used to guide model building and, in particular, for non-proportional hazards models. It turns out that the regression effect process essentially contains all the information we need and can steer us in the direction of better models. These resulting models will necessarily have good fits, and optimal predictive power as measured by the R^2 coefficient given the available information. We will see how goodness of fit tests can be structured under the heading of the regression effect process. This will apply at all levels of model complexity, from the simple proportional hazards model for a binary covariate to higher dimensional models in which effects may be a mixture of proportional and non-proportional ones. Confidence bands for the process corresponding to a constant effect can provide some initial building blocks. As described in the introductory chapter, goodness of fit and predictive ability are two very different aspects of statistical models, and the link between the two is not always clear. In this chapter, we consider more formally how improving the fit of a non-proportional hazards model will increase its predictive power.

The purpose of a statistical model when representing data is twofold: first to simplify as much as possible the several relationships and sub-relationships that may be present in the data, and secondly, to use such a structure as an inferential tool. An important question raised by the use of a particular model on observed data is its goodness of fit; i.e., is it plausible to consider that the data can be viewed as having been generated from that model? Procedures for evaluating the goodness of fit can be used to study the quality of a model as a whole, or certain parts only. For example, in a clinical trial context, we may question whether the working hypotheses concerning differences in treatment are satisfied in the presence of other covariates, without considering how well these covariates are modeled. The goodness of fit of a model can be assessed with hypothesis testing or using graphical methods. The interpretation of goodness of fit tests must be performed with care. Most of them are powerful for detecting specific alternative hypotheses, e.g., an effect that increases or decreases over time. We need to have an idea of the type of alternative hypothesis to consider when looking at a dataset, but this is not always realistic. Moreover, the result of a test is often limited to reading off the p -value. This value can decrease simply by increasing the number of observations, and not with a necessarily better fit.

Rejecting the null hypothesis of a good fit of the model to the data does not generally get us very far. We need some kind of pointers on where to go to improve fit. Graphical methods, for their part, often have the advantage of

suggesting visually the type of alternative hypothesis, i.e., more realistic models to consider. Even here, results can be confusing or difficult to summarize if we are not used to interpreting such plots. In our view the regression effect process, processes in the multivariate case, can ease interpretation and play a fundamental role in model construction. Before looking more closely at the regression effect process and especially the time-transformed regression effect process, we present a brief overview of some of the less recent proposals in the literature.

10.3 Classical graphical methods

The first graphical method proposed for testing the proportional hazards hypothesis for categorical variables was proposed by Kay (1977). The method is based on the following formulation of the proportional hazards model:

$$\log \{-\log(S(t|Z))\} = \beta^T Z + \log \left(\int_0^t \lambda_0(s) ds \right).$$

Thus, if the covariate Z has categories z_1, \dots, z_M , the lines $\log(-\log(S(t|Z = z_m)))$ plotted as a function of time for each category $m = 1, \dots, M$, are parallel. In practice, the conditional survival function $S(\cdot|Z)$ is replaced by its Kaplan-Meier estimator $\hat{S}(\cdot|Z)$ (Equation 3.1). There are certain disadvantages to this method. For instance, continuous covariates have to be discretized, and the choice of points at which to cut is not always obvious (Andersen, 1982). Second, the parallel feature we are looking for between the estimated, noisy functions is not always readily detectable. Another way to proceed is to plot the difference between the functions and look at whether this remains constant over time. Thus, with two categories, we can plot one as a function of the other at different times, which should be reasonably well approximated by a straight line (Andersen, 1982). The larger variance in the Kaplan-Meier estimators later in the study where often there will be few remaining subjects can sometimes obscure a picture of simple linearity. We may suspect a deviation from linearity at these later time points that is no more than a manifestation of increased noise.

Other procedures are essentially based on studying residuals and these, quite naturally, bring us closer to our own approach based on the regression effect process. The first such residuals used in graphical goodness of fit methods were those presented in Cox and Snell (1968). Their distribution under the Cox model was studied by Kay (1977). Lagakos (1981), O'Quigley (1982) and Crowley and Storer (1983) have underlined their weak power for detecting poor fits. Furthermore, these residuals also reflect variability in the covariate (Baltazar-Aban and Peña, 1995). Graphical methods for examining the goodness of fit of proportional hazards models based on residuals can be divided into two categories according to whether the residuals are cumulative or not. Within the non-cumulative methods, a large class of martingale residuals from Barlow

and Prentice (1988) can be examined by plotting its elements as a function of observed times of deaths or their ranks. The martingale residual of individual $j = 1, \dots, n$, at time t is defined by

$$M_j(t) = N_j(t) - \hat{\Lambda}_j(t), \quad \text{with} \quad \hat{\Lambda}_j(t) = \int_0^t \pi_j(\hat{\beta}, s) d\bar{N}(s),$$

where $\hat{\beta}$ is the partial maximum likelihood estimator of β . If the model is valid, $M_1(t), \dots, M_n(t)$ are the outputs of independent martingales (Gill, 1980). The class of martingale residuals in Barlow and Prentice (1988) is then defined by

$$\int_0^t \phi(s) M_j(s) ds, \quad j = 1, \dots, n, \quad 0 \leq t \leq T,$$

where ϕ is a deterministic or random and predictable function of time. The integral of a predictable process with respect to the martingale residuals is a martingale with known properties, including being centered and uncorrelated. This can be seen on the residuals plot if the model is valid. The residuals in Schoenfeld (1982), the weighted ones of Schoenfeld (Lin, 1991), and those in Kay (1977), belong to this class of residuals. The Schoenfeld ones have an important role in goodness of fit methods. Recall that these are defined at each time of death t by

$$r_\beta(t) = Z_j(t) - \mathcal{E}_\beta(Z|t), \quad (10.1)$$

where j is the individual who dies at time t . Grambsch and Therneau (1994) have suggested plotting the Schoenfeld residuals standardized over time in order to examine the validity of the proportional hazards model, and if it is rejected, obtain some idea of the time-dependent effect. Essentially, they have shown that for each time of death t :

$$E \left(\left[\mathcal{V}_{\hat{\beta}}(Z|t)^{-1} r_{\hat{\beta}}(t) \right]_j \right) + \hat{\beta}_j \approx \beta_j(t), \quad j = 1, \dots, p,$$

where $\hat{\beta} = (\hat{\beta}_1, \dots, \hat{\beta}_p)$ is the partial maximum likelihood estimator of the proportional hazards model and $\beta(t) = (\beta_1(t), \dots, \beta_p(t))$ the true time-dependent model of the non-proportional hazards model. This approach is used often, and is implemented in the `survival` package in R. Furthermore, the authors have proposed a confidence band for the regression effect, but this is not reliable if the proportional hazards model is not satisfied. More recently, Sasieni and Winnett (2003) have suggested using the smoothed residuals of martingale differences. These can be plotted as a function of time for a fixed value of the covariate, allowing a glimpse at temporal dependency. These can also be plotted as a function of a covariate at a fixed time, to study how to introduce it into the model. In the former, several plots need to be output for the various values of the covariates,

and in the latter, for various times. In practice, it may be difficult to interpret all of these. Also, these methods—based on non-cumulative residuals—require smoothing functions to join up the residuals. As underlined by Lin et al. (1993), results are often sensitive to the choice of smoothing technique employed. Different smoothings can lead to very different conclusions. Examining the quality of proportional hazards models with such methods can be tricky.

To get around these kinds of problems, several authors have proposed using cumulative martingale residuals. Arjas (1988) suggests plotting the expected number of deaths under the proportional hazards model as a function of the ranks of the times of death. The model fits the data well if the resulting curve is close to the diagonal. A score process, inspired by the goodness of fit test of Wei (1984), was introduced by Therneau et al. (1990) and is denoted $\{U(\beta, t), 0 \leq t \leq \mathcal{T}\}$, where

$$U(\beta, t) = \sum_{i=1}^n \int_0^t \{\mathcal{Z}_i(s) - \mathcal{E}_\beta(Z|s)\} dN_i(s), \quad 0 \leq t \leq \mathcal{T}. \quad (10.2)$$

Under the proportional hazards model, a standardized version of $\{U(\hat{\beta}, t), 0 \leq t \leq \mathcal{T}\}$ converges to a Brownian bridge. A test statistic corresponding to the supremum of the absolute value of the standardized process can then be applied. The limit distribution of the statistic is the Kolmogorov distribution. Wei (1984) proposed this test for the binary covariate case and Therneau et al. (1990) extended it to a time-dependent covariate. In the presence of correlated covariates, the limit distribution of the statistic is no longer a Brownian bridge. Lin et al. (1993) showed that the non-standardized score process (10.2) converges to a centered Gaussian variable whose variance-covariance matrix depends on the data. Hence, a comparison of the observed process over time with a large number of processes simulated from the Gaussian limit distribution can give an idea of the asymptotic distribution of the process's supremum. In practice, interpretation of such plots is not always easy, notably because the envelope of the process is not deterministic and contains noise.

The classic case of linear regression with Gaussian noise and covariates corresponds to a very special kind of regression model. Any sub-model, i.e., all conditional and marginal models, are also linear and the residual noise remains Gaussian albeit with a potentially different variance. This type of model structure is, however, just about unique, and with very few conditions, can even be taken as a characterization of the multinormal model. For non-linear models, and specifically the proportional hazards models with several covariates, the sub-models will not belong to the same class. Thus, evaluating the goodness of fit of a multivariate proportional hazards models by looking individually at the univariate models is not really addressing the relevant question. In fact, in the absence of available tools for checking the overall validity of the model, most methods for testing the proportional hazards hypothesis with one covariate suppose that it is true for the others, an assumption that is generally false (Scheike and Martinussen, 2004).

Many methods depend on the covariance between the covariates, such as the ones proposed by Grambsch and Therneau (1994) and Lin et al. (1993). To deal with this problem, Scheike and Martinussen (2004), looking at a non-proportional hazards model, developed estimation techniques and goodness of fit tests for the proportional hazards model for each covariate, while allowing for the possibility that other covariates have time-dependent effects. Their simulations show that the method works well compared to the often-used ones of Grambsch and Therneau (1994) and Lin et al. (1993) when the proportional hazards hypothesis is invalid and/or there are correlated covariates. Their test statistics depend on the estimation of the regression parameter, which requires a hard-to-understand algorithm based on kernel smoothing. Also, the form of the regression parameter estimator is not a smooth and explicit function of time. The goodness of fit tests are of the Kolmogorov-Smirnov and Cramér-von Mises type and the p -values are calculated by simulation since the limiting distributions of the statistics are not known.

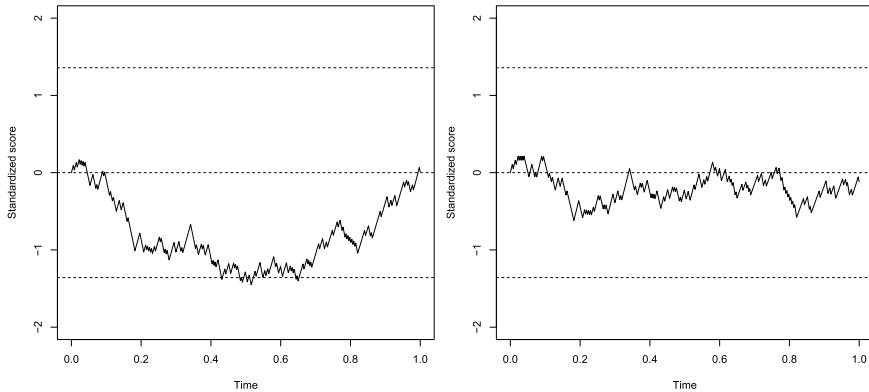
For more details on these less recent goodness of fit tests for proportional hazards models, the reader may wish to consult the works by Therneau and Grambsch (2000), Klein and Moeschberger (2003) and Martinussen and Scheike (2006). Our preference as outlined below is to make full use of known and established properties of the regression effect process which, depending on the circumstances, we will approximate by a Brownian motion or a Brownian bridge. We make use of Theorem 9.2 indicating how a correctly specified model corresponds to a Brownian motion with linear drift. This theorem, together with an elementary conditioning argument, leads to

Corollary 10.1. *Under the non-proportional hazards model with parameter $\beta(t)$,*

$$U_n^*(\beta(t), t) - t \times U_n^*(\beta(t), 1) \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} \mathcal{W}^0(t),$$

where $\mathcal{W}^0(t)$, $t \in (0, 1)$, is the Brownian bridge.

As a corollary, the result is slightly misplaced in the text, and if we were guided only by logical flow, then it would follow immediately from Theorem 9.2. We present it here because this is where we need it, as a key support to assessing goodness of fit. It tells us that the above elementary transformation will look like a Brownian bridge in practice, when our chosen model can be considered well specified and one that could plausibly have generated the observations. We will see this in later real examples, but as an appetizer, the reader might consider Figure 10.1 where—leaving aside any formal testing procedure—the visual impression is enough to tell us that, in this case, a non-proportional hazards model with time-dependent regression effects provides a much better fit to the observations than does the simple proportional hazards model. In a multivariate setting we can obtain several such graphs; examples being the process for the prognostic index, the process for particular covariates where others are kept fixed and processes for stratified models.



(a) Goodness of fit for receptor status in breast cancer study in PH model with constant effects
 (b) Goodness of fit for receptor status in breast cancer study in NPH model and non-constant effects using Corollary 10.1

Figure 10.1: Dotted lines show upper 5% limit of the supremum of a Brownian bridge process. Left-hand process for PH model encroaches on this limit. Right-hand process for the NPH model with time-dependent effects shows a much improved fit. The process appears to be not unlike a Brownian bridge

10.4 Confidence bands for regression effect process

The regression effect process provides us with insight into both the strength of effect as well as the fit of the assumed model. No less importantly, when the mechanism generating the observations differs from the assumed model, the regression effect process will indicate to us in which direction to look for a better fitting model. Here, we consider confidence bands for the process. These offer some help in seeing whether the model's assumed functional form for the parameter can be deemed sufficiently plausible.

Let us consider the non-proportional hazards model with regression coefficient $\beta(t) = (\beta_1(t), \dots, \beta_p(t))$, and suppose that conditions **B1**, **B2**, **B3** and **B4** of Chapter 9 hold. We suppose furthermore than the sequence $(M_n)_n$ in hypothesis **B3** is such that $\sqrt{n}M_n \rightarrow 0$ as $n \rightarrow \infty$. For $i = 1, \dots, p$, consider the following null and alternative hypotheses:

$$H_{0,i} : \exists b_i, \forall t, \beta_i(t) = b_i, \quad H_{1,i} : \nexists b_i, \forall t, \beta_i(t) = b_i.$$

Under the hypothesis $H_{0,i}$, the covariate $Z^{(i)}$ is consistent with the proportional hazards model, while under $H_{1,i}$, its coefficient changes over time. The proposition below allows us to define the limit behavior of the i th component of the regression effect process under $H_{0,i}$. This then allows us to construct confidence bands for the constant nature of a parameter around the corresponding compo-

ment of the process, as well as a goodness of fit test for the proportional hazards model. Recall that the limit distribution of the supremum of the absolute value of a Brownian bridge is the Kolmogorov distribution. For $\alpha = 5\%$, the upper quantile of order α of this distribution is $a(\alpha) = 1.358$.

Corollary 10.2. (Confidence bands). *Let $i \in \{1, \dots, p\}$ and $\beta_0 \in \mathbb{B}'$. Under $H_{0,i}$, i.e., under the proportional hazards model with a constant parameter b not necessarily equal to β_0 , and denoting $a(\alpha)$ the upper quantile of order α of the Kolmogorov distribution, we have:*

$$\lim_{n \rightarrow +\infty} P\left(\forall t \in [0, 1], \left[\hat{\Sigma}^{-1/2} U_n^*(\beta_0, t)\right]_i \in BC_i(\alpha)\right) = 1 - \alpha,$$

where $BC_i(\alpha)$ is a level $1 - \alpha$ confidence band given by:

$$\left[t \left[\hat{\Sigma}_{\cdot, i}^{-1/2} U_n^*(\beta_0, 1)\right]_i - \left\| \hat{\Sigma}_{\cdot, i}^{-1/2} \right\|_2 a(\alpha); t \left[\hat{\Sigma}_{\cdot, i}^{-1/2} U_n^*(\beta_0, 1)\right]_i + \left\| \hat{\Sigma}_{\cdot, i}^{-1/2} \right\|_2 a(\alpha)\right].$$

Corollary 10.3. (Goodness of fit test). *Let $i \in \{1, \dots, p\}$ and $\beta_0 \in \mathbb{B}'$. Under the non-proportional hazards model with parameter $\beta(t)$ not necessarily equal to β_0 , the hypothesis $H_{0,i}$ is rejected at asymptotic level α if:*

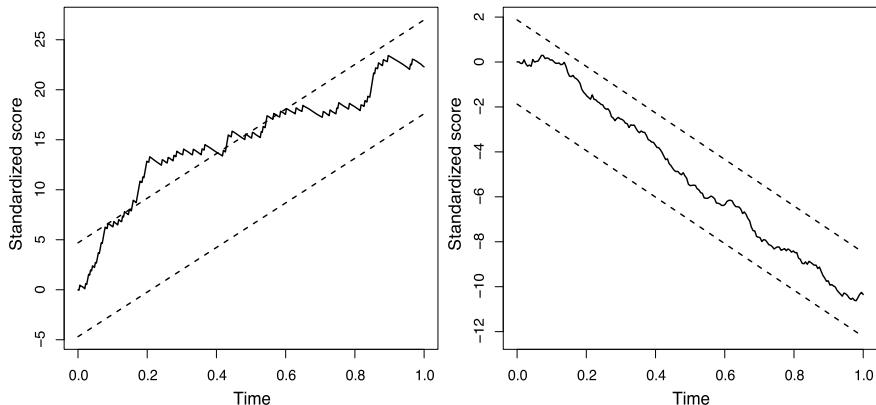
$$\left\| \hat{\Sigma}_{\cdot, i}^{-1/2} \right\|_2^{-1} \sup_{t \in [0, 1]} \left| \left(\hat{\Sigma}^{-1/2} \{U_n^*(\beta_0, t) - tU_n^*(\beta_0, 1)\} \right)_i \right| \geq a(\alpha),$$

where $a(\alpha)$ is the order α upper quantile of the Kolmogorov distribution. Denoting K a random variable from this distribution, the test's asymptotic p -value is:

$$P\left(K \geq \left\| \hat{\Sigma}_{\cdot, i}^{-1/2} \right\|_2^{-1} \sup_{t \in [0, 1]} \left| \left(\hat{\Sigma}^{-1/2} \{U_n^*(\beta_0, t) - tU_n^*(\beta_0, 1)\} \right)_i \right|\right).$$

If the i th component of the process $\hat{\Sigma}^{-1/2} U_n^*(\beta_0, \cdot)$ exits the confidence band $IC(\alpha)_i$ or the p -value is less than α , the hypothesis $H_{0,i}$ that the effect $\beta_i(t)$ is constant is rejected at level α . However, by simultaneously testing several hypotheses $H_{0,i}$ for different covariates, we see inflation in the overall level of the test. This means that a process can exit the confidence band even if the corresponding effect is constant over time, with an overall level greater than α (Figure 10.2).

This could be a problem if our definitive conclusion on the goodness of fit of the proportional hazards model is based on these confidence bands or the goodness of fit test. This is not the case however since such formal testing does not have any central role in model building. Confidence bands are just one tool in model construction. The non-detection of a constant effect may be corrected



(a) Regression effect process for tumor size in breast cancer based on constant effect
 (b) Regression effect process for gastric cancer data using two slope changepoint model

Figure 10.2: Process and confidence intervals under PH assumption. Left-hand figure shows clear evidence of non proportionality whereas right-hand figure, for a fitted piecewise model, appears to provide an adequate fit.

in later model selection steps respecting the proportional hazards hypothesis, as described in the next section. Note that the proportional hazards hypothesis is evaluated for each of the covariates, taking into account the correlation between them and allowing for the possibility of other time-dependent effects. Simulations comparing the performance of this goodness of fit test for proportional hazards models with other common tests are presented in Section 10.9.

10.5 Structured tests for time dependency

In his article introducing the proportional hazards model, Cox (1972) suggested replacing the constant regression coefficient by a non-constant one, then testing this postulated time dependency. Formally, this amounts to considering the class of tests based on the non-proportional hazards model with parameter

$$\beta(t) = \beta_0 + \beta_1 g(t),$$

where β_0 and β_1 are constant parameters and g is a function of time. This means testing the null hypothesis concerning the validity of the proportional hazards model $H_0 : \beta_1 = 0$ against the alternative $H_1 : \beta_1 \neq 0$ (Grambsch and Therneau, 1994; Therneau and Grambsch, 2000). A wide range of g are possible (O'Quigley, 2008). If g is a known function of time, this reverts to running a score test as proposed in Cox (1972). The function g can also be a predictable process. For example, $g(t) = \bar{N}(t^-)$ corresponds to the test in Breslow et al. (1984) based on the ranked times of events. One last example is a piecewise constant function

g with pre-defined change-points (Moreau et al., 1985; O'Quigley and Pessione, 1989; Schoenfeld, 1980).

Wei (1984) has developed a goodness of fit test for one binomial variable, with test statistic $\sup_t |U(\hat{\beta}, t)|$, where $\hat{\beta}$ is the maximum partial likelihood estimator. This corresponds to applying the test for the supremum of a Brownian bridge, and the asymptotic distribution of the statistic is the Kolmogorov one. This test has inspired several graphical methods, which were mentioned in the previous section (Lin et al., 1993, 1996; Therneau et al., 1990). Schoenfeld (1980) has developed a χ^2 test which partitions the covariate and time space and compares the observed values in the various partitions to the expected ones under the model. Lin (1991) has proposed a score test based on a weighted maximum log-likelihood estimator with regression coefficient β , noting that under this model, the limiting distribution of the difference between this estimator and the usual partial maximum likelihood one is a centered normal distribution.

In addition, Murphy and Sen (1991) and Gray (1992) have proposed goodness of fit tests for the proportional hazards model, modeling the regression coefficient $\beta(t)$ in a non-proportional hazards model by projection onto basis functions, then testing whether the projections are time-dependent.

The collection of goodness of fit procedures that we recommend is based on the results in Theorems 9.2 and 9.3 which indicate that the form of time dependency for the regression coefficient expresses itself via the regression effect process. Thus, testing the goodness of fit of proportional hazards models corresponds to examining the linearity of the process's drift. Before looking at this aspect more closely, let us first consider the notion of explained variance in the non-proportional hazards model, as well as its estimator, the R^2 coefficient which allows us to measure the predictive strength of any model. Again, all of the predictive information is contained in the regression effect process and so, structuring model building around a careful study of this process allows us to address all of the key questions that arise.

10.6 Predictive ability of a regression model

EXPLAINED VARIANCE

Let X be a random variable whose second-order moment exists. The Bienaymé-Chebyshev inequality:

$$\forall \varepsilon > 0, \quad P(|X - E(X)| \geq \varepsilon) \leq \varepsilon^{-2} \operatorname{Var}(X)$$

says that the spread of X around its expectation can be bounded in terms of its variance. The smaller the variance, the greater the probability that X is closer, on average, to its expectation. Using a statistical model, the idea of evaluating the proximity of a response variable to its predicted value, in terms of the variance, would appear natural. The smaller the variance, the better the quality of the

prediction. For random variables X and Y whose second-order moments exist, the variance can be always expanded as:

$$\text{Var}(Y) = \text{Var}(E(Y|X)) + E(\text{Var}(Y|X)), \quad (10.3)$$

which then allows us to define the explained variance parameter. Suppose that we want to model a real-valued response variable Y with the help of a vector of explanatory variables X . If the chosen model and the covariate information, X , allow a good characterization of Y , then the expected values of Y given X will show relatively large dispersion when compared to the simple imprecision (noise) associated with any given fixed value of X . Whenever $\text{Var}(E(Y|X))$ is large compared to the average of $\text{Var}(Y|X)$, then prediction becomes easy. Conversely, whenever the opposite holds, then the residual noise will mask to a greater or lesser extent any underlying signal. It makes sense to describe $\text{Var}(E(Y|X))$ as the signal and $E(\text{Var}(Y|X))$ as the noise, so that we can write: $\text{Var}(Y) = \text{signal} + \text{noise}$. We can formally introduce a parameter that can be viewed as the explained variation in Y given X via,

Definition 10.1. *The proportion of the variance of Y explained by X and the model, denoted $\Omega_{Y|X}^2$, is:*

$$\Omega_{Y|X}^2 = \frac{\text{Var}(E(Y|X))}{\text{Var}(Y)} = \frac{\text{Var}(Y) - E\text{Var}(Y|X)}{\text{Var}(Y)} = \frac{\text{signal}}{\text{signal} + \text{noise}}. \quad (10.4)$$

This $\Omega_{Y|X}^2$ is commonly known as the explained variance, and corresponds to the ratio of the variance of the expected values of the response variable given the variables X and the model, over the marginal variance of the response variable. In a non-proportional hazards model, two coefficients can be defined, depending on whether T is the response variable and Z the explanatory variable, or vice versa.

THE EXPLAINED VARIANCE OF T GIVEN Z .

Replacing Y by T and X by Z in formula (10.4), we could define the explained variance of T given Z by $\Omega_{T|Z}^2 = \text{Var}E(T|Z)/\text{Var}(T)$. Intuitively, this seems to be what we should use to represent the predictive ability of the proportional hazards model. Indeed, when we construct a survival model using data, one of our goals is to predict an individual's survival, taking into account the values of some of their characteristics given in the vector Z . We would therefore like to calculate the predictive power of Z on T . The following result throws some light on why this coefficient may not be the most useful measure of predictive ability in the proportional hazards model context.

Proposition 10.1. *Under the proportional hazards model with one covariate $Z \in \mathbb{R}$ and a constant risk $\lambda_0(t) = a$, $a > 0$, we have $\Omega_{T|Z}^2 \leq 1/2$.*

If the regression coefficient β in the proportional hazards model is large, this means that the association between T and Z is strong and the probability that an individual dies earlier or later will be strongly influenced by the value of Z . As this value influences survival to a greater extent as the value of the regression coefficient increases (in absolute value), the predictive power of Z on T should increase when $|\beta|$ does, and not be bounded by a value strictly less than 1. Proposition 10.1 prevents this happening.

Furthermore, using formula (10.14), we can show that if Z is Bernoulli with parameter 1/2, then $\Omega_{T|Z}^2$ is bounded above by 1/3, and that this upper bound changes as the Bernoulli parameter does. Thus, the comparison of two values of $\Omega_{T|Z}^2$ for different marginal distributions of Z is difficult since the coefficients are not on the same scale. This is illustrated in Figure 10.3 with conditional survival curves under the proportional hazards model with $\beta = 1.5$ or $\beta = 4$, $\lambda_0(t) = 1$, and a binary covariate Z . Notice that the conditional survival curves are further away from each other when β is larger. This should improve the prediction quality since the covariate is more strongly linked to survival. However, using equation (10.14), we can show that for $\beta = 1.5$ and $Z \sim \mathcal{B}(0.5)$, and for $\beta = 4$ and $Z \sim \mathcal{B}(0.3)$, the value of the explained variance is the same: $\Omega_{T|Z}^2 = 0.22$.

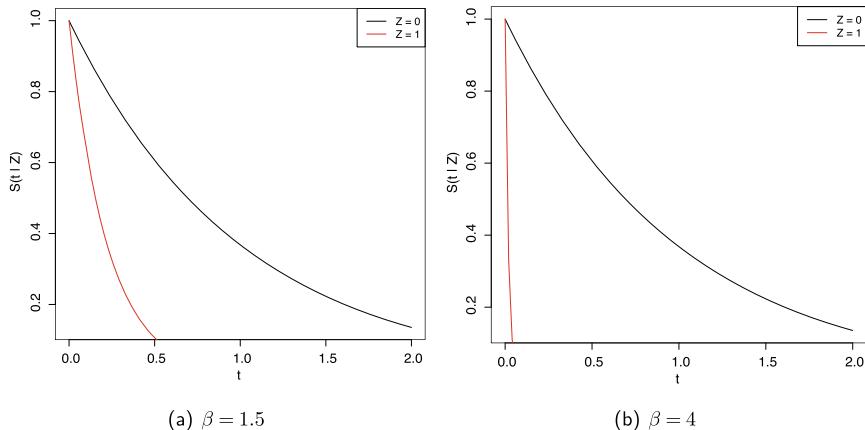


Figure 10.3: Conditional survival curves $S(\cdot|Z)$ as a function of time under the proportional hazards model with $\lambda_0(t) = 1$, $\beta = 1.5$ or 4, and binary covariates.

Also, O'Quigley and Flandre (1994) have shown that a non-linear transform which changes the time axis but keeps the order of deaths the same alters the explained variance of T given Z . For proportional hazards models however, changing the time scale in this way does not change the estimation of the regression coefficient β . The predictive ability of the model should not change. The explained variance of T given Z would therefore seem to be a poor indicator of

the predictive power of variables in proportional hazards models, and even less so in non-proportional ones. Let us now look at the explained variance of Z given T .

THE EXPLAINED VARIANCE OF Z GIVEN T

Recall that $Z = Z(t)$ is a vector containing p possibly time-dependent covariates. The explained variance of Z given T under the non-proportional hazards model is not defined in the same way for the univariate case ($p = 1$) as for the multivariate ($p > 1$) one. In what follows, we will only look at the explained variance based on the distribution of Z conditional on T , and to simplify notation, we will not always specify the conditional nature of this distribution.

Definition 10.2. *In the univariate non-proportional hazards model, the explained variance, given as a function of the time-dependent regression coefficient $\beta(t)$, is defined as:*

$$\Omega^2(\beta(t)) = \frac{\text{Var}(E_{\beta(t)}(Z|T=t))}{\text{Var}(Z)} = 1 - \frac{E(\text{Var}_{\beta(t)}(Z|T=t))}{\text{Var}(Z)}. \quad (10.5)$$

When $Z(t)$ is a vector containing $p > 1$ covariates, at time t an individual i is characterized by its real-valued prognostic index $\eta_i(t) = \beta(t)^T Z_i(t)$, which is a realization of the random variable $\eta(t) = \beta(t)^T Z(t)$. Hence, the model's predictive ability can equally be calculated using either Z or η (Altman and Andersen, 1986; Andersen et al., 1983). Note also that any time dependency in $\beta(t)$ becomes very quickly intractable when working with the conditional distribution of T given Z whereas, for the conditional distribution of $Z(t)$ given $T = t$, no additional concerns are raised.

Definition 10.3. *The explained variance of the non-proportional hazards model with more than one covariate can be defined as a function of β :*

$$\Omega^2(\beta(t)) = \frac{\text{Var}(E(\eta|T))}{\text{Var}(\eta)} = 1 - \frac{E(\text{Var}(\eta|T))}{\text{Var}(\eta)}, \quad \eta(t) = \beta(t)^T Z(t). \quad (10.6)$$

We can define the expectation with respect to the distribution of T of the variance of the covariate or the prognostic index evaluated with the coefficient $\alpha_2(t)$ under the non-proportional hazards model with p -dimensional parameter $\alpha_1(t)$. In most cases we will have, $\alpha_2(t) = \alpha_1(t) = \hat{\beta}(t)$ and the extra flexibility is not needed. However, for testing some null hypotheses and for building particular tests having certain optimality properties it is useful to have this added generality. With this in mind, we write:

$$Q(F, \alpha_1(t), \alpha_2(t)) = \int_0^1 E_{\alpha_1(t)} \left(\{Z(t) - E_{\alpha_2(t)}(Z|T=t)\}^2 \middle| T=t \right) dF(t) \quad (10.7)$$

where F is the cumulative distribution function of T . It is always easier to think in terms of a univariate “response” variable, $Z(t)$, as in the above expression. In practice we will often be focused on multivariate outcomes and this can come in one of two forms. The first is again just a univariate “response” where the other variables have been accounted for via the model. The second form is where we summarize the vector of outcome variables as a single variable, in particular a linear sum, $\alpha_2(t)^T Z(t)$ and, usually, the prognostic index itself so that $\alpha_2(t) = \hat{\beta}(t)$. We then have:

$$\Omega^2(\beta(t)) = 1 - Q^{-1}(F, 0, \beta(t)) Q(F, \beta(t), \beta(t)).$$

This way of writing the explained variance will be used later to construct an estimator. The difficulties we had with the explained variance based on the distribution of T given Z in the previous section disappear when considering the explained variance of Z conditional on T , thanks to the following result.

Proposition 10.2. *Under the non-proportional hazards model with parameter $\beta(t)$, and for a p -dimensional covariate vector ($p \in \mathbb{N}^*$), the following three statements hold:*

1. $\Omega^2(0) = 0$ and $0 \leq \Omega_{Z|T}^2(\beta(t)) \leq 1$,
2. Ω^2 is invariant for transformations that are strictly increasing in T as well as for transformations that are linear in Z ,
3. If $\beta(t) = \beta_0$ is constant and $p = 1$, Ω^2 is an increasing function of $|\beta_0|$ and $\lim_{|\beta| \rightarrow \infty} \Omega^2(\beta) = 1$.

Proof of Proposition 10.2. In Chapter 13 of O’Quigley (2008) and in O’Quigley and Xu (2001), Ω^2 is a function of a constant regression parameter, corresponding to the proportional hazards model. Extending this to the time-dependent parameter case is trivial. For the multivariate case, see Xu (1996). \square

Generalizing the third statement above requires a little thought. Since $\beta(t)$ is no longer constant it may not be immediately clear what we mean when we talk of increasing β . The following is however true. For any model based on $\beta(t)Z$, there exists an equivalent model based on $\beta^*Z^*(t)$ in which the time dependency has now been absorbed by $Z^*(t)$. This is a purely formal manipulation but allows us to place any non-proportional hazards model under a proportional hazards heading. As a result we see that, given $Z^*(t)$, it is true that Ω^2 is an increasing function of $|\beta^*|$. What we have is that part 3 of Proposition 10.2 holds in a particular direction of the functional coefficient space, the direction that remains within the functional form of the model. The explained variance coefficient remains unchanged when applying strictly increasing transformations in time. Only the ranks matter. In consequence, we can continue to work with the standardized

time scale described in Chapter 9. We view the explained variance as a population parameter that, typically, needs to be estimated.

10.7 The R^2 estimate of Ω^2

Numerous measures of the predictive ability of proportional hazards models have been proposed in the literature (Choodari-Oskooei et al., 2012). Some of these are called “ R^2 coefficients”, which classically refer to estimators of Ω^2 or “explained variance”, without actually being estimators of the explained variance as defined in Section 10.6. Most of these measures are entirely ad hoc and constructed without any theoretical justifications. Furthermore, they are almost always defined for proportional hazards models, but for the most part their behavior in the presence of poorly-specified models has not been studied. Corresponding measures for the situation of non-proportional hazards are not usually developed and it is not clear how any of these measures work in the situation of non-proportional hazards. For example, when considering rival candidate non-proportional hazards models, we would like the model that is closer to the mechanism generating the observations to result in a greater value of R^2 . It is not our goal here to thoroughly review all such estimators. Instead, we will concentrate on a specific estimator of Ω^2 : an R^2 coefficient constructed with the same residuals as the regression effect process, introduced by O’Quigley and Flandre (1994).

After defining Ω^2 , we will extend the definition to non-proportional hazards models. We also show Ω^2 to have all of the desirable properties that we would hope for such a measure. Some simple but strong results indicate why the R^2 coefficient, a consistent estimator of Ω^2 , based on the regression effect process turns out to be a fundamental tool in model construction. It guides us away from poorly fitting models toward better fitting models, and ultimately, will assume its greatest value when the chosen model coincides with the mechanism that can be viewed as having truly generated the observations.

In order to obtain our needed theorems we extend the Schoenfeld residuals, defined in Equation (7.43), to the non-proportional hazards model with parameter $\beta(t)$. Furthermore, we keep in mind that this is carried out on the transformed scale of the previous chapter so that:

$$r_{\beta(t)}(t) = \mathcal{Z}(t) - \mathcal{E}_{\beta(t)}(Z|t), \quad 0 \leq t \leq 1,$$

where $\mathcal{Z}(t)$ is the value of the covariate of the individual who has died at t and $\mathcal{E}_{\beta(t)}(Z|t)$ its expectation under the model. Let us denote \tilde{F} the estimator of the empirical cumulative distribution function of T on the transformed scale:

$$\tilde{F}(t) = \frac{1}{k_n} \bar{N}^*(t) = \frac{1}{k_n} \sum_{i=1}^n \mathbf{1}_{\phi_n(X_i) \leq t, \delta_i = 1}, \quad 0 \leq t \leq 1.$$

In the presence of uncensored data, \tilde{F} corresponds to the usual estimator of the empirical cumulative distribution function of T on the transformed scale. The value $Q(F, \alpha_1(t), \alpha_2(t))$ in equation (10.7) can be estimated in the univariate case by:

$$\tilde{Q}(\tilde{F}, \alpha_1(t), \alpha_2(t)) = \int_0^1 \{r_{\alpha_1(s)}(s)\}^2 d\tilde{F}(s) = \frac{1}{k_n} \sum_{i=1}^{k_n} \{r_{\alpha_1(t_i)}(t_i)\}^2. \quad (10.8)$$

where, although no role is being played by $\alpha_2(t)$, we keep it as an argument to $\tilde{Q}(\tilde{F}, \alpha_1(t), \alpha_2(t))$ in order to keep the notation simple. The reason for this is that, in the multivariate case, where we work with the same equation but replace $r_{\alpha_1(s)}(s)$ by $\alpha_2(s)^T r_{\alpha_1(s)}(s)$, we can put both cases under the same heading.

We now have the ingredients we need for the main definition of this section.

Definition 10.4. *The R^2 is defined as $R^2 = R^2(\hat{\beta}(t))$, where $\hat{\beta}(t)$ is a convergent estimator of $\beta(t)$, and for any function α from $[0, 1]$ to \mathbb{R}^p ,*

$$R^2(\alpha(t)) = 1 - \tilde{Q}^{-1}(\tilde{F}, 0, \alpha(t)) \tilde{Q}(\tilde{F}, \alpha(t), \alpha(t)) \quad (10.9)$$

With one covariate in the model, the R^2 coefficient is the ratio of the sum of the squared Schoenfeld residuals for the model with parameter $\hat{\beta}(t)$, over the sum of the squared residuals for the model whose parameter is zero. When instead the model has several covariates, the coefficient is still the ratio of sums of squared residuals, but the residuals are now evaluated using the prognostic indicator $\hat{\beta}(t)^T Z(t)$ rather than the covariate $Z(t)$ directly. A quantity very closely related to $R^2(\alpha(t))$ is $R_S^2(\alpha(t))$ defined as:

$$R_S^2(\alpha(t)) = 1 - \hat{Q}^{-1}(\hat{F}, 0, \alpha(t)) \hat{Q}(\hat{F}, \alpha(t), \alpha(t)), \quad (10.10)$$

the only change from Equation 10.9 being that \tilde{F} is replaced by \hat{F} . In practical work this is only very rarely of interest and we can safely ignore $R_S^2(\alpha(t))$ and only keep in mind $R^2(\alpha(t))$. The purpose of $R_S^2(\alpha(t))$ is for theoretical study. In the absence of censoring $R_S^2(\alpha(t))$ and $R^2(\alpha(t))$ coincide, but when censoring is present, we will need work with $R_S^2(\alpha(t))$ in order to demonstrate large sample consistency. The explanation for this is that the estimator \tilde{F} does not converge to the cumulative distribution function of T . As a consequence the R^2 coefficient will not generally converge to the explained variance Ω^2 . Noting that $d\hat{F} = -d\hat{S}$ then

$$d\hat{S}(t_i) = \hat{S}(\phi_n^{-1}(t_i)) - \hat{S}(\phi_n^{-1}(t_i)^-)$$

is the jump in the estimator at time of death t_i for $i = 1, \dots, k_n$ (Equation 3.13). This weighted modification to the R^2 coefficient with jumps from the Kaplan-Meier estimator allows us to obtain convergence of $R_S^2(\hat{\beta}(t))$ to the proportion of explained variance Ω^2 when $\hat{\beta}(t)$ is a convergent estimator of $\beta(t)$ (O'Quigley

and Xu, 2001; Xu, 1996). However, in real-world settings, it has been observed that R^2 depends very weakly on the censoring, even for rates as high as 90% and that, in the great majority of applications, the standard deviation of R^2 will be an order of magnitude greater than any bias. Simulation work by Choodari-Oskooei et al. (2012) has confirmed the independence of the unweighted R^2 coefficient with respect to censoring, and a higher variance for $R_{\hat{S}}^2$ in return for its lack of bias. For these reasons, in our work we mostly work with the unweighted R^2 coefficient in real applications.

The coefficient of explained variance $\Omega^2(\beta(t))$ can be estimated with $R^2(\hat{\beta}(t))$, where $\hat{\beta}(t)$ is a consistent estimator of the true regression parameter $\beta(t)$. The following theorem justifies the use of the R^2 coefficient when evaluating the goodness of fit of the non-proportional hazards model, and gives its asymptotic behavior under poorly-specified models.

Theorem 10.1. *Under the non-proportional hazards model with parameter $\beta(t)$, if $p = 1$ and conditions A1, A2, A3 and A4 of Section 9.4 hold, then there exists a constant $C(\beta) > 0$ such that for all $\alpha \in \mathbb{B}$:*

$$\lim_{n \rightarrow \infty} R^2(\alpha(t)) = 1 - \frac{C(\beta) + \int_0^1 (e(\alpha(t), t) - e(\beta(t), t))^2 dt}{C(\beta) + \int_0^1 (e(0, t) - e(\beta(t), t))^2 dt}, \quad (10.11)$$

and

$$\arg \max_{b \in \mathbb{B}} \lim_{n \rightarrow +\infty} R^2(b(t)) = \beta \quad \text{a.s.}$$

If $p > 1$ and conditions B1, B2, B3 and B4 of Section 9.5 hold, for all $\alpha \in \mathbb{B}'$ with $\alpha \neq 0$, we have:

$$\lim_{n \rightarrow \infty} R^2(\alpha(t)) = 1 - \frac{\int_0^1 \alpha(t)^T \Sigma \alpha(t) dt + \int_0^1 (\alpha(t)^T \{e(\beta(t), t) - e(\alpha(t), t)\})^2 dt}{\int_0^1 \alpha(t)^T \Sigma \alpha(t) dt + \int_0^1 (\alpha(t)^T \{e(\beta(t), t) - e(0, t)\})^2 dt}.$$

For $t \in [0, 1]$ and γ a function from $[0, 1]$ to \mathbb{R}^p , recall the definition:

$$e(\gamma(t), t) = s^{(1)}(\gamma(t), t)/s^{(0)}(0, t).$$

Theorem 10.1 says that in the univariate case, if $\beta(t)$ is the true regression coefficient and if the sample size is large enough, the maximum of the R^2 function is reached for the true parameter $\beta(t)$. Generalizing this to the multivariate case is not a simple task. Consider the two covariate case $Z^{(1)}$ and $Z^{(2)}$ with effects $\beta_1(t)$ and $\beta_2(t)$. If we suppose that $\beta_2(t)$ is known, this becomes equivalent to estimating $\beta_1(t)$ in the univariate model. Theorem 10.1 then applies and $\beta_1(t)$ reaches the maximum for R^2 in the limit. In practice, $\beta_2(t)$ is not known but we can get close to it using a convergent estimator. By conditioning on this

estimator, i.e., treating $\beta_2(t)$ as a known constant, the result again holds for $\beta_1(t)$. We can therefore consider than the result of Theorem 10.1 applies for each variable taken separately, conditional on convergent estimators for all of the other regression coefficients.

The R^2 coefficient and the regression effect process are constructed using the same residuals. The regression effect process allows us to check the goodness of fit of the non-proportional hazards model, while the R^2 coefficient is a measure of the model's predictive ability. Though these aspects are different, their construction using the same quantities is natural. This means that in essence all of the relevant information regarding the predictive power of a model as well as model adequacy is contained in the regression effect process. It is natural to expect that introducing additional covariates into any model will result in some increase, if only an apparent one, in predictive power. Here, we see that for any given set of covariables, improvements in fit will result in improvements in predictive power. Formal theorems enable us to safely rely on this result as a means to guide model construction.

10.8 Using R^2 and fit to build models

Using the results of Theorem 9.3 and its corollary, the regression effect process U_n^* can be used to determine the form of the multivariate regression coefficient $\beta(t)$. The process's drift will mirror the form of $\beta(t)$. No other techniques like local smoothing, projection onto basis functions, or kernel estimation are necessary (Cai and Sun, 2003; Hastie and Tibshirani, 1990; Scheike and Martinussen, 2004). For example, as illustrated in Figure 9.4a, an effect which is constant until time τ , followed by a zero-valued effect is easy to spot, even for small sample sizes. This simple notion generalizes immediately. Suppose that the p components of a time-dependent regression parameter

$$\beta(t) = (\beta_1(t), \dots, \beta_p(t))$$

can be written $\beta_j(t) = \beta_{0,j} h_j(t)$ for $j = 1, \dots, p$, with

- $\beta_0 = (\beta_{0,1}, \dots, \beta_{0,p}) \in \mathbb{R}^p$ an unknown constant (over time) regression parameter,
- $h = (h_1, \dots, h_p)$ a known function from $[0, 1]$ to \mathbb{R}^p .

Chauvel (2014) obtains a convergent estimator of $\beta(t)$ as $\widehat{\beta}(t) = (\widehat{\beta}_1(t), \dots, \widehat{\beta}_p(t))$ with $\widehat{\beta}_j(t) = \widehat{\beta}_{0,j} h_j(t)$ ($j = 1, \dots, p$), where $\widehat{\beta}_0 = (\widehat{\beta}_{0,1}, \dots, \widehat{\beta}_{0,p})$ is the maximum partial likelihood estimator of β_0 . This estimate can be obtained by transferring the temporal part h of the regression coefficient into the covariate (Cox, 1972). This function h can be determined graphically using the regression effect process (see the examples in Section 10.9). For example, if the j th component of the

drift is a linear function, then the function h_j is constant. If the drift is concave, h_j is decreasing, and if convex, increasing. Also, the confidence bands defined in Section 10.4 can help to evaluate the plausibility of a constant effect over time for $\beta_j(t)$, resulting in a constant function h_j for $j = 1, \dots, p$.

In the presence of non-proportional hazards, we need a more complex strategy, one that takes on board both the predictive strength of the covariates as well as the accuracy of the modeling of the time dependency. We propose using the R^2 coefficient which—on top of indicating the predictive abilities of the model—also converges to a function whose maximum is attained when the true function h is chosen (Theorem 10.1). When several models provide plausible suggestions for h , the estimators $\hat{\beta}(t)$ and corresponding R^2 coefficients should be evaluated. The time-dependent function h retained is the one that maximizes the R^2 . In this way, we obtain a non-proportional hazards model with a good fit and maximal predictive ability. The measure of predictive ability is maximized over the set \mathbf{B} of m preselected time-dependent regression effects. More formally, denote $\mathbf{B} = \{\beta_1(t), \dots, \beta_m(t)\}$ the set of m functions from $[0, 1]$ to \mathbb{R}^p . The regression coefficient $\beta^*(t)$ selected is the one for which

$$\beta^*(t) = \arg \max_{\alpha(t) \in \mathbf{B}} R^2(\alpha(t)).$$

Note that the set \mathbf{B} may be arbitrarily large and not even limited to countable sets, albeit such a level of generalization would not bring any practical benefit. The following theorem shows the equivalence between this maximization problem when $n \rightarrow \infty$ and one in which we minimize an L^2 norm.

Theorem 10.2. (Chauvel and O’Quigley 2017). *Let $p = 1$. Suppose that conditions A1, A2, A3 and A4 hold, $\beta(t) \in \mathbb{B}$ and $\mathbf{B} \subset \mathbb{B}$. Under the non-proportional hazards model with regression parameter $\beta(t)$, where $\beta(t)$ is not necessarily in \mathbf{B} , the coefficient*

$$\beta^*(t) = \arg \max_{\alpha(t) \in \mathbf{B}} \lim_{n \rightarrow \infty} R^2(\alpha(t))$$

is a solution to

$$\beta^*(t) = \arg \min_{\alpha(t) \in \mathbf{B}} \|\beta(t) - \alpha(t)\|_2,$$

where for $\alpha(t) \in \mathbb{B}$, $\|\beta(t) - \alpha(t)\|_2^2 = \int_0^1 (\beta(t) - \alpha(t))^2 dt$.

In other words, for a large enough sample size, selecting the regression coefficient which maximizes the R^2 is the same as choosing the coefficient closest to the true regression function in terms of L^2 distance. It is possible that the true regression coefficient, if it exists, is not in the set \mathbf{B} . We can, however, make it arbitrarily close. In any event, models are chosen to represent data either because they fit well, or for their predictive performance. Many models may correspond to

one or both of these aspects, not just the “true model”. Here, priority is given to goodness of fit, with the selection of candidate regression coefficients, and only then is a model’s predictive performance considered. We have chosen to work with the R^2 coefficient, but any measure of predictive performance that satisfies Theorem 10.1 could be considered in principle.

When the trend in the process’s drift is a concave function, the effect decreases with time, and if it is convex, the effect increases. In order to get the largest possible R^2 , it would be possible to construct a time-dependent effect which is closer and closer to the observed drift of the process, like for example a piecewise constant effect with many change-points. In general, this will lead to overfitting, and interpretation of the coefficient will not be straightforward. A compromise needs to be made between high predictive ability and simplicity of the coefficient, notably for interpretation purposes. This is comparable to the linear model situation in which the estimated explained variance—the R^2 coefficient—is positively biased and increasingly so as the model’s dimension increases. A balance needs to be found between the goal of improving predictive strength and the danger of poor prediction due to overfitting. This boils down to respecting the elementary principles of parsimonious model building.

10.9 Some simulated situations

Chauvel and O’Quigley (2017) present two distinct simulated situations. The first looks at the goodness of fit of proportional hazards models, and the second illustrates methods for constructing non-proportional hazards models with the help of the regression effect process and the R^2 coefficient. We generate samples under the non-proportional hazards model as described in the appendix. Let us consider the behavior of the goodness of fit test for the proportional hazards model presented in Section 10.4. This will be compared to standard tests which are also based on graphical methods, allowing visualization of the form of the time-dependent effect in the case of poor model fitting.

The first comparison test—proposed by Lin et al. (1993)—is based on the standardized score process defined in equation (10.2). This is different with the regression effect process defined and studied here because the increments of the process of Lin et al. (1993) are non-standardized Schoenfeld residuals summed with respect to times of death in the initial time scale, and an overall standardization is applied to all increments. The test statistic is the supremum over time of the absolute value of the globally standardized process. If the covariates are non-correlated, the statistic’s limit distribution is the Kolmogorov one (Therneau et al., 1990). If they are correlated, the limit distribution is not known but can be evaluated numerically using Monte Carlo simulation. Thus, the confidence envelope of the process and the goodness of fit test are obtained by simulating N processes according to the process’s Gaussian limit distribution. We have chosen $N = 10^3$ here. The R code for evaluating the process, the simulated envelope,

and the goodness of fit test can be found in the `timereg` package (Martinussen and Scheike, 2006). Below we refer to this test with the abbreviation LWY.

A second goodness of fit test for the proportional hazards model is due to Grambsch and Therneau (1994). They consider an effect of the form $\beta(t) = \beta_0 + \beta_1 \log(t)$ and propose a score test for the null hypothesis $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$. This test can be found in the `survival` package in R. We label this test GT in the following.

The final goodness of fit test compared here was proposed by Scheike and Martinussen (2004). Under the non-proportional hazards model, they propose a Cramér-von Mises-type test based on the cumulative value of the estimator of $\beta(t)$. This estimate is made using an algorithm involving kernel smoothing. Code for this test can be found in the `timereg` package in R. This test is labeled SM below. These authors also proposed a Kolmogorov-Smirnov test, but based on their simulations it appears less powerful, so we have not retained it here. The test based on the regression effect process introduced in Section 10.4 is denoted U_n^* in the tables taken from Chauvel (2014), Chauvel and O'Quigley (2017).

SIMPLE TWO COVARIATE EXAMPLE

Data were simulated under the proportional hazards model with two centered normal covariates $Z^{(1)}$ and $Z^{(2)}$ with variance 2 and correlation $\rho(Z^{(1)}, Z^{(2)})$ of 0, 0.3, 0.5 or 0.7. The sample size was either $n = 50$, $n = 100$ or $n = 200$. The mean rate of censoring was 25%: data were censored at the time 1.5 on the initial scale. The non-proportional hazards model is

$$\lambda(t|Z^{(1)}, Z^{(2)}) = \exp\left(0.5Z^{(1)} + \beta_2(t)Z^{(2)}\right),$$

where $\beta_2(t)$ takes the values 0, 0.5, $1_{t \leq 0.5}$, $1.51_{t \leq 0.5}$ or $0.51_{t \leq 0.5} - 0.51_{t \geq 0.5}$ on the transformed time scale. The covariate $Z^{(1)}$ respects the proportional hazards hypothesis. In the first two cases, $Z^{(2)}$ does too. In the next two, the hypothesis is not satisfied because the effect is constant at the start and becomes zero when half of the deaths have been observed. In the final case, the hypothesis is not satisfied and the type of effect changes half-way through the study (the variable's effect on survival is positive at the start and negative at the end). This situation corresponds to risks which cross over each other. For each trial setting, 3000 samples were simulated in order to evaluate the level and empirical power of the tests.

SOME SUMMARY RESULTS

Tables 10.1 and 10.2 show the results for the two proportional hazards models. We see that the level is well-estimated at 5% for each situation and each test, except for the Scheike and Martinussen test, which has a level above 5% for small sample sizes ($n = 50$ and $n = 100$).

n	$\rho(Z^{(1)}, Z^{(2)})$	LWY		GT		SM		U_n^*	
		$\beta_1(t)$	$\beta_2(t)$	$\beta_1(t)$	$\beta_2(t)$	$\beta_1(t)$	$\beta_2(t)$	$\beta_1(t)$	$\beta_2(t)$
50	0.0	6.3	5.3	5.5	4.4	7.4	7.0	3.3	2.9
50	0.3	5.4	4.8	5.2	4.6	8.1	7.0	2.8	3.1
50	0.5	5.0	4.7	5.2	4.5	6.8	6.7	3.0	2.6
50	0.7	5.7	6.5	5.0	5.1	7.2	8.8	3.1	3.2
100	0.0	5.4	5.7	4.8	5.0	5.7	6.5	3.4	3.7
100	0.3	5.5	5.2	4.9	4.7	5.9	5.9	3.4	3.6
100	0.5	5.7	5.9	4.5	4.8	5.4	6.6	3.4	3.5
100	0.7	5.4	5.4	5.7	6.1	6.5	6.7	3.9	3.7
200	0.0	4.6	5.6	4.3	5.1	5.0	5.7	3.9	3.9
200	0.3	5.5	5.7	4.9	4.8	5.2	5.4	4.5	4.2
200	0.5	6.2	5.7	4.4	4.9	4.9	5.1	3.8	3.9
200	0.7	5.5	5.2	5.7	5.0	5.3	5.0	4.6	3.8
400	0.0	5.4	5.7	4.7	4.7	4.6	4.7	4.5	4.6
400	0.3	5.6	4.5	5.1	4.4	5.0	5.2	4.4	3.6
400	0.5	4.8	5.6	4.8	5.2	5.3	5.2	4.4	5.2
400	0.7	6.0	5.2	5.1	5.3	5.7	5.9	5.1	5.0

Table 10.1: Empirical level of tests (in %) for $\beta_2(t) = 0$. Taken from Chauvel (2014).

n	$\rho(Z^{(1)}, Z^{(2)})$	LWY		GT		SM		U_n^*	
		$\beta_1(t)$	$\beta_2(t)$	$\beta_1(t)$	$\beta_2(t)$	$\beta_1(t)$	$\beta_2(t)$	$\beta_1(t)$	$\beta_2(t)$
50	0.0	5.3	5.9	4.8	4.9	7.3	6.7	3.0	2.8
50	0.3	6.1	5.4	5.0	4.5	7.8	7.0	3.2	3.0
50	0.5	5.1	5.1	5.0	4.5	7.0	7.2	2.9	2.6
50	0.7	5.4	5.5	5.1	5.0	8.1	7.7	3.1	3.0
100	0.0	5.9	5.3	4.3	4.3	5.6	5.3	3.6	3.3
100	0.3	5.0	5.7	4.9	4.3	5.3	5.0	3.5	3.8
100	0.5	5.4	5.5	4.8	4.9	6.1	6.7	3.5	3.4
100	0.7	5.4	5.8	5.6	5.4	5.4	6.0	3.6	3.6
200	0.0	5.2	5.5	3.9	4.6	5.3	5.7	3.5	3.4
200	0.3	6.0	5.6	4.8	4.1	5.7	4.9	4.1	3.6
200	0.5	5.7	5.8	4.9	4.5	5.3	5.0	4.0	4.1
200	0.7	4.6	4.9	5.1	4.4	5.2	4.7	3.4	4.2
400	0.0	4.7	5.6	4.7	5.6	4.6	5.4	3.9	4.6
400	0.3	4.2	5.0	4.5	4.4	5.2	4.6	4.2	4.5
400	0.5	6.6	5.3	5.8	4.9	5.1	5.4	4.6	5.0
400	0.7	5.7	6.3	5.2	4.9	5.1	5.3	4.9	5.1

Table 10.2: Empirical level of tests (in %) for $\beta_2(t) = 0.5$. Taken from Chauvel (2014).

n	$\rho(Z^{(1)}, Z^{(2)})$	LWY		GT		SM		U_n^*	
		$\beta_1(t)$	$\beta_2(t)$	$\beta_1(t)$	$\beta_2(t)$	$\beta_1(t)$	$\beta_2(t)$	$\beta_1(t)$	$\beta_2(t)$
50	0.0	5.3	14.9	4.4	13.6	7.2	7.8	2.6	11.3
50	0.3	6.3	14.7	5.0	12.8	7.4	7.6	3.3	9.6
50	0.5	6.1	15.1	4.6	11.9	6.8	7.6	2.7	8.3
50	0.7	9.3	16.8	5.4	11.0	6.3	7.4	2.8	7.9
100	0.0	4.6	28.3	4.8	21.2	5.2	9.0	2.4	22.4
100	0.3	6.4	27.9	5.2	20.7	5.8	8.3	4.0	21.7
100	0.5	9.4	30.6	5.0	20.3	5.5	9.1	4.1	19.8
100	0.7	14.6	30.0	5.5	15.1	5.8	7.2	3.4	13.5
200	0.0	5.5	55.5	4.6	37.9	5.4	14.7	3.7	50.4
200	0.3	8.2	57.5	4.7	38.4	4.9	13.2	4.1	50.7
200	0.5	13.9	59.2	5.0	34.0	4.4	12.8	4.6	42.7
200	0.7	29.3	59.9	4.9	23.8	5.3	9.2	3.8	28.3
400	0.0	5.3	89.6	4.7	60.8	4.4	23.4	4.3	87.6
400	0.3	11.5	90.4	5.1	62.5	5.4	25.9	4.2	86.0
400	0.5	27.7	91.2	5.6	56.0	5.1	21.5	4.6	79.6
400	0.7	56.3	92.5	5.5	46.2	4.9	16.8	4.7	63.3

Table 10.3: Empirical level ($\beta_1(t)$ column) and power ($\beta_2(t)$ column) of several tests (in %) for $\beta_2(t) = \mathbf{1}_{t \leq 0.5}$. Taken from Chauvel (2014).

In the non-proportional hazards setting (Tables 10.3, 10.4 and 10.5), an increase in correlation between covariates corresponds to an increase in the estimated level of the LWY test. This has already been pointed out by Scheike and Martinussen (2004), meaning that this test is not necessarily a good choice when covariates are correlated. Such correlation will occur often in practice. The other tests do not have this kind of problem. However, in their simulations, Scheike and Martinussen (2004) do find this type of inflation in the GT test level in the presence of correlation between covariates. This is likely because the non-constant effects $\beta_2(t)$ considered are relatively simple. The GT test supposes that the coefficient $\beta_2(t)$ is constant while the test is run for $\beta_1(t)$, so it must be used with caution. Note that the power of the GT test is equivalent or inferior to that of the goodness of fit test based on the regression effect process.

Lastly, the SM test is less powerful than the goodness of fit test based on the regression effect process in the situations considered by Chauvel (2014). Further, for smaller sample sizes ($n = 50$), the SM test provides inadequate control on the 5% level. In conclusion, the simulations show that the goodness of fit test based on the regression effect process is the most powerful among those considered in the situations looked at. This test is not affected by correlation between covariates, and its level is well-calibrated, even for small sample sizes.

		LWY		GT		SM		U_n^*	
n	$\rho(Z^{(1)}, Z^{(2)})$	$\beta_1(t)$	$\beta_2(t)$	$\beta_1(t)$	$\beta_2(t)$	$\beta_1(t)$	$\beta_2(t)$	$\beta_1(t)$	$\beta_2(t)$
50	0.0	7.0	51.6	5.6	57.4	8.9	53.4	2.9	56.3
50	0.3	5.8	44.9	5.4	53.1	7.4	47.5	3.1	49.5
50	0.5	10.0	41.9	5.9	49.2	8.0	43.1	3.4	41.9
50	0.7	14.6	37.3	5.5	37.2	7.9	31.4	2.3	27.4
100	0.0	7.5	82.2	4.9	85.4	6.0	81.9	3.4	88.0
100	0.3	7.0	80.9	5.5	83.9	6.2	80.0	4.3	86.6
100	0.5	13.9	76.5	5.6	78.6	6.4	72.4	3.8	77.9
100	0.7	26.1	70.4	5.7	63.6	5.4	56.5	2.8	58.8
200	0.0	8.7	99.0	5.3	98.9	5.1	98.6	4.6	99.7
200	0.3	8.4	98.5	6.0	98.9	5.4	97.9	4.4	99.5
200	0.5	22.3	97.5	5.9	97.4	5.1	95.8	4.5	98.6
200	0.7	53.4	96.6	5.7	92.4	5.1	88.3	3.6	92.9
400	0.0	13.7	100.0	5.0	100.0	4.9	100.0	4.6	100.0
400	0.3	11.4	100.0	5.4	100.0	4.7	100.0	4.3	100.0
400	0.5	43.5	100.0	6.1	100.0	4.6	99.9	5.0	100.0
400	0.7	84.5	100.0	5.7	99.6	4.9	99.4	5.0	100.0

Table 10.4: Empirical level ($\beta_1(t)$ column) and power ($\beta_2(t)$ column) of several tests (in %) for $\beta_2(t) = 1.5\mathbf{1}_{t \leq 0.5}$. Taken from Chauvel (2014).

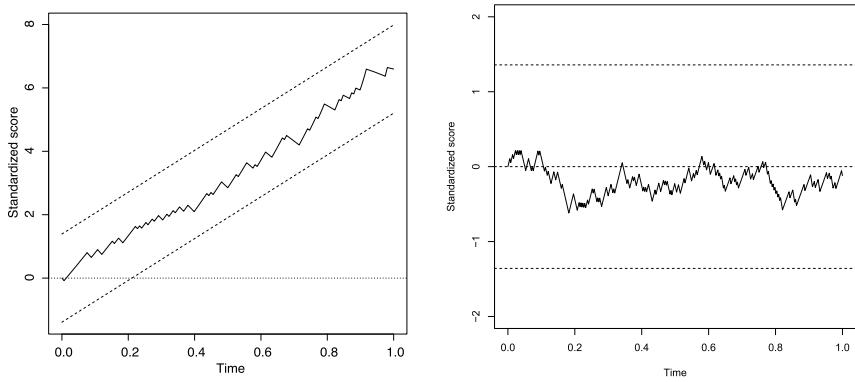
		LWY		GT		SM		U_n^*	
n	$\rho(Z^{(1)}, Z^{(2)})$	$\beta_1(t)$	$\beta_2(t)$	$\beta_1(t)$	$\beta_2(t)$	$\beta_1(t)$	$\beta_2(t)$	$\beta_1(t)$	$\beta_2(t)$
50	0.0	5.4	14.4	4.6	12.6	6.8	7.8	2.8	9.8
50	0.3	5.8	14.9	4.9	13.0	7.0	8.1	2.7	9.8
50	0.5	8.0	15.2	5.4	12.6	7.1	8.4	3.2	9.4
50	0.7	8.4	15.2	5.3	9.8	6.6	6.7	3.0	6.7
100	0.0	6.2	26.8	5.4	21.6	5.9	9.1	3.4	21.3
100	0.3	6.8	29.0	5.0	21.3	5.5	8.7	3.6	21.6
100	0.5	9.8	29.6	5.2	20.1	5.8	8.2	4.3	19.9
100	0.7	15.2	31.1	5.3	16.0	5.4	7.5	3.1	14.6
200	0.0	6.0	52.8	5.2	35.8	5.5	12.4	4.2	47.7
200	0.3	8.3	56.7	5.3	37.1	5.1	13.0	4.3	48.6
200	0.5	14.9	59.7	4.7	34.0	5.1	11.6	3.6	43.0
200	0.7	29.9	61.8	6.2	26.0	5.2	10.2	4.3	29.7
400	0.0	5.4	87.3	4.5	62.4	5.6	25.0	4.3	86.2
400	0.3	11.1	90.0	5.3	61.6	5.3	22.4	4.3	85.2
400	0.5	27.0	91.7	4.5	55.4	4.1	18.7	4.1	79.6
400	0.7	56.6	93.2	4.9	46.6	4.6	15.4	4.0	64.5

Table 10.5: Empirical level ($\beta_1(t)$ column) and power ($\beta_2(t)$ column) of several tests (in %) for $\beta_2(t) = 0.5\mathbf{1}_{t \leq 0.5} - 0.5\mathbf{1}_{t \geq 0.5}$. Taken from Chauvel (2014).

SIMPLE DEPARTURES FROM PROPORTIONAL HAZARDS MODELS

Via a number of examples, Chauvel and O'Quigley (2017) and Chauvel (2014) illustrate how to construct non-proportional hazards models starting with the regression effect process and the R^2 coefficient. To begin with, we will illustrate the method in the univariate setting before moving on to the multivariate one. We will draw confidence bands for each process as described in Section 10.4 in order to get an idea of the validity of the proportional hazards hypothesis. The R package PHeval¹ developed by Chauvel (2014) allows us to evaluate and plot the regression effect process, as well as calculate the R^2 coefficient. We will proceed with a simulated data set of size $n = 200$ and one covariate Z following a Bernoulli distribution with parameter 0.5. Censoring times are simulated independently of event times from a uniform distribution on $[0, 3]$. We first simulate a sample under the proportional hazards model with parameter 1:

$$\lambda(t|Z) = \exp(Z). \quad (10.12)$$



(a) The process $U_n^*(0, t)$ (solid line) and its confidence band (dotted lines) (b) The process $U_n^*(0, t) - tU_n^*(0, 1)$ and the supremum of Brownian bridge shown in dots.

Figure 10.4: The regression effect process $U_n^*(0, \cdot)$ (solid line) and its confidence band (dotted lines) for the simulated dataset with $\beta(t) = 1$ (left-hand) and transformed process (right-hand).

The rate of censoring is 18%. The regression effect process $U_n^*(0, \cdot)$ and its confidence band under the proportional hazards model are plotted as a function of time in Figure 10.4. We observe drift, which implies a non-zero effect. The drift would appear to be linear and the process does not exit the 95% confidence bands; thus, the proportionality hypothesis appears reasonable. As the drift is increasing, the coefficient will be positive. The estimator based on the partial maximum likelihood of the proportional hazards models gives 1.08 and the R^2

¹Available at <http://cran.r-project.org/web/packages/PHeval/index.html>

coefficient is 0.21. Note that the transformation to the Brownian bridge is valid when applied to Brownian motion with drift, as long as the drift is linear. The slope term disappears in the transformation.

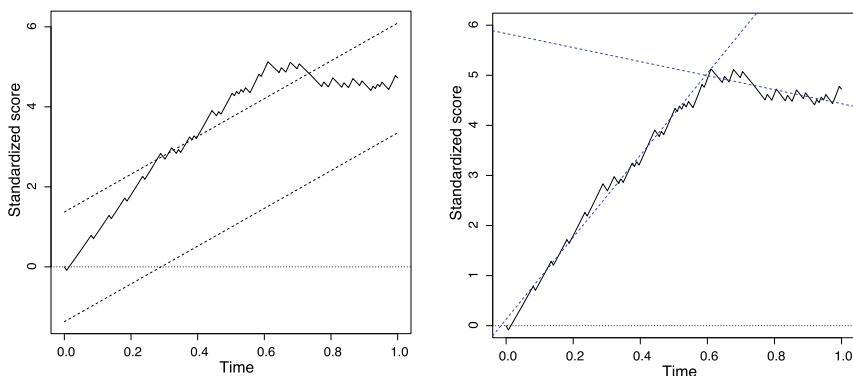
A SIMPLE CHANGE-POINT MODEL

Suppose we simulate a dataset under the change-point model

$$\lambda(t|Z) = \exp(\beta(t)Z); \quad \beta(t) = \mathbf{1}_{t \leq 0.4}$$

with t on the initial time scale. The coefficient is equal to 1 for the first part of the study, then zero afterwards. The simulated sample has 26% censoring. Figure 10.5(a) shows the regression effect process $U_n^*(0, \cdot)$ as a function of the transformed time between 0 and 1. Drift can be seen, indicating the presence of a non-zero effect. The drift does not seem to be linear and the process exits the 95% confidence band. There is evidence of an effect that changes over time. The drift appears linear until time 0.6 on the transformed time scale, and the effect appears to be weaker afterwards. This suggests that a single regression coefficient alone may not be adequate to provide a good summary of the observations. A more involved model would postulate a constant effect until time 0.6 on the transformed time scale, then zero after, denoted $\beta_{0.6a}(t) = \beta_0 \mathbf{1}_{t \leq 0.6}$. Here, the observed drift after time 0.6 corresponds to noise. The effect after 0.6 could also be non-zero, so we consider the effect

$$\beta_{0.6b}(t) = \beta_0 \mathbf{1}_{\{t \leq 0.6\}} + C_{0.6} \mathbf{1}_{\{t > 0.6\}},$$



(a) The process $U_n^*(0, \cdot)$ (solid line) and its confidence band (dotted lines)

(b) $U_n^*(0, \cdot)$ and fitted piecewise constant coefficients with a change-point at $t = 0.6$

Figure 10.5: The regression effect process $U_n^*(0, \cdot)$ for the dataset simulated according to a change-point model.

with unknown β_0 and $C_{0.6}$. The value $C_{0.6}$ is the one multiplied by the regression coefficient in the second part of the study. In Figure 10.5(b), two straight lines have been fitted to the process using linear regression, before and after $t = 0.6$. The ratio of the second one's slope to the first's gives $C_{0.6} = -0.17$.

We also consider other change-point times at $t = 0.4$, $t = 0.5$, and $t = 0.7$, and in particular the coefficients $\beta_{0.4a}(t) = \beta_0 \mathbf{1}_{t \leq 0.4}$, $\beta_{0.5a}(t) = \beta_0 \mathbf{1}_{t \leq 0.5}$ and $\beta_{0.7a}(t) = \beta_0 \mathbf{1}_{t \leq 0.7}$, which are zero in the second part of the study. The piecewise constant coefficients are not zero in the second part of the study; $\beta_{0.4b}$, $\beta_{0.5b}$ and $\beta_{0.7b}$ were considered and plotted over the process in Figure 10.6. Fitting was performed using linear regression.

The shape of the regression effect process indicates that the effect $\beta(t)$ decreases with time. Several models with continuously decreasing effects $\beta(t) = \beta_0 h(t)$ were selected together with the already mentioned coefficients. For each model, the maximum likelihood estimator $\hat{\beta}_0$ and the R^2 coefficient was calculated. The coefficients $\beta_{0.6a}(t)$ and $\beta_{0.6b}(t)$ give the largest value of R^2 here, 0.25, which corresponds to an increase of 80% in predictive ability with respect to the proportional hazards model, which had $R^2 = 0.14$. Of the two models with excellent predictive power, we retain the one with regression parameter $\beta_{0.6a}(t) = \beta_0 \mathbf{1}_{t \leq 0.6}$. Indeed, as the two models fit the data well and have good predictive ability, we would choose to retain the one with the simplest coefficient. The time $t = 0.6$ on the transformed scale corresponds to the time 0.39 on the initial scale. Recall that on the initial scale, the change-point time used to simulate the data was 0.4. The regression coefficient selected is therefore close to the one used to simulate the data (Table 10.6).

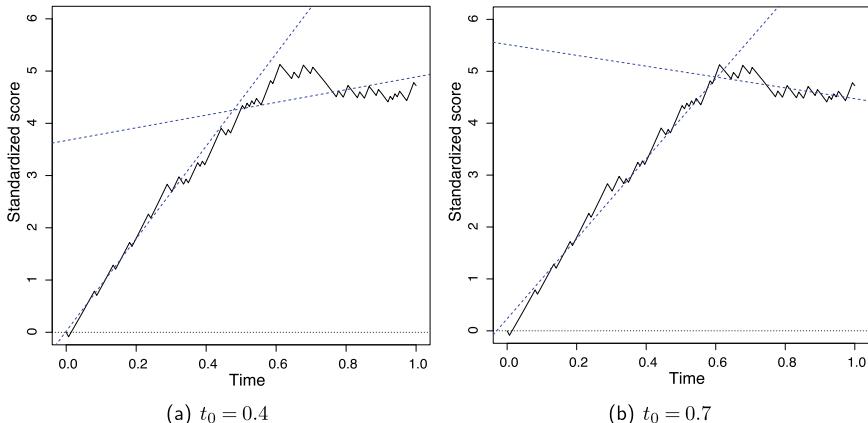


Figure 10.6: Regression effect process $U_n^*(0, \cdot)$ (solid line) and piecewise constant coefficient fits (dotted lines) with a change-point at time t_0 , for a dataset simulated according to a change-point model.

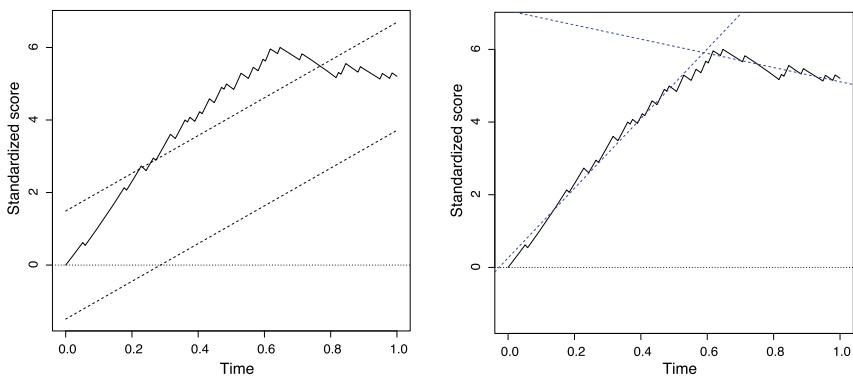
R^2	0.25	0.25	0.24	0.23	0.23	0.23
$\beta(t)$	$\beta_{0.6a}(t)$	$\beta_{0.6b}(t)$	$\beta_0(1-t)^2$	$\beta_{0.5a}(t)$	$\beta_{0.5b}(t)$	$\beta_0(1-t)$
$\hat{\beta}_0$	1.46	1.42	2.78	1.58	1.57	1.90
R^2	0.22	0.22	0.21	0.19	0.17	0.14
$\beta(t)$	$\beta_{0.7b}(t)$	$\beta_{0.7a}(t)$	$\beta_0(1-t^2)$	$\beta_{0.4b}(t)$	$\beta_{0.4a}(t)$	β_0
$\hat{\beta}_0$	1.20	1.21	1.38	1.57	1.51	0.74

Table 10.6: R^2 coefficients and partial maximum likelihood estimators $\hat{\beta}_0$ for the dataset simulated under a change-point model.

STEADILY DECREASING REGRESSION EFFECTS

Consider now the case of a continuously decreasing effect. We simulate data using $\beta(t) = 3(1-t)^2$ on the transformed scale $[0, 1]$ with $n = 200$ individuals and uniform censoring on $[0, 3]$.

The regression effect process $U_n^*(0, \cdot)$ is plotted as a function of time in Figure 10.7(a) with a confidence band within which the process would lie with high probability under a proportional hazards hypothesis. However, the process exits the 95% confidence band, indicating that the proportional hazards assumption is not wholly satisfactory. The trend in the drift is concave, with the effect appearing to decrease over time. Among other possibilities, the effect could be linear, quadratic, or even piecewise constant. For the latter, the trend would seem to be linear up to $t = 0.6$, corresponding to a constant coefficient. Then, the drift moves to smaller values, leading us to consider the coefficient $\beta_{06a}(t) = \beta_0(\mathbf{1}_{t \leq 0.6} + C_{06}\mathbf{1}_{t > 0.6})$, where β_0 and C_{06} are unknown. C_{06} is the value by which the coefficient is multiplied in the second part of the study. Using linear



(a) The process $U_n^*(0, \cdot)$ (solid line) and its confidence band (dotted lines)

(b) $U_n^*(0, \cdot)$ and fitted piecewise constant coefficients with a change-point at $t = 0.6$

Figure 10.7: The regression effect process $U_n^*(0, \cdot)$ for a dataset simulated with a continuously decreasing effect.

regression, two lines have been fitted to the process, before and after the change-point at $t = 0.6$ (Figure 10.7(b)). The ratio of the second slope to the first gives the value $C_{06} = -0.2$. We also considered the coefficient $\beta_{06b}(t) = \beta_0 \mathbf{1}_{t \leq 0.6}$, which equals zero after $t = 0.6$. Using the same methodology, we looked at piecewise constant coefficients with change-points at $t = 0.4, 0.5$ or 0.7 (see the definitions of the coefficients in the previous example). Several models with continuously decreasing effects over time were investigated.

R^2	0.36	0.34	0.34	0.34	0.31	0.31	0.31
$\beta(t)$	$\beta_0(1-t)^2$	$\beta_{0.6b}(t)$	$\beta_{0.6a}(t)$	$\beta_0(1-t)$	$\beta_{0.5a}(t)$	$\beta_0(1-t^2)$	$\beta_{0.5b}(t)$
$\hat{\beta}_0$	3.54	1.64	1.65	2.34	1.78	1.69	1.77
R^2	0.31	0.31	0.30	0.29	0.28	0.21	
$\beta(t)$	$\beta_{0.7b}(t)$	$\beta_{0.7a}(t)$	$\beta_{0.4b}(t)$	$\beta_0 \exp(-t)$	$\beta_{0.4a}(t)$	β_0	
$\hat{\beta}_0$	1.44	1.43	2.01	1.79	1.98	0.95	

Table 10.7: R^2 coefficients and maximum partial likelihood estimators $\hat{\beta}_0$ for the dataset simulated under some different NPH models. Taken from Chauvel (2014).

The R^2 coefficients and partial maximum likelihood estimates $\hat{\beta}_0$ of β_0 are shown in Table 10.7. The weakest R^2 corresponds to the proportional hazards model. The model selected is the one with the largest R^2 , corresponding here to the model associated with the coefficient $\beta(t) = \beta_0(1-t)^2$, where β_0 is estimated as 3.54. The R^2 of this model is 0.36, corresponding to an increase of 70% in predictive ability. Again, we see that the data were simulated under the model that was correctly selected.

GRAPHS IN THE MULTIVARIATE CASE

Here, we simulated two centered normal covariates $Z^{(1)}$ and $Z^{(2)}$ with variance 1 and correlation 0.5 according to:

$$\lambda(t|Z) = \exp\left(\beta_1(t)Z^{(1)} + \beta_2 Z^{(2)}\right),$$

with $\beta_1(t) = \mathbf{1}_{t \leq 0.7}$ and $\beta_2 = -1$. For this example, we chose to simulate, without censoring, $n = 200$ individuals. Each component of the bivariate process $\hat{\Sigma}^{-1/2}U_n^*(0, \cdot)$ and its corresponding 95% confidence bands are plotted as a function of the transformed time between 0 and 1 in Figures 10.8(a) and 10.8(b). The proportional hazards hypothesis for covariate $Z^{(1)}$ is rejected at the 5% level since the process exits the confidence bands. The shape of the process's drift suggests a piecewise constant effect with a change-point at $t_0 = 0.6$ on the transformed time scale. As in the univariate case, two fits were implemented using linear regression, one before $t_0 = 0.6$ and one after. The ratio of the slopes is -0.12 , so we consider the regression effect $\beta_1(t) = \beta_1 B_{0.6}(t)$, with $B_{0.6}(t) = \mathbf{1}_{t \leq 0.6} - 0.12 \mathbf{1}_{t \geq 0.6}$. Other piecewise constant regression coefficients $\beta(t) = \beta_1 B_{t_0}(t)$ were also inves-

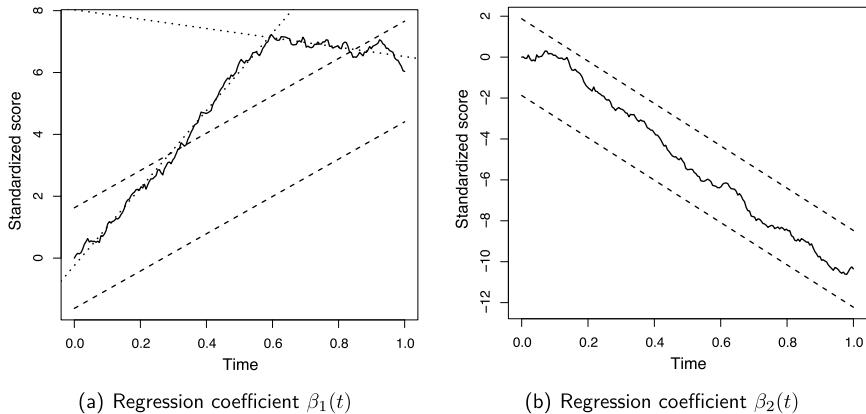


Figure 10.8: The regression effect process $\hat{\Sigma}^{-1/2}U_n^*(0, \cdot)$ (solid lines), confidence bands (dashed lines) and piecewise constant coefficient fits (dotted lines) for simulated multivariate data.

tigated, with change-points at times $t_0 \in \{0.45, 0.5, \dots, 0.7\}$ on the transformed time scale. For each time t_0 , the ratio of the slopes was calculated in order to determine the value to be multiplied by the coefficient in the second part of the study. The second covariate, $Z^{(2)}$, appears to have a constant regression effect since the process stays well inside the 95% confidence bands, and the drift appears close to being linear (Figure 10.8b). Therefore, we would choose to work with the regression coefficient $\beta_2(t) = \beta_2$ (Table 10.8).

$\beta_1(t)$	β_1	$\beta_1 B_{0.45}(t)$	$\beta_1 B_{0.5}(t)$	$\beta_1 B_{0.55}(t)$	$\beta_1 B_{0.6}(t)$	$\beta_1 B_{0.7}(t)$
$\hat{\beta}_1$	0.45	0.93	0.96	0.89	0.95	0.72
$\hat{\beta}_2$	-0.73	-0.72	-0.73	-0.74	-0.79	-0.77
R^2	0.24	0.35	0.37	0.35	0.39	0.32

Table 10.8: Partial maximum likelihood estimates $\hat{\beta}(t)$ and R^2 coefficients for simulated multivariate data. Best fitting model has $R^2 = 0.39$.

Estimation results are shown in Table 10.8. The proportional hazards model gives an R^2 of 0.24. The largest R^2 is obtained with $\beta_1(t) = \beta_1 B_{0.6}(t)$, increasing the predictive ability by 60% with respect to the proportional hazards model. In conclusion, we retain the model with $\beta_1(t) = \beta_1 B_{0.6}(t)$ and $\beta_2(t) = \beta_2$. Note that the time $t_0 = 0.6$ on the transformed scale corresponds to the time 0.68 on the initial one, which is close to the 0.7 used to simulate the data.

10.10 Illustrations from clinical studies

FREIREICH LEUKEMIA STUDY

The Freireich dataset is often referred to in the survival analysis setting (Acute Leukemia Group B et al., 1963; Cox, 1972). These observations are generally taken to respect the proportional hazards hypothesis. Half of the 42 leukemia patients in the study received a new treatment: 6-Mercaptopurine, while the other half received the control. The regression effect process $U_n^*(0, \cdot)$ and its confidence band are plotted as a function of the transformed time scale $[0, 1]$ in Figure 10.9. As expected, the process does not exit the 95% confidence band; the proportional hazards model fits the data well.

CLINICAL TRIAL IN BREAST CANCER

In a clinical trial context, data were collected from 1504 breast cancer patients at the Curie Institute, Paris. After initially successful treatment, patients were followed over a 15-year period. The rate of censoring in the study was 24%. One of the study goals was to build a descriptive survival model leading to a better understanding of the influence of various prognostic factors on post-treatment survival. As the effect of some of the factors might change over time, the resulting prognostic rules, i.e., linear combinations of effects and the prognostic factors, should reflect this.

The prognostic factors in the database were the presence of the progesterone receptor, tumor size above 60 mm, and cancer stage above 2. The multivariate regression effect process $\hat{\Sigma}^{-1/2} U_n^*(0, \cdot)$ and its confidence bands are plotted as

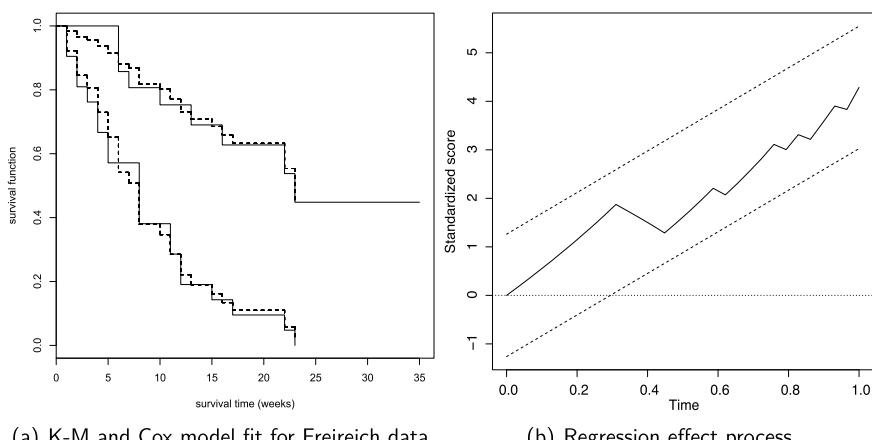
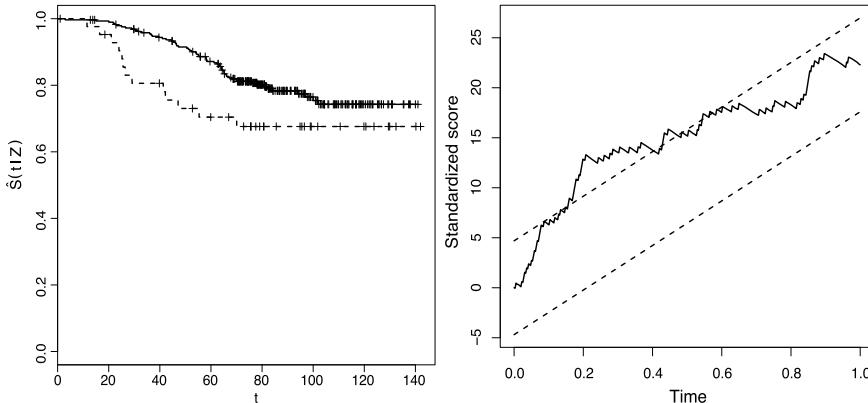


Figure 10.9: Kaplan-Meier curves and model based curves for Freireich data. Regression effect process indicates strong effects and good model fit.



(a) K-M plots for tumor size groups for breast cancer data suggestive of diminishing effects.
 (b) Regression effect process for the variable tumor size from the breast cancer study.

Figure 10.10: Kaplan-Meier curves for variable tumor size for Curie Institute breast cancer study. Regression effect process indicates clear effects that diminish gradually with time.

a function of the transformed time $[0, 1]$ in Figure 10.10. The process exits the 95% confidence band and the effect appears not to be constant, with a gradually decreasing slope over time. Hence, one model that fits the data better than the proportional hazards one has a piecewise constant effect for the tumor size variable, with a change-point at $t = 0.2$ on the transformed time scale, and constant effects for the other two prognostic factors. As in the simulations, two straight lines have been fitted to the process, before and after $t = 0.2$, which leads us to consider a regression effect for the tumor size variable of $\beta_{size}(t) = \beta_0(\mathbf{1}_{t \leq 0.2} + 0.24\mathbf{1}_{t > 0.2})$. The predictive ability of this model with respect to the proportional hazards one corresponds to an increase of more than 30% in the R^2 , moving from 0.29 to 0.39 (Table 10.9).

Figure 10.11 shows the process for the presence of progesterone receptor effect as a function of the transformed time scale. The drift is not entirely linear but the process stays within the confidence bands corresponding to a constant effect over time. We considered several potential regression effects: a change-point model with a jump at $t = 0.5$ on the transformed time scale, i.e., $\beta_{rec0}(t) = \beta_0(\mathbf{1}_{t \leq 0.5} + 0.39\mathbf{1}_{t > 0.5})$, and several continuous effects: $\beta_{rec1}(t) = \beta_0(1 - t)$, $\beta_{rec2}(t) = \beta_0(1 - t)^2$, $\beta_{rec3}(t) = \beta_0(1 - t^2)$ and $\beta_{rec4}(t) = \beta_0 \log(t)$. Figure 10.12 shows the process for the cancer grade effect. It touches the lower bound in the constant effect confidence band, but does not breach it. The drift does not seem to be linear, and gives the impression of a negative effect that decreases over time. The simplest effect possible, i.e., constant over time, would nevertheless appear to be a good candidate; however, in the model construction context, we

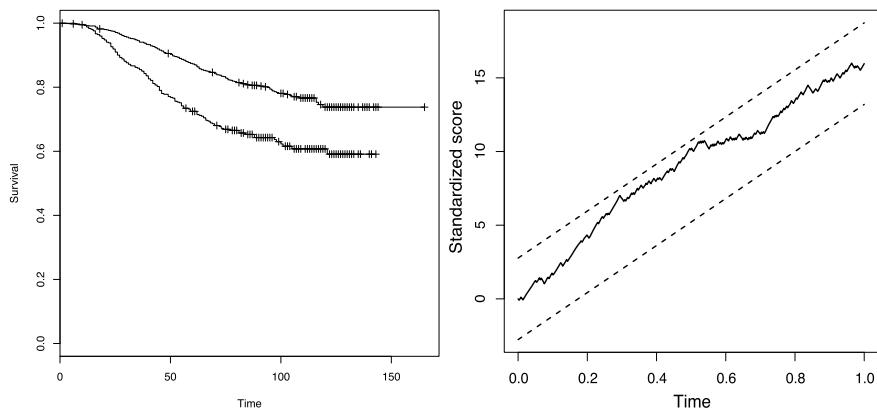
will also consider time-dependent effects. In particular, we examine the piecewise constant effect $\beta_{grade}(t) = \beta_0(t)(\mathbf{1}_{t \leq 0.4} + 0.69\mathbf{1}_{t > 0.4})$.

We looked at all combinations of these effects, as well as the proportional hazards model. For each combination, the R^2 was calculated and the regression effects were estimated by maximizing the partial likelihood.

Tumor size	Progesterone receptor	Stage	R^2
0.84	1.03	-0.68	0.29
$1.77(\mathbf{1}_{t \leq 0.2} + 0.24\mathbf{1}_{t > 0.2})$	1.03	-0.66	0.39
0.85	$-1.02\log(t)$	-0.67	0.39
$1.74(\mathbf{1}_{t \leq 0.2} + 0.24\mathbf{1}_{t > 0.2})$	$-1.02\log(t)$	-0.66	0.51
$1.72(\mathbf{1}_{t \leq 0.2} + 0.24\mathbf{1}_{t > 0.2})$	$-1.02\log(t)$	$-0.82(\mathbf{1}_{t \leq 0.4} + 0.69\mathbf{1}_{t > 0.4})$	0.52

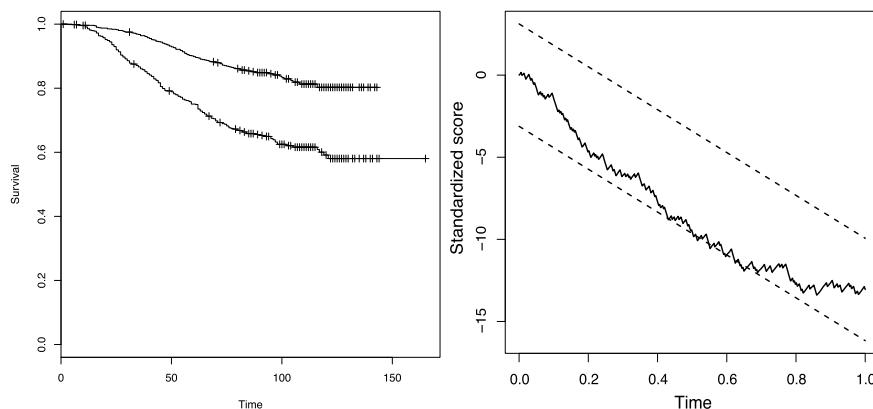
Table 10.9: Partial likelihood estimates and R^2 coefficients for the clinical trial data.

Some results are shown in Table 10.9. The proportional hazards model gives an R^2 of 0.29. As we have mentioned previously, a model allowing the tumor size effect to change in a simple way over time gives a great improvement in predictive ability (in the order of 30%). The largest R^2 is obtained with piecewise constant effects for tumor size and grade, and an effect proportional to $\log(t)$ for the progesterone receptor. The predictive ability of this model is 80% better than the proportional hazards one for the three prognostic factors. If we instead introduce a time-dependent effect for the cancer grade while also allowing the tumor size and progesterone effects to change over time, we get an increase in



(a) K-M plots for tumor groups based on pro- (b) Mildly concave regression effect re- gesterone status for breast cancer study maining within 95% limits

Figure 10.11: Kaplan-Meier curves for variable progesterone status for breast cancer study. Regression effect process is suggestive of some weak time dependency.



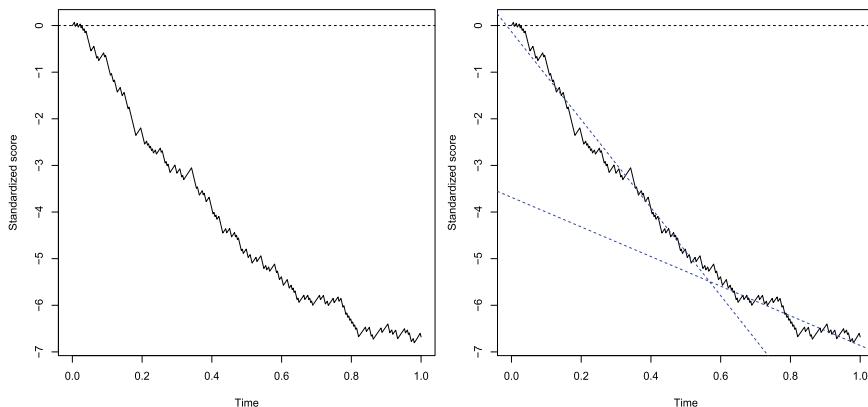
(a) K-M plots for groups based on grade for breast cancer study
(b) Regression effect process for covariate grade. PH assumption borderline.

Figure 10.12: Kaplan-Meier curves for variable tumor grade for breast cancer study. Process indicates the presence of effects that appear to diminish with time.

the R^2 from 0.51 to 0.52. Such a small improvement cannot justify the added complexity of this model; therefore, we retain the one with constant cancer stage effect and non-constant effects for tumor size and progesterone.

BUILDING MORE COMPLEX MODELS ONE STEP AT A TIME

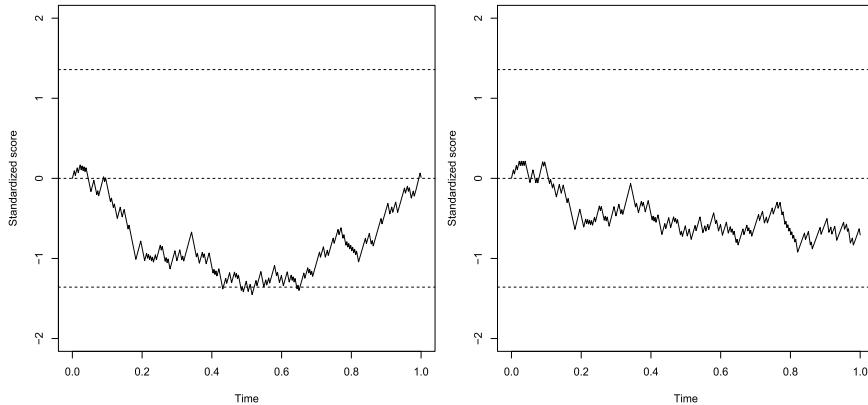
The regression effect process is a tool that allows us to construct non-proportional hazards models. The idea of the process is to use the raw data without estimating any parameters, and graphically evaluate the temporal form of the regression effect, as is the case for scatterplots in two dimensions for uncensored data. Note that processes developed until now for investigating the goodness of fit of proportional hazards models use an estimate of the regression parameter. Another difference between the regression effect process and others found in the literature is in terms of standardization. Two types are applied: one to the time, the other to increments of the process. We have chosen to standardize each increment rather than apply an overall standardization. This allows us to analytically obtain the process's limit distribution, as well as decorrelate covariates. Knowledge of the process's asymptotic distribution means that it is possible to derive a goodness of fit test and confidence bands for a constant effect in the proportional hazards model for each component of the process. These tests are not affected by correlations between covariates since this has been taken into account in the construction of the process (Figure 10.13). Related goodness of fit and predictive ability procedures that complement the above techniques for statistical modeling provide us with a coherent model construction methodology. Intuitively, models



(a) Regression effect process for covariate receptor status. Effects diminish with time.
 (b) Changepoint model for the process for receptor status. Ratio of β coefficients = 2.8.

Figure 10.13: Fitting a simple change-point model to the regression effect process for the variable receptor in the breast cancer study in order to obtain a better fit.

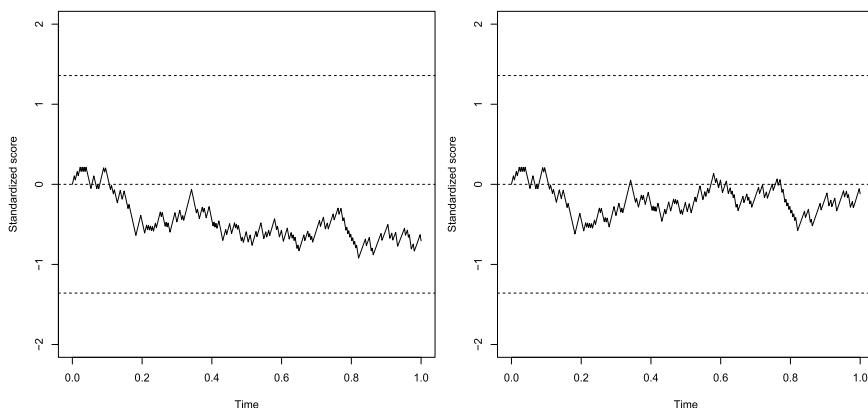
constructed in this way should have better predictive abilities, as is confirmed by the theoretical and practical results of the chapter. Change-point models, that is, non-proportional hazards (NPH) models with piecewise constant effects, are more sophisticated than proportional hazards ones. The model construction method presented here could potentially be used to estimate the change-point locations, as well as the various values of the regression parameter. In some ways, analogous to the linear model in which the addition of a quadratic term is the first step in a direction that moves away from a linearity assumption, we can see a model with a single change-point as the first step away from a PH model in the direction of NPH models. In the examples we have considered, we looked at the one change-point case, and the ratio of the two regression coefficients was estimated by linear regression using the slope of the regression effect process. It may instead be possible to estimate the two parts of the coefficient by maximizing the partial likelihood (Figure 10.14). In the method we have proposed, users play an important role in determining the regression coefficients that lead to the data being modeled satisfactorily. We chose not to estimate the parameter using basis functions like splines or histograms, so as to be able to interpret the regression coefficients more easily. The retained model will fit the data well and have good predictive abilities. There may be other candidate models, and for these, we could go through the same steps. Our viewpoint is that the model is no more than an efficient way of summarizing the infinitely complex mechanisms that can be considered to have generated the data. Necessarily, there can be more than a single candidate model, not only in terms of the inclusion of certain regressors, but also in terms of the particular model structure. The tools used here allow us



(a) Centered regression effect process for co-
variante receptor status (b) Centered regression effect process for
change-point model

Figure 10.14: Right-hand figures indicates an improvement of fit for a simple change-point model when compared to a basic PH model.

to move away from models showing poor adequacy in the direction of models with improved adequacy. The sense of adequacy is both in terms of providing a model based description that can plausibly be considered to have generated the observations, and in terms of the strength of prediction. These two senses feed off one another so that model construction is to some degree of an iterative nature. In Figure 10.15 we can see a further small improvement in the model fit



(a) Centered regression effect process for co-
variante receptor status (b) Centered regression effect process for
change-point model

Figure 10.15: Right-hand figures indicates improvement of fit for a double change-point model when compared to a simple change-point model.

by increasing the complexity of the model from a single change-point to a model with two change-points. The improvement in fit is very modest and probably not worth the extra complexity that is called for. The predictive capability of the two models barely differs. We can use the difference between the regression effect process and twice the area under the curve to evaluate the degree of deviation from the proportional hazards hypothesis (Chapter 11). If the proportional hazards model is valid, this difference ought to be small and a statistical test can readily be built to test it. When there is a single covariate, knowledge of the limit distribution of this statistic leads to the construction of a goodness of fit test. Extending this to the multivariate case is not obvious and needs further study.

Under the non-proportional hazards model with parameter $\beta(t)$, the regression effect process $\{U_n^*(\beta(t), t), 0 \leq t \leq 1\}$ converges to standard Brownian motion. It would be interesting to study the limit behavior of the process $\{\widehat{U_n^*(\beta(t), t)}, 0 \leq t \leq 1\}$, where $\widehat{\beta(t)}$ is a convergent estimator of $\beta(t)$. Knowledge of the asymptotic distribution of this process would allow us to examine whether the estimator of $\beta(t)$ selected by the model construction method is close to the true parameter of the model. If the process is evaluated at $\widehat{\beta(t)}$, where $\widehat{\beta(t)}$ is the partial maximum likelihood estimator of $\beta(t)$, then the process is no longer (\mathcal{F}_t^*) -predictable. Though our asymptotic results are based on this property, it is not simple to derive the asymptotic distribution. Nevertheless, note that the goodness of fit of the non-proportional hazards model with constant parameter $\beta_0(t)$ can be evaluated using the test with null hypothesis $H_0 : \beta(t) = \beta_0(t)$ studied in Chapter 11.

10.11 Classwork and homework

1. It is not uncommon in the applied scientific literature to encounter statements such as " R^2 indicated an excellent fit at over 70%". Explain why this statement is incorrect. Find examples from linear regression and logistic regression where poor fit is associated with a large value of R^2 and where a good fit is associated with small values of R^2 .
2. Describe the advantages of graphical goodness of fit procedures over formal goodness of fit tests.
3. The fit for a proportional hazards model with a continuous covariate may be poor in two distinct ways: a regression coefficient that turns out to be not independent of time or a poorly-specified functional form for the continuous covariate. Show that no formal test can distinguish either situation from the other.
4. To fit the drift parameter for a Brownian motion with linear drift it suffices to draw a straight line between the origin and the arrival point. Show how this idea extends to non-linear drift.

5. Show how we can approximate the drift parameter function for an observed Brownian motion with non-linear drift by using multiple change-points. How would you make use of R^2 to obtain the best possible fit while minimizing the number of change-points?
6. Given two models: one has a higher R^2 than the other but shows a significantly poorer fit. Which of the two models would you recommend? Give reasons.
7. On the basis of a real data set make use of stepwise model building techniques to obtain the model that appears to be the most satisfactory. At each step argue the basis underlying any specific decision. On what basis would we conclude that, given the information we have, we are unlikely to find a better performing model.
8. Consider a model with a single continuous covariate. For the purposes of interpretability it may be desired to use the continuous covariate to establish two distinct groups. How would the techniques of this chapter help do that? How would you decide if the two group model was preferable to the model with a continuous covariate, or vice versa?
9. Consider a model with a binary covariate. Describe the impact of increasing censoring on the regression effect process under; (1) independent censoring and (2) covariate dependent censoring, otherwise known as conditionally independent censoring.

10.12 Outline of proofs

Proposition 10.1 Under the hypotheses of Proposition 10.1, the distribution of T conditional on Z is an exponential $\mathcal{E}(\alpha \exp(\beta Z))$. Thus, $E(T|Z) = \exp(-\beta Z)/\alpha$ and $E(T^2|Z) = 2\exp(-2\beta Z)/\alpha^2$. We have therefore that:

$$\begin{aligned}\text{Var}(T) &= E(T^2) - E(T)^2 = E(E(T^2|Z)) - E(E(T|Z))^2 \\ &= \frac{1}{\alpha^2} (2E(\exp(-2\beta Z)) - E(\exp(-\beta Z))^2) = \frac{1}{\alpha^2} (2\phi_Z(-2\beta) - \phi_Z(-\beta)^2),\end{aligned}$$

where ϕ_Z is the moment generating function of Z . Also:

$$\text{Var}(T|Z) = \frac{1}{\alpha^2} \exp(-2\beta Z),$$

and

$$E(\text{Var}(T|Z)) = \frac{1}{\alpha^2} E(\exp(-2\beta Z)) = \frac{1}{\alpha^2} \phi_Z(-2\beta). \quad (10.13)$$

Using these equations and the definition of $\Omega_{T|Z}^2$, we get:

$$\Omega_{T|Z}^2 = 1 - \frac{\phi_Z(-2\beta)}{2\phi_Z(-2\beta) - \phi_Z(-\beta)^2} = 1 - \frac{1}{2 - \phi_Z(-\beta)^2 \phi_Z^{-1}(-2\beta)}. \quad (10.14)$$

As the moment generating function is positive, the result follows.

Theorem 10.2 Let $t \in [0, 1]$. Under the conditions of the theorem, a Taylor expansion of $e(\alpha(t), t)$ gives:

$$e(\alpha(t), t) = e(\beta(t), t) - (\alpha(t) - \beta(t)) v(\gamma(t), t),$$

where γ is in the ball with center α and radius $\sup_{t \in [0, 1]} |\alpha(t) - \beta(t)|$. Under A3, there exists a constant $C(\gamma)$ such that $v(\gamma(t), t) = C(\gamma)$. Using this expansion in equation (10.11), we get:

$$\lim_{n \rightarrow \infty} R^2(\alpha(t)) = 1 - \frac{C(\beta) + C(\gamma)^2 \int_0^1 (\alpha(t) - \beta(t))^2 dt}{C(\beta) + \int_0^1 (e(0, t) - e(\beta(t), t))^2 dt}.$$

As a consequence, maximizing the function $\alpha \mapsto \lim_{n \rightarrow \infty} R^2(\alpha)$ is equivalent to minimizing the function $\alpha \mapsto \int_0^1 (\alpha(t) - \beta(t))^2 dt$.



Chapter 11

Hypothesis tests based on regression effect process

11.1 Chapter summary

We revisit the standard log-rank test and several modifications of it that come under the heading of weighted log-rank tests. Taken together these provide us with an extensive array of tools for the hypothesis testing problem. Importantly, all of these tests can be readily derived from within the proportional and non-proportional hazards framework. Given our focus on the regression effect process, it is equally important to note that these tests can be based on established properties of this process under various assumptions. These properties allow us to cover a very broad range of situations. With many different tests, including goodness-of-fit tests, coming under the same heading, it makes it particularly straightforward to carry out comparative studies on the relative merits of different choices. Furthermore, we underline the intuitive value of the regression effect process since it provides us with a clear visual impression of the possible presence of effects as well as the nature of any such effects. In conjunction with formal testing, the investigator has a powerful tool to study dependencies and co-dependencies in survival data.

11.2 Context and motivation

Significance tests have an important role to play in formal decision making but also as a way to provide the shortest possible confidence intervals in a model fitting and estimation setting. There are several possible statistics for testing a null hypothesis of no difference in survival experience between groups against some alternative hypothesis. The alternative may be of a general nature or may be more specific. Specific directions moving away from the null will be mirrored,

in the regression effect process. If the form of this process can be anticipated ahead of time this will allow us to derive tailor made tests with high power for certain specific alternatives. Mostly, the null hypothesis corresponds to an absence of regression effects for two or more distinct groups, i.e., $\beta(t) = 0$ for all t , and, in this case, the test statistics often assume a particularly simple form. A classical example arises in the context of a clinical trial when we wish to compare the survival experience of patients from different treatment groups, the goal being to assess whether or not a new treatment improves survival. For the sake of simplicity, and clarity of presentation, we mostly consider the case with two groups defined by a single binary variable. Extensions to several groups raise both operational and conceptual issues. We may wish to test the impact of all variables taken together—an example is provided of a clinical trial with 3 treatment arms defined via 2 binary covariates—or we may wish to test the impact of one variable having controlled, in some way, for the effects of one or more of the other covariables. Often we will restrict the alternative hypothesis to one that belongs to the proportional hazards family.

11.3 Some commonly employed tests

Consider a straightforward clinical trial. Patients in the first group receive the new treatment, indexed by T , and patients in the second group receive a placebo, indexed by P . A comparison of the two survival functions is carried out by testing a null hypothesis H_0 against an alternative H_1 , expressed more formally as:

$$H_0 : \forall t, S_P(t) = S_T(t), \text{ versus } H_1 : \exists \mathcal{T} \text{ such that } \forall t \in \mathcal{T}, S_P(t) \neq S_T(t),$$

where $\int_{\mathcal{T}} dt > 0$. This is a little cumbersome but is needed to rule out (non-detectable and thereby not consistent) alternatives defined on a null set, i.e., a time interval with zero measure. We can write this more succinctly in terms of the parameter function $\beta(t)$ so that we consider H_0 and H_1 where:

$$H_0 : \forall t, \beta(t) = 0, \text{ versus } H_1 : \exists \mathcal{T} \text{ such that } \int_{\mathcal{T}} dt > 0 \text{ and } \int_{\mathcal{T}} \beta^2(t) dt > 0.$$

The above alternative is very general and while it will have power against any kind of departure from the null, we can increase this power, in some cases greatly increase this power, by considering more specific alternative hypotheses. This is a constant theme throughout this chapter. We focus more on two-sided tests, but one-sided tests can be structured once we figure out how to order different possible functions for $\beta(t)$. This is simple when β is constant and does not depend on t , otherwise the question itself is not always easily framed. A comparison of survival functions can also be useful in epidemiological studies, to compare for example the survival of groups of subjects exposed to different environmental

factors. In these kinds of studies we would usually expect that, under the alternative hypothesis, at any time (age), t , the difference between groups would be in the same direction.

LOG-RANK AND WEIGHTED LOG-RANK TESTS

When trying to detect a difference between survival functions, the log-rank test (Mantel, 1966) is the most commonly used. The test can be expressed as a likelihood based score test under a proportional hazards model in which the covariate is discrete (Cox, 1972). Denote $U(0)$ (resp. $-\mathcal{I}(0)$) the first (resp. second) derivative of Cox's partial log-likelihood with respect to the parameter β , calculated at $\beta = 0$. On the standardized time scale from 0 to 1, these quantities are respectively

$$U(0) = \int_0^1 \{\mathcal{Z}(s) - \mathcal{E}_0(Z | s)\} d\bar{N}^*(s) = \sum_{j=1}^{k_n} \{\mathcal{Z}(t_j) - \mathcal{E}_0(Z | t_j)\},$$

and

$$\mathcal{I}(0) = \int_0^1 \mathcal{V}_0(Z | s) d\bar{N}^*(s) = \sum_{i=1}^{k_n} \mathcal{V}_0(Z | t_i).$$

By definition, the test with statistic $U(0)^2/\mathcal{I}(0)$ is a score test for testing $H_0 : \beta = 0$ in the proportional hazards model with parameter β .

Definition 11.1. *The log-rank test rejects H_0 with Type 1 risk α if $|L_n| \geq z^{\alpha/2}$, where $z^{\alpha/2}$ is the upper quantile of order $\alpha/2$ of the standardized normal distribution, and*

$$L_n = \frac{U(0)}{\sqrt{\mathcal{I}(0)}} = \sum_{j=1}^{k_n} \frac{\mathcal{Z}(t_j) - \mathcal{E}_0(Z | t_j)}{\sqrt{\sum_{i=1}^{k_n} \mathcal{V}_0(Z | t_i)}} \quad (11.1)$$

is the test statistic.

When $\beta \in \mathbb{R}^p$, the statistic L_n^2 is defined by $L_n^2 = U(0)^T \mathcal{I}(0)^{-1} U(0)$ and converges to a χ^2 distribution with $p-1$ degrees of freedom under the null hypothesis. Written in this way, the log-rank test can be extended to continuous variables (e.g., age) to test for their effect on survival. The test statistic is easy to implement, and the p -value of the test can be accurately approximated since the asymptotic distribution is known. The great popularity of this test is in particular due to it being the most powerful (Peto and Peto, 1972) under local alternative hypotheses of the proportional hazards type, which are written

$$H_1: \quad \exists \beta_0 \neq 0, \forall t, \beta(t) = \beta_0 \neq 0.$$

This test is powerful for detecting the presence of an effect β that is constant in time between the groups, under the proportional hazards model. On the other hand, in the presence of non-proportional hazards, i.e., an effect $\beta(t)$ which varies over time, the log-rank test is no longer optimal in the sense that there exist alternative consistent tests with greater power. This is the case, for instance, in the presence of hazards that intersect, which happens when the coefficient changes sign during the study period (Leurgans, 1983, 1984).

Definition 11.2. *The weighted log-rank test has the statistic LW_n given by*

$$LW_n = \sum_{j=1}^{k_n} W_n(t_j) \frac{\mathcal{Z}(t_j) - \mathcal{E}_0(Z | t_j)}{\sqrt{\sum_{i=1}^{k_n} W_n(t_i)^2 \mathcal{V}_0(Z | t_i)}}, \quad (11.2)$$

where W_n is a function with positive weights, defined on $[0, 1]$.

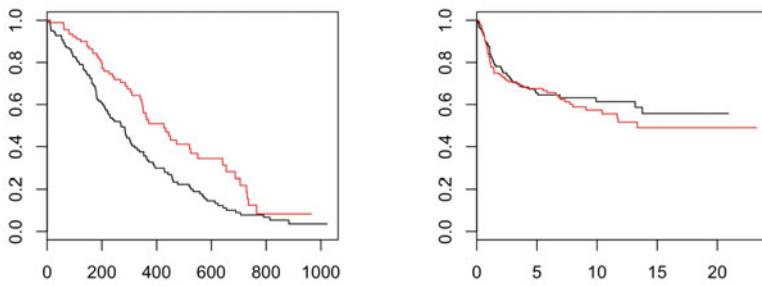
If we wish to use the techniques of martingale theory to study these kinds of weights then we will usually need some added restrictions, the most notable being that W_n be \mathcal{F}_t^* -predictable. In words, at time t we only use information available up to that time point and we do not let $W_n(t)$ depend on outcomes that can take place at values greater than or equal to t . Specific choices of weights lead to the ability to detect different types of temporal effect $\beta(t)$. Recall that the transformation ϕ_n^{-1} , defined on $[0, 1]$, lets us map times $t \in [0, 1]$ back to the initial time scale (see Section 9.3). Several suggestions for weights for particular situations have been presented, among which:

- $W_n(t) = \sum_{i=1}^n Y_i(\phi_n^{-1}(t))$, the number of individuals at risk at time t (Breslow, 1970; Gehan, 1965),
- $W_n(t) = \hat{S}(\phi_n^{-1}(t))$, the Kaplan-Meier estimator of survival at time t (Prentice, 1978),
- $W_n(t) = \tilde{S}(\phi_n^{-1}(t))$, where \tilde{S} is a modified version of the Kaplan-Meier estimator (Prentice, 1978),
- $W_n(t) = g\left\{\sum_{i=1}^n Y_i(\phi_n^{-1}(t))\right\}$, for a given function g (for example, $g(x) = \sqrt{x}, \forall x$) (Tarone and Ware, 1977),
- $W_n(t) = \hat{S}(\phi_n^{-1}(t))^p \left\{1 - \hat{S}(\phi_n^{-1}(t))\right\}^q$, $p, q \geq 0$, which are weights of the family described in Fleming and Harrington (1991, 2005).

Many other weights have been proposed (Lagakos et al., 1990; Mantel and Stablein, 1988; Wu and Gilbert, 2002). Note that the test with the weights proposed in Gehan (1965) corresponds to a modification of the Wilcoxon (1945) test that takes censored data into account. This test is equivalent to the two sample Mann-Whitney test and to the Kruskal-Wallis test when there are more than two

samples (Breslow, 1970). Aalen (1978) and Gill (1980) have shown that weighted log-rank statistics converge to centered normal distributions. Jones and Crowley (1989, 1990) have defined a broader class of tests that includes weighted and unweighted log-rank tests. This class also includes tests for comparing more than two survival functions, as well as specific constructions such as a trend across groups.

All of these weighted log-rank tests were developed on the initial time scale. However, since they are rank tests, they remain unchanged on the transformed time scale $[0, 1]$ using the mapping ϕ_n from Equation (9.2). When the weights are larger for earlier times, we will put more emphasis there so that such a test ought to be more sensitive to the detection of early effects. This can be expressed as the coefficients $\beta(t)$ decreasing with time (Fig. 11.1).



(a) Apparently strong effects diminish through time causing significant losses in statistical power.
 (b) No apparent effects before median. Either there are no effects or effects delayed in time.

Figure 11.1: Examples of non-proportionality detectable from the Kaplan-Meier curves. Correct weighting will significantly improve the power of the log-rank test.

Many other choices are possible. The weights proposed by Tarone and Ware (1977) and Fleming and Harrington (1991) involve classes of weights for which a good choice of function g or (p, q) can help detect late effects, i.e., coefficients $\beta(t)$ which increase with time. In immunotherapy trials there is often an early period in which no effects can be seen, followed by a period in which effects become manifest. A judicious choice of the weights described by Tarone and Ware (1977) and Fleming and Harrington (1991) enables us to increase the power of the basic log-rank test. The optimal choice of g or (p, q) depends on the data, and may not be easy to choose. Peckova and Fleming (2003) have proposed a weighted log-rank test whose weights are chosen adaptively as a function of the data. Such tests will generally display good behavior, although their properties, both under the null and the alternative, may be difficult to anticipate.

While these tests can exhibit good power under non-proportional hazards-type alternative hypotheses, they would be expected to lose power with respect to unweighted log-rank tests under proportional hazards-type alternative hypothe-

ses. Moreover, in order to set suitable weights, if the alternative hypothesis involves non-proportional hazards, it is necessary to know the type of variation that is likely to arise (early effects, late effects, larger effects in the middle of the study, etc.). Yang and Prentice (2010) have proposed an adaptive test which the authors claim will simplify the log-rank test when dealing with proportional hazards. Although the model upon which the test is based (see Equation 6.26) may simplify under some parametric constraint, it does not follow that the same holds for the test based upon this model. Indeed, unlike the log-rank test, the test based upon this model is not invariant to the coding of the covariate. The class is broad and does include the proportional hazards and proportional odds models. The test statistic is the maximum difference in the absolute value of the two weighted log-rank statistics $LW_{1,n}$ and $LW_{2,n}$, with respective weights $W_n^1(t) = \hat{\lambda}_T(t)/\hat{\lambda}_P(t)$ and $W_n^2(t) = \hat{\lambda}_P(t)/\hat{\lambda}_T(t)$, where the instantaneous hazard estimators in the groups receiving the placebo, $\hat{\lambda}_P$, and the treatment, $\hat{\lambda}_T$, are based on model (6.26). This test has unacceptable behavior and we would not recommend its use.

COMBINING LOG-RANK AND WEIGHTED LOG-RANK TESTS

Regulations for clinical trials require that the choice of statistical analysis be defined in advance, and consequently, a choice of the test to be used needs to be made before data are collected. It is often difficult to predict whether the effect of the treatment group on survival time is constant or whether it changes over time, and if so, what type of change is involved. The log-rank test is the most commonly used test, due to its optimal local behavior under proportional hazards-type alternative hypotheses. However, when faced with non-proportional hazards situations, the low power of the test may prevent detection of the effect of groups on survival, even when differences are strong and manifest from the Kaplan-Meier curves themselves. For such reasons, developing a test that behaves like the log-rank under proportional hazards but that continues to maintain good power, and indeed overtaking the log-rank, in non-proportional hazards situations appears as a worthwhile goal. Several attempts in this direction have been made. The first approach is to combine several weighted log-rank statistics, usually as a linear combination. For example, to detect the presence of an effect that appears after a certain time lag, Zucker and Lakatos (1990) have studied a test whose statistic is the sum of a log-rank statistic and a weighted log-rank one, chosen to be good at detecting late effects. Lee (1996) has looked at a linear combination of four weighted log-rank statistics with weights from the family given in Fleming and Harrington (1991). Denoting by $G^{p,q}$ the weighted log-rank statistic with weights $\hat{S}(t)^p(1 - \hat{S}(t))^q$, $p, q \geq 0$, the four statistics considered by Lee are:

- $G^{0,0} = L_n$, for detecting constant effects (proportional hazards),
- $G^{2,0}$, for detecting early effects,

- $G^{0,2}$, for detecting late effects,
- $G^{2,2}$, for detecting effects in the middle of the study.

Similarly, Wu and Gilbert (2002) have proposed a linear combination of several weighted log-rank statistics with weights of the form

$$W_n^a(t) = \left\{ \hat{S}(t) - (a\hat{S}(\tau_0) + 1 - a) \right\}^2, \quad (11.3)$$

where \hat{S} is the Kaplan-Meier estimator, τ_0 the maximum follow-up duration, and a a parameter between 0 and 1 that allows for detection of different types of time-dependent effect. The disadvantage of all these linear procedures is to choose the optimal weights to give to each statistic. A poor choice of weights could result in a significant loss of power, even in situations normally covered well by one of the included statistics (Lee, 1996). On the other hand, such tests are of interest if investigators already have an idea of the type of effect expected.

Another possibility for combining statistics is to keep the one with the maximum absolute value. Tarone (1981) has studied the test whose statistic is the maximum of the log-rank statistic and the weighted log-rank statistic with the weights given in Gehan (1965). Self (1991) has investigated a supremum test over the class of weighted log-rank statistics with weights of the form $t^{\theta_1}(1-t)^{\theta_2}$, $\theta_1, \theta_2 \geq 0$. Fleming and Harrington (1991) have studied the properties of the test whose statistic is the maximum of several statistics $G^{p,q}$. Lee (1996) studied this test using simulation, looking at the class of four statistics $G^{0,0}$, $G^{2,0}$, $G^{0,2}$ and $G^{2,2}$. Wu and Gilbert (2002) have also studied the theoretical properties of the test whose statistic is the maximum of several weighted log-rank tests with weights $W_n^a(t)$ (Equation 11.3). Kosorok and Lin (1999) have looked at the supremum of a class of statistics indexed by functions, extending the results of Jones and Crowley (1990) to the multiple covariate case. Finally, Breslow et al. (1984) have studied the statistic given by the maximum of the log-rank statistic and a statistic representing the score associated with a covariate, to test the null hypothesis of absence of group and covariate effects. Rather than considering the previous statistics at the end of the study, one idea is to work with their supremum over time. The log-rank statistic can then be seen as a process tracing the history of the effect's strength. If the effect depends on time, deviation from the null hypothesis can be more easily detected. The introduction of weights—the choice of which is always delicate—is possible but not necessary for detecting time-dependent effects. These tests are called Rényi-type tests, i.e., modified Kolmogorov-Smirnov tests for comparing distributions of censored data. Their asymptotic and small-sample behavior have been studied by several authors, both theoretically and via simulation (Fleming and Harrington, 1984; Fleming et al., 1980, 1987; Fleming and Harrington, 1991; Gill, 1980; Schumacher, 1984). Eng and Kosorok (2005) provide the formula for the number of subjects necessary for the supremum version of the test. Note that

Schumacher (1984) also looked at Cramér-von Mises-type tests (Appendix C). Kosorok and Lin (1999) have also contributed to the study of supremum-type tests over time. They looked at both the supremum and infimum over time, leading to a powerful test for order-based alternative hypotheses, like for example $H_1 : S_P(t) < S_T(t), \forall t$.

WEIGHTED KAPLAN-MEIER STATISTICS.

Tests other than log-rank-type ones have also been introduced for comparing survival functions. Pepe and Fleming (1989) have proposed a class of tests based on the differences between Kaplan-Meier estimators of survival functions. The differences can either be weighted or unweighted. Shen and Cai (2001) have proposed an adaptive version of these tests, where the statistic is the maximum over several statistics from the class, with various weights. Lee (2011) has studied the asymptotic behavior of these tests using simulation. Note however that such tests are not rank tests, and can be affected by increasing monotone transformations with respect to time. Moreover, even if such tests are powerful under alternative hypotheses with ordered survival functions, e.g., $H_1 : S_T(t) > S_P(t), \forall t$, their power is low if the functions intersect. For these two reasons, we do not look further at this type of testing here.

In the following section, we consider a class of tests that we believe essentially solves all of the problems described above. These tests are constructed via the regression effect process U_n^* . Several tests are looked at in greater detail, each based on particular characteristics of the process. In principle there is almost no limit to the variety of tests that we can create. This great level of generality can be used to our advantage when we wish, and are able, to introduce into the problem any knowledge we may have concerning the nature of the departure from a null hypothesis of absence of effects. Combined tests are easily constructed as well, and basing them on different aspects of the regression effect process and its known limiting distribution, makes it very convenient when seeking to establish joint distributions. It is also reassuring to know that well-known popular tests, such as the log-rank and weighted log-rank tests, are readily obtained as special cases of tests based on the regression effect process. Some finite sample behavior by simulation Chauvel (2014), Chauvel and O'Quigley (2014) is recalled in Section 11.8 and some real data are analyzed in Section 11.9.

11.4 Tests based on the regression effect process

The asymptotic behavior of the regression effect process U_n^* , given by Theorems 9.1 and 9.2, allows us to construct hypothesis tests on the value of the regression parameter $\beta(t)$. The null hypothesis H_0 and alternative hypothesis H_1 that we wish to test are

$$H_0 : \forall t, \beta(t) = \beta_0, \text{ versus } H_1 : \int_{\mathcal{T}} \{\beta(t) - \beta_0\}^2 dt > 0,$$

where β_0 is a fixed constant and $\int_{\mathcal{T}} dt > 0$. The reason for the slightly involved expression for H_1 is to overcome situations in which the test could be inconsistent. It enables us to cater for cases where, for example, $\beta(t)$ differs from β_0 only on a finite set of points or, to be more precise, on a set of time points of measure zero. Under H_0 , the drift of the regression effect process $U_n^*(\beta_0, \cdot)$ at β_0 is zero, and the process converges to standard Brownian motion. We can therefore base tests on the properties of Brownian motion. When $\beta_0 = 0$, this amounts to testing for the absence of an effect between the various values of the covariate Z . In particular, if the covariate is categorical and represents different treatment groups, this means testing the presence of a treatment effect on survival. The $\beta_0 = 0$ case thus corresponds to the null hypothesis of the log-rank test.

Several possibilities are mentioned in the book of O'Quigley (2008), including tests of the distance from the origin at time t , the maximal distance covered in a given time interval, the integral of the Brownian motion, the supremum of a Brownian bridge, reflected Brownian motion, and the arcsine law. We will take a closer look at the first two of these, followed by a new proposition. In this section we look at the univariate case with one covariate and one effect to test ($p = 1$), before extending the results to the case of several coefficients.

The following, almost trivial, result turns out to be of great help in our theoretical investigations. It is that every non-proportional hazard model, $\lambda(t|Z(t)) = \lambda_0(t) \exp\{\beta(t)Z(t)\}$, is equivalent to a proportional hazards model.

Lemma 11.1. *For given $\beta(t)$ and covariate $Z(t)$ there exists a constant β_0 and time-dependent covariate $Z^*(t)$ so that $\lambda(t|Z(t)) = \lambda_0(t) \exp\{\beta(t)Z(t)\} = \lambda_0(t) \exp\{\beta_0 Z^*(t)\}$.*

There is no loss of generality if we take $\beta_0 = 1$. The result is immediate upon introducing $\beta_0 \neq 0$ and defining a time-dependent covariate $Z^*(t) \equiv Z(t)\beta(t)\beta_0^{-1}$. The important thing to note is that we have the same $\lambda_0(t)$ either side of the equation and that, whatever the value of $\lambda(t|Z(t))$, for all values of t , these values are exactly reproduced by either expression, i.e., we have equivalence. Simple though the lemma is, it has strong and important ramifications. It allows us to identify tests that are unbiased, consistent, and indeed, most powerful in given situations. A non-proportional hazards effect can thus be made equivalent to a proportional hazards one simply by multiplying the covariate Z by $\beta(t)$. The value of this may appear somewhat limited in that we do not know the form or magnitude of $\beta(t)$. However in terms of theoretical study, this simple observation is very useful. It will allow us to identify the uniformly most powerful test, generally unavailable to us, and to gauge how close in performance comes any of those tests that are available in practice.

Our structure hinges on a useful result of Cox (1975). Under the non-proportional hazards model with $\beta(t)$, the increments of the process U_n^* are

centered with variance 1. From Proposition 9.3 we also have that these increments are uncorrelated, a key property in the derivation of the log-rank statistic. Based on the regression effect process, two clear candidate statistics stand out for testing the null hypothesis of no effect. The first is the “distance-traveled” at time t test and the second is the area under the curve test. Under proportional hazards the correlation between these two test statistics approaches one as we move away from the null. Even under the null this correlation is high and we consider this below. Under non-proportional hazards the two tests behave very differently. A combination of the two turns out to be particularly valuable for testing the null against various non-proportional hazards alternatives, including declining effects, delayed effects, and effects that change direction during the study.

DISTANCE TRAVELED FROM THE ORIGIN AT TIME t

The distance traveled from the origin at time t , and in particular when $t = 1$, appears as a logical, and intuitive, candidate for a test. Much more can be said as we see in this section. Under the null, this distance, $U_n^*(0,1)$, has expectation equal to zero. The area under the curve test is the area, positive and negative, between the process and the x-axis. Again, under the null hypothesis this has expectation equal to zero. Under proportional hazards, the test based on $U_n^*(0,1)$ turns out to be the best available test for the following reasons:

Lemma 11.2. *For $0 < s < 1$ and under the null hypothesis, $H_0 : \beta(t) = 0$, $U_n^*(0,s)/\sqrt{s}$ converges in distribution to $\Phi(z)$ where $\Phi(\cdot)$ is the standard normal distribution.*

Corollary 11.1. *Set z^α such that $\alpha = 2(1 - \Phi(z^\alpha))$. The distance traveled at time t test rejects H_0 with a type I error α if $|U^*(\beta_0, t)|/\sqrt{t} \geq z^{\alpha/2}$. The p-value for this test is given by $2[1 - \Phi(|U^*(\beta_0, t)|/\sqrt{t})]$.*

Corollary 11.2. *The distance traveled test is unbiased against proportional hazards alternatives on the interval $(0,t)$.*

Corollary 11.3. *Under the null hypothesis, $H_0 : \beta(t) = 0$, the distance traveled test is consistent against proportional hazards alternatives.*

The lemma is an immediate consequence of the Donsker-Prokhorov invariance principle. The corollaries follow from the Neyman-Pearson lemma as applied to the standard normal distribution and one with the same variance but a translated mean.

Corollary 11.4. *Under $H_1 : \beta > 0$, assuming that $k_n/n \rightarrow C$ as $n \rightarrow \infty$, then, for $t > s$*

$$\lim_{n \rightarrow \infty} EU_n^*(0,t) - EU_n^*(0,s) = \mathcal{R}(k_n, \beta)(t-s), \quad \lim_{n \rightarrow \infty} \text{Var}\{U_n^*(0,t) - U_n^*(0,s)\} = t-s$$

where \mathcal{R} is monotonic increasing in k_n for fixed β and in β for fixed k_n .

Corollary 11.5. Under $H_1 : \beta > 0$, suppose that $\mathbb{P}(s)$ is the p-value for $U_n^*(0, s)/\sqrt{s}$, then, assuming that $k_n/n \rightarrow C$ as $n \rightarrow \infty$, then, for $t > s$, $E\mathbb{P}(t) < E\mathbb{P}(s)$.

By applying the Neyman-Pearson lemma to the normal distribution, the above two corollaries tell us that the distance traveled test is the uniformly most powerful test of the null hypothesis, $H_0 : \beta = 0$ against the alternative, $H_1 : \beta > 0$. The log-rank test, L_n shares this property and this is formalized in Proposition 11.2. Lemma 11.1 allows us to make a stronger conclusion. Assume the non-proportional hazards model and consider the null hypothesis, $H_0 : \beta(t) = 0$ for all t . Then:

Lemma 11.3. The uniformly most powerful test of $H_0 : \beta(t) = 0$ for all t , is the distance traveled test applied to the proportional hazards model $\lambda(t|Z(t)) = \lambda_0(t)\exp\{\beta_0 Z^*(t)\}$ in which $Z^*(t) = \beta_0^{-1}\beta(t)Z(t)$ and where we take 0/0 to be equal to 1.

Since, under the alternative, we may have no information on the size or shape of $\beta(t)$, this result is not of any immediate practical value. It can though act as a bound, not unlike the Cramer-Rao bound for unbiased estimators and let us judge how well any test is performing when compared with the optimal test in that situation.

The following proposition states that the distance test statistic is consistent; that is, the larger the sample, the better we will be able to detect the alternative hypothesis of a proportional hazards nature. The result continues to hold for tests of the null versus non-proportional hazards alternatives but restrictions are needed, the most common one being that any non-null effect is monotonic through time.

Proposition 11.1. Under the alternative hypothesis $H_1 : \beta(t) \neq \beta_0$, that is, under the non-proportional hazards model with parameter $\beta(t) \neq \beta_0$, we have:

$$\lim_{n \rightarrow \infty} P\left(\frac{1}{\sqrt{t}}|U_n^*(\beta_0, t)| \geq z^{\alpha/2}\right) = 1.$$

For a proof see the proof of Proposition 11.5.

Remark. The test can be applied at any time point t , making it flexible. For example, if we have prior knowledge that the effect disappears beyond some time point τ , we can apply the test over the time interval $(0, \tau)$. When such information is not available, we take $t = 1$, which, under proportional hazards, will maximize the power.

LOG-RANK TEST AS A DISTANCE TRAVELED TEST

Instead of standardizing by $\mathcal{V}_0(Z|t_i)$ at each t_i we might, instead, use a simple average of such quantities over the observed failures. Note that $\int \mathcal{V}_0(Z|t)d\hat{F}(t)$

is consistent for $E(\text{var}(Z|t))$ under the proportional hazards model with $\beta(t) = 0$. Rather than integrate with respect to $d\hat{F}$ it is more common, in the counting process context, to integrate with respect to $d\bar{N}$, the two coinciding in the absence of censoring. If, we replace $\mathcal{V}_0(Z|s)$ by $\bar{\mathcal{V}}_0(Z)$ then the distance traveled test coincides exactly with the log-rank test. The following proposition gives an asymptotic equivalence under the null between the log-rank and the distance traveled tests:

Proposition 11.2. *Let L_n denote the log-rank statistic and let $U_n^*(0,1)$ be the distance statistic $U_n^*(\beta_0,1)$ evaluated at $\beta_0 = 0$. We have that*

$$|U_n^*(0,1) - L_n| \xrightarrow[n \rightarrow \infty]{P} 0.$$

In order to obtain the result, we consider an asymptotic decomposition of the statistic L_n in a Brownian motion with an added drift. The decomposition parallels that of the regression effect process at time $t = 1$. The conclusion is then immediate. A convergence in probability result implies a convergence in distribution result so that, under the conditions, the log-rank test and the distance from the origin test can be considered to be equivalent, this equivalence holding under the null and under local alternatives. Results for distant alternatives are likely to be rather less straightforward, in particular since the censoring, whether independent or conditionally independent, is likely to be involved.

Under the conditions we have stated, the distance test statistic converges to that of the log-rank, guaranteeing maximum power under alternative hypotheses of the proportional hazards type. This result is not surprising since the expressions for the log-rank statistic L_n (Equation 11.1) and the distance test (Equation 9.5) differ only in the standardization used. For the former, standardization takes place globally for all times at the end of the experiment, while for the latter, standardization happens at each time of death. However, under the homoscedasticity condition A3, the standardization does not change in expectation over time. Recall that this condition, the use of which we justified in Section 9.4, has been used by other authors (Grambsch and Therneau, 1994; Xu, 1996). The result given in Proposition 11.2 also implies that the distance test is not optimal under alternative hypotheses of a non-proportional hazards since the power properties of the log-rank test also hold here. Many other tests can be based on the regression effect process U_n^* and properties of Brownian motion. We look at several of these in the rest of this chapter.

KOLMOGOROV TYPE TESTS

Since we are viewing the process $U^*(\hat{\beta}, \beta_0, t)$ as though it were a realization of a Brownian motion, we can consider some other well-known functions of Brownian motion. Consider then the bridged process $U_0^*(\hat{\beta}, \beta_0, t)$:

Definition 11.3. *The bridged process is defined by the transformation*

$$U_0^*(\hat{\beta}, \beta_0, t) = U^*(\hat{\beta}, \beta_0, t) - t \times U^*(\hat{\beta}, \beta_0, 1)$$

Lemma 11.4. *The process $U_0^*(\hat{\beta}, \beta_0, t)$ converges in distribution to the Brownian bridge, in particular $E U_0^*(\hat{\beta}, \beta_0, t) = 0$ and $\text{Cov}\{U_0^*(\hat{\beta}, \beta_0, s), U_0^*(\hat{\beta}, \beta_0, t)\} = s(1-t)$.*

The Brownian bridge is also called tied-down Brownian motion for the obvious reason that at $t = 0$ and $t = 1$ the process takes the value 0. Carrying out a test at $t = 1$ will not then be particularly useful and it is more useful to consider, as a test statistic, the greatest distance of the bridged process from the time axis. We can then appeal to:

Lemma 11.5.

$$\Pr \left\{ \sup_u |U_0^*(\hat{\beta}, \beta_0, u)| \geq a \right\} \approx 2 \exp(-2a^2), \quad (11.4)$$

which follows as a large sample result since

$$\Pr \left\{ \sup_u |U_0^*(\hat{\beta}, \beta_0, u)| \leq a \right\} \rightarrow 1 - 2 \sum_{k=1}^{\infty} (-1)^{k+1} \exp(-2k^2 a^2), \quad a \geq 0.$$

This is an alternating sign series and therefore, if we stop the series at $k = 2$ then the error is bounded by $2 \exp(-8a^2)$ which for most values of a that we will be interested in will be small enough to ignore. For alternatives to the null hypothesis ($\beta = 0$) belonging to the proportional hazards class, the Brownian bridge test is not generally a consistent test. The reason for this is that the transformation to the Brownian bridge will have the effect of canceling the linear drift term. The large sample distribution under both the null and the alternative will be the same. Nonetheless, the test has value under alternatives of a non-proportional hazards nature, in particular an alternative in which $U^*(\hat{\beta}, \beta_0, 1)$ is close to zero, a situation we might anticipate when the hazard functions cross over. If the transformation to a Brownian bridge results in a process that would arise with small probability under the assumption of a Brownian bridge then this is indicative of the effects of a non-proportional hazards nature.

REFLECTED BROWNIAN MOTION

An interesting property of Brownian motion described in Appendix B is the following. Let $\mathcal{W}(t)$ be Brownian motion, choose some positive value r and define the process $\mathcal{W}_r(t)$ in the following way: If $\mathcal{W}(t) < r$ then $\mathcal{W}_r(t) = \mathcal{W}(t)$. If $\mathcal{W}(t) \geq r$ then $\mathcal{W}_r(t) = 2r - \mathcal{W}(t)$. It turns out that the reflected process $\mathcal{W}_r(t)$ is also Brownian motion (Appendix B). Choosing r to be negative and defin-

ing $\mathcal{W}_r(t)$ accordingly we have the same result. The process $\mathcal{W}_r(t)$ coincides exactly with $\mathcal{W}(t)$ until such a time as a barrier is reached. We can imagine this barrier as a mirror and beyond the barrier the process $\mathcal{W}_r(t)$ is a simple reflection of $\mathcal{W}(t)$. So, consider the process $U^r(\hat{\beta}, \beta_0, t)$ defined to be $U^*(\hat{\beta}, \beta_0, t)$ if $|U^*(\hat{\beta}, \beta_0, t)| < r$ and to be equal to $2r - U^*(\hat{\beta}, \beta_0, t)$ if $|U^*(\hat{\beta}, \beta_0, t)| \geq r$.

Lemma 11.6. *The process $U^r(\hat{\beta}, \beta_0, t)$ converges in distribution to Brownian motion, in particular, for large samples, $E U^r(\hat{\beta}, \beta_0, t) = 0$ and $\text{Cov}\{U^r(\hat{\beta}, \beta_0, s), U^r(\hat{\beta}, \beta_0, t)\} = s$.*

Under proportional hazards there is no obvious role to be played by U^r . However, imagine a non-proportional hazards alternative where the direction of the effect reverses at some point, the so-called crossing hazards problem. The statistic $U^*(\hat{\beta}, 0, t)$ would increase up to some point and then decrease back to a value close to zero. If we knew this point, or had some reasons for guessing it in advance, then we could work with $U^r(\hat{\beta}, \beta_0, t)$ instead of $U^*(\hat{\beta}, \beta_0, t)$. A judicious choice of the point of reflection would result in a test statistic that continues to increase under such an alternative so that a distance from the origin test might have reasonable power. In practice we may not have any ideas on a potential point of reflection. We could then consider trying a whole class of points of reflection and choosing that point which results in the greatest test statistic. We require different inferential procedures for this.

A bound for a supremum-type test can be derived by applying the results of Davies (1977, 1987). Under the alternative hypothesis we could imagine increments of the same sign being added together until the value r is reached, at which point the sign of the increments changes. Under the alternative hypothesis the absolute value of the increments is strictly greater than zero. Under the null, r is not defined and, following the usual standardization, this set-up fits in with that of Davies. We can define γ_r to be the time point satisfying $U^*(\hat{\beta}, \beta_0, \gamma_r) = r$. A two-sided test can then be based on the statistic $M = \sup_r \{|U^r(\hat{\beta}, \beta_0, 1)| : 0 \leq \gamma_r \leq 1\}$, so that:

$$\Pr \{ \sup |U^r(\hat{\beta}, \beta_0, 1)| > c : 0 \leq \gamma_r \leq 1 \} \leq \Phi(-c) + \frac{\exp(-c^2/2)}{2\pi} \int_0^1 \{-\rho_{11}(\gamma)\}^{\frac{1}{2}} d\gamma$$

where Φ denotes the cumulative normal distribution function,

$$\rho_{11}(\gamma) = \{\partial^2 \rho(\phi, \gamma) / \partial \phi^2\}_{\phi=\gamma}$$

and where $\rho(\gamma_r, \gamma_s)$ is the autocorrelation function between the processes $U^r(\hat{\beta}, \beta_0, 1)$ and $U^s(\hat{\beta}, \beta_0, 1)$. In general, the autocorrelation function $\rho(\phi, \gamma)$, needed to evaluate the test statistic is unknown. However, it can be consistently estimated using bootstrap resampling methods (O'Quigley and Pessione, 1989, 1991). For γ_r and γ_s taken as fixed, we can take bootstrap samples from which

several pairs of $U^r(\hat{\beta}, \beta_0, 1)$ and $U^s(\hat{\beta}, \beta_0, 1)$ can be obtained. Using these pairs, an empirical, i.e., product moment, correlation coefficient can be calculated. Under the usual conditions (Efron, 1981a,b,c, 1987; Efron and Stein, 1981), the empirical estimate provides a consistent estimate of the true value. This sampling strategy is further investigated in related work by O'Quigley and Natarajan (2004). A simpler approximation is available (O'Quigley, 1994) and this has the advantage that the autocorrelation is not needed. This may be written down as

$$\Pr \{ \sup |U^r(\hat{\beta}, \beta_0, 1)| > M \} \approx \Phi(-M) + 2^{-3/2} V_\rho \exp(-M^2/2)/\sqrt{\pi}, \quad (11.5)$$

where $V_\rho = \sum_i |U^r(\hat{\beta}, \beta_0, 1) - U^s(\hat{\beta}, \beta_0, 1)|$, the γ_i , ranging over $(\mathcal{L}, \mathcal{U})$, are the turning points of $T(0, \hat{\beta}; \cdot)$ and M is the observed maximum of $T(0, \hat{\beta}; \cdot)$. Turning points only occur at the k_n distinct failure times and, to keep the notation consistent with that of the next section, it suffices to take $\gamma_i, i = 2, \dots, k_n$, as being located half way between adjacent failures. To this set we can add $\gamma_1 = 0$ and γ_{k_n+1} to be any value greater than the largest failure time, both resulting in the usual constant estimator.

AREA UNDER THE CURVE

We now define the area under the regression effect process at time t .

Definition 11.4. For any $\gamma \in \mathbb{B}$, the process representing the area under the standardized score function at time t , denoted $\{J_n(\gamma, t), 0 \leq t \leq 1\}$, is defined by

$$J_n(\gamma, t) = \int_0^t U_n^*(\gamma, u) du, \quad 0 \leq t \leq 1.$$

Theorem 9.1 and properties of Brownian motion (Bhattacharya and Waymire, 1990) allow us to establish the following result.

Proposition 11.3. Under the proportional hazards model with parameter β_0 , the process representing the area under the standardized score function converges in distribution to an integrated Brownian motion:

$$J_n(\beta_0, t) \xrightarrow[n \rightarrow \infty]{\mathcal{L}} \int_0^t \mathcal{W}(s) ds, \quad t > 0.$$

Furthermore, the covariance of the process is given by

$$\lim_{n \rightarrow \infty} \text{Cov} \{ J_n(\beta_0, s), J_n(\beta_0, t) \} = s^2 \left(\frac{t}{2} - \frac{s}{6} \right), \quad 0 \leq s \leq t \leq 1. \quad (11.6)$$

Recall that the random variable $\int_0^t \mathcal{W}(s)ds$ has a centered normal distribution with variance $t^3/3$. The proposition above allows us to define a test using the area under the score process at time t .

Proposition 11.4. *Let $t \in [0, 1]$. The statistic for the area under the standardized score process at time t is $J_n(\beta_0, t)$. The test for the area under the curve at time t rejects H_0 at asymptotic level α if*

$$(3t^{-3})^{1/2} |J_n(\beta_0, t)| \geq z^{\alpha/2}.$$

The asymptotic p-value of the test is $2 \left[1 - \Phi \left((3t^{-3})^{1/2} |J_n(\beta_0, t)| \right) \right]$.

Remark. As for the distance from origin traveled, we are at liberty to choose values for t other than $t = 1$. This opens up the possibility of a very wide range of potential tests. For now, we only consider $t = 1$, and to keep the text simple, we will simply call the test the “area under the curve test”. The following proposition indicates that the test statistic for the area under the curve is consistent.

Proposition 11.5. *Under the alternative hypothesis $H_1 : \beta(t) \neq \beta_0$, that is, under the non-proportional hazards model with parameter $\beta(t) \neq \beta_0$, we have:*

$$\lim_{n \rightarrow \infty} P \left((3t^{-3})^{1/2} |J_n(\beta_0, t)| \geq z^{\alpha/2} \right) = 1.$$

Under proportional hazards alternative hypotheses, the distance test has a power slightly greater than the area under the curve test since it is asymptotically equivalent to the log-rank test. On the other hand, when the effect $\beta(t)$ decreases with time, the area under the curve test will become the more powerful of the two. This can be seen in an example involving a change-point model in which the effect is constant up to a certain time τ and vanishes afterwards (see Figure 9.4(a)). Until time τ , the drift of the process $U_n^*(\beta_0, \cdot)$ is linear with a strictly positive slope, but after τ , the drift is approximately parallel to the time axis, i.e., close to zero. A test based on the distance from the origin will only take into account the last value of the process, and not the strength of the effect before τ . This will lessen the power of the distance test compared to that of the area under the curve test, in which all the information collected over the evolution of the process is retained.

Remark. According to Propositions 11.1 and 11.5, the distance and area under the curve statistics are both consistent. Using the covariance function given below we can work out the Pitman asymptotic efficiency (see Appendix D.2). This is, for all t , given by $\sqrt{3}/2 = 0.87$. A comparison of the test’s powers for finite-sized samples confirms this finding in Section 11.8. This result could be used to guide sample size calculations in a clinical trial where there is concern that, if the null does not hold, there is a high chance that the

alternative will display non-proportional hazards behavior. We could protect ourselves against losing power by increasing the sample size by $1/0.87 = 1.15$, i.e., by 15% and using the area under the curve test. If, indeed the alternative manifests proportional hazards behavior then our increased sample size assures that we keep the same power we would have had with the original sample and the original test. If on the other hand the alternative is of a non-proportional hazards nature then we could anticipate large gains in power when we witness effects waning with time. The following lemma provides the result needed for the above to hold. (Ross, 1996):

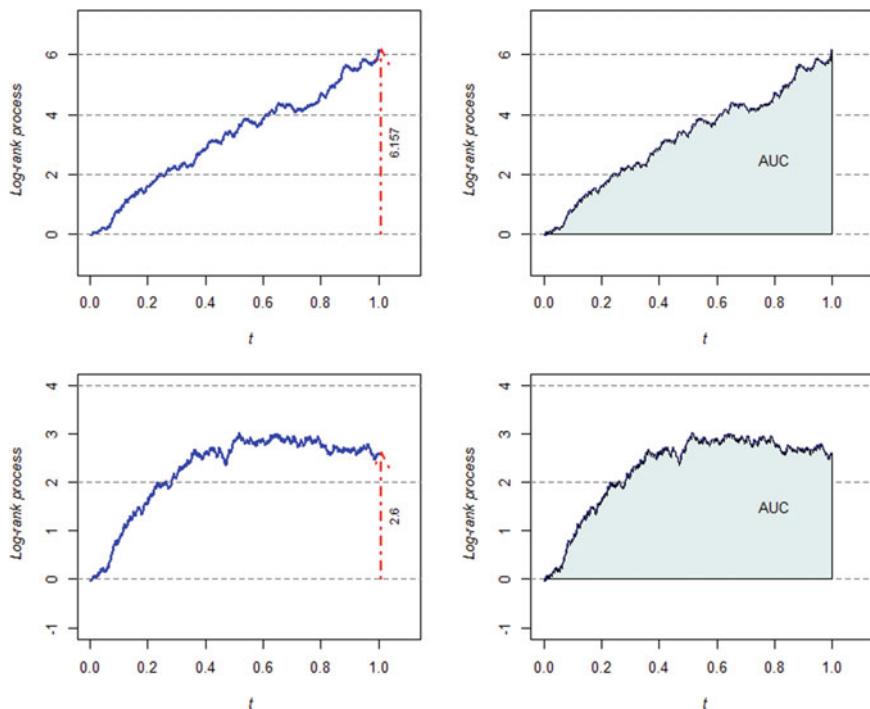


Figure 11.2: Under proportional hazards the upper two figures indicate very close agreement between log-rank and AUC. Lower curves indicate substantial gains of AUC over log-rank when effects diminish with time.

Lemma 11.7. Let $t \in [0, 1]$. Under the proportional hazards model with parameter β_0 , the covariance function of $J_n(\beta_0, t)$ and $U_n^*(\beta_0, t)$ is such that

$$\text{Cov} \{J_n(\beta_0, t), U_n^*(\beta_0, t)\} \xrightarrow[n \rightarrow \infty]{P} t^2/2.$$

The result opens up the possibility for whole classes of tests based on different ways of combining the component contributions. We can build tests based on different combinations of the two that can have good power under both types of alternative hypothesis (proportional and non-proportional hazards). Below, we define a class of adaptive tests that we describe as restrictive adaptive tests. They combine the distance statistic $U_n^*(\beta_0, t)$ with the area under the curve statistic $J_n(\beta_0, t)$. The user is allowed to bias the choice, hence the term restrictive. The resulting statistic takes the value of one or the other of the tests. The class is adaptive in the sense that users do not directly choose which statistic to apply; instead, it is chosen according to the data, with a parameter set by the user. We also present an unrestricted adaptive test that does not depend on any constraining parameters. We can control the degree to which we want the data to, in some sense, have their say. At one extreme, no amount of data will convince us to move away from the distance traveled (log-rank) test, whereas, on the other hand, a more even handed approach might specify no preference for either test.

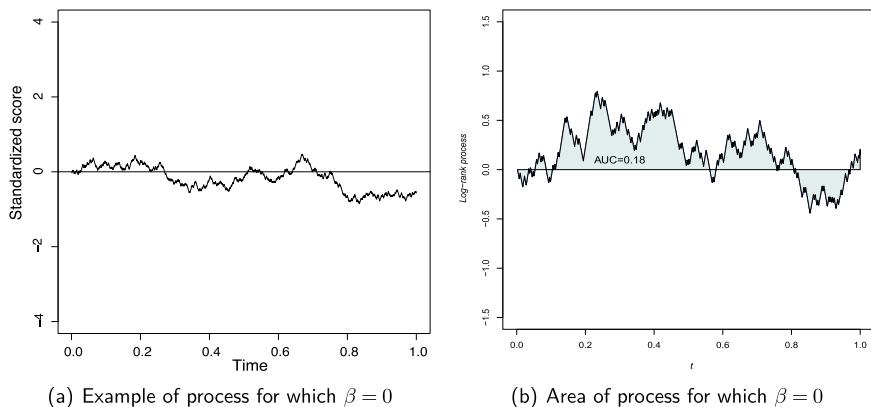


Figure 11.3: Two examples of the regression effect process $U_n^*(0, \cdot)$ under the null, $\beta = 0$. When reading such a figure, note the different scales.

INTEGRATED LOG-RANK TEST

Proposition 11.2 shows the large sample equivalence between the log-rank test and the distance from origin test. An analogous result can be shown in the same way to establish the equivalence between the area under the curve test and one based on integrating the log-rank process over the interval $(0,1)$. The distinction between these two tests is a minor one but worth having in order to complete the picture of available tests. A test based on integrating the log-rank process over the interval $(0,1)$ is referred to as the integrated log-rank test. It will offer no advantage, nor disadvantage, over the area under the curve test under the null

and in the vicinity of the null and so is not studied deeply in its own right. Since the absolute difference between the area under the curve test and the integrated log-rank test converges in probability to zero, under the usual conditions on n , k_n , and the censoring, we may sometimes refer to them as essentially the same thing. The term “integrated log-rank test” may be used informally in either case.

We can glean a lot from the visual impression given by Figure 11.2. The upper two processes correspond to a case arising under proportional hazards. The most efficient estimate of the slope of a Brownian motion with linear drift is given by connecting the origin $(0,0)$ to the point of arrival of the process at $t = 1$. The triangular area under this slope corresponds to precisely 0.5 multiplied by the distance from origin (log-rank test). The variance of this would be 0.25 times the variance of the log-rank statistic and so we see, immediately, that the area of the triangle can itself be taken to be the log-rank test. Now, under proportional hazards, we have linear drift so that the integrated log-rank test—the area under the process—will be almost identical in value to the log-rank test. The triangular area and the area under the process almost coincide. This is well confirmed in simulations and, even when the drift is equal to zero, i.e., under the null hypothesis, the correlation between the log-rank test and the integrated log-rank test is very high. Let us now consider the lower two graphs of Figure 11.2. We have no difficulty in seeing that the triangular area in this case will be quite different from the area under the process. Not only that, but in cases like these, where this area corresponds to the test statistic, it is clear that we will lose a lot of power by working with the triangular area (log-rank test) rather than the area under the process (integrated log-rank test). The differences in these two areas could be large. The lower graphs indicate an effect that diminishes with time. A yet stronger illustration is presented in Figure 11.4(b) where we can see that a very great loss of power would be consequent upon the choice of a statistic based on the triangular area (log-rank) rather than all of the area under the curve (Fig. 11.3).

Note also that it is easy to encounter such a diminishing effect, even under proportional hazards. Suppose, for example, that some percentage of the treated group fails to stick with the treatment. This may be due to undesirable secondary effects or a lack of conviction concerning the potential of the treatment. Either way, we will see a dilution of effect through time resulting in a heterogeneous rather than a homogeneous sample. The impact of this can be a regression process where the effect appears to diminish with time. When so, it is very likely that the integrated log-rank would be a more powerful choice of test than the log-rank test. As a result, in very many practical situations, it can be more logical to take the integrated log-rank test, rather than the log-rank test, as our working standard. There are several other possibilities, many of which remain open to exploration. Combining the log-rank test with the integrated log-rank test can also provide a valuable tool, in particular in situations where we anticipate a proportion of non-responders. We look at this in the following section.

AREA OVER THE CURVE

Whereas the area under the curve is easy to visualize, the area over the curve needs a bit more thought. And, just as the area under the curve will provide the basis for a test close to optimal when regression effects dissipate with time, the area over the curve shows itself to be particularly effective when regression effects are delayed. As well as $U_n^*(\beta(t), t)$, and the drift of this process under the null and alternative hypotheses, we are also interested in the behavior of $\int_0^t U_n^*(\beta(u), u) du$.

Definition 11.5. *The area over the curve (AOC) test statistic $T_n(\beta_0, 1)$ is given by;*

$$T_n(\beta(t), t) = U_n^*(\beta(t), t) - \int_0^t U_n^*(\beta(u), u) du \quad (11.7)$$

Note that the time scale is on the interval $(0,1)$ so that $U_n(0,1)$ corresponds to the rectangular area formed by the coordinates $(0,1)$ and $(0, U_n(0,1))$. The area below the regression effect process is given by $\int_0^1 U_n^*(0, u) du$ so that subtracting this from $U_n(0,1)$ results in the shaded area. This area is our test statistic. Before considering more precisely the statistical properties of the test statistic we can build our intuition by making some observations. The expected value of the test statistic under $H_0 : \beta(t) = 0, t \in (0, 1)$, will be zero. This is because the expected value of the regression effect process and the area under its path are both zero under H_0 . These two quantities are correlated but that does not impact the

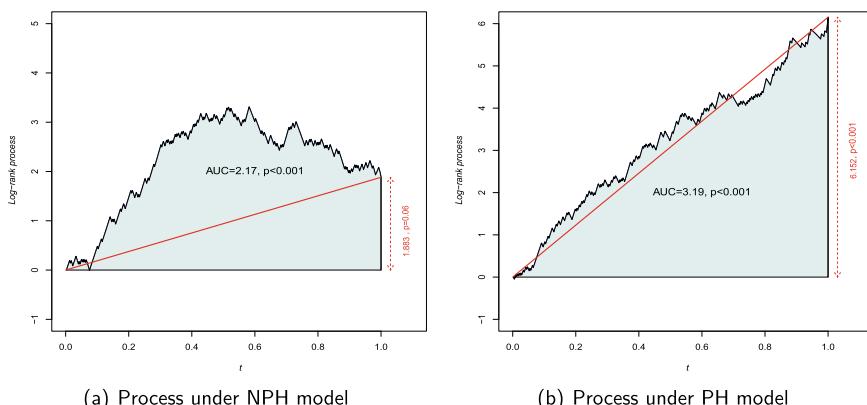
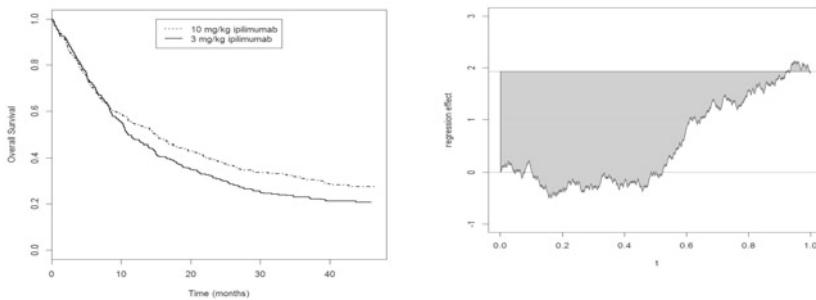


Figure 11.4: The regression effect process $U_n^*(0, \cdot)$ under PH and strong, non-proportional hazards (NPH) effect. It is clear that, under PH, log-rank test (triangular area) will be close to AUC. Under NPH, the area under the curve will provide a more powerful test statistic.

expectation. It will impact the variance and we consider this in the next section. Under a proportional hazards departure from H_0 we anticipate a linear drift in the regression effect process. The test statistic will then be well approximated by one half of the area given by $U_n(0, 1)$ which is, in turn, equivalent to the log-rank statistic. The test will then be very close to the log-rank test and a small amount of power is lost. For a delayed effect, as long as it is not too weak and allows $U_n(0, 1)$ to drift ultimately away from the origin, we will obtain a large value of the test statistic, all the more so as $U_n(0, 1)$ finally climbs away from the region close to the axis.

Lemma 11.8. *Under the model, with parameter $\beta_0(t)$, $T_n(\beta_0(t), t)$ is a Gaussian process for which we have; $E T_n(\beta_0(t), t) = 0$ and $\text{Var} T_n(\beta_0(t), t) = t(1-t) + t^3/3$. In particular, we have that $\text{Var} T_n(\beta_0(t), 1) = 1/3$.*

Tests based upon $T_n(\beta_0(t), t)$ would control for Type I error and would be anticipated to have good power, albeit suboptimal, in proportional hazards situations and to have high power for delayed effect alternatives.



(a) Kaplan-Meier curves for 2 immunological treatments, taken from Hoos et al (2010) (b) Area over curve of the regression effect process for the study of Hoos et al (2010)

Figure 11.5: Kaplan-Meier curves and corresponding regression effect process $U_n^*(0, \cdot)$. The area over the curve is used as the test statistic.

Figure 11.5 is taken from Flandre and O’Quigley (2019) and shows how the test statistic might look under an alternative of a delayed effect. This study is fully described in Hodi et al. (2018). For a period not far short of the median survival times there appears to be no effect. Beyond this the effect appears to be a significant one that would tend to rule out the plausibility of the null hypothesis of absence of effect.

DELAYED TRANSIENT EFFECTS

The area under the curve and the area over the curve tests, described above, will exhibit good power under situations of waning effects and delayed effects

respectively. These tests would be competitors for the Harrington-Fleming class of tests investigated by Lee (1996). These were described under the above heading “Combining log-rank and weighted log-rank tests”. Lee considered a linear combination of four weighted log-rank statistics with weights from the family given in Fleming and Harrington (1991). Recall that $G^{p,q}$ was the weighted log-rank statistic with weights $\hat{S}(t)^p(1 - \hat{S}(t))^q$, $p, q \geq 0$ and that $G^{0,0} = L_n$, whereas $G^{2,0}$ was geared to picking up early effects, $G^{0,2}$ was geared to picking up late effects and $G^{2,2}$ for effects in the “middle of the study.”

This is somewhat vague but let’s suppose that it corresponds to transient effects that do not kick in immediately. The circumstance may not be such a common one and it is not one that we have studied. We would anticipate the regression effect process to start out looking like standard Brownian motion with no drift, followed by a period during which a regression effect may become manifest and then a later period looking again like Brownian motion, although, now, shifted with respect to the time axis. Given the anticipated form of the regression effect process we speculate that a test based on the maximum distance traveled of the Brownian bridge process, i.e., the process obtained following the transformation to a tied-down process, would exhibit good power properties. This has not been studied, and in the absence of a compelling application, may continue to remain neglected. Certainly the situation of waning effects or of delayed effects is much more likely to be encountered in practice.

DISTRIBUTION OF SIGNED PROCESS

Although mostly of theoretical interest the time that the regression effect process remains on one side of the time axis is very readily observed. Under the null hypothesis of no effect, the fact that the regression effect process converges in law to Brownian motion allows us to immediately deduce the large sample behavior of this quantity. Let us call $\Delta(u)$ the amount of time that the process is above (below) the time axis by time $t = u$. We have the following simple proposition:

Proposition 11.6. *Suppose that $\Delta(u)$ corresponds to the percentage of time that the process, $U_n^*(\beta(t), t)$ is greater than zero. Then:*

$$\lim_{n \rightarrow \infty} P[\{\Delta(u) : U_n^*(\beta(t), t) > 0\} > w] = 2 \times \pi^{-1} \sin^{-1} \sqrt{w}.$$

This follows as an immediate application of the arcsine result for Brownian motion. Despite its simplicity the test does not appear to be that useful. It is readily seen to be unbiased and consistent but, in practice, distinguishing the null from the alternative is not easy. There are a couple of reasons for this. The power curve is very steep very close to zero and one but, otherwise, power is low. Under the null it is quite possible and not rare to spend a very large percentage of time on one side of the time axis.

By plugging in a couple of numbers, we see for example, that, if only one percent of the time the process finds itself the other side of the time axis then the probability for this, under the null, is around 0.06, not even achieving the 5% limit of significance. The second difficulty is that this is, again, a large sample result. For around one hundred failures, only one of which corresponds to a part of the curve being the other side of the time axis, we could not claim a significant result. The approximation is likely to be poor then for sample sizes (failures) not well into the hundreds. We know, of course, by standard theory, that, under the null, the process will return to cross the time axis with probability one. We may have to wait a long time though. What could be said, in practical cases, is that, if the whole of the process is located on one side of the time axis, then the test would have a p -value of zero, and without wishing to put too much weight on that, there does appear to be evidence of some kind of effect.

RESTRICTED ADAPTIVE TESTS

Chauvel and O'Quigley (2014) consider restricted adaptive tests in which we limit the class of possibilities under consideration, more or less favoring particular members of the class. For instance, we may prefer to use the log-rank test unless there are strong indications of non-proportionality of the hazards. We can introduce a parameter γ to favor one test over the others. In our case, a large value of γ favors the log-rank test, resulting in a negligible loss of power under the proportional hazards hypothesis compared to direct application of the log-rank test. The downside is low power for detecting effects that wane or disappear through time. Smaller values of γ make it easier to detect effects when hazards are not proportional.

Under proportional hazards-type alternative hypotheses $H_{1,PH}$, the process $U_n^*(\beta_0, \cdot)$ can be approximated by a Brownian motion with linear drift, according to Theorem 9.2. Thus, at time t , twice the integrated process $J_n(\beta_0, t)$ will be close in value to the distance from the origin $U_n^*(\beta_0, t)$. Next, denote by $\Delta_n(\beta_0, t)$ the difference between these two processes:

$$\Delta_n(\beta_0, t) = 2J_n(\beta_0, t) - U_n^*(\beta_0, t).$$

If the absolute value of $\Delta_n(\beta_0, t)$ is fairly small we can expect the hazards to be proportional, and the distance test or, equally so, the log-rank test (Proposition 11.2) to be powerful. If not, the area under the curve test is likely to be more powerful. In this way, we define a class of restricted adaptive tests that depend on the threshold $\gamma \geq 0$.

Definition 11.6. *The class of statistics for restricted adaptive tests at time t is denoted $\{M_n^\gamma(\beta_0, t), \gamma \geq 0\}$, where for $\gamma \geq 0$,*

$$M_n^\gamma(\beta_0, t) = \left\{ \sqrt{3}|J_n(\beta_0, t)| - |U_n^*(\beta_0, t)| \right\} \mathbf{1}_{|\Delta_n(\beta_0, t)| \geq \gamma + |U_n^*(\beta_0, t)|}.$$

We have choice in which value of t to work with, and if effects are known to drop to zero beyond some point τ then it makes sense to choose $t = \tau$. In most situations, and wishing to make full use of all of the observations, we will typically choose $t = 1$.

Remark. The p -value of the test can be calculated numerically by simulating N independent pairs $(X_1, Y_1), \dots, (X_N, Y_N)$ of centered normal distributions in \mathbb{R}^2 with variance-covariance matrix

$$\begin{pmatrix} t & t^2/2 \\ t^2/2 & t^3/3 \end{pmatrix}.$$

The p -value can then be taken to be:

$$\frac{1}{N} \sum_{i=1}^N \mathbf{1}_{f(X_i, Y_i) \geq M_n^\gamma(\beta_0, t)},$$

where for $i = 1, \dots, N$,

$$f(X_i, Y_i) = \left\{ \sqrt{3}|X_i| - |Y_i| \right\} \mathbf{1}_{|2X_i - Y_i| \geq \gamma + |Y_i|}.$$

When $\gamma = 0$, $M_n^\gamma(\beta_0, t)$ is the absolute value of the area under the curve. When $\gamma \rightarrow \infty$, $M_n^\gamma(\beta_0, t)$ tends to the absolute value of the distance from the origin. Between the two, $M_n^\gamma(\beta_0, t)$ is equal to the absolute value of one or the other. The optimal choice of γ depends on the unknown form of the alternative hypothesis $H_{1,PH}$ or $H_{1,NPH}$. One strategy is to work with a large value of γ , which essentially amounts to applying the log-rank test and keeping some power in reserve for unanticipated alternative hypotheses in which the effect changes strongly over time (for example, a change of sign). However, this test will be less powerful than the log-rank in detecting the presence of a non-zero constant effect. If we have no wish to give priority to either of the alternative hypotheses, rather than estimating the optimal parameter γ and losing power to detect the alternative hypothesis, we could simply just take the test that provides the smallest p -value. Of course this then requires adjustment in order to maintain control on Type 1 error.

Definition 11.7. The statistic for a simple unrestricted adaptive test, in which no parameter γ is specified, can be written as:

$$M_n(\beta_0, t) = \max \left\{ |U_n^*(\beta_0, t)|, \sqrt{3}|J_n(\beta_0, t)| \right\}.$$

Proposition 11.7. (Chauvel and O'Quigley 2014). Denote $\phi\{\cdot; 0, \Sigma(t)\}$ the density of the centered normal distribution in \mathbb{R}^2 with variance-covariance matrix

$$\Sigma(t) = \begin{pmatrix} t & \sqrt{3}t^2/2 \\ \sqrt{3}t^2/2 & t^3 \end{pmatrix}.$$

The unrestricted adaptive test rejects H_0 at asymptotic level α if $M_n(\beta_0, t) \geq q^\alpha$, where q^α is the upper quantile of $\mathcal{N}(0, \Sigma(t))$ such that:

$$1 - 2 \int_0^{q^\alpha} \int_0^{q^\alpha} \phi\{u, v; 0, \Sigma(t)\} dudv = \alpha.$$

The p -value of the test is

$$1 - 2 \int_0^{M_n(\beta_0, t)} \int_0^{M_n(\beta_0, t)} \phi\{u, v; 0, \Sigma(t)\} dudv.$$

The p -value can be calculated numerically using simulation by adapting the method described earlier in the section. The test is adaptive in the sense that the data will determine which of the two statistics will be used. The possibilities are limited—only two choices are available—but could be readily extended to a broader range of situations if required. The performance of the test will lie between that of the distance test and the area under the curve test. Thus, the alternative hypothesis $H_1 : \beta(t) \neq \beta_0$ is detected well in general, whether the effect depends on time or not. Naturally, if we truly are under proportional hazards then we anticipate a small loss in power when compared with the distance traveled or log-rank test. As hazards become more non-proportional, and specifically if the hazard ratio declines with time, then, again, the adaptive test will lose a small amount of power compared to the area under the curve test. It is not possible to find a test that is the best under all circumstances and the way to view the adaptive test here is as a good compromise test that covers several possible situations. Unless we have very strong prior knowledge about which of these situations most likely prevails, then, on average, the adaptive test will fare very well. We next consider some details concerning the multivariate setting.

TESTS BASED ON MULTIVARIATE REGRESSION EFFECT PROCESSES

Consider a p -dimensional vector of covariates:

$$Z(t) = (Z^{(1)}(t), Z^{(2)}(t), \dots, Z^{(p)}(t))$$

and recall that we have already studied the multivariate standardized score process in Section 9.5. Here, we suppose that [B1](#), [B1](#), [B3](#) and [B4](#) hold, and that $\beta, \beta_0 \in \mathbb{B}'$. The standardized score process is a random function from $[0, 1]$ to \mathbb{R}^p , written

$$U_n^*(\beta_0, t) = (U_n^*(\beta_0, t)_1, U_n^*(\beta_0, t)_2, \dots, U_n^*(\beta_0, t)_p), \quad 0 \leq t \leq 1.$$

By integrating each component of this process with respect to time, we can also define the area under the curve process of the multivariate standardized score process.

Definition 11.8. *The area under the curve process of the multivariate standardized score process is a random function from $[0, 1]$ to \mathbb{R}^p such that*

$$J_n(\beta_0, t) = \left(\int_0^t U_n^*(\beta_0, s)_1 ds, \int_0^t U_n^*(\beta_0, s)_2 ds, \dots, \int_0^t U_n^*(\beta_0, s)_p ds \right), \quad 0 \leq t \leq 1.$$

Under $H_0 : \beta(t) = \beta_0$, Theorem 9.3 implies the following convergence:

$$U_n^*(\beta_0, \cdot) \xrightarrow[n \rightarrow \infty]{\mathcal{L}} \mathcal{W}_p, \quad J_n(\beta_0, \cdot) \xrightarrow[n \rightarrow \infty]{\mathcal{L}} \int \mathcal{W}_p(s) ds.$$

To simplify notation, without loss of generality we consider $p = 2$. Equations 11.6 and Lemma 11.7, which describe the asymptotic behavior of covariance functions between various random quantities, imply the following result.

Proposition 11.8. (Chauvel and O'Quigley 2014). *Let $t \in [0, 1]$. Under the null hypothesis $H_0 : \beta = \beta_0$, that is, under the PH model with parameter β_0 , the vector*

$$\left(U_n^*(\beta_0, t)_1, U_n^*(\beta_0, t)_2, \sqrt{3}J_n(\beta_0, t)_1, \sqrt{3}J_n(\beta_0, t)_2 \right)^T$$

converges in distribution to a centered Gaussian vector with variance-covariance matrix $\tilde{\Sigma}_t$ given by:

$$\tilde{\Sigma}_t = \begin{pmatrix} t & 0 & \sqrt{3}t^2/2 & 0 \\ 0 & t & 0 & \sqrt{3}t^2/2 \\ \sqrt{3}t^2/2 & 0 & t^3 & 0 \\ 0 & \sqrt{3}t^2/2 & 0 & t^3 \end{pmatrix}.$$

We can extend this multivariate result to the distance from the origin, area under the curve, and restricted, or unrestricted, adaptive tests, in order to test the null hypothesis $H_0 : \beta(t) = \beta_0$ against $H_1 : \beta(t) \neq \beta_0$. The first corollary extends the test to the distance from the origin.

Corollary 11.6. *The distance from the origin statistic is given by:*

$$\mathbb{U}_n(\beta_0, t) = t^{-1} \{ U_n^*(\beta_0, t)_1^2 + U_n^*(\beta_0, t)_2^2 \}.$$

Under $H_0 : \beta(t) = \beta_0$, $\mathbb{U}_n(\beta_0, t)$ converges in distribution to a chi-squared distribution with two degrees of freedom χ_2^2 when $n \rightarrow \infty$. The distance from the origin test rejects H_0 at asymptotic level α if $\mathbb{U}_n(\beta_0, t) \geq Q_2(\alpha)$, where $P(\chi_2^2 \geq Q_2(\alpha)) = \alpha$.

The next corollary deals with the multivariate area under the curve.

Corollary 11.7. *The area under the curve statistic is given by:*

$$\mathbb{J}_n(\beta_0, t) = 3t^{-3} \{ J_n(\beta_0, t)_1^2 + J_n(\beta_0, t)_2^2 \}.$$

Under $H_0 : \beta(t) = \beta_0$, $\mathbb{J}_n(\beta_0, t)$ converges in distribution to a χ_2^2 when $n \rightarrow \infty$. The area under the curve test rejects H_0 at asymptotic level α if $\mathbb{J}_n(\beta_0, t) \geq Q_2(\alpha)$, where $P(\chi_2^2 \geq Q_2(\alpha)) = \alpha$.

Lastly, we also extend the unrestricted adaptive test to the multivariate case.

Corollary 11.8. *Denote $\mathbb{M}_n(\beta_0, t)$ the statistic for the multivariate unrestricted adaptive test, where*

$$\mathbb{M}_n(\beta_0, t) = \max \{ \mathbb{U}_n(\beta_0, t), \mathbb{J}_n(\beta_0, t) \}.$$

This test rejects H_0 at asymptotic level α if $\mathbb{M}_n(\beta_0, t) \geq Q(\alpha)$, where

$$1 - \int_{\mathbb{R}^4} \mathbf{1}_{p^2+q^2 \leq Q(\alpha), r^2+s^2 \leq Q(\alpha)} \phi(p, q, r, s; 0, \tilde{\Sigma}_1) dp dq dr ds = \alpha,$$

and $\phi(\cdot; 0, \tilde{\Sigma}_1)$ is the density of the normal distribution $\mathcal{N}_4(0, \tilde{\Sigma}_1)$. The p-value of the test is:

$$1 - \int_{\mathbb{R}^4} \mathbf{1}_{p^2+q^2 \leq \mathbb{M}_n(\beta_0, t), r^2+s^2 \leq \mathbb{M}_n(\beta_0, t)} \phi(p, q, r, s; 0, \tilde{\Sigma}_1) dp dq dr ds.$$

Conceptually the multivariate situation is no more complex than the univariate one. The difficulty lies in deciding how to treat the information from the several univariate processes. This will often boil down to a consideration of the kinds of assumptions we wish to make. For instance, testing the impact of one variable while controlling for a second variable through stratification makes a weaker assumption on the whole model structure than assuming that both regression effect processes would be linear, whether under the null or the alternative. When the assumption is a fairly accurate one then we would anticipate gaining a little in power by working under the assumption than not. How much the gain may need to be studied on a case by case basis.

11.5 Choosing the best test statistic

In a planned experiment we will work out our sample size and other design features in line with statistical requirements such as control of Type 1 error and power. These, in turn, lean upon some postulated possibilities, both under the null and under the alternative. As we have seen, if the alternative differs significantly from that postulated we can lose a lot of power. We may be working with a promising treatment but our inability to see just how the patient population will respond to the treatment can easily result in a failed study. We would like

to minimize the risk of such failures. This is a goal that the results of this chapter show to be achievable if we have information on time dependencies under the alternative hypothesis. Unfortunately our knowledge here may be imprecise. Nonetheless, even if we fall a little short of some optimal goal, we can still reduce the risk of trial failure by choosing our test to accommodate effectively plausible outcomes. We look at some broad considerations in this section.

PROPORTIONAL HAZARDS ALTERNATIVES

When the relative risk does not depend on time we know from the theory of earlier chapters that the log-rank test—equivalently the distance from the origin test—will be the uniformly most powerful unbiased test; the test of choice in situations of proportional hazards. The question then is, how plausible is it to assume proportional hazards. In many cases this would be a reasonable assumption.

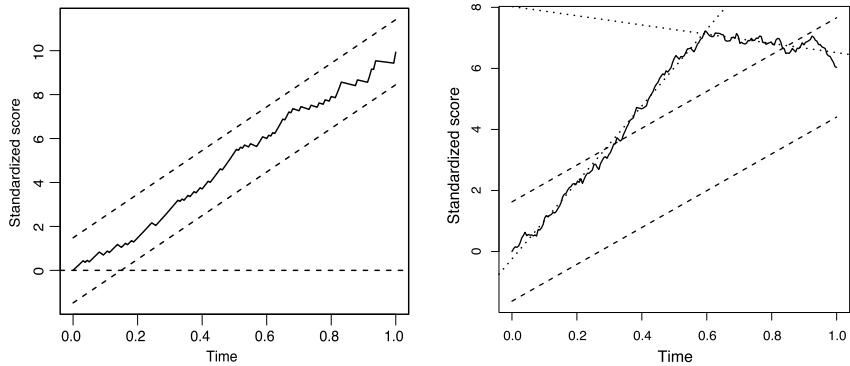
The physiological impact of a treatment may confer some advantage to all of those being treated, and in such a way, that whenever we consider a comparative study of two groups, one of which has received the treatment while the other group, otherwise homogeneous, received some other treatment, the advantage remains a constant one. This ought to be a not uncommon occurrence, the advantage, in some sense, being long lasting. If such a set-up appears to describe the study in hand, at least approximately, then the log-rank test will, most likely, be the best choice we could make. This partly explains its popularity. In epidemiological applications it is very common for the strongest risk factor by far to be age. Other risk factors may be an order of magnitude weaker and interact, if at all, very weakly with age. Once more, a sensible working assumption would be one of proportional hazards and the best test available to us the log-rank or distance from the origin test (Fig. 11.6).

It is generally admitted that the data used to illustrate Cox's method, the Freireich data, come from a clinical trial where a proportional hazards constraint could reasonably be assumed to hold. This is confirmed by looking at the Brownian bridge transform of the regression effect process shown in Figure 11.7. For these data the several ties were randomly split, two such splits being illustrated in Figure 11.7.

Potential departures from this basic set-up need careful consideration. One example would be where we have a treatment group that contains a mixture of responders and non-responders. The responders would behave like subjects in the reference group. The ratio of the responders to all members of the treatment group will change through time. This will tend to express itself as non-proportional hazards of the convex kind outlined just below.

CONCAVE NON-PROPORTIONAL HAZARDS ALTERNATIVES

In our formulation, a proportional hazards model is simply a special case of a non-proportional hazards one. It might be considered to be a proportional



(a) Test based on PH assumption would have near optimal power
 (b) Test based on single cutpoint model would have high power

Figure 11.6: Test based on PH assumption near optimal for left hand figure, and near optimal after recoding via a cutpoint model for right hand figure.

and a non-proportional hazards model simultaneously. Informally though we will suppose a non-proportional hazards model to be one where the log risk function, $\beta(t)$, is not constant in t . This non-proportional class is indeed very large. We can nonetheless identify at least two very important, and substantial, subclasses within this general class. We refer to these as the concave and the convex classes. Let's deal first with the concave class. The concave class will correspond to situations where the regression effect diminishes with time. Figure 11.8 shows an example taken from the Curie Institute breast cancer study. The non-linearity of the regression effect process is clear and the process lies entirely above the line connecting the arrival point to the origin. The area under the curve test—equivalently the integrated log-rank test—is likely to outperform the log-rank test. This is apparent by considering the area between the process and the straight line. If negligible then we are close to proportional hazards anyway and little, if anything, will be lost. If non-negligible then making use of the log-rank test means that we may well miss effects that are of importance although of a non-proportional hazards nature.

CONVEX NON-PROPORTIONAL HAZARDS ALTERNATIVES

Another large class of non-proportional hazard effects can be described as convex effects. In a way very much analogous to concave effects, if we linearly connect the arrival point of the process to the origin and find that all of the points on the line lie above the regression effect process then this is a case of convex effects. The area under the curve will be a test with behavior yet worse than the log-rank test since the area between the line and the process is removed, rather than added, to our test. On the other hand, the area over the curve will be, in general,

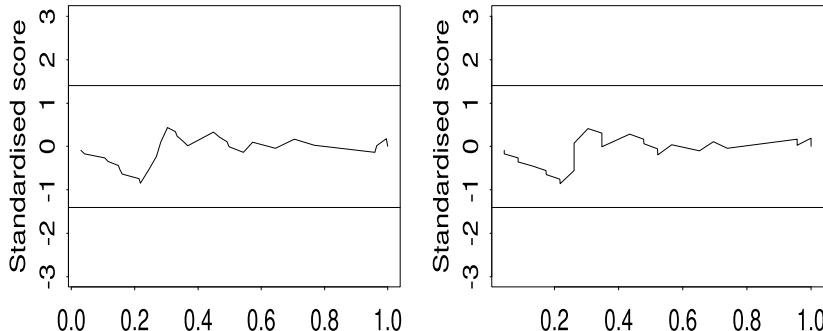


Figure 11.7: Regression effect process transformed to a bridge process for the Freireich data via $W_n^*(t) = U_n^*(\hat{\beta}, t) - tU_n^*(\hat{\beta}, 1)$. The process lies well within the limits of the supremum of the limiting process, the Brownian bridge.

a more powerful test than either of these two. We consider this below. Before looking at some particular cases note that the concepts convex and concave, in this context, while inspired by geometrical concepts can differ slightly. Small departures from the assumption will imply small departures from test behavior so, if a curve is mostly concave, then, by and large, it will still be advantageous to make use of the area under the curve test.

DELAYED TREATMENT EFFECTS

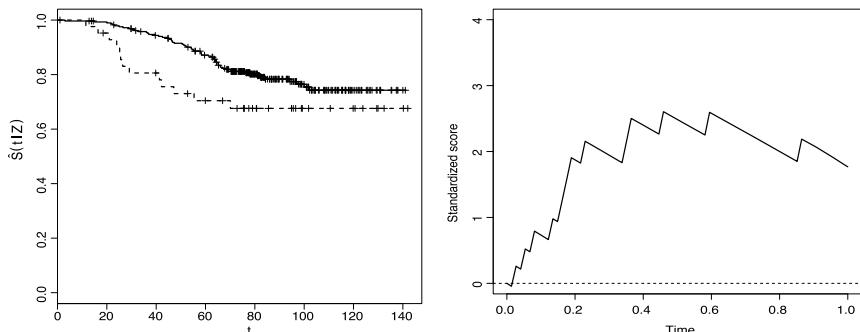
Suppose, as is frequently observed in immunotherapy trials, that each subject has a period of varying length during which time no effect is observed. The effect would kick in at some point once triggered for each individual. We could postulate, for patient i , an effect $\beta_i(t)$ such that:

$$\lambda_i(t|Z) = \lambda_0(t) \exp\{\beta_i(t)Z\} = \lambda_0(t) \exp\{W_i(t)\beta_0(t)Z\}, \quad (11.8)$$

where $W_i(t) = I(t > \tau_i)$. Each patient has their own specific time point τ_i at which the effect begins to take a hold. At time t , the percentage of “trigger points” τ_i , less than t , will be an increasing function of t , say $G(t)$ and, for the effect $\beta(t)$, governing the whole study group:

$$\beta(t) = E\beta_i(t) = E[W_i(t)\beta_0] = \beta_0 \times \Pr(\tau_i \leq t) = \beta_0 G(t).$$

The overall impression will be that of a non-proportional hazards effect whereby the between time points, $t = 0$ and $t = \min \tau_i, i = 1, \dots, n$, the treated group does not differ to the control group but, beyond which, we would observe a gradual increase in overall effect. If we knew the form of $G(t)$, we could derive an optimal test. While this is not available to us, we can obtain a test that will generally be powerful using the area over the curve statistic. Figure 11.9 is somewhat



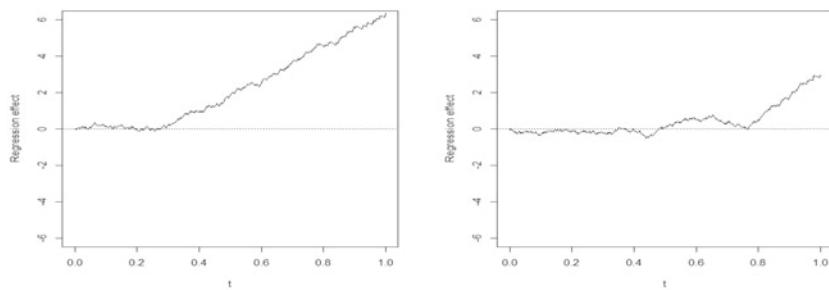
(a) Kaplan-Meier curves based on tumor size for Curie Institute breast cancer study. (b) Regression effect process for tumor size for Curie Institute breast cancer study.

Figure 11.8: When the points linearly connecting the arrival point of the process to the origin lie wholly below the regression effect process we call this concave effects.

idealized in supposing an initial time period of no effect, common to the whole group, and beyond which all of the treated group are triggered. The model just described is no doubt more realistic in postulating a distribution for the triggered times. The anticipated form of the regression effect process comes under the heading of convex non-proportional hazards. Again, the area over the curve will provide a statistic generally more powerful than that of the log-rank test since the triangular area corresponding to the log-rank test is most likely smaller.

EROSION OF REGRESSION EFFECT

Suppose that $\beta(t) > 0$ for all t . Furthermore, suppose that $\beta(t)$ is a non-increasing regression function whereby late in the study it takes on values that are non-negligibly less than the early values. This will lead to a regression effect process that is concave in appearance. The kind of situation we will encounter will often correspond to that illustrated in Figure 11.4(b), possibly with a less pronounced effect. Nonetheless, as long as the triangular area is notably smaller than the area under the curve, then the AUC test is to be preferred. We may often anticipate that effects will diminish through time and, if so, a test based on AUC is likely to be a better choice than the log-rank test. All of this remains true when $\beta(t) < t$ for all t , only now we take the area which is negative to be that between the process and the time axis, i.e., the area under the curve although, in practice, above the regression process. Concave processes are more likely to be encountered in retrospective studies focused on risk factors than in randomized clinical trials. Consider the Curie breast cancer study (Figure 11.10) where the risk factors grade and tumor size both show non-proportional hazards effects. One explanation would be that there is non-negligible error in these kinds of classifications so that



(a) Effects do not kick-in until the region of the 25th percentile
 (b) Effects do not kick-in until the region of the 75th percentile

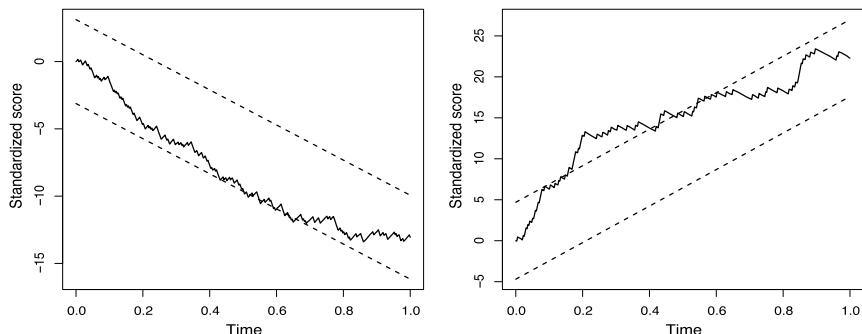
Figure 11.9: An idealized model for immunotherapy effects, absent initially and then manifesting themselves at and beyond some approximate time point.

the relative balance between groups stabilizes through time. In the long run the groups will look more similar than they did initially.

PRESENCE OF NON-RESPONDERS

Consider a clinical trial for which under the usual arguments, such as those laid out above, we could take a proportional hazards assumption to be a reasonable one. In this case the distance from the origin (log-rank) test would be close to optimal. This assumes however that we are dealing with homogeneous groups; a treated group and a, for the sake of argument, placebo group. Suppose though that among those treated there exists a subset of patients who will not be susceptible to the treatment's action. For these patients they behave similarly to patients from the placebo group. The treatment group is not then homogeneous and is a mixture of both types of patients. Under the alternative hypothesis, the prognostic impact of belonging to the treatment group will grow in strength. The reason for this is that the composition of the treatment group will change through time, losing more of the non responding patients within this group with respect to the responders.

The relative balance within the treatment group of responders to non-responders increases with time. The regression effect process will also appear to increase with time. The regression effect process will look similar to that we obtain with delayed treatment effect, an example being immunotherapy trials. Knowing, even if only roughly, how many non-responders we may have at the outset, would allow us to construct a close to optimal test. Without such knowledge we can still improve on the log-rank test by noting that we will be observing effects of a convex nature. The area over the curve test would be a good choice in such cases, in particular when the percentage of non-responders is likely to be quite large.



(a) Regression effect process for grade for Curie Institute breast cancer study. Regression coefficient negative and effects wane over time.
 (b) Regression effect process for tumor size for Curie Institute breast cancer study. Regression coefficient positive and effects wane over time

Figure 11.10: Curie Institute breast cancer study. Prognostic factors grade and tumor size both show evidence of waning effects. Grade lies just within the 95% PH limits. Size clearly violates the PH assumption.

CATEGORIZATION ERRORS AND MISCLASSIFICATION

While it is not exceptionally rare for a patient in a clinical trial to end up in a group different from what he/she was initially assigned to, such cases would have a negligible effect on subsequent tests. For any given trial there are likely to be no more than one or very few such occurrences. For descriptive studies this can be quite different. In prognostic studies we will often find ourselves dealing with categorizations that are not particularly sharp. Tumor grade, tumor size, stage, and several other indicators may well be clear cut on either end of the spectrum, but for those middle categories, it can be very difficult to classify accurately. As a result, prognostic studies will often include variables that will have a non-negligible subset of subjects who may not be assigned the most appropriate group. This situation is not altogether unlike that for non-responders although, here, we would anticipate seeing concave as opposed to convex effects. It could indeed go either way and each case ought to be considered on its own merits.

An example is seen in the Curie Institute breast cancer study. The variable grade, when classified into two groups shows clear effects that appear to be of a concave nature. The reason is that the poorer prognoses will be among the shorter survival times and, while these may include some misclassified patients, the percentage of truly labeled poor grades will be higher earlier on than later. As time unfolds, the initial classification has less relevance. The physical manifestation of this in the regression effect process is one where we see concave effects. The area under the curves will generally provide a more powerful test than the log-rank test.

11.6 Relative efficiency of competing tests

We now know that, under proportional hazards alternatives, the distance from the origin test—equivalently the log-rank test—is not only unbiased and consistent but is the uniformly most powerful unbiased test (UMPU). What happens though when these optimality conditions are not met, specifically when the non-proportionality shown by the regression effect takes a particular form. We might first consider different codings for $Z(t)$ in the model: $\lambda(t|Z(t)) = \lambda_0(t) \exp\{\beta(t)\psi[Z(t)]\}$, and in view of Lemma 11.3, if $\psi[Z(t)] = C_0\beta(t)Z(t)$ for any strictly positive C_0 , then this results in the most powerful test of $H_0 : \beta(t) \equiv 0$. We can make use of results from Lagakos et al. (1984) in order to obtain a direct assessment of efficiency that translates as the number of extra failures that we would need to observe in order to nullify the handicap resulting from not using the most powerful local test. Proposition 11.9 provides the results we need and its validity only requires the application of Lemma 11.3.

Exploiting the formal equivalence between time-dependent effects and time dependency of the covariate, we can make use of a result of O'Quigley and Prentice (1991) which extends a basic result of Lagakos et al. (1984). This allows us to assess the impact of using the coding $\psi[Z(t)]$ in the model in place of the optimal coding $\psi^*[Z(t)] = C_0\beta(t)Z(t)$. See also Schoenfeld (1981) and Lagakos (1988).

Proposition 11.9. *An expression for asymptotic relative efficiency is given by:*

$$e^2(\psi, \psi^*) = \frac{[(\int \mathbb{V}[\psi\{Z(t)\}, \psi^*\{Z(t)\}]\lambda_0(t)dt)]^2}{\int \mathbb{V}[\psi\{Z(t)\}, \psi\{Z(t)\}]\lambda_0(t)dt \int \mathbb{V}[\psi^*\{Z(t)\}, \psi^*\{Z(t)\}]\lambda_0(t)dt},$$

with $\mathbb{V}(a(t), b(t)) = E\{Y(t)a(t)b(t)\} - E\{Y(t)a(t)\}E\{Y(t)b(t)\}/E\{Y(t)\}$.

The integrals are over the interval $(0,1)$ and the expectations are taken with respect to the distribution of \mathcal{F}_t again on the interval $(0,1)$. We can use this expression to evaluate how much we lose by using the log-rank statistic when, say, regression effects do not kick in for the first 25% of the study (on the $\phi_n(t)$ timescale). It can also inform us on how well we do when using, say, the area over the curve test as opposed to the optimal (unknown in practice) test. In the absence of censoring this expression simplifies to that of the correlation between the different processes. In some cases, an example being the log-rank process, and the AUC process, the known properties of the limiting processes are readily available and provide an accurate guide.

11.7 Supremum tests over cutpoints

The above tests will be optimal, i.e., consistent and uniformly most powerful unbiased tests, in given circumstances. Moving away from particular circumstances, we will continue to obtain tests with near optimal properties. This requires us to carefully consider the type of behavior that is likely to be exhibited. When the alternative mirrors the form of $\beta(t)$, we have an optimal test. Typically we look for something broader since $\beta(t)$ is not known. The three classes considered above—linear, concave, and convex—are precise enough to obtain powerful tests, albeit falling short of the optimal test for specific situations. We might then conclude that we have an array of tools from which we can find tests with very good performance in a wide array of cases. This is true. Yet, there are still other ways of tackling the problem.

For example, we may not wish to favor any class of alternatives; linear, concave or convex. The alternative is then very broad. An appropriate class of tests for this situation, a class that we would anticipate to have good power properties, might be based on cutpoints. The several examples of Chapter 10 suggest that change-point models can closely approximate the regression effect process. In most cases a single cutpoint will be good enough and provide a significantly improved fit to that attained by a model with a constant regression effect. Extending to more than a single cutpoint is, in principle, quite straightforward, raising no additional methodological considerations, although more cumbersome computationally. Under the null, the cutpoint would not be defined. Under the alternative, we would have as many cutpoints as we choose to model and for $s - 1$ cutpoints we would obtain s regression effect processes, each one, under the null and conditional upon its starting point, being uncorrelated with the others. A single cutpoint for example leads to two regression effect processes; one from zero up to the cutpoint, the other from the cutpoint to the point 1 on the transformed time scale (Fig. 11.11).

Tests within such a structure fall under the heading studied by Davies (1977, 1987) whereby the nuisance parameter (cutpoint) is defined only under the alternative and not under the null. In the case of a single cutpoint, we write:

$$T_n^2(\beta(t), \theta) = \theta^{-1}\{U_n^*(\beta(t), \theta)\}^2 + (1 - \theta)^{-1}\{U_n^+(\beta(t), \theta)\}^2 \quad (11.9)$$

where $U_n^+(\beta(t), t) = U_n^*(\beta(t), 1) - U_n^*(\beta(t), \theta)$. The ingredients we need are provided by the variance-covariance matrix of the unsquared quantities. Writing these as:

$$\text{Var} \begin{pmatrix} U_n^*(\beta(t), \theta) / \sqrt{\theta} \\ U_n^+(\beta(t), 1) / \sqrt{1-\theta} \\ DU_n^*(\beta(t), \theta) / \sqrt{\theta} \\ DU_n^+(\beta(t), 1) / \sqrt{1-\theta} \end{pmatrix} = \begin{pmatrix} I & A(\theta) \\ A^T(\theta) & B(\theta) \end{pmatrix}$$

where $DU(\theta) = \partial U / \partial \theta$, we have most of what we need. Finally, suppose that $\eta(\theta)$ are zero-mean normal random variables with variances given by λ_1 and λ_2 where $\lambda_i, i = 1, 2$, are the eigenvalues of $B(\theta) - A^T(\theta)A(\theta)$. Then:

$$\mathbb{P}\{\sup T_n^2(\beta(t), \theta) > u; 0 < \theta < 1\} \leq \mathbb{P}(\chi_2^2 > u) + \frac{1}{\pi} \int_0^1 \sqrt{u} e^{-u/2} E\|\eta(\theta)\| d\theta$$

Clearly, the calculations are not that straightforward although, with modern software such as R, they are readily programmed. In addition, Davies (1987) makes a number of suggestions for simpler approximations. Some study of these suggestions, in the context of survival data with changepoints, has been carried out by O'Quigley and Pessione (1991), O'Quigley (1994), and O'Quigley and Natarajan (2004) and indicate that the accuracy of the approximations is such that they can be reliably used.

A full study of tests based on taking the supremum over an interval of cutpoints has yet to be carried out. In terms of power and test optimality any advantages will be necessarily small since we already know how to obtain near optimal tests. These optimal, and near optimal, tests apply in a range of situations and include cutpoints. So, despite the extra work, there may not be much to be gained. One useful advantage is that any such supremum test immediately furnishes both interval and point estimates of the changepoint itself. This can be valuable in moving forward from the rejection of a null hypothesis to the construction of a more plausible hypothesis.

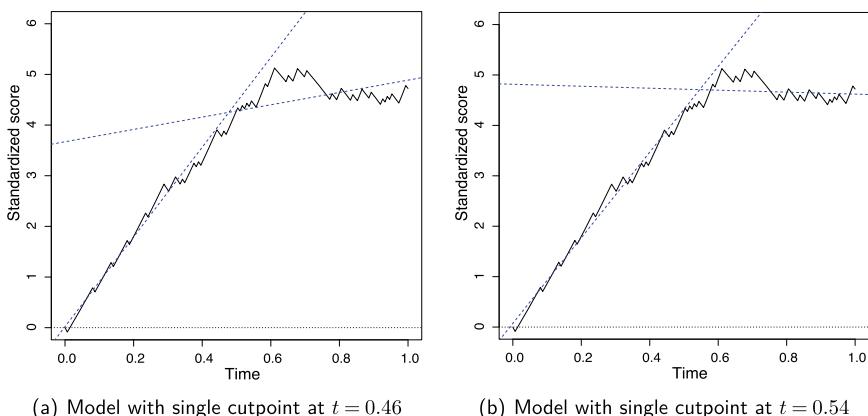


Figure 11.11: Two possible cutpoints for a single cutpoint model. Note the impact of small changes in cutpoint on the slopes of the regression effect process.

11.8 Some simulated comparisons

A large simulation study was carried out by Chauvel (2014), some of which was reproduced in Chauvel and O'Quigley (2014) and Chauvel and O'Quigley (2017). The behavior of the tests described above for finite sample sizes was compared with a number of other established tests. The advantages are readily highlighted.

FRAMEWORK FOR THE SIMULATIONS

The times of death T are simulated according to an exponential distribution with parameter 1. The censoring times C , independent of T , also follow an exponential distribution, whose fixed parameter takes a range of values, chosen to calibrate the rate of censoring, as shown in Table 11.1. The rate of censoring in the sample is thus 30%, 50% or 70%. Data is simulated with one covariate ($p = 1$). This covariate Z follows either a Bernoulli distribution with parameter 1/2, a normal distribution with mean 1/2 and variance 1/4, a continuous uniform distribution on $0.5 \times [1 - \sqrt{3}, 1 + \sqrt{3}]$, or an exponential distribution with parameter 1/2. These parameters have been chosen in order to make the first two moments identical. The sample size is either 60, 100 or 200. The tests have been calibrated so that $\alpha = 0.05$. Several regression coefficients were considered. The $\beta(t) = 0$ for all $t \in [0, 1]$ case helps us check the level of the tests. The $\beta(t) = 0.5$ (or $\beta(t) = 0.8$) for all $t \in [0, 1]$ cases are the two proportional hazards situations we choose to look at. Other cases we consider involve non-proportional hazards, with either a continuous decreasing effect ($\beta(t) = 1.5(1-t)$ or $\beta(t) = 2(1-t)^2$, for $t \in [0, 1]$) or hazards that cross over: ($\beta(t) = -1.5\mathbf{1}_{t \leq 0.5} + 1.5\mathbf{1}_{t \geq 0.5}$ for $t \in [0, 1]$, or with swapped signs). We also look at three change-point models ($\beta(t) = \mathbf{1}_{t \leq 0.3}$, $\beta(t) = \mathbf{1}_{t \leq 0.5}$ and $\beta(t) = \mathbf{1}_{t \leq 0.7}$, $t \in [0, 1]$). For each regression coefficient and each rate of censoring considered, 3000 samples are generated as described above and used to evaluate the empirical level and power of each test.

λ	0.5	1	2
$P(C \leq T)$	0.3	0.5	0.7

Table 11.1: Link between the rate of censoring $P(C \leq T)$ and parameter λ of the censoring distribution $\mathcal{E}(\lambda)$.

Remark. The theorems of Section 9 were proven for uniformly bounded covariates. These theorems form the basis of the tests presented in this chapter. However, here we run simulations with normal and exponential covariates since these cases are often found in applications, even though they are indeed not uniformly bounded. More examples can be found in Chauvel (2014).

$\beta(t)$	Log-rank	$J_n(0,1)$	$M_n^\gamma(0,1)$			
			$\gamma = 0.5$	$\gamma = 1$	$\gamma = 1.5$	$\gamma = 2$
0	4.9	5.2	5.0	5.1	4.8	4.6
0.5	50.9	40.9	42.7	46.3	49.5	50.3
0.8	88.1	76.8	79.2	83.5	86.6	87.5
$\mathbf{1}_{t \leq 0.3}$	21.8	38.5	35.4	32.0	25.3	21.2
$\mathbf{1}_{t \leq 0.5}$	54.1	72.6	70.1	66.7	58.3	52.8
$\mathbf{1}_{t \leq 0.7}$	83.4	89.0	88.3	85.5	82.3	80.9
$-1.5\mathbf{1}_{t \leq 0.5} + 1.5\mathbf{1}_{t \geq 0.5}$	16.8	85.3	82.7	83.4	84.4	80.0
$1.5\mathbf{1}_{t \leq 0.5} - 1.5\mathbf{1}_{t \geq 0.5}$	15.5	85.1	82.3	83.2	84.4	79.5
$1.5(1-t)$	84.7	91.6	91.0	88.8	85.2	83.0
$2(1-t)^2$	73.8	90.6	89.2	86.0	78.7	72.8

Table 11.2: Empirical level of significance and power of each test (in %) based on 3000 simulated datasets for each $\beta(t)$. Each dataset has 100 elements (50 subjects per group). The rate of censoring is fixed at 30%.

SETTING THE CALIBRATION PARAMETER γ

Simulations were first run to calibrate γ for the restricted adaptive test with statistic $M_n^\gamma(0,1)$. Note that the advantage of this test is that it loses very little power compared with the log-rank test under the proportional hazards hypothesis, while gaining power under non-proportional hazards-type alternative hypotheses. The sample size was set to $n = 100$, with the covariate Z simulated following a Bernoulli with parameter 1/2 and a 30% rate of censoring. For each regression coefficient $\beta(t)$, the log-rank test, area under the curve test with statistic $J_n(0,1)$, and restricted adaptive test $M_n^\gamma(0,1)$ were run, with $\gamma = 0.5, 1, 1.5$, and 2. The results are given in Table 11.2. When $\gamma = 2$, the power of the test with statistic $M_n^\gamma(0,1)$ is very close to that of the log-rank test, except when the risks intersect, i.e., when the regression coefficient changes sign over time. In such situations, the test with statistic $M_n^2(0,1)$ is much more powerful than the log-rank. In other situations, the test is weak in detecting the presence of an effect that changes with time, inheriting the behavior of the log-rank. When $\gamma = 0.5$, the power of the test with statistic $M_n^\gamma(0,1)$ is now instead very close to that of the area under the curve test. Hence, we see a large loss in power (around 10%) with respect to the log-rank test under proportional hazards-type alternative hypotheses. Lastly, the $\gamma = 1$ and $\gamma = 1.5$ cases are between these two extremes. The power of the test with statistic $M_n^1(0,1)$ is closer to that of the area under the curve test, while the behavior under the alternative hypothesis of the test with statistic $M_n^{1.5}(0,1)$ is closer to that of the log-rank. Unlike the pre-

vious cases, these tests have acceptable power under all alternative hypotheses considered. On this basis we find $\gamma = 1.5$ provides a good all-round compromise and tends to keep to a minimum any loss in power with respect to the log-rank test under proportional hazards-type alternative hypotheses.

PERFORMANCE OF THE TESTS

Chauvel and O'Quigley (2014) examined the performance at $t = 1$ of the distance from the origin test, area under the regression effect process test, and restricted adaptive tests, with respective statistics $U_n^*(0,1)$, $J_n(0,1)$, $M_n(0,1)$, and $M_n^{1.5}(0,1)$. We also included the log-rank test (Mantel, 1966) and two other adaptive tests which were briefly described in Section 11.3 and are summarized below. More results are given in Chauvel (2014). The first test is that of Lee (1996), whose statistic is the maximum of four weighted log-rank statistics with weights from the family $G^{\rho,\gamma}$ given in Fleming and Harrington (1991). These weights are useful to detect respectively early effects, effects in the middle of a study, delayed effects between the survival curves of the two groups, and a proportional hazards situation. The procedure is similar to the test proposed by Self (1991). The class of tests in Fleming and Harrington (1991) is known to have good power for detecting time-dependent effects.

The other test statistic, proposed by Yang and Prentice (2010), is the maximum between two weighted log-rank statistics, with weights based on the inference of the quantities defined by the model (6.26) of Yang and Prentice (2005). The authors argue that, asymptotically, this statistic would be equivalent to the log-rank when hazards are proportional.

Consider the one-covariate case with a Bernoulli distribution with parameter 1/2. Results are presented in Table 11.3 for the $n = 100$ case, Table 11.4 for $n = 60$, and Table 11.5 for $n = 200$. The conclusion is the same for each sample size considered.

With the exception of the tests of Lee and Yang and Prentice, test levels are controlled well at 5%. Inflation of the size of the Yang and Prentice test confirms their own simulation results for comparable sample sizes and rates of censoring. As expected (see Proposition 11.2), the power of the distance test is very close to that of the log-rank's under proportional and non-proportional hazards-type alternative hypotheses.

When the hazards are proportional, the area under the curve test is less powerful than the log-rank, while the power of the restricted adaptive test M_n lies between the two, with an average loss compared to the log-rank of only 3%. The powers of the restricted adaptive test M_n^γ with $\gamma = 1.5$ and the Lee test are comparable to those of the log-rank. The most powerful test would seem to be the Yang and Prentice one. This result is surprising since the log-rank test is the most powerful in this situation (Peto and Peto, 1972). This apparent contradiction can be explained by poor control of the type I error of the test.

The first non-proportional hazards case involves a piecewise-constant effect, corresponding to a change-point model. We vary the times at which the changes occur. The area under the curve test is more powerful than the log-rank and distance tests. The power of the restricted adaptive test M_n is closest to that of the area under the curve test. As for the power of the test with statistic $M_n^{1.5}$, it is low and close to the log-rank's. The tests described in this chapter have better power than the Lee test. The Yang and Prentice test appears to be a powerful test but the drawbacks of this test, one being inadequate control over Type 1 error, rule out its use in practice.

In cases where the hazards intersect, the powers of the log-rank, distance, and Lee tests are low. The Yang and Prentice test undergoes a severe reduction in power, up to 30%, between an effect having a negative and then a positive value and a positive followed by a negative one. As for the restricted adaptive test and the area under the curve test, they have reasonable power in situations where hazards intersect. In the last set of cases studied, the effect $\beta(t)$ decreases continuously with time. All of the test's powers were similar, with the area under the curve test being the most powerful.

To summarize the simulations, in general situations in which no hypothesis is made about the type of effect, use of the restricted adaptive test with statistic M_n is recommended. The simulation results show good power under alternative hypotheses of both the proportional and non-proportional hazard types. The Yang and Prentice test should be used with caution because of its lack of control of type I errors and dependency on the chosen coding. If the goal is to lose as little power as possible compared to the log-rank test, we recommend using one of the tests from the restricted adaptive class with statistics M_n^γ . However, these tests are less powerful than the M_n one for non-proportional hazards. The Lee and the Yang and Prentice tests were developed to compare survival in two groups of patients. Thus, they can only be applied when there are binary covariates. We also studied the behavior of the tests when there are continuous covariates. The parameters of the distributions were chosen so that the first two moments were the same. The sample size was fixed at $n = 100$ and the rate of censoring at 30%. Some of the findings are reproduced in Table 11.6. For a given test and regression coefficients, the respective powers of each test are similar for the three covariate distributions.

11.9 Illustrations

We consider several publicly available data sets. Also included are data collected by the Curie Institute for studying breast cancer and a number of data sets from the University of Leeds, U.K. focused on the impact of markers on survival in cancer. Note that the analysis of a multivariate problem often comes down to considering several univariate regression effect processes. This is because the relevant questions typically concern the impact of some variable, often treat-

ment, having taken into account several potentially confounding variables. Once these confounding variables have been included in the model then the process of interest to us is a univariate one where the expectation and variance have been suitably adjusted via model estimation. This is also the case when using stratification or, indeed, when considering the prognostic index which would be a linear combination of several variables.

BREAST CANCER STUDY AT THE INSTITUT CURIE, PARIS

The dataset is made up of 1504 patients with breast cancer. The data is complex with several subgroups; for our illustration, we consider a specific subgroup of 332 patients. Among the prognostic factors, we focus our attention on the influence of tumor size in millimeters (mm) on patient survival. We separate the patients into two groups depending on whether their tumor size is above or below 60 mm. Patients were monitored for 12 years. The rate of censoring in the dataset was 22%. Kaplan-Meier estimators of survival are shown in Figure 11.12(a). During the first 60 months, the estimated survival rates are quite different, corresponding to a strong effect between the patient groups. After 60 months, the curves get closer together, reflecting a decrease in effect. We see that 80% of patients whose tumor size is less than 60 mm survive at least 75 months. This drops to 28 months for patients with tumor size above 60 mm.

The corresponding standardized score process is shown in Figure 11.12(b). As this process is not centered around 0, the presence of drift and thus an effect is clear. As the drift is not linear, the effect decreases over time. This decrease has an impact on the log-rank ($p = 0.06$) and distance from the origin ($p = 0.08$) tests, which are not significant. Nevertheless, the area under the curve and restricted adaptive tests are highly significant, with respective p -values of 0.001 and 0.002. The conclusion from these two tests is consistent with our analysis of the Kaplan-Meier estimators.

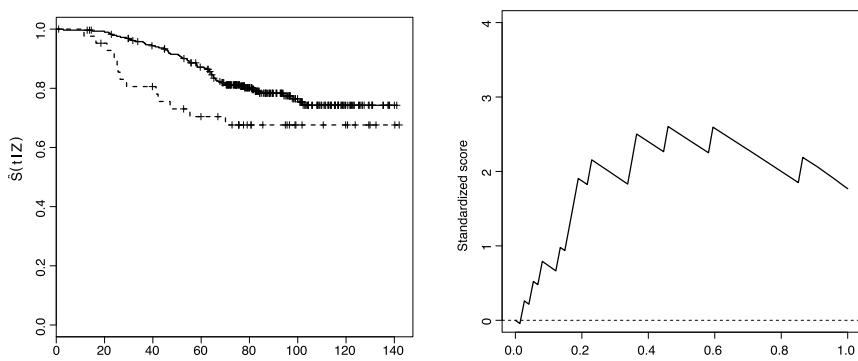
A LARGE RANDOMIZED CLINICAL TRIAL

In a recent clinical trial, 1636 patients were randomized to 3 treatment groups and followed for 65 months. The goal of the trial was to compare the influence of different treatments on patient survival for individuals with the same illness. The Kaplan-Meier estimators of the survival curves are shown in Figure 11.13(a). Patient survival in randomized groups 2 and 3 appears similar, and different to that of group 1. This is immediately interesting since groups 2 and 3 correspond to slightly different administrations of the treatment. Independently then, given the observed survival of group 1, each group appears to differ from this group in the same way. The survival experiences of the two treated groups all but coincide. Using the treatment covariate, two binary covariates were constructed with “treatment group 1” as a reference. The corresponding multivariate standardized score process $U_n^*(0, \cdot)$ is shown as a function of time in Figure 11.13(b).

Despite the noise, the drifts of the processes are clearly visible, corresponding to the alternative hypothesis of the presence of an effect—which appears to be nonlinear. Its value decreases with time in such a way that the multivariate log-rank test is not significant, with a p -value of 0.10. The multivariate distance test has a p -value of 0.09. In contrast, the multivariate area under the curve and multivariate restricted adaptive tests are highly significant, with respective p -values of 0.005 and 0.008. These tests are less affected by the decrease in effect, and for this reason they detect it. They thus correspond more closely to our intuition when looking at the Kaplan-Meier curves which do seem to suggest the presence of true differential effects.

11.10 Some further thoughts

In order to compare the survival of several groups, the log-rank test is commonly used. This test is the most powerful for detecting a fixed effect over time. On the other hand, several authors have shown its weakness in the presence of an effect which varies. The purpose of Chauvel and O'Quigley (2014) was to develop a test with good power under both types of alternative hypothesis. The asymptotic properties of the regression effect process studied in Chapter 9 lead to the construction of several new tests. The first of these is the distance from the origin test, which is asymptotically equivalent to the log-rank. There are many other possibilities including the area under the curve test. This test, slightly less powerful than the log-rank test in the presence of a constant effect, allows for significant power gains when there is an effect that changes with time.



(a) Kaplan-Meier estimators of patient survival with tumor size below (solid line) or above (dashed line) 60 mm.

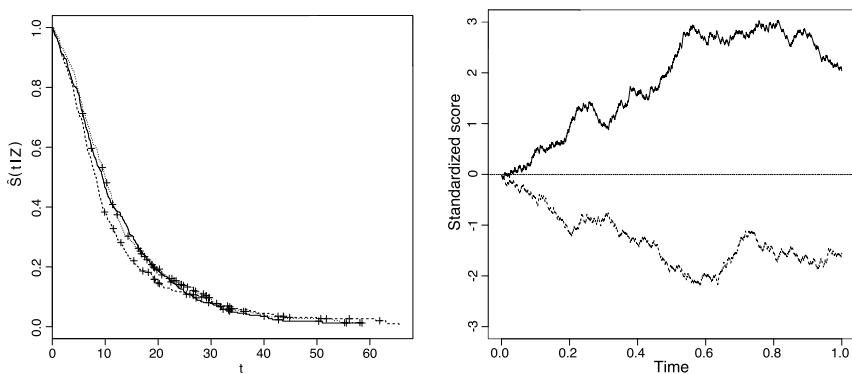
(b) Standardized score process for tumor size.

Figure 11.12: Kaplan-Meier estimators and standardized score process for the Curie Institute data.

The class of adaptive tests considered will assume either the value of the distance traveled or the value of the area under the curve. The parameter allows us to calibrate the proximity of the tests to the log-rank. However, fixing this parameter may be tricky, so if there is no compelling reason to privilege the log-rank test, Chauvel and O'Quigley (2014) proposed the use of an unrestricted adaptive test. This corresponds to a statistic that is the maximum absolute value of the distance from the origin and the area under the curve.

Simulations show that these adaptive tests perform well compared to other tests in the literature. In addition to the tests' control of the level and their good power properties, they also have the advantage of being easily interpretable and simple to plot. In applications, at the same time as we look at the p -values, parameter estimates, and confidence intervals, it would certainly be helpful to consider plots of the processes in order to get a visual impression of how close or far we appear to be from the conditions under which we know we would achieve something close to an optimal test. This tells us something not only about the potential lack of plausibility of the null hypothesis but also something about the nature of a more plausible alternative hypothesis.

By calculating the standardized score process U_n^* at β_0 , the results can immediately be extended to the null hypothesis $H_0 : \beta(t) = \beta_0$, $\forall t$, against the alternative $H_1 : \exists t_0, \beta(t_0) \neq \beta_0$. If β_0 is non-zero, we are no longer testing for the presence of a difference in survival between groups, but rather whether the relative risk between groups is equal to β_0 or not. Although we have mostly considered testing a null against an alternative of a constant value for the regression parameter, it is however possible to consider testing the null hypothesis $H_0 : \beta(t) = \beta_0(t)$ against $H_1 : \beta(t) \neq \beta_0(t)$, where $\beta_0(t)$ changes with time.



(a) Kaplan-Meier estimators of survival in treatment groups 2 (solid line), 3 (dashed line), contrasted to that of group 1 (dotted line).

(b) Standardized score process of groups 2 vs 1 (solid line) and 3 vs 1 (dashed line). Both comparisons suggest NPH effects.

Figure 11.13: Kaplan-Meier estimators and standardized score processes for clinical trial data with 3 treatment groups.

Other combinations of the distance from the origin and area under the curve tests could also be studied, rather than the linear combination looked at here:

$$\theta U_n^*(\beta_0, t) + (1 - \theta) J_n(\beta_0, t), \quad 0 \leq \theta \leq 1.$$

However, depending on whether the hazards are proportional or not, simulations tend to show that the greatest power is reached for $\theta = 1$ or $\theta = 0$, respectively. We then find ourselves back with the restricted adaptive test. Aside from this, it would also be interesting to study other tests based on the properties of different functions of Brownian motion, which will approximate the limit process of $U_n^*(\beta_0, \cdot)$ under varying assumptions.

We have applied the tests described here with $t = 1$. It would be possible to apply them at any time τ if we knew, for example, that the effect became zero after some time point, τ . This would also allow for increases in the power of the tests. It may also be of interest to study tests whose statistic is the supremum over time of the distance or area under the curve tests. In the multivariate case, the tests we have introduced allow us to simultaneously test whether all parameters are equal to zero. An extension to this could be to find a way to use the process to test whether specific parameters only were equal to zero. Suppose we had two variables Z_1 and Z_2 in a non-proportional hazards model with respective parameters $\beta_1(t)$ and $\beta_2(t)$. To test $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$, one idea would be to stratify with respect to the covariate Z_2 by estimating $\beta_2(t)$ and examining the process whose construction is based on the probabilities $\{\pi_i(\hat{B}(t), t), i = 1, \dots, n\}$, with $\hat{B}(t) = (0, \hat{\beta}_2(t))$.

11.11 Classwork and homework

1. Suppose we have a study in which, before some time point τ , treatment A has an advantage over treatment B but, beyond that point, this advantage becomes a disadvantage. Without any formal analysis how might you obtain a quick rough-and-ready estimate of τ .
2. For the previous question consider a null hypothesis that deems effects after τ to be equal to those before τ and both equal to zero. How might you make use of the known properties of reflected Brownian motion to construct a consistent test of this hypothesis?
3. Continuing the previous question, how might we make use of Kolmogorov's theorem and the Brownian bridge to construct an alternative test to one based on reflected Brownian motion.
4. Discuss the relative advantages and drawbacks of the two tests described in the previous two questions. Are you able to make any clear recommendations for given situations?

5. On the basis of different data sets, calculate the distance from the origin test, the log-rank test, and the arcsine test. On the basis of these results what does your intuition tell you?
6. Under proportional hazards alternatives to the null hypothesis of no difference between two groups, show that the log-rank test, the distance from the origin test, the area under the curve test are all unbiased and consistent but that Kolmogorov's test will not generally be unbiased nor consistent.
7. For some data, construct $100(1 - \alpha)\%$ confidence intervals for β in a proportional model based on the distance from the origin test and the area under the curve test. For a non-proportional hazards model with coefficient $\beta(t)$, how might you best go about constructing test based confidence intervals for $\beta(t)$.
8. One of the conditions for consistency relates to the rate of censoring; specifically we need to have $k_n/n \rightarrow C$ as $n \rightarrow \infty$ where $0 < C < 1$. Suppose, for some true situation, that $k_n/n \rightarrow 0$ as $n \rightarrow \infty$. Indicate why all of the tests, including the log-rank test, would be inconsistent in such a situation.
9. Under proportional hazards, with coefficient β_0 , show that as $|\beta_0|$ moves away from zero, the area under the curve test tends to a test that is fully efficient.
10. For the situation of either proportional hazards or concave non-proportional hazards, we can calculate two areas: the area under the process and the area of the triangle defined by the arrival point of the process. The ratio of these two areas provides a goodness-of-fit statistic. Under proportional hazards, these two areas will be very close in size. As we move away from proportional hazards the difference between them will grow. Construct a formal test based on these two areas. Use the relative difference in these two areas to construct an index of fit. Discuss the advantages and drawbacks of such an index in practice.
11. Repeat the previous question, changing the word concave to convex. What would the impact of this be? Can you generalize the arguments to embrace the more general setting in which one of three situations may prevail: proportional hazards, concave non-proportional hazards, and convex non-proportional hazards.
12. For a clinical trial the presence of a percentage of non-responders suggests a particular choice of test in order to maximize power. Suppose we have a different situation that involves heterogeneity; that is one where the effect of treatment differs from patient to patient. This is the case described by frailty models. Which test would be the most suitable for this situation.

13. Consider a two-group clinical trial in which, beyond the median of the combined groups there is no further effect. Use the expression for efficiency to calculate how many more patients would be needed to maintain the power of the most powerful test when the design is based on the log-rank test. Hint—use the convergence in distribution result between the log-rank and the distance traveled tests.
14. Guidance on the choice of tests has so far made little reference to censoring beyond the fact that it is assumed either independent or conditionally independent of the failure mechanism. The impact of a conditionally independent censoring mechanism is potentially greater on the distance from origin test than it is on the log-rank test. Explain why this would be so.

11.12 Outline of proofs

Propositions 11.1 and 11.5 First, recall the result of Theorem 9.2:

$$U_n^*(\beta_0, t) - \sqrt{k_n} A_n(t) \xrightarrow[n \rightarrow \infty]{\mathcal{L}} C_1(\beta, \beta_0) \mathcal{W}(t), \quad 0 \leq t \leq 1,$$

where

$$A_n(t) \xrightarrow[n \rightarrow \infty]{P} C_2 \int_0^t \{\beta(s) - \beta_0\} ds, \quad 0 \leq t \leq 1.$$

Let $t \in [0, 1]$. The functions $U_n^*(\beta_0, \cdot)$ and A_n are almost surely bounded so that we can write $J_n(\beta_0, t) - \sqrt{k_n} \int_0^t A_n(s) ds$ as:

$$\int_0^t U_n^*(\beta_0, s) ds - \sqrt{k_n} \int_0^t A_n(s) ds \xrightarrow[n \rightarrow \infty]{\mathcal{L}} C_1(\beta, \beta_0) \int_0^t \mathcal{W}(s) ds.$$

In consequence, we have convergence in probability:

$$\frac{1}{\sqrt{k_n}} U_n^*(\beta_0, t) \xrightarrow[n \rightarrow \infty]{P} C_2 \int_0^t \{\beta(s) - \beta_0\} ds,$$

and

$$\frac{1}{\sqrt{k_n}} J_n(\beta_0, t) \xrightarrow[n \rightarrow \infty]{P} C_2 \int_0^t \int_0^s (\beta(u) - \beta_0) du ds.$$

Recall that $\lim_{n \rightarrow \infty} \sqrt{k_n} = \infty$, and for any $s \in [0, 1]$, $\int_0^s \{\beta(u) - \beta_0\} du \neq 0$, which together imply:

$$\lim_{n \rightarrow \infty} P \left(\frac{1}{\sqrt{t}} |U_n^*(\beta_0, t)| \geq z^{\alpha/2} \right) = 1, \quad \lim_{n \rightarrow \infty} P \left((3t^{-3})^{1/2} |J_n(\beta_0, t)| \geq z^{\alpha/2} \right) = 1.$$

$\beta(t)$	cens. (%)	Log- rank	$U_n^*(0, 1)$	$J_n(0, 1)$	$M_n(0, 1)$	$M_n^{1.5}(0, 1)$	Lee	Yang Prentice
0	30	4.8	4.5	4.9	5.0	4.4	6.1	6.4
	50	5.2	5.0	5.4	5.0	5.0	6.3	6.8
	70	5.1	4.7	4.8	4.8	4.4	5.8	6.2
0.5	30	50.9	49.9	39.6	47.3	48.8	50.4	54.2
	50	51.0	50.2	39.0	47.4	48.9	51.0	53.8
	70	51.7	50.8	40.9	48.7	49.7	50.8	55.5
0.8	30	87.1	86.6	74.6	84.1	85.3	86.3	89.0
	50	87.2	86.7	75.8	85.3	85.7	86.8	88.4
	70	88.0	87.1	76.8	85.8	85.9	87.3	89.0
$\mathbf{1}_{t \leq 0.3}$	30	23.7	22.2	40.1	34.6	26.2	19.8	40.4
	50	18.4	17.6	30.8	26.9	21.0	16.1	34.4
	70	13.5	12.7	21.7	18.5	14.4	11.1	26.6
$\mathbf{1}_{t \leq 0.5}$	30	55.5	52.4	73.2	69.4	59.0	49.5	69.1
	50	44.3	41.9	61.0	55.9	46.7	37.3	57.7
	70	32.8	30.3	44.4	40.9	32.6	27.2	43.6
$\mathbf{1}_{t \leq 0.7}$	30	83.1	80.6	88.2	86.9	82.1	79.0	87.3
	50	72.4	69.4	77.3	76.4	71.1	66.6	78.1
	70	56.0	51.9	60.2	59.0	53.3	49.2	61.2
$-1.5\mathbf{1}_{t \leq 0.5} + 1.5\mathbf{1}_{t \geq 0.5}$	30	17.1	15.1	85.7	81.3	84.8	27.5	61.5
	50	13.5	11.7	72.7	67.7	69.6	39.5	42.3
	70	12.0	11.4	56.0	49.5	49.0	37.0	25.8
$1.5\mathbf{1}_{t \leq 0.5} - 1.5\mathbf{1}_{t \geq 0.5}$	30	15.9	13.8	85.3	81.4	84.3	26.0	87.8
	50	14.0	12.4	74.0	68.0	71.0	40.6	77.1
	70	11.0	10.2	55.2	49.1	48.7	36.7	61.6
$1.5(1-t)$	30	85.0	82.9	90.7	90.0	84.7	81.3	90.1
	50	74.5	71.0	81.5	79.9	73.2	68.6	81.2
	70	58.0	54.9	64.7	62.8	56.3	52.0	65.8
$2(1-t)^2$	30	73.5	70.8	89.7	87.0	78.2	67.8	88.1
	50	60.7	57.4	78.7	74.7	64.6	54.0	77.9
	70	46.7	43.0	62.1	58.1	47.2	39.4	62.8

Table 11.3: Empirical level of significance and power of each test (in %) based on 3000 simulated datasets for each $\beta(t)$ and each rate of censoring. Each dataset has 100 elements (50 subjects per group). The tests have nominal levels of 5%.

$\beta(t)$	cens. (%)	Log- rank	$U_n^*(0, 1)$	$J_n(0, 1)$	$M_n(0, 1)$	$M_n^{1.5}(0, 1)$	Lee	Yang Prentice
0	30	4.9	4.7	4.7	4.3	4.7	5.7	6.6
	50	5.4	5.5	4.9	5.1	5.3	6.5	7.1
	70	4.9	4.7	4.3	4.4	4.7	5	4.9
0.5	30	35.6	34.6	26.9	32.5	33.5	36.4	38.1
	50	27.6	26.7	19.5	25.4	26.0	27.7	29.7
	70	19.9	19.9	14.8	18.6	19.3	21.1	21.1
0.8	30	68.2	67.6	53.7	64.8	66.0	68.9	70.4
	50	53.7	52.8	40.8	50.6	51.3	54.6	56.2
	70	38.9	38.2	28.7	36.6	37.1	40.0	39.4
$\mathbf{1}_{t \leq 0.3}$	30	16.5	15.2	25.4	22.4	17.6	14.9	29.3
	50	13.5	12.7	19.9	17.6	13.7	12.5	23.6
	70	10.5	10.2	15.0	13.2	10.5	9.2	16.6
$\mathbf{1}_{t \leq 0.5}$	30	36.8	34.1	50.5	46.3	38.2	33.0	49.2
	50	29.3	26.7	40.2	36.6	28.9	25.0	40.1
	70	23.1	21.0	28.7	26.5	21.6	18.8	29.0
$\mathbf{1}_{t \leq 0.7}$	30	61.9	58.3	66.4	64.7	59.1	57.0	67.4
	50	51.6	47.9	53.8	54.4	48.8	47.2	57.5
	70	38.4	34.9	37.9	38.4	34.8	33.4	40.2
$-1.5\mathbf{1}_{t \leq 0.5}$ $+1.5\mathbf{1}_{t \geq 0.5}$	30	11.9	10.2	65.8	59.4	61.0	15.6	42.7
	50	10.7	9.3	51.8	45.8	43.3	22.4	29.6
	70	9.4	8.4	36.6	30.8	25.5	20.2	18.9
$1.5\mathbf{1}_{t \leq 0.5}$ $-1.5\mathbf{1}_{t \geq 0.5}$	30	13	11.2	66.6	59.9	61.7	16.8	67.6
	50	11.2	9.7	51.3	44.8	43.4	22.6	51.3
	70	8.7	8.1	36.7	31.2	26.0	19.1	38.6
$1.5(1-t)$	30	65.8	62.4	71.6	70.7	63.8	61.4	73.8
	50	53.1	48.6	57.7	57.1	50.3	46.9	60.9
	70	40.8	37.6	42.4	42.1	37.8	36.3	44.7
$2(1-t)^2$	30	53.3	49.4	69.8	65.6	55.5	48.4	71.0
	50	44.9	41.3	57.4	53.4	44.2	39.3	58.3
	70	29.7	27.6	37.5	34.7	29.2	25.2	37.5

Table 11.4: Empirical level of significance and power of each test (in %) based on 3000 simulated datasets for each $\beta(t)$ and each rate of censoring. Each dataset has 60 elements (30 subjects per group). The tests have nominal levels of 5%.

$\beta(t)$	cens. (%)	Log- rank	$U_n^*(0, 1)$	$J_n(0, 1)$	$M_n(0, 1)$	$M_n^{1.5}(0, 1)$	Lee	Yang Prentice
0	30	5.5	5.4	4.6	5.2	5.1	6.7	6.3
	50	5.2	5.3	4.9	5	5.2	6.2	6.6
	70	5.2	5.2	5.5	5.3	5.1	5.8	6.4
0.5	30	79.8	79.3	68.4	76.7	78.8	78.4	81.7
	50	67	66.3	54.8	64.2	65.4	64.2	69
	70	50.3	48.9	41.1	47.9	48.2	47.2	52.2
0.8	30	99.2	99.1	97.1	98.9	98.8	99.1	99.3
	50	96.9	96.5	91.7	96.3	95.7	95.9	97.1
	70	86.7	86.3	76.8	84.8	85	84.6	87.8
$\mathbf{1}_{t \leq 0.3}$	30	40.3	38.9	69.6	63.2	52.4	33.7	62.7
	50	29.9	28.8	55.6	49	38.4	24.3	49.5
	70	22.3	21.5	39.8	35.2	26.4	17.6	39.9
$\mathbf{1}_{t \leq 0.5}$	30	82.9	81.6	96.4	95	89.5	78	93.3
	50	69.8	68.2	88.7	85.4	76.2	63.4	83.6
	70	53.7	51	72.2	67.9	57.4	45.1	67.5
$\mathbf{1}_{t \leq 0.7}$	30	98.2	97.9	99.3	99.1	98.3	97.2	99
	50	94.2	93	96.7	96.1	94	91.3	96.2
	70	83.8	80.9	88.1	87.3	82.7	77.8	87.8
$-1.5\mathbf{1}_{t \leq 0.5} + 1.5\mathbf{1}_{t \geq 0.5}$	30	23.9	19.9	98.9	98.4	98.8	56.4	94
	50	20.7	17.2	95.3	93.5	95.1	73.3	70
	70	15.6	13	84.2	80.4	83.1	63.3	41.8
$1.5\mathbf{1}_{t \leq 0.5} - 1.5\mathbf{1}_{t \geq 0.5}$	30	23.7	19.7	98.7	98.3	98.7	53.9	98.8
	50	20.6	17	94.8	92.9	94.3	74	96.9
	70	15.7	13.7	83.1	79.3	81.6	61	88.5
$1.5(1-t)$	30	98.7	98.4	99.8	99.7	98.8	98.3	99.6
	50	95.5	94.2	98.4	97.9	95.8	93	97.8
	70	86.5	84.3	92.1	91.1	86.6	81.7	91.4
$2(1-t)^2$	30	95.5	94.8	99.7	99.5	98	93.3	99.5
	50	89	87.1	98.1	97.3	93.7	84.1	97.3
	70	73.1	70.1	89.7	86.9	78.3	65.7	87.8

Table 11.5: Empirical level of significance and power of each test (in %) based on 3000 simulated datasets for each $\beta(t)$ and each rate of censoring. Each dataset has 200 elements (100 subjects per group). The tests have nominal levels of 5%.

$\beta(t)$	Covariate	$U_n^*(0,1)$	$J_n(0,1)$	$M_n(0,1)$	$M_n^{1.5}(0,1)$
0	Normal	4.9	4.5	4.9	4.9
	Uniform	5.4	4.8	5.2	5.5
	Exponential	5.0	5.0	5.0	5.3
0.5	Normal	46.9	40.6	45.9	46.4
	Uniform	48.7	41.0	46.5	48.2
	Exponential	48.3	38.5	45.4	47.9
0.8	Normal	84.6	76.4	83.5	84.1
	Uniform	84.5	76.3	82.1	84.0
	Exponential	84.7	76.1	83.0	83.9
$\mathbf{1}_{t \leq 0.3}$	Normal	22.4	41.7	37.2	28.2
	Uniform	21.2	41.1	35.7	27.4
	Exponential	21.6	42.1	36.1	27.7
$\mathbf{1}_{t \leq 0.5}$	Normal	49.1	71.2	66.7	56.4
	Uniform	50.6	71.5	66.7	56.9
	Exponential	49.0	71.5	67.5	56.9
$\mathbf{1}_{t \leq 0.7}$	Normal	77.0	87.3	84.9	79.4
	Uniform	75.9	85.6	83.6	77.5
	Exponential	75.7	85.9	83.7	78.1
$-1.5\mathbf{1}_{t \leq 0.5} + 1.5\mathbf{1}_{t \geq 0.5}$	Normal	9.8	71.9	66.2	70.1
	Uniform	9.1	73.3	67.9	71.4
	Exponential	9.6	73.1	67.9	71.2
$1.5\mathbf{1}_{t \leq 0.5} - 1.5\mathbf{1}_{t \geq 0.5}$	Normal	9.4	74.8	69.6	73.0
	Uniform	9.5	73.8	68.2	71.9
	Exponential	9.6	72.5	67.6	70.5
$1.5(1-t)$	Normal	81.5	91.4	89.4	84.0
	Uniform	79.9	91.3	89.9	83.2
	Exponential	79.7	90.5	88.7	83.
$2(1-t)^2$	Normal	70.9	90.5	87.9	79.7
	Uniform	68.7	89.6	86.4	78.6
	Exponential	70.2	90.2	87.1	79.1

Table 11.6: Empirical levels of significance and test powers (in %) based on 3000 datasets for each $\beta(t)$ and covariate distribution. The rate of censoring is set at 30%. Each dataset is of size 100.

Appendix A

Probability

A.1 Essential tools for survival problems

We recall some of the fundamental tools used to establish the inferential basis for our models. The main ideas of stochastic processes, in particular Brownian motion and functions of Brownian motion, are explained in terms that are not overly technical. The background to this, i.e., distribution theory and large sample results, is recalled. Rank invariance is an important concept, i.e., the ability to transform some variable, usually time, via monotonic increasing transformations without having an impact on inference. These ideas hinge on the theory of order statistics and the basic notions of this theory are recalled. An outline of the theory of counting processes and martingales is presented, once again without leaning too heavily upon technical measure-theoretic constructions. The important concepts of explained variation and explained randomness are outlined in elementary terms, i.e., only with reference to random variables, and at least initially, making no explicit appeal to any particular model. This is important since the concepts are hardly any less fundamental than a concept such as variance itself. They ought therefore stand alone, and not require derivation as a particular feature of some model. In practice, of course, we may need to estimate conditional distributions and making an appeal to a model at this point is quite natural.

A.2 Integration and measure

The reader is assumed to have some elementary knowledge of set theory and calculus. We do not recall here any of the basic notions concerning limits, continuity, differentiability, convergence of infinite series, Taylor series, and so on and the rusty reader may want to refer to any of the many standard calculus

texts when necessary. One central result which is frequently called upon is the mean value theorem. This can be deduced as an immediate consequence of the following result known as Rolle's theorem:

Theorem A.1. *If $f(x)$ is continuously differentiable at all interior points of the interval $[a, b]$ and $f(a) = f(b)$, then there exists a real number $\xi \in (a, b)$ such that $f'(\xi) = 0$.*

A simple sketch would back up our intuition that the theorem would be correct. Simple though the result appears to be, it has many powerful implications including:

Theorem A.2. *If $f(x)$ is continuously differentiable on the interval $[a, b]$, then there exists a real number $\xi \in (a, b)$ such that*

$$f(b) = f(a) + (b - a)f'(\xi).$$

When $f(x)$ is monotone then ξ is unique. This elementary theorem can form the basis for approximation theory and series expansions such as the Edgeworth and Cornish-Fisher (see Section A.10). For example, a further immediate corollary to the above theorem obtains by expanding in turn $f'(\xi)$ about $f'(a)$ whereby:

Corollary A.1. *If $f(x)$ is at least twice differentiable on the interval $[a, b]$ then there exists a real number $\xi \in (a, b)$ such that*

$$f(b) = f(a) + (b - a)f'(a) + \frac{(b - a)^2}{2}f''(\xi).$$

The ξ of the theorems and corollary would not typically be the same and we can clearly continue the process, resulting in an expansion of $m + 1$ terms, the last term being the m th derivative of $f(x)$, evaluated at some point $\xi \in (a, b)$ and multiplied by $(b - a)^m/m!$. An understanding of Riemann integrals as limits of sums, definite and indefinite integrals, is mostly all that is required to follow the text. It is enough to know that we can often interchange the limiting processes of integration and differentiation. The precise conditions for this to be valid are not emphasized. Indeed, we almost entirely avoid the tools of real analysis. The Lebesgue theory of measure and integration is on occasion referred to, but a lack of knowledge of this will not hinder the reader. Likewise we will not dig deeply into the measure-theoretic aspects of the Riemann-Stieltjes integral apart from the following extremely useful construction:

Definition A.1. *The Riemann integral of the function $f(x)$ with respect to x , on the interval $[a, b]$, is the limit of a sum $\sum \Delta_i f(x_{i-1})$, where $\Delta_i = x_i - x_{i-1} > 0$, for an increasing partition of $[a, b]$ in which $\max \Delta_i$ goes to zero.*

The limit is written $\int_a^b f(x)dx$ and can be seen to be the area under the curve $f(x)$ between a and b . If $b = \infty$ then we understand the integral to exist if the limit exists for any $b > 0$, the result itself converging to a limit as $b \rightarrow \infty$.

Similarly for $a = -\infty$. Now, instead of only considering small increments in x , i.e., integrating with respect to x , we can make use of a more general definition. We have:

Definition A.2. *The Riemann-Stieltjes integral of the function $f(x)$ with respect to $g(x)$ is the limit of a sum $\sum\{g(x_i) - g(x_{i-1})\}f(x_{i-1})$, for an increasing partition of $[a, b]$ in which, once again, $\max \Delta_i$ goes to zero.*

The limit is written $\int_a^b f(x)dg(x)$ and, in the special case where $g(x) = x$, reduces to the usual Riemann integral. For functions, necessarily continuous, whereby $g(x)$ is an antiderivative of, say, $h(x)$ and can be written $g(x) = \int_{-\infty}^x h(u)du$ then the Stieltjes integral coincides with the Riemann integral $\int f(x)h(x)dx$. On the other hand whenever $g(x)$ is a step function with a finite or a countable number of discontinuities then $\int f(x)dg(x)$ reduces to a sum, the only contributions arising at the discontinuities themselves. This is of great importance in statistical applications where step functions naturally arise as estimators of key functions. A clear example of a step function of central importance is the empirical distribution function, $F_n(x)$ (this is discussed in detail in Appendix D). We can then write the sample mean $\bar{x} = \int udF_n(u)$ and the population mean $\mu = \int udF(u)$, highlighting an important concept, that fluctuations in the sample mean can be considered a consequence of fluctuations in $F_n(x)$ as an estimate of $F(x)$. Consider the following theorem, somewhat out of sequence in the text but worth seeing here for its motivational value. The reader may wish to take a glance ahead at Appendix A.4 and Appendix C.5.

Theorem A.3. *For every bounded continuous function $h(x)$, if $F_n(x)$ converges in distribution to $F(x)$, then $\int h(x)dF_n(x)$ converges in distribution to $\int h(x)dF(x)$.*

This is the Helly-Bray theorem. The theorem will also hold (see the Exercises) when $h(x)$ is unbounded provided that some broad conditions are met. A deep study of $F_n(x)$ as an estimator of $F(x)$ is then all that is needed to obtain insight into the sample behavior of the empirical mean, the empirical variance and many other quantities. Of particular importance for the applications of interest to us here, and developed, albeit very briefly, in Appendix B.3, is the fact that, letting $M(x) = F_n(x) - F(x)$, then

$$E \left\{ \int h(x)dM(x) \right\} = \int h(x)dF(x) - \int h(x)dF(x) = 0, \quad (\text{A.1})$$

a seemingly somewhat innocuous result until we interchange the order of integration (expectation, denoted by E being an integral operator), and under some very mild conditions on $h(x)$ described in Appendix B.3, we obtain a formulation of great generality and into which can be fit many statistical problems arising in the context of stochastic processes (see Appendix B.3).

A.3 Random variables and probability measure

The possible outcomes of any experiment are called events where any event represents some subset of the sample space. The sample space is the collection of all events, in particular the set of elementary events. A random variable X is a function from the set of outcomes to the real line. A probability measure is a function of some subset of the real line to the interval $[0,1]$. Kolmogorov (2018) provides axioms which enable us to identify any measure as being a probability measure. These axioms appear very reasonable and almost self-evident, apart from the last, which concerns assigning probability measure to infinite collections of events. There are, in a well defined sense, many more members in the set of all subsets of any infinite set than in the original set itself, an example being the set of all subsets of the positive integers which has as many members as the real line. This fact would have hampered the development of probability without the inclusion of Kolmogorov's third axiom which, broadly says that the random variable is measurable, or, in other words, that the sample space upon which the probability function is defined is restricted in such a way that the probability we associate with the sum of an infinite collection of mutually exclusive events is the same as the sum of the probabilities associated with each composing event. We call such a space a measurable space or a Borel space, the core idea being that the property of additivity for infinite sums of probabilities, as axiomatized by Kolmogorov, holds. The allowable operations on this space are referred to as a sigma-algebra. Subsets of a sigma-algebra—the most common case being under some kind of conditioning—are referred to as sub sigma-algebras and inherit the axiomatic properties defined by Kolmogorov.

A great deal of modern probability theory is based on measure-theoretic questions, questions that essentially arise from the applicability or otherwise of Kolmogorov's third axiom in any given context. This is an area that is highly technical and relatively inaccessible to non-mathematicians, or even to mathematicians lacking a firm grounding in real analysis. The influence of measure theory has been strongly felt in the area of survival analysis over the last 20 or so years and much modern work is now of a very technical nature. Even so, none of the main statistical ideas, or any of the needed demonstrations in this text, require such knowledge. We can therefore largely avoid measure-theoretic arguments, although some of the key ideas that underpin important concepts in stochastic processes are touched upon whenever necessary. The reader is expected to understand the meaning of the term *random variable* on some level.

Observations or outcomes as random variables and, via models, the probabilities we will associate with them are all part of a theoretical, and therefore artificial, construction. The hope is that these probabilities will throw light on real applied problems and it is useful to keep in mind that, in given contexts, there may be more than one way to set things up. Conditional expectation is a recurring central topic but can arise in ways that we did not originally anticipate.

We may naturally think of the conditional expected survival time given that a subject begins the study under, say, some treatment. It may be less natural to think of the conditional expectation of the random variable we use as a treatment indicator given some value of time after the beginning of treatment. Yet, this latter conditional expectation, as we shall see, turns out to be the more relevant for many situations.

A.4 Convergence for random variables

Simple geometrical constructions (intervals, balls) are all that are necessary to formalize the concept of convergence of a sequence in real and complex analysis. For random variables there are a number of different kinds of convergence, depending upon which aspect of the random variable we are looking at. Consider any real value Z and the sequence $U_n = Z/n$. We can easily show that $U_n \rightarrow 0$ as $n \rightarrow \infty$. Now let U_n be defined as before except for values of n that are prime. Whenever n is a prime number then $U_n = 1$. Even though, as n becomes large, U_n is almost always arbitrarily close to zero, a simple definition of convergence would not be adequate and we need to consider more carefully the sizes of the relevant sets in order to accurately describe this. Now, suppose that Z is a uniform random variable on the interval $(0,1)$. We can readily calculate the probability that the distance between U_n and 0 is greater than any arbitrarily small positive number ϵ and this number goes to zero with n . We have convergence in probability. Nonetheless there is something slightly erratic about such convergence, large deviations occurring each time that n is prime. When possible, we usually prefer a stronger type of convergence. If, for all integer values m greater than n and as n becomes large, we can assert that the probability of the distance between U_m and 0 being greater than some arbitrarily small positive number goes to zero, then such a mode of convergence is called strong convergence. This stronger convergence is also called convergence with probability one or almost sure convergence. Consider also $(n+3)U_n$. This random variable will converge almost surely to the random variable Z . But, also, we can say that the distribution of $\log_e(n+3)U_n$, at all points of continuity z , becomes arbitrarily close to that of a standard exponential distribution. This is called convergence in distribution. The three modes of convergence are related by:

Theorem A.4. *Convergence with probability one implies convergence in probability. Convergence in probability implies convergence in distribution.*

Note also that, for a sequence that converges in probability, there exists a subsequence that converges with probability one. This latter result requires the tools of measure theory and is not of wide practical applicability since we may not have any obvious way of identifying such a subsequence. Added conditions can enable the direction of the “implies” arrow to be inverted. For example convergence in distribution implies convergence in probability when the limiting

random variable is constant. In theoretical work it can sometimes be easier to obtain results for weak rather than strong convergence. However, in practical applications, we usually need strong (almost sure, “with probability one”) convergence since this corresponds in a more abstract language to the important idea that, as our information increases, our inferences becomes more precise.

Convergence of functions of random variables

In constructing models and establishing inference for them we will frequently appeal to two other sets of results relating to convergence. The first of these is that, for a continuous function $g(z)$, if Z_n converges in probability to c , then $g(Z_n)$ converges in probability to $g(c)$ and, if Z_n converges in distribution to Z , then $g(Z_n)$ converges in distribution to $g(Z)$. The second set, Slutsky’s theorem (a proof is given in Randles and Wolfe (1979)), enables us to combine modes of convergence. In particular, for modeling purposes, if a convergence in distribution result holds when the parameters are known, then it will continue to hold when those same parameters are replaced by consistent estimators. This has great practical value.

A.5 Topology and distance measures

Here, we describe some further tools that are helpful in determining large sample behavior. Such behavior, in particular almost sure convergence and the law of the iterated logarithm allow us to anticipate what we can expect in moderate to large sample sizes. Small sample behavior is considered separately.

Compact topological spaces

For readers interested in the proofs given at the end of the chapters, we will need to appeal on occasion to more formal concepts of convergence. These require that the way in which we measure distance be made formal and the spaces in which sit the functions of interest need to be defined. The main results we need are obtained for spaces that are bounded and closed and these are called compact spaces.

Topology of the spaces $D[0, 1]$ and $C[0, 1]$

We denote by $(C[0, 1], E)$ the space of continuous functions defined on the closed interval $[0, 1]$ assuming values in the space E and $(D[0, 1], E)$ the space of right continuous functions with left limits (called cadlag, a French acronym of this definition). These are defined on $[0, 1]$ with values in E . The context in which we use these concepts relates to random elements of these two spaces with $E = \mathbb{R}^p$, $p \in \mathbb{N}^*$. In our context, the interval $[0, 1]$ corresponds to a transformation of time to this interval, and the context usually makes it clear whether we are referring

to transformed time or time on the original scale. Moving between the scales is described precisely.

The interval $[0,1]$ arises in a very natural way when dealing with distribution functions and it also appears natural to work with uniform distance and to consider uniform convergence as the basic concept behind our metric definition. With this in mind we have:

Definition A.3. *The uniform distance d between two elements of $(C[0,1], \mathbb{R})$ is defined by:*

$$d(f,g) = \sup_{0 \leq t \leq 1} |f(t) - g(t)|, \quad f, g \in (C[0,1], \mathbb{R}). \quad (\text{A.2})$$

In order to get around the problem of functions with jumps at the discontinuities, Skorokhod developed a more suitable metric providing the basis for the Skorokhod topology. Specifically, we have the definition:

Definition A.4. *Skorokhod distance δ is defined by:*

$$\delta(f,g) = \inf_{\lambda \in \Lambda} \{d(f, g(\lambda)) \vee d(\lambda, I)\}, \quad f, g \in (D[0,1], \mathbb{R}), \quad (\text{A.3})$$

where I is the identity transform on $[0,1]$, Λ indicates the class of structure preserving homomorphisms of $[0,1]$ into itself such that $\lambda(0) = 0$ and $\lambda(1) = 1$ for all $f\lambda \in \Lambda$ and $a \vee b = \max(a, b)$.

The idea is to not allow any jumps to dominate and is more fully explored and explained in Billingsley (1999). Note also that

$$\delta(f,g) \leq d(f,g), \quad f, g \in (D[0,1], \mathbb{R}). \quad (\text{A.4})$$

which implies the following result:

Proposition A.1. (Billingsley, 1999). *Consider the sequence of functions, $(f_n)_{n \in \mathbb{N}}$ of $(D[0,1], \mathbb{R})$ and $f \in (D[0,1], \mathbb{R})$. If*

$$\lim_{n \rightarrow +\infty} d(f_n, f) = 0, \quad \text{then} \quad \lim_{n \rightarrow +\infty} \delta(f, g_n) = 0.$$

In other words, in the space $(D[0,1], \mathbb{R})$, convergence with respect to a topology of uniform convergence implies convergence with respect to the topology of Skorokhod. At the same time, if the limit is a continuous function then the two forms of convergence are equivalent.

Proposition A.2. (Billingsley, 1999). *For all $f \in (C[0,1], \mathbb{R})$; $(g_n)_{n \in \mathbb{N}} \in (D[0,1], \mathbb{R})$,*

$$\lim_{n \rightarrow +\infty} d(f, g_n) = 0 \iff \lim_{n \rightarrow +\infty} \delta(f, g_n) = 0.$$

Some simple properties and the fact that cadlag functions arise naturally when considering empirical distribution functions, enable us to find results easily.

The uniform limit of a sequence of cadlag functions is cadlag and the proposition shows that it is equivalent to show convergence of a sequence of cadlag functions to a continuous function with a uniform convergence topology or the Skorokhod topology.

Topologies of $(D[0,1], \mathbb{R}^p)$

When we are specifically interested in higher dimensions then we can consider the space of cadlag functions of $[0,1]$ in \mathbb{R}^p where p is an integer strictly greater than 1. In this case, several distances can be considered and, for the proofs in the multivariate setting given here, we make use of the product distance of Skorokhod (Chauvel, 2014).

Definition A.5. Let $f = (f_1, \dots, f_p)$ and $g = (g_1, \dots, g_p)$ be two functions on $(D[0,1], \mathbb{R}^p)$, with f_i and g_i functions on $(D[0,1], \mathbb{R})$, for $i = 1, \dots, p$. The Skorokhod product distance between f and g is defined by:

$$\delta_p(f, g) = \sum_{i=1}^p \delta(f_i, g_i).$$

In some cases it is worth considering the Skorokhod distance (Equation A.3) in which the absolute value in Definition A.2 of uniform distance is replaced by the sup norm such that $\|a\| = \max_{i=1, \dots, p} |a_i|$, for $a \in \mathbb{R}^p$. The induced topology corresponding to this distance is referred to as having the strong Skorokhod property. The following proposition also turns out to be a useful tool.

Proposition A.3. For all functions in $f \in (C[0,1], \mathbb{R}^p)$ and for all function sequences, $(g_n)_{n \in \mathbb{N}^*} \in (D[0,1], \mathbb{R}^p)$, if $(g_n)_{n \in \mathbb{N}^*}$ converges to f with respect to uniform convergence topology in \mathbb{R}^p , i.e.

$$\lim_{n \rightarrow \infty} \sup_{0 \leq t \leq 1} \|f(t) - g_n(t)\| = 0, \text{ then } \lim_{n \rightarrow \infty} \delta_p(f, g_n) = 0.$$

To summarize, if a sequence of functions of $(D[0,1], \mathbb{R}^p)$ converges to a function of $(C[0,1], \mathbb{R}^p)$ with respect to the topology of uniform convergence, then we can say that convergence is also obtained with respect to the product topology of Skorokhod. We will use these results in the chapters concerning the regression effect process in order to anticipate the properties for large samples of certain empirical processes.

A.6 Distributions and densities

We anticipate that most readers will have some familiarity with the basic ideas of a distribution function $F(t) = \Pr(T < t)$, a density function $f(t) = dF(t)/dt$, expectation and conditional expectation, the moments of a random variable, and

other basic tools. Nonetheless we will go over these elementary notions in the context of survival in the next chapter. We write

$$E \psi(T) = \int \psi(t) f(t) dt = \int \psi(t) dF(t)$$

for the expected value of the function $\psi(T)$. Such an expression leaves much unsaid, that $\psi(t)$ is a function of t and therefore $\psi(T)$ itself random, that the integrals exist, the domain of definition of the function being left implicit, and that the density $f(t)$ is an antiderivative of the cumulative distribution $F(t)$ (in fact, a slightly weaker mathematical construct, absolute continuity, is enough but we do not feel the stronger assumption has any significant cost attached to it). There is a wealth of solid references for the rusty reader on these topics, among which Billingsley (1999), Rao et al. (1973) and Serfling (2009) are particularly outstanding. It is very common to wish to consider some transformation of a random variable, the simplest situation being that of a change in origin or scale. The distribution of sums of random variables arises by extension to the bivariate and multivariate cases.

Theorem A.5. Suppose that the distribution of X is $F(x)$ and that $F'(x) = f(x)$. Suppose that $y = \phi(x)$ is a monotonic function of x and that $\phi^{-1}(y) = x$. Then, if the distribution of Y is $G(y)$ and $G'(y) = g(y)$,

$$G(y) = F\{\phi^{-1}(y)\}; \quad g(y) = f\{\phi^{-1}(y)\} \left| \frac{d\phi(x)}{dx} \right|_{x=\phi^{-1}(y)}^{-1} \quad (\text{A.5})$$

Theorem A.6. Let X and Y have joint density $f(x,y)$. Then the density $g(w)$ of $W = X + Y$ is given by

$$g(w) = \int_{-\infty}^{\infty} f(x, w-x) dx = \int_{-\infty}^{\infty} f(w-y, y) dy. \quad (\text{A.6})$$

A result for $W = X - Y$ follows immediately and, in the case of X and Y being independent, the corresponding expression can also be written down readily as a product of the two respective densities. Similar results hold for the product or ratio of random variables (see Rohatgi and Saleh (2015), Chapter 8) but, since we have no call for them in this work, we do not write them down here. An immediate corollary that can give an angle on small sample behavior of statistics that are written as sums is:

Corollary A.2. Let X_1, \dots, X_n be independent, not always identically distributed, continuous random variables with densities $f_1(x)$ to $f_n(s)$ respectively. Let $S_n = \sum_{j=1}^n X_j$. Then the density, $g_n(s)$, of S_n is given by

$$g_n(s) = \int_{-\infty}^{\infty} g_{n-1}(s-x) f_n(x) dx.$$

This result can be used iteratively for building up successive solutions by carrying out the integration. The integration itself will mostly be not particularly tractable and can be evaluated using numerical routines. Note the difference between making a large sample statistical approximation to the sum and that of a numerical approximation to the integral. The integral expression itself is an exact result.

Normal distribution

A random variable X is taken to be a normal variate with parameters μ and σ when we write $X \sim \Phi(\mu, \sigma^2)$. The parameters μ and σ^2 are the mean and variance, respectively, so that $\sigma^{-1}(X - \mu) \sim \Phi(0, 1)$. The distribution $\Phi(0, 1)$ is called the standard normal. The density of the standard normal variate, that is, having mean zero and variance one, is typically denoted $\phi(x)$ and the cumulative distribution $\Phi(x)$. The density $f(x)$, for $x \in (-\infty, \infty)$ is given by

$$f(x) = \phi(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right].$$

For stochastic processes described below, Brownian motion relates to a Gaussian process, that is, it has been standardized, in an analogous way that the standard normal relates to any other normal distribution. For the normal distribution, all cumulants greater than 2 are equal to zero. Simple calculations show that, for $X \sim \mathcal{N}(0, 1)$, then $E(X^r) = (r-1)(r-3)\dots 3.1$. Thus, all odd moments are equal to zero and all even moments are expressible in terms of the variance. The normal distribution is of very great interest in view of it frequently being the large sample limiting distribution for sums of random variables. These arise naturally via simple estimating equations. These topics are looked at in greater detail below.

The multivariate normal can be characterized in various ways. If and only if all marginal distributions and all conditional distributions are normal then we have multivariate normality. If and only if all linear combinations are univariate normal then we have multivariate normality. It is only necessary to be able to evaluate the standard normal integral, $\Phi(x) = 1 - \int_x^\infty \phi(u)du$, since any other normal distribution, $f(x)$, can be put in this form via the linear transformation $(X - \mu)/\sigma$. Tables, calculator, and computer routines can approximate the numerical integral. Otherwise, it is worth bearing in mind the following:

Lemma A.1. *Upper and lower bounds for the normal integral can be obtained from*

$$\frac{x}{1+x^2} e^{-x^2/2} < \int_x^\infty e^{-u^2/2} du < \frac{1}{x} e^{-x^2/2}.$$

The lemma tells us that we expect $1 - \Phi(x)$ to behave like $\phi(x)/x$ as x increases. The ratio $\phi(x)/x$ is known as Mill's ratio. Approximate calculations are then possible without the need to resort to sophisticated algorithms, although, in modern statistical analysis, it is now so commonplace to routinely use computers that the value of the lemma is rather limited. The normal distribution plays an important role in view of the central limit theorem described below but also note the interesting theorem of Cramér (2004) whereby, if a finite sum of independent random variables is normal, then each variable itself is normal. Cramer's theorem might be contrasted with central limit theorems whereby sums of random variables, under broad conditions, approach the normal as the sum becomes infinitely large. The normal distribution is important since it provides the basis for Brownian motion and this is the key tool that we will use for inference throughout this text.

Uniform distribution and the probability integral transform

For the standard uniform distribution in which $u \in [0, 1]$, $f(u) = 1$ and $F(u) = u$. Uniform distributions on the interval $[a, b]$ correspond to the density $f(u) = 1/(b - a)$ but much more important is the fact that for any continuous distribution, $G(t)$, we can say:

Theorem A.7. *For the random variable T , having distribution $G(t)$, letting $U_1 = G(T)$ and $U_2 = 1 - G(T)$, then both U_1 and U_2 have a standard uniform distribution.*

This central result, underpinning a substantial body of work on simulation and resampling, is known as the probability integral transform. Whenever we can invert the function G , denoted G^{-1} , then, from a single uniform variate U we obtain the two variates $G^{-1}(U)$ and $G^{-1}(1 - U)$ which have the distribution G . The two variates are of course not independent but, in view of the strong linearity property of expectation (the expectation of a linear function of random variables is the same linear function of the expectations), we can often use this to our advantage to improve precision when simulating. Another interesting consequence of the probability integral transform is that there exists a transformation of a variate T , with any given distribution, into a variate having any other chosen distribution. Specifically, we have:

Corollary A.3. *For any given continuously invertible distribution function H , and continuous distribution $G(t)$, the variate $H^{-1}\{G(T)\}$ has distribution H .*

In particular, it is interesting to consider the transformation $\Phi^{-1}\{G_n(T)\}$ where G_n is the empirical estimate (discussed below) of G . This transformation, which preserves the ordering, makes the observed distribution of observations as close to normal as possible. Note that since the ordering is preserved, use of the transformation makes subsequent procedures nonparametric in as much as the original distribution of T has no impact. For the problems of interest to

us in survival analysis we can use this in one of two ways: firstly, to transform the response variable time in order to eliminate the impact of its distribution, and secondly, in the context of regression problems, to transform the distribution of regressors as a way to obtain greater robustness by reducing the impact of outliers.

Exponential distribution and cumulative hazard transformation

The standard exponential distribution is defined on the positive real line $(0, \infty)$. We have, for $u \in (0, \infty)$, $f(u) = \exp(-u)$ and $F(u) = 1 - \exp(-u)$. An exponential distribution with mean $1/\alpha$ and variance $1/\alpha^2$ has density $f(u) = \alpha \exp(-\alpha u)$ and cumulative distribution $F(u) = 1 - \exp(-\alpha u)$. The density of a sum of m independent exponential variates having mean $1/\alpha$, is an Erlang density whereby $f(u) = \alpha(\alpha u)^{m-1} \exp(-\alpha u)/\Gamma(m)$ and where $\Gamma(m) = \int_0^\infty \exp(-u)u^{m-1} du$. The gamma distribution has the same form as the Erlang although, for the gamma, the parameter m can be any real positive number and is not restricted to being an integer. An exponential variate U can be characterized as a power transformation on a Weibull variate in which $F(t) = 1 - \exp[(-\alpha t)^k]$. Finally, we have the important result:

Theorem A.8. *For any continuous positive random variable T , with distribution function $F(t)$, the variate $U = \int_0^T f(u)/[1 - F(u)]du$ has a standard exponential distribution.*

This result is important in survival modeling and we appeal to it frequently. The function $f(t)/[1 - F(t)]$ is known as the hazard function and $\int_0^t f(u)/[1 - F(u)]du$ as the cumulative hazard function. The transformation is called the cumulative hazard transformation.

A.7 Multivariate and copula models

The multivariate normal distribution, like its univariate counterpart, is completely specified by two compound quantities, the vector of means and the variance-covariance matrix. However, the assumptions of the multivariate situation are much more restrictive than anything needed for the simple univariate case. Indeed, if we use these restrictions to characterize the multivariate distribution, then we see straight away how strong are the assumptions we make. For example, if all marginal distributions are assumed normal, and all conditional distributions are assumed normal, this is a minimal characterization of the multivariate normal distribution. If all linear combinations of the components or sub-components of a multivariate distribution are univariate normal then this again characterizes multivariate normality. In consequence an assumption of multivariate normality is a very strong one. If we are comfortable in making this assumption the benefits are great. All of the regressions are linear with constant expressions for the variances

and covariances. Measures of explained variation are obtained immediately in terms of the covariances and variances, i.e., the correlation coefficients.

Finding useful models with weaker assumptions is not easy, in particular if we wish to have some easily interpretable parameter that quantifies dependency such as the correlation coefficient. Multivariate normal distributions have normal marginals but, as mentioned above, having normal marginals is not enough to result in a multivariate normal distribution. An infinity of distinct multivariate distributions can have the same univariate marginals. Indeed, for any pair of univariate marginals there exists an infinite number of multivariate distributions. Even something as simple as having uniform marginals, unless independent, is associated with an unlimited choice of candidate multivariate distributions having those same marginals.

One approach to constructing multivariate distributions from univariate marginals is described as a copula model. Again, the possibilities are infinite. For example, suppose we have distribution functions, $F(x)$ and $G(y)$, with respective densities, $f(x)$ and $g(y)$ for the variables X and Y . Consider the bivariate density function, $h(x,y;\theta)$ where;

$$h(x,y;\theta) = \Phi_{\theta}^{(2)} \{ \Phi^{-1}(u_x), \Phi^{-1}(u_y) \} \quad (\text{A.7})$$

where $\Phi_{\theta}^{(2)}$ indicates the standardized bivariate normal distribution with correlation coefficient θ , $\Phi(\cdot)$ the univariate normal distributions and u_x, u_y , independent uniform variates. The probability integral transform, described just above, tells us that, for continuous $F(x)$ and $G(y)$, we have a one-to-one correspondence with the uniform distribution. We can break down the steps by starting with X and Y , transforming these to the uniform interval via $u_x = F(x)$ and $u_y = G(y)$, subsequently transforming once more to normal marginals via $\Phi^{-1}(\cdot)$, and finally, via the bivariate normal distribution with parameter θ , creating the bivariate model with normal (0,1) marginals and association parameter θ . We can then use F^{-1} and G^{-1} to return to our original scale. This is referred to as the normal copula and the creation of a bivariate model with a particular association parameter, θ , is, in this case, quite transparent. Also, the dependency parameter, θ , has a concrete interpretation as a measure of explained variation on the transformed scales. Clearly, this set-up is readily generalized by replacing the first and second occurrences of Φ in Equation A.7 by any other distribution functions, not necessarily the same, and by replacing $\Phi_{\theta}^{(2)}$ by a different joint distribution with a different dependency structure.

There are very many more, in fact an infinite number of alternative ways to create bivariate models with some association structure from given marginals. Many of these have been used in the survival context, in particular when dealing with the problem of competing risks, or when considering surrogate endpoints. These are mentioned in the main text. Building and exploiting the structure of

copula models is a field in its own right discussed in many texts and articles. A thorough and clear discussion is given in Trivedi and Zimmer (2007).

A.8 Expectation

It is worth saying a word or two more about expectation as a fundamental aspect of studies in probability. Indeed it is possible for the whole theory to be constructed with expectation as a starting point rather than the now classical axiomatic structure to probability. For a function of a random variable T , $\psi(T)$ say, as stated at the beginning of the previous section, we write, $E(\psi(T))$ of this function via

$$E\psi(T) = \int \psi(t)f(t)dt = \int \psi(t)dF(t),$$

where the integrals, viewed as limiting processes, are all assumed to converge. The normal distribution function for a random variable X is completely specified by $E(X)$ and $E(X^2)$. In more general situations we can assume a unique correspondence between the moments of X , $E(X^r)$, $r = 1, 2, \dots$, and the distribution functions as long as these moments all exist. While it is true that the distribution function determines the moments the converse is not always true. However, it is almost always true (Kendall et al., 1987) and, for all the distributions of interest to us here, the assumption can be made without risk. It can then be helpful to view each moment, beginning with $E(X)$, as providing information about $F(x)$. This information typically diminishes quickly with increasing r . We can use this idea to improve inference for small samples when large sample approximations may not be sufficiently accurate. Moments can be obtained from the moment generating function, $M(t) = E\{\exp(tx)\}$ since we have:

Lemma A.2. *If $\int \exp(tx)f(x)dx < \infty$ then*

$$E(X^r) = \left\{ \frac{\partial^r M(t)}{\partial t^r} \right\}_{t=0}, \text{ for all } r.$$

We often focus on the variance function which is also an expectation and is of particular interest to one of our central goals here, that of constructing useful measures of the predictive strength of any model. At the root of the construction lie two important inequalities, the Chebyshev-Bienaym   inequality and Jensen's inequality described below. For this we first need:

Definition A.6. *The real-valued function $w(x)$ is called "convex" on some interval I (an infinite set and not just a point) whenever, for $x_1, x_2 \in I$ and for $0 \leq \lambda \leq 1$, we have*

$$w[\lambda x_1 + (1 - \lambda)x_2] \leq \lambda w(x_1) + (1 - \lambda)w(x_2).$$

It is usually sufficient to take convexity to mean that $w'(x)$ and $w''(x)$ are greater than or equal to zero at all interior points of I since this is a consequence of the definition. We have (Jensen's inequality):

Lemma A.3. *If w is convex on I then, assuming expectations exist on this interval, $w[E(X)] \leq E[w(X)]$. If w is linear in X throughout I , that is, $w''(x) = 0$ when twice differentiable, then equality holds.*

For the variance function we see that $w(x) = x^2$ is a convex function and so the variance is always positive. The further away from the mean, on average, the observations are to be found, then the greater the variance. We return to this in Chapter 10. Although very useful, the moment-generating function, $M(t) = E\{\exp(tX)\}$ has a theoretical weakness in that the integrals may not always converge. It is for this, mainly theoretical, reason that it is common to study instead the characteristic function, which has an almost identical definition, the only difference being the introduction of complex numbers into the setting. The characteristic function, denoted by $\phi(t)$, always exists and is defined as:

$$\phi(t) = M(it) = \int_{-\infty}^{\infty} \exp(itx)dF(x), \quad i^2 = -1.$$

Note that the contour integral in the complex plane is restricted to the whole real axis. Analogous to the above lemma concerning the moment-generating function we have:

$$E(X^r) = (-i)^r \left\{ \frac{\partial^r \phi(t)}{\partial t^r} \right\}_{t=0}, \quad \text{for all } r.$$

This is important in that it allows us to anticipate the cumulative generating function which turns out to be of particular importance in obtaining improved approximations to those provided by assuming normality. We return to this below in Section A.10. If we expand the exponential function then we can write:

$$\phi(t) = \int_{-\infty}^{\infty} \exp(itx)dF(x) = \exp \left\{ \sum_{r=1}^{\infty} \kappa_r (it)^r / r! \right\}$$

and, identifying κ_r as the coefficient of $(it)^r / r!$ in the expansion of $\log \phi(t)$. The function $\psi(t) = \log \phi(t)$ is called the cumulative generating function. When this function can be found then the density $f(x)$ can be defined in terms of it. We have the important relation

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \phi(t) dt, \quad \phi(t) = \int_{-\infty}^{\infty} e^{itx} f(x) dx.$$

It is possible to approximate the density $f(x)$ by working with i.i.d. observations X_1, \dots, X_n and the empirical characteristic function $\phi(t) = n^{-1} \sum_{i=1}^n \exp(itx_i)$ which can then be inverted. It is also possible to approximate the integral using a

method of numerical analysis, the so-called method of steepest descent, to obtain a saddlepoint approximation (Daniels, 1983, 1954, 1980, 1987). We return to this approximation below in Section A.10.

A.9 Order statistics and their expectations

The normal distribution and other parametric distributions described in the next chapter play a major role in survival modeling. However, the robustness of any inferential technique to particular parametric assumptions is always a concern. Hopefully, inference is relatively insensitive to departures from parametric assumptions or is applicable to whole families of parametric assumptions. The most common way to ensure this latter property is via the theory of order statistics which we recall here. Consider the n independent identically distributed (i.i.d.) random variables: X_1, X_2, \dots, X_n and a single realization of these that we can order from the smallest to the largest: $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$. Since the X_i are random, so also are the $X_{(i)}$, and the interesting question concerns what we can say about the probability structure of the $X_{(i)}$ on the basis of knowledge of the parent distribution of X_i . In fact, we can readily obtain many useful results which, although often cumbersome to write down, are in fact straightforward. Firstly we have:

Theorem A.9. *Taking $P(x) = \Pr(X \leq x)$ and $F_r(x) = \Pr(X_{(r)} \leq x)$ then:*

$$F_r(x) = \sum_{i=r}^n \binom{n}{i} P^i(x) [1 - P(x)]^{n-i}. \quad (\text{A.8})$$

This important result has two immediate and well-known corollaries dealing with the maximum and minimum of a sample of size n .

Corollary A.4.

$$F_n(x) = P^n(x), \quad F_1(x) = 1 - [1 - P(x)]^n \quad (\text{A.9})$$

In practice, in order to evaluate $F_r(x)$ for other than very small n , we exploit the equivalence between partial binomial sums and the incomplete beta function. Thus, if, for $a > 0, b > 0$, $B(a, b) = \int_0^1 t^{a-1} (1-t)^{b-1} dt$ and $I_\pi(a, b) = \int_0^\pi t^{a-1} (1-t)^{b-1} dt / B(a, b)$, then putting $P(x) = \pi$, we have that $F_r(x) = I_\pi(r, n-r+1)$. These functions are widely tabulated and also available via numerical algorithms to a high level of approximation. An alternative, although less satisfying, approximation would be to use the DeMoivre-Laplace normal approximation to the binomial sums. Differentiation of (A.8) provides the density which can be written as

$$f_r(x) = \frac{1}{B(r, n-r+1)} P^{r-1}(x) [1 - P(x)]^{n-r} p(x). \quad (\text{A.10})$$

Since we have a relatively straightforward expression for the distribution function itself, then this expression for the density is not often needed. It can come in handy in cases where we need to condition and apply the law of total probability. Expressions for $f_1(x)$ and $f_n(x)$ are particularly simple and we have:

Corollary A.5.

$$f_1(x) = n[1 - P(x)]^{n-1}p(x), \quad f_n(x) = nP^{n-1}(x)p(x). \quad (\text{A.11})$$

More generally it is also straightforward to obtain

Theorem A.10. *For any subset of the n order statistics: $X_{n_1}, X_{n_2}, \dots, X_{n_k}$, $1 \leq n_1 \leq \dots \leq n_k$, the joint distribution $f(x_1, \dots, x_k)$ is expressed as*

$$f(x_1, \dots, x_k) = n! \left[\prod_{j=1}^k p(x_j) \right] \prod_{j=0}^k \left\{ \frac{[P(x_{j+1}) - P(x_j)]^{n_{j+1} - n_j - 1}}{(n_{j+1} - n_j - 1)!} \right\} \quad (\text{A.12})$$

in which $p(x) = P'(x)$. This rather involved expression leads to many useful results including the following corollaries:

Corollary A.6. *The joint distribution of $X_{(r)}$ and $X_{(s)}$ is*

$$F_{rs}(x, y) = \sum_{j=s}^n \sum_{i=r}^j \frac{n!}{i!(j-i)!(n-j)!} P^i(x) [P(y) - P(x)]^{j-i} [1 - P(y)]^{n-j}.$$

The joint distribution of $X_{(r)}$ and $X_{(s)}$ is useful in establishing a number of practical results such as the distribution of the range, the distribution of the interquartile range and an estimate for the median among others. Using the result (Appendix A.6) for the distribution of a difference, a simple integration then leads to the following:

Corollary A.7. *Letting $W_{rs} = X_{(s)} - X_{(r)}$ then: in the special case of a parent uniform distribution we have*

$$f(w_{rs}) = \frac{1}{B(s-r, n-s+r+1)} w_{rs}^{s-r-1} (1-w_{rs})^{n-s+r}. \quad (\text{A.13})$$

Taking $s = n$ and $r = 1$, recalling that $B(\alpha, \beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha + \beta)$ and that $\Gamma(n) = n!$, then we have the distribution of the range for the uniform.

Corollary A.8. *Letting $w = U_{(n)} - U_{(1)}$ be the range for a random sample of size n from the standard uniform distribution, then the cumulative distribution is given by*

$$F_U(w) = nw^{n-1} - (n-1)w^n. \quad (\text{A.14})$$

Straightforward differentiation gives $f_U(w) = n(n-1)w^{n-2}(1-w)$, a simple and useful result. For an arbitrary distribution, $F(\cdot)$ we can either carry out the

same kind of calculations from scratch or, making use once more of the probability integral transform (see Section A.3), use the above result for the uniform and transform into arbitrary F . Even this is not that straightforward since, for some fixed interval (w_1, w_2) , corresponding to $w = w_2 - w_1$ from the uniform, the corresponding $F^{-1}(w_2) - F^{-1}(w_1)$ depends not only on $w_2 - w_1$ but on w_1 itself. Again we can appeal to the law of total probability, integrating over all values of w_1 from 0 to $1-w$. In practice, it may be good enough to divide the interval $(0, 1-w)$ into a number of equally spaced points, ten would suffice, and simply take the average. Interval estimates for any given quantile, defined by $P(\xi_\alpha) = \alpha$, follow from the basic result and we have:

Corollary A.9. *In the continuous case, for $r < s$, the pair $(X_{(r)}, X_{(s)})$ covers ξ_α with probability given by $I_\pi(r, n-r+1) - I_\pi(r, n-s+1)$.*

Theorem A.11. *For the special case in which $n_1 = 1$, $n_2 = 2$, ... $n_n = n$, then*

$$f(x_1, \dots, x_n) = n! \prod_{j=1}^n p(x_j). \quad (\text{A.15})$$

A characterization of order statistics: Markov property

The particularly simple results for the exponential distribution lead to a very useful and powerful characterization of order statistics. If Z_1, \dots, Z_n are i.i.d. exponential variates with parameter λ , then an application of Corollary A.4 shows that the minimum of Z_1 to Z_n has itself an exponential distribution with parameter $n\lambda$. We can define the random variable Y_1 to be the gap time between 0 and the first observation, $Z_{(1)}$. The distribution of Y_1 (equivalently $Z_{(1)}$) is exponential with parameter $n\lambda$. Next, we can define Y_2 to be the gap $Z_{(2)} - Z_{(1)}$. In view of the lack of memory property of the exponential distribution, once $Z_{(1)}$ is observed, the conditional distribution of each of the remaining $(n-1)$ variables, given that they are all greater than the observed time $Z_{(1)}$, remains exponential with parameter λ . The variable Y_2 is then the minimum of $(n-1)$ i.i.d. exponential variates with parameter λ . The distribution of Y_2 is therefore, once again, exponential, this time with parameter $(n-1)\lambda$. More generally we have the following lemma:

Lemma A.4. *If $Z_{(1)}, \dots, Z_{(n)}$ are the order statistics from a sample of size n of standard exponential variates, then, defining $Z_{(0)} = 0$,*

$$Y_i = Z_{(i)} - Z_{(i-1)}, \quad i = 1, \dots, n$$

are n independent exponential variates in which $E(Y_i) = 1/(n-i+1)$.

This elementary result is very important in that it relates the order statistics directly to sums of simple independent random variables which are not themselves order statistics. Specifically we can write:

$$Z_{(r)} = \sum_{i=1}^r \{Z_{(i)} - Z_{(i-1)}\} = \sum_{i=1}^r Y_i ,$$

leading to the immediate further lemma:

Lemma A.5. *For a sample of size n from the standard exponential distribution and letting $\alpha_i = 1/(n-i+1)$, we have:*

$$E[Z_{(r)}] = \sum_{i=1}^r E(Y_i) = \sum_{i=1}^r \alpha_i , \quad \text{Var}[Z_{(r)}] = \sum_{i=1}^r \text{Var}(Y_i) = \sum_{i=1}^r \alpha_i^2 .$$

The general flavor of the above result applies more generally than just to the exponential and, applying the probability integral transform (Section A.6), we have:

Lemma A.6. *For an i.i.d. sample of size n from an arbitrary distribution, $G(x)$, the r th largest order statistic, $X_{(r)}$ can be written:*

$$X_{(r)} = G^{-1}\{1 - \exp(-Y_1 - Y_2 - \cdots - Y_r)\},$$

where the Y_i are independent exponential variates in which $E(Y_i) = 1/(n-i+1)$.

One immediate conclusion that we can make from the above expression is that the order statistics from an arbitrary distribution form a Markov chain. The conditional distribution of $X_{(r+1)}$ given $X_{(1)}, X_{(2)}, \dots, X_{(r)}$ depends only on the observed value of $X_{(r)}$ and the distribution of Y_{r+1} . This conditional distribution is clearly the same as that for $X_{(r+1)}$ given $X_{(r)}$ alone, hence the Markov property. If needed we can obtain the joint density, f_{rs} , of $X_{(r)}$ and $X_{(s)}$, ($1 \leq r < s \leq n$) by a simple application of Theorem A.10. We then write:

$$f_{rs}(x, y) = \frac{n! P^{r-1}(x)p(x)p(y)[P(y) - P(x)]^{s-r-1}[1 - P(y)]^{n-s}}{(r-1)!(s-r-1)!(n-s)!} .$$

From this we can immediately deduce the conditional distribution of $X_{(s)}$ given that $X_{(r)} = x$ as:

$$f_{s|r}(y|x) = \frac{(n-r)!}{(s-r-1)!(n-s)!} \frac{p(y)[P(y) - P(x)]^{s-r-1}[1 - P(y)]^{n-s}}{[1 - P(x)]^{n-r}} .$$

A simple visual inspection of this formula confirms again the Markov property. Given that $X_{(r)} = x$ we can view the distribution of the remaining $(n-r)$ order statistics as an ordered sample of size $(n-r)$ from the conditional distribution $P(u|u > x)$.

Expected values of order statistics

Given the distribution of any given order statistic we can, at least in principle, calculate any moments, in particular the mean, by applying the basic definition. In practice, this may be involved and there may be no explicit analytic solution. Integrals can be evaluated numerically but, in the majority of applications, it can be good enough to work with accurate approximations. The results of the above subsection, together with some elementary approximation techniques are all that we need. Denoting the distribution of X as $P(x)$, then the probability integral transform (Appendix A.6) provides that $U = P(X)$ has a uniform distribution. The moments of the order statistics from a uniform distribution are particularly simple so that: $E\{U_{(r)}\} = p_r = r/(n+1)$. Denoting the inverse transformation by $Q = P^{-1}$, then

$$X_{(r)} = P^{-1}\{U_{(r)}\} = Q\{U_{(r)}\}.$$

Next, we can use a Taylor series development of the function $X_{(r)}$ about the p_r so that:

$$X_{(r)} = Q(p_r) + \{U_{(r)} - p_r\}Q'(p_r) + \{U_{(r)} - p_r\}^2 Q''(p_r)/2 + \dots$$

and, taking expectations, term by term, we have:

$$E\{X_{(r)}\} \approx Q(p_r) + \frac{p_r q_r}{2(n+2)} Q''(p_r) + \frac{p_r q_r}{(n+2)^2} \left\{ \frac{1}{3}(q_r - p_r)Q'''(p_r) + \frac{1}{8}p_r q_r Q''''(p_r) \right\}$$

and

$$\begin{aligned} \text{Var}\{X_{(r)}\} &= \frac{p_r q_r}{2(n+2)} [Q'(p_r)]^2 + \\ &\quad \frac{p_r q_r}{(n+2)^2} \left\{ 2(q_r - p_r)Q'(p_r)Q''(p_r) + p_r q_r (Q'(p_r)Q'''(p_r) + [Q''(p_r)]^2) \right\}. \end{aligned}$$

It is straightforward to establish some relationships between the moments of the order statistics and the moments from the parent distribution. Firstly note that:

$$E \left\{ \sum_{r=1}^n X_{(r)}^k \right\}^m = E \left\{ \sum_{r=1}^n X_r^k \right\}^m,$$

so that, if μ and σ^2 are the mean and variance in the parent population, then $\sum_{r=1}^n \mu_r = n\mu$ and $\sum_{r=1}^n E\{X_{(r)}^2\} = nE(X^2) = n(\mu^2 + \sigma^2)$.

Normal parent distribution

For the case of a normal parent the expected values can be evaluated precisely for small samples and the approximations themselves are relatively tractable for larger sample sizes. One approach to data analysis in which it may be desirable to

have a marginal normal distribution in at least one of the variables under study is to replace the observations by the expectations of the order statistics. These are sometimes called normal scores, typically denoted by $\xi_{rn} = E(X_{(r)})$ for a random sample of size n from a standard normal parent with distribution function $\Phi(x)$ and density $\phi(x)$. For a random sample of size n from a normal distribution with mean μ and variance σ^2 we can reduce everything to the standard case since $E(X_{(r)}) = \mu + \xi_{rn}\sigma$. Note that, if n is odd, then, by symmetry, it is immediately clear that $E(X_{(r)}) = 0$ for all r that are odd. We can see that $E(X_{(r)}) = -E(X_{(n-r+1)})$. For n as small as, say, 5 we can use integration by parts to evaluate ξ_{r5} for different values of r . For example, $\xi_{55} = 5 \int 4\Phi^3(x)\phi^2(x)dx$ which then simplifies to: $\xi_{55} = 5\pi^{-1/2}/4 + 15\pi^{-3/2}\sin^{-1}(1/3)/2 = 1.16296$. Also, $\xi_{45} = 5\pi^{-1/2}/2 - 15\pi^{-3/2}\sin^{-1}(1/3) = 0.49502$ and $\xi_{35} = 0$. Finally, $\xi_{15} = -1.16296$ and $\xi_{25} = -0.49502$. For larger sample sizes in which the integration becomes too fastidious we can appeal to the above approximations using the fact that

$$Q'(p_r) = \frac{1}{\phi(Q)}, \quad Q''(p_r) = \frac{Q}{\phi^2(Q)}, \quad Q'''(p_r) = \frac{1+2Q^2}{\phi^3(Q)}, \quad Q''''(p_r) = \frac{Q(7+6Q^2)}{\phi^4(Q)}.$$

The above results arise from straightforward differentiation. Analogous calculations can be used to obtain exact or approximate expressions for $\text{Cov}\{X_{(r)}, X_{(s)}\}$.

A.10 Approximations

Approximations to means and variances for functions of T (δ -method)

Consider some differentiable monotonic function of X , say $\psi(X)$. Our particular concern often relates to parameter estimates in which case the random variable X would be some function of the n i.i.d. data values, say θ_n as an estimator of the parameter θ . In the cases of interest, θ_n converges with probability one to θ and so also does $\psi(\theta_n)$ to $\psi(\theta)$. Although θ_n may not be unbiased for θ , for large samples, the sequence $E(\theta_n)$ converges to $E(\theta) = \theta$. Similarly $E[\psi(\theta_n)]$ converges to $\psi(\theta)$. The mean value theorem (Section A.2) enables us to write

$$\phi(\theta_n) = \psi(\theta) + (\theta_n - \theta)\phi'(\theta) + \frac{(\theta_n - \theta)^2}{2}\psi''(\xi) \quad (\text{A.16})$$

for $\xi \in (\theta \pm \theta_n)$. Rearranging this expression, ignoring the third term on the right-hand side, and taking expectations we obtain

$$\text{Var}\{\psi(\theta_n)\} \approx E\{\psi(\theta_n) - \psi(\theta)\}^2 \approx \{\psi'(\theta)\}^2 \text{Var}(\theta_n) \approx \{\psi'(\theta_n)\}^2 \text{Var}(\theta_n)$$

as an approximation to the variance. The approximation, once obtained in any given setting, is best studied on a case-by-case basis. It is an exact result for linear functions. For these, the second derivative is equal to zero and, more

generally, the smaller the absolute value of this second derivative, the better we might anticipate the approximation to be. For θ_n close to θ the squared term will be small in absolute value when compared with the linear term, an additional motivation to neglecting the third term. For the mean, the second term of Equation (A.16) is zero when θ_n is unbiased, otherwise close to zero and, this time, ignoring this second term, we obtain

$$E\{\psi(\theta_n)\} \approx \psi(\theta_n) + \frac{1}{2}\text{Var}(\theta_n)\psi''(\theta_n) \quad (\text{A.17})$$

as an improvement over the rougher approximation based on the first term alone of the above expression. Extensions of these expressions to the case of a consistent estimator $\psi(\theta_n) = \psi(\theta_{1n}, \dots, \theta_{pn})$ of $\psi(\theta)$ proceeds in the very same way, only this time based on a multivariate version of Taylor's theorem. These are:

$$\begin{aligned} \text{Var}\{\psi(\theta_n)\} &\approx \sum_{j=1}^p \sum_{m \geq j}^p \frac{\partial\psi(\theta)}{\partial\theta_j} \frac{\partial\psi(\theta)}{\partial\theta_m} \text{Cov}(\theta_{jn}, \theta_{mn}), \\ E\{\psi(\theta_n)\} &\approx \psi(\theta_{1n}, \dots, \theta_{pn}) + \frac{1}{2} \sum_j \sum_m \frac{\partial^2\psi(\theta_n)}{\partial\theta_j \partial\theta_m} \text{Cov}(\theta_{jn}, \theta_{mn}). \end{aligned}$$

When $p = 1$ then the previous expressions are recovered as special cases. Again, the result is an exact one in the case where $\psi(\cdot)$ is a linear combination of the components θ_j and this helps guide us in situations where the purpose is that of confidence interval construction. If, for example, our interest is on ψ and some strictly monotonic transformation of this, say ψ^* , is either linear or close to linear in the θ_j , then it may well pay, in terms of accuracy of interval coverage, to use the delta-method on ψ^* , obtaining the end points of the confidence interval for ψ^* and subsequently inverting these, knowing the relationship between ψ and ψ^* , in order to obtain the interval of interest for ψ . Since ψ and ψ^* are related by one-to-one transformations then the coverage properties of an interval for ψ^* will be identical to those of its image for ψ . Examples in this book include confidence intervals for the conditional survivorship function, given covariate information, based on a proportional hazards model as well as confidence intervals for indices of predictability and multiple coefficients of explained variation.

Cornish-Fisher approximations

In the construction of confidence intervals, the δ -method makes a normality approximation to the unknown distribution and then replaces the first two moments by local linearization. A different approach, while still working with a normal density $\phi(x) = (2\pi)^{-1/2} \exp(-x^2/2)$, in a way somewhat analogous to the construction of a Taylor series, is to express the density of interest, $f(x)$, in terms of a linear combination of $\phi(x)$ and derivatives of $\phi(x)$. Normal distributions with non-zero means and variances not equal to one are obtained by the

usual simple linear transformation and, in practical work, the simplest approach is to standardize the random variable X so that the mean and variance corresponding to the density $f(x)$ are zero and one, respectively.

The derivatives of $\phi(x)$ are well-known, arising in many fields of mathematical physics and numerical approximations. Since $\phi(x)$ is simply a constant multiplying an exponential term it follows immediately that all derivatives of $\phi(x)$ are of the form of a polynomial that multiplies $\phi(x)$ itself. These polynomials (apart from an alternating sign coefficient $(-1)^i$) are the Hermite polynomials, $H_i(x), i = 0, 1, \dots$, and we have:

$$H_0 = 1, \quad H_1 = x, \quad H_2 = x^2 - 1, \quad H_3 = x^3 - 3x, \quad H_4 = x^4 - 6x^2 + 3,$$

with H_5 and higher terms being calculated by simple differentiation. The polynomials are of importance in their own right, belonging to the class of orthogonal polynomials and useful in numerical integration. Indeed, we have that:

$$\int_{-\infty}^{\infty} H_i^2(x) \phi(x) dx = i!, \quad i = 0, \dots : \quad \int_{-\infty}^{\infty} H_i(x) H_j(x) \phi(x) dx = 0, \quad i \neq j.$$

This orthogonality property is exploited in order for us to obtain explicit expressions for the coefficients in our expansion. Returning to our original problem we wish to determine the coefficients c_i in the expansion

$$f(x) = \sum_{i=0}^{\infty} c_i H_i(x) \phi(x) \tag{A.18}$$

and, in order to achieve this we multiply both sides of equation (A.18) by $H_j(x)$, subsequently integrating to obtain the coefficients

$$c_j = \frac{1}{j!} \int_{-\infty}^{\infty} f(x) H_j(x) dx. \tag{A.19}$$

Note that the polynomial $H_j(x)$ is of order j so that the right-hand side of equation (A.19) is a linear combination of the moments, (up to the j th), of the random variable X having associated density $f(x)$. These can be calculated step-by-step. For many standard densities several of the lower-order moments have been worked out and are available. Thus, it is relatively straightforward to approximate some given density $f(x)$ in terms of a linear combination of $\phi(x)$.

The expansion of Equation (A.18) can be used in theoretical investigations as a means to study the impact of ignoring higher-order terms when we make a normal approximation to the density of X . We will use the expansion in an attempt to obtain more accurate inference for proportional hazards models fitted using small samples. Here the large sample normal assumption may not be sufficiently accurate and the approximating equation is used to motivate potential improvements obtained by taking into account moments of higher order than

just the first and second. When dealing with actual data, the performance of any such adjustments need to be evaluated on a case-by-case basis. This is because theoretical moments will have to be replaced by observed moments and the statistical error involved in that can be of the same order, or greater, than the error involved in the initial normal approximation. If we know or are able to calculate the moments of the distribution, then the c_i are immediately obtained. When the mean is zero we can write down the first four terms as:

$$c_0 = 1, c_1 = 0, c_2 = (\mu_2 - 1)/2, c_3 = \mu_3/6, c_4 = (\mu_4 - 6\mu_2 + 3)/24,$$

from which we can write down an expansion in terms of $\phi(x)$ as

$$f(x) = \phi(x) \{1 + (\mu_2 - 1)H_2(x)/2 + \mu_3 H_3(x)/6 + (\mu_4 - 6\mu_2 + 3)H_4(x)/24 + \dots\}.$$

This series is known as the Gram-Charlier series, and stopping the development at the fourth term corresponds to making corrections for skewness and kurtosis. In the development of the properties of estimators in the proportional hazards model we see that making corrections for skewness can help make inference more accurate, whereas, at least in that particular application, corrections for kurtosis appear to have little impact (Chapter 7).

Saddlepoint approximations

A different, although quite closely related, approach to the above uses saddlepoint approximations. Theoretical and practical work on these approximations indicate them to be surprisingly accurate for the tails of a distribution. We work with the inversion formula for the cumulant generating function, a function that is defined in the complex plane, and in this two-dimensional plane, around the point of interest (which is typically a mean or a parameter estimate) the function looks like a minimum in one direction and a maximum in an orthogonal direction: hence the name “saddlepoint.” Referring back to Section A.8 recall that we identified κ_r as the coefficient of $(it)^r/r!$ in the expansion of the cumulant generating function $K(t) = \log \phi(t)$ where $\phi(t)$ is the characteristic function. We can exploit the relationship between $\phi(t)$ and $f(x)$; that is:

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \phi(t) dt, \quad \phi(t) = \int_{-\infty}^{\infty} e^{itx} f(x) dx.$$

to approximate $f(x)$ by approximating the integral. The numerical technique that enables this approximation to be carried out is called the method of steepest descent and is described in Daniels (1954). The approximation to $f(x)$ is simply denoted as $f_s(x)$, and carrying through the calculations, we find that

$$f_s(x) = \left\{ \frac{n}{2\pi K''(\lambda_x)} \right\}^{1/2} \exp[n\{K(\lambda_x) - x\lambda_x\}] \quad (\text{A.20})$$

in which the solution to the differential equation in λ , $K'(\lambda) = x$ is given by λ_x . Our notation here of x as a realization of some random variable X is not specifically referring to our usual use of X as the minimum of survival time T and the censoring time C . It is simply the variable of interest and that variable, in our context, will be the score statistic arising from the estimating equation (Chapter 7). For now, we assume the score to be composed of n contributions so that we view x as a mean based on n observations. Since, mostly, we are interested in the tails of the distribution, it can often help to approximate the cumulative distribution directly rather than make a subsequent appeal to numerical integration. Denoting the saddlepoint approximation to the cumulative distribution by $F_s(x)$, we write

$$F_s(x) = \Phi(u_x) + \phi(u_x)(u_x^{-1} + v_x^{-1}) \quad (\text{A.21})$$

where $\phi(x)$ indicates the standard normal density, $\Phi(x) = \int_{-\infty}^x \phi(u)du$, the cumulative normal, $u_x = [2n\{x\lambda_x - K(\lambda_x)\}]^{1/2}\text{sgn}(\lambda_x)$, and $v_x = \lambda_x\{nK''(\lambda_x)\}^{1/2}$. Since we are only concerned with tail probabilities we need not pay attention to what occurs around the mean. If we do wish to consider $F_s(x)$, evaluated at the mean, the approximation is slightly modified and the reader is referred to Daniels (1987).

Appendix B

Stochastic processes

B.1 Broad overview

We define a stochastic process to be a collection of random variables indexed by $t \in T$. We write these as $X(t)$, or X_t , and take t to be fixed. If the set T has only a finite or a countably infinite number of elements then $X(t)$ is referred to as a discrete-time process. We will be most interested in continuous-time processes. In applications we can standardize by the greatest value of t in the set T that can be observed, and so we usually take $\sup\{t : t \in T\} = 1$. We also take $\inf\{t : t \in T\} = 0$. We will be especially interested in observations on any given process between 0 and t . We call this the sample path.

Independent increments and stationarity

Consider some partition of $(0,1)$ in which $0 = t_0 < t_1 < t_2 < \dots < t_n = 1$. If the set of random variables $X(t_i) - X(t_{i-1})$ $i = 1, \dots, n$ are independent then the stochastic process $X(t)$ is said to have independent increments. Another important property is that of stationarity. We say that a stochastic process $X(t)$ has stationary increments if $X(s+t) - X(s)$ has the same distribution for all values of s . Stationarity indicates, in as much as probabilistic properties are concerned, that when we look forward, from the point s , a distance t , the only relevant quantity is how far forward t we look. Our starting point itself is irrelevant. As we progress through time, everything that we have learned is summarized by the current position. It can also be of value to consider a process with a slighter weaker property, the so-called second-order stationarity. Rather than insist on a requirement for the whole distribution we limit our attention to the first two moments and the covariance between $X(s+t)$ and $X(s)$ which depends only upon $|t|$. Our main focus is on Gaussian processes which, when they have the property of second-order stationarity, will in consequence be stationary processes.

Also, simple transformations can produce stationary processes from nonstationary ones, an example being the transformation of the Brownian bridge into an Ornstein-Uhlenbeck process.

Gaussian processes

If for every partition of $(0,1)$, $0 = t_0 < t_1 < t_2 < \dots < t_n = 1$, the set of random variables $X(t_1), \dots, X(t_n)$ has a multivariate normal distribution, then the process $X(t)$ is called a Gaussian process. Brownian motion, described below, can be thought of as simply a standardized Gaussian process. A Gaussian process being uniquely determined by the multivariate means and covariances it follows that such a process will have the property of stationarity if for any pair $(s, t : t > s)$, $\text{Cov}\{X(s), X(t)\}$ depends only on $(t - s)$. In practical studies we will often deal with sums indexed by t and the usual central limit theorem will often underlie the construction of Gaussian processes.

B.2 Brownian motion

Consider a stochastic process $X(t)$ on $(0, 1)$ with the following three properties:

1. $X(0) = 0$, i.e., at time $t = 0$ the starting value of X is fixed at 0.
2. $X(t), t \in (0, 1)$ has independent stationary increments.
3. At each $t \in (0, 1)$ the distribution of $X(t)$ is $\mathcal{N}(0, t)$.

This simple set of conditions completely describes a uniquely determined stochastic process called Brownian motion. It is also called the Wiener process or Wiener measure. It has many important properties and is of fundamental interest as a limiting process for a large class of sums of random variables on the interval $(0, 1)$. An important property is described in Theorem B.1 below. Firstly we make an attempt to describe just what a single realization of such a process might look like. Later we will recognize the same process as being the limit of a sum of independent random contributions. The process is continuous and so, approximating it by any drawing, there cannot be any gaps. At the same time, in a sense that can be made more mathematically precise, the process is infinitely jumpy. Nowhere does a derivative exist. Figure B.1 illustrates this via simulated approximations. The right-hand figure could plausibly be obtained from the left-hand one by homing in on any small interval, e.g., $(0.20, 0.21)$, subtracting off the value observed at $t = 0.20$, and rescaling by a multiple of ten to restore the interval of length 0.01 to the interval $(0, 1)$. The point we are trying to make is that the resulting process itself looks like (and indeed is) a realization of Brownian motion. Theoretically, this could be repeated without limit which allows us to understand in some way how infinitely jumpy is the process. In practical examples we can only ever approximate the process by linearly connecting up adjacent simulated points.

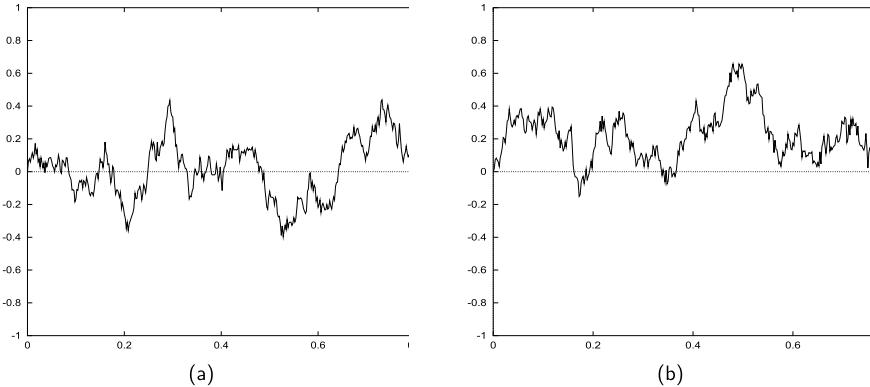


Figure B.1: Two independently simulated Brownian motions on the interval $(0,1)$.

Theorem B.1. *Conditioning on a given path we have:*

$$\begin{aligned} \Pr \{X(t+s) > x | X(s) = x_s, X(u), 0 \leq u < s\} \\ &= \Pr \{X(t+s) > x | X(s) = x_s\} \end{aligned}$$

So, when looking ahead from time point s to time point $t+s$, the previous history indicating how we arrived at s is not relevant. The only thing that matters is the point at which we find ourselves at time point s . This is referred to as the Markov property. The joint density of $X(t_1), \dots, X(t_n)$ can be written as:

$$f(x_1, x_2, \dots, x_n) = f_{t_1}(x_1) f_{t_1-t_2}(x_2 - x_1) \cdots f_{t_n-t_{n-1}}(x_n - x_{n-1})$$

This follows from the independent stationary increment condition. A consequence of the above result is that we can readily evaluate the conditional distribution of $X(s)$ given some future value $X(t)$ ($t > s$). Applying the definition for conditional probability we have the following:

Corollary B.1. *The conditional distribution of $X(s)$ given $X(t)$ ($t > s$) is normal with a mean and a variance given by:*

$$E\{X(s)|X(t) = w\} = ws/t, \quad \text{Var}\{X(s)|X(t) = w\} = s(t-s)/t.$$

This result helps provide insight into another useful process, the Brownian bridge described below. Other important processes arise as simple transformations of Brownian motion. The most obvious to consider is where we have a Gaussian process satisfying conditions (1) and (2) for Brownian motion but where, instead of the variance increasing linearly, i.e., $\text{Var } X(t) = t$, the variance increases either too quickly or too slowly so that $\text{Var } X(t) = \phi(t)$ where $\phi(\cdot)$ is some monotonic increasing function of t . Then we can transform the time axis using $\phi(\cdot)$ to

produce a process satisfying all three conditions for Brownian motion. Consider also the transformation

$$V(t) = \exp(-\alpha t/2) X\{\exp(\alpha t)\}$$

where $X(t)$ is Brownian motion. This is the Ornstein-Uhlenbeck process. It is readily seen that:

Corollary B.2. *The process $V(t)$ is a Gaussian process in which $E\{V(t)\} = 0$ and $\text{Cov}\{V(t), V(s)\} = \exp\{-\alpha(t-s)/2\}$.*

Note that $\text{Var}V(t)$ is a constant not depending on t , so that $V(t)$ is often referred to as variance stabilizing.

Time-transformed Brownian motion

Consider a process, $X^\psi(t)$, defined via the following three conditions, for some continuous ψ such that, $\psi(t') > \psi(t)$ ($t' > t$); (1) $X^\psi(0) = 0$ (2) $X^\psi(t), t \in (0, 1)$ has independent stationary increments; (3) at each $t \in (0, 1)$ the distribution of $X^\psi(t)$ is $\mathcal{N}\{0, \psi(t)\}$. The usual Brownian motion described above is exactly this process when $\psi(t) = t$. However, in view of the continuity and monotonicity of ψ , there exists an inverse function ψ^{-1} such that $\psi^{-1}\{\psi(t)\} = t$. Clearly, we can transform the process $X^\psi(t)$ by multiplying, at each t , by $\sqrt{t/\psi(t)}$, and, defining $\sqrt{0/\psi(0)} = 0$. The resulting process we can call $X(t)$ and it is readily seen that this process is standard Brownian motion. Thus, the only crucial assumption in Brownian motion is that of independent increments. Once we can assert this to be the case, it is only a question of scale and location to obtain standard Brownian motion.

Brownian bridge

Let $\mathcal{W}(t)$ be Brownian motion. We know that $\mathcal{W}(0) = 0$. We also know that with probability one the process $\mathcal{W}(t)$ will return at some point to the origin. Let's choose a point, and in particular the point $t = 1$ and consider the conditional process $\mathcal{W}^0(t)$, defined to be Brownian motion conditioned by the fact that $\mathcal{W}(1) = 0$. For small t this process will look very much like the Brownian motion from which it is derived. As t goes to one the process is pulled back to the origin since at $t = 1$ we have that $\mathcal{W}^0(1) = 0$ and $\mathcal{W}(t)$ is continuous. Also $\mathcal{W}^0(0) = \mathcal{W}(0) = 0$. Such a process is called tied down Brownian motion or the Brownian bridge. We will see below that realizations of a Brownian bridge can be viewed as linearly transformed realizations of Brownian motion itself, and vice versa.

From the results of above the section we can investigate the properties of $\mathcal{W}^0(t)$. The process is a Gaussian process so we only need to consider the mean and covariance function for the process to be completely determined. We have:

$$E\{\mathcal{W}(s)|\mathcal{W}(1) = 0\} = 0 \quad \text{for } s < t.$$

This comes immediately from the above result. Next we have:

Theorem B.2.

$$\text{Cov}(\mathcal{W}(s), \mathcal{W}(t)|\mathcal{W}(1) = 0) = s(1-t). \quad (\text{B.1})$$

This provides a simple definition of the Brownian bridge as being a Gaussian process having mean zero and covariance function $s(1-t)$, $s < t$. An alternative way of constructing the Brownian bridge is to consider the process defined as:

$$\mathcal{W}^0(t) = \mathcal{W}(t) - t\mathcal{W}(1), \quad 0 \leq t \leq 1.$$

Clearly $\mathcal{W}^0(t)$ is a Gaussian process. We see that

$$E\{\mathcal{W}(0)\} = \mathcal{W}(0) = E\{\mathcal{W}(1)\} = \mathcal{W}(1) = E\{\mathcal{W}(t)\} = 0$$

so that the only remaining question is the covariance function for the process to be completely and uniquely determined. The following corollary is all we need.

Corollary B.3. *The covariance function for the process defined as $\mathcal{W}^0(t)$ is,*

$$\text{Cov}\{\mathcal{W}^0(s), \mathcal{W}^0(t)\} = s(1-t) \quad s < t.$$

This is the covariance function for the Brownian bridge developed above and, by uniqueness, the process is therefore itself the Brownian bridge. Such a covariance function is characteristic of many observed phenomena. The covariance decreases linearly with distance from s . As for Brownian motion, should the covariance function decrease monotonically rather than linearly, then a suitable transformation of the time scale enables us to write the covariance in this form. At $t = s$ we recover the usual binomial expression $s(1-s)$.

Notice that not only can we go from Brownian motion to a Brownian bridge via the simple transformation

$$\mathcal{W}^0(t) = \mathcal{W}(t) - t\mathcal{W}(1), \quad 0 \leq t \leq 1,$$

but the converse is also true, i.e., we can recover Brownian motion, $X(t)$, from the Brownian bridge, $Z(t)$, via the transformation

$$X(t) = (t+1)Z\left(\frac{t}{t+1}\right). \quad (\text{B.2})$$

To see this, first note that, assuming $Z(t)$ to be a Brownian bridge, then $X(t)$ is a Gaussian process. It will be completely determined by its covariance process $\text{Cov}\{X(s), X(t)\}$. All we then require is the following lemma:

Lemma B.1. *For the process defined in (B.2), $\text{Cov}\{X(s), X(t)\} = s$.*

The three processes: Brownian motion, the Brownian bridge, and the Ornstein-Uhlenbeck are then closely related and are those used in the majority of applications. Two further related processes are also of use in our particular applications: integrated Brownian motion and reflected Brownian motion.

Integrated Brownian motion

The process $Z(t)$ defined by: $Z(t) = \int_0^t \mathcal{W}(u)du$, where $\mathcal{W}(t)$ is Brownian motion is called integrated Brownian motion. Note that: $dZ(t)/dt = \mathcal{W}(t)$ so that, for example, in the context of a model of interest, should we be able to construct a process converging in distribution to a process equivalent to Brownian motion, then the integrated process will converge in distribution to a process equivalent to integrated Brownian motion. We can see (by interchanging limits) that $Z(t)$ can be viewed as the limit of a sum of Gaussian processes and is therefore Gaussian. Its nature is completely determined by its mean and covariance function. We have that:

$$E\{Z(t)\} = E\left\{\int_0^t \mathcal{W}(u)du\right\} = \int_0^t E\{\mathcal{W}(u)\}du = 0. \quad (\text{B.3})$$

For $s < t$ we have:

Lemma B.2. *The covariance function for $Z(s)$ and $Z(t)$ is*

$$\text{Cov}\{Z(s), Z(t)\} = s^2(t/2 - s/6). \quad (\text{B.4})$$

Lemma B.3. *The covariance function for $Z(t)$ and $\mathcal{W}(t)$ is:*

$$\text{Cov}\{Z(t), \mathcal{W}(t)\} = t^2/2. \quad (\text{B.5})$$

For a model in which inference derives from cumulative sums, this would provide a way of examining how reasonable are the underlying assumptions if repetitions are available. Repetitions can be obtained by bootstrap resampling if only a single observed process is available. Having standardized, a plot of the log-covariance function between the process and the integrated process against log-time ought to be linear with slope of two and intercept of minus log 2 assuming that model assumptions hold.

Reflected Brownian motion

Suppose we choose some positive value r and then define the process $\mathcal{W}_r(t)$ as a function of Brownian motion, $\mathcal{W}(t)$, in the following way: If $\mathcal{W}(t) < r$ then $\mathcal{W}_r(t) = \mathcal{W}(t)$. If $\mathcal{W}(t) \geq r$ then $\mathcal{W}_r(t) = 2r - \mathcal{W}(t)$. We have:

Lemma B.4. $\mathcal{W}_r(t)$ is a Gaussian process, $E\mathcal{W}_r(t) = 0$, $\text{Cov}\{\mathcal{W}_r(s), \mathcal{W}_r(t)\} = s$ when $s < t$.

Thus, $\mathcal{W}_r(t)$ is also Brownian motion. Choosing r to be negative and defining $\mathcal{W}_r(t)$ so that, when $\mathcal{W}(t) > r$ then $\mathcal{W}_r(t) = \mathcal{W}(t)$. If $\mathcal{W}(t) \leq r$ then $\mathcal{W}_r(t) = 2r - \mathcal{W}(t)$. accordingly we have the same result. The process $\mathcal{W}_r(t)$ coincides exactly with $\mathcal{W}(t)$ until such a time as a barrier is reached. We can imagine this barrier as a mirror, and beyond the barrier the process $\mathcal{W}_r(t)$ is a simple reflection of $\mathcal{W}(t)$. The interesting thing is that the resulting process is itself Brownian motion. One way of conceptualizing the idea is to imagine a large number of realizations of a completed Brownian motion process sampled independently. Imagine then these same realizations with a reflection applied. Then, whatever the point of reflection, if we consider the two collected sets of realizations, our overall impression of the behavior of the two processes will be the same. The value of this construction is to be seen in situations where, at some point in time, corresponding to some expected point of reflection under a hypothesis of drift, the drift changes direction. Under the hypothesis of Brownian motion, both Brownian motion, and Brownian motion reflected at some point, will look alike and will obey the same probability laws. Under an alternative hypothesis of drift, however (see below), the behaviors will look quite different. This observation enables a simple construction with which to address the problem of crossing hazards.

Maximum of a Brownian motion on $(0,1)$

A useful further result can be immediately obtained from the preceding one dealing with reflected Brownian motion. Suppose that $\mathcal{W}(t)$ is a Brownian motion. We might wish to consider the process $M(t) = \sup_{u \in (0,t)} \mathcal{W}(u)$, which is the greatest value obtained by the process $\mathcal{W}(u)$ in the interval $(0,t)$. The greatest absolute distance is also of interest but, by symmetry arguments, this can be obtained immediately from the distribution of $M(t)$. Another related question, useful in interim analyses, is the distribution of $\mathcal{W}(t)$ given the maximum $M(t)$ obtained up until that time point. We have the following:

Lemma B.5. If $\mathcal{W}(t)$ is standard Brownian motion on $(0,1)$ and $M(t)$ the maximum value attained on the interval $(0,t)$, i.e., $M(t) = \sup_{u \in (0,t)} \mathcal{W}(u)$, then

$$\Pr\{M(t) > a\} = 2\Pr\{\mathcal{W}(t) > a\}.$$

This is a simple and elegant result and enables us to make simultaneous inference very readily. Sometimes, when using a Brownian motion approximation

for a process, we may want to, for example, describe an approximate confidence interval for the whole process rather than just a confidence interval at a single point t . In such a case the above result comes into play. The joint distribution is equally simple and we make use of the following:

Lemma B.6. *If $\mathcal{W}(t)$ is standard Brownian motion and $M(t)$ the maximum value attained on the interval $(0, t)$, i.e., $M(t) = \sup_{u \in (0, t)} \mathcal{W}(u)$, then*

$$\Pr\{\mathcal{W}(t) < a - b, M(t) > a\} = \Pr\{\mathcal{W}(t) > a + b\}.$$

The conditional distribution $\Pr\{\mathcal{W}(t) < a - b | M(t) > a\}$ can then be derived immediately by using the results of the two lemmas.

Brownian motion with drift

We will see that simple Brownian motion provides a good model for describing score statistics, or estimating equations, once standardized. This is because we can visualize these sums as approximating a limiting process arising from summing increments, for which the expected value is equal to zero. The setting in which we study such sums is typically that of evaluating some null hypothesis, often one of some given effect, $H_0 : \beta = \beta_0$, but sometimes a less obvious one, in the goodness-of-fit context, for example, whereby we can have, $H_0 : \beta(t) = \hat{\beta}$. Almost invariably, when we consider a null hypothesis, we have an alternative in mind, frequently a local or first alternative to the null. For a null hypothesis of Brownian motion, a natural and immediate alternative is that of Brownian motion with drift. Consider then the stochastic process $X(t)$ defined by:

$$X(t) = \mathcal{W}(t) + \mu t$$

where $\mathcal{W}(t)$ is Brownian motion. We can immediately see that $E\{X(t)\} = \mu t$ and $\text{Var}\{X(t)\} = t$. As for Brownian motion $\text{Cov}\{X(s), X(t)\} = s, s < t$. Alternatively we can define the process in a way analogous to our definition for Brownian motion as a process having the following three properties:

1. $X(0) = 0$.
2. $X(t), t \in (0, 1)$ has independent stationary increments.
3. At each $t \in (0, 1)$, $X(t)$ is $\mathcal{N}(\mu t, t)$.

Clearly, if $X(t)$ is Brownian motion with drift parameter μ , then the process $X(t) - \mu t$ is standard Brownian motion. Also, for the more common situation in which the mean may change non-linearly with time, provided the increments are independent, we can always construct a standard Brownian motion by first subtracting the mean at time t , then transforming the timescale in order to achieve a linearly increasing variance. Note that for non-linear differentiable functions of drift, these can always be approximated locally by a linear function so that the

essential nature of the process, whether the drift is linear or smoothly non-linear, is the same. We will make use of this idea in those chapters devoted to fit and model building.

Probability results for Brownian motion

There are a number of well-established and useful results for Brownian motion and related processes. The arcsine law can be helpful in comparing processes. Defining $X^+(t)$ to be the time elapsed from the origin that the Brownian process remains positive, i.e., $\sup\{t : X(s) > 0 : 0 < s < t\}$ then $\Pr(X^+ < x) = (2/\pi)\sin^{-1}\sqrt{x}$. This law can be helpful in comparing processes and also in examining underlying hypotheses. For the Brownian bridge the largest distance from the origin in absolute value has a known distribution given in a theorem of Kolmogorov:

$$\Pr \left\{ \sup_t |\mathcal{W}_0(t)| \leq \alpha \right\} \rightarrow 1 - 2 \sum_{k=1}^{\infty} (-1)^{k+1} \exp(-2k^2\alpha^2), \quad \alpha \geq 0. \quad (\text{B.6})$$

The sum can be seen to be convergent since this is an alternating sign series in which the k th term goes to zero. Furthermore, the error in ignoring all terms higher than the n th is less, in absolute value, than the size of the $(n+1)$ th term. Given that the variance of $\mathcal{W}_0(t)$ depends on t it is also of interest to study the standardized distribution $B_0(t) = W_0(t)/\sqrt{t(1-t)}$. This is, in fact, the Ornstein-Uhlenbeck process. Simple results for the supremum of this are not possible since the process becomes unbounded at $t = 0$ and $t = 1$. Nonetheless, if we are prepared to reduce the interval from $(0,1)$ to $(\varepsilon_1, \varepsilon_2)$ where $\varepsilon_1 > 0$ and $\varepsilon_2 < 1$ then we have an approximation due to Miller and Siegmund (1982):

$$\Pr \left\{ \sup_t |B_0(t)| \geq \alpha \right\} \approx \frac{4\phi(\alpha)}{\alpha} + \phi(\alpha) \left(\alpha - \frac{1}{\alpha} \right) \log \left\{ \frac{\varepsilon_2(1-\varepsilon_1)}{\varepsilon_1(1-\varepsilon_2)} \right\}, \quad (\text{B.7})$$

where $\phi(x)$ denotes the standard normal density. This enables us to construct confidence intervals for a bridged process with limits themselves going to zero at the endpoints. To obtain these we use the fact that $\Pr\{W_0(t) > \alpha\} = \Pr\{\sqrt{t(1-t)}B_0(t) > \alpha\}$. For most practical purposes though it is good enough to work with Equation B.6 and approximate the infinite sum by curtailing summation for values of k greater than 2.

B.3 Counting processes and martingales

While the basic ideas behind counting processes and martingales are both simple and natural, if we wish to maintain the fullest generality—in practice allowing the time interval to stretch without limit, as well as allowing the unboundedness for covariates—then things become difficult and very technical. Rebolledo's multivariate central limit theorem for martingales requires great care in its application.

Our preference is to assume for both time and the covariates, at least for the purposes of obtaining large sample results, support on finite bounded intervals. This enables us to work with standard well-known central limit theorems. Specifically we put time on the interval $(0,1)$. In this appendix we motivate the use of counting processes and martingales in the context of survival analysis problems and we describe their application in real situations. This also helps establish the links with other works in the field that do make an appeal to Rebollo's martingale central limit theorem. The goal of this appendix is to provide some understanding of the probability structure upon which the theory is based.

Martingales and stochastic integrals

Recalling the discussion of Section A.2 and that, for a bounded function $H(x)$ and the empirical distribution function $F_n(x)$, we have, by virtue of the Helly-Bray theorem, that $\int H(x)dF_n(x)$ converges in distribution to $\int H(x)dF(x)$. If we define $M(x) = F_n(x) - F(x)$ and change the order of integration, i.e., move the expectation operator, E , outside the integral, then

$$E \left\{ \int H(x)dM(x) \right\} = 0.$$

This expression is worth dwelling upon. We think of E as being an integral operator or as defining some property of a random variable, specifically a measure of location. The random variable of relevance is not immediately apparent but can be seen to be $F_n(x)$, an n -dimensional function from the observations to the interval $[0, 1]$. We can suppose, at least initially, the functions $F(x)$ and $H(x)$ to be fixed and known. Our conceptual model allows the possibility of being able to obtain repetitions of the experiment, each time taking n independent observations. Thus, for some fixed given x , the value of $F_n(x)$ will generally vary from one experiment to the next. We view x as an argument to a function, and $F_n(x)$ as being random having a distribution studied below in Section C.2. Recalling Section A.2 on integration, note that we can rewrite the above equation as:

$$E \lim_{\max \Delta_i \rightarrow 0} \sum \{M(x_i) - M(x_{i-1})\} H(x_{i-1}) = 0, \quad (\text{B.8})$$

where $\Delta_i = x_i - x_{i-1} > 0$ and where, as described in Section A.2 the summation is understood to be over an increasing partition in which $\Delta_i > 0$ and $\max \Delta_i$ goes to zero. Now, changing the order of taking limits, the above expression becomes

$$\lim_{\max \Delta_i \rightarrow 0} \sum E \{ [M(x_i) - M(x_{i-1})] H(x_{i-1}) \} = 0, \quad (\text{B.9})$$

a result which looks simple enough but that has a lot of force when each of the infinite number of expectations can be readily evaluated. Let's view Equation B.9 in a different light, one that highlights the sequential and ordered nature of the partition. Rather than focus on the collection of $M(x_i)$ and $H(x_i)$, we can focus our attention on the increments $M(x_i) - M(x_{i-1})$ themselves, the increments being multiplied by $H(x_{i-1})$, and, rather than work with the overall expectation implied by the operator E , we will set up a sequence of conditional expectations. Also, for greater clarity, we will omit the term $\lim_{\max \Delta_i \rightarrow 0}$ altogether. We will put it back when it suits us. This lightens the notation and helps to make certain ideas more transparent. Later, we will equate the effect of adding back in the term $\lim_{\max \Delta_i \rightarrow 0}$ to that of replacing finite differences by infinitesimal differences. Consider then

$$U = \sum \{M(x_i) - M(x_{i-1})\} H(x_{i-1}), \quad (\text{B.10})$$

and unlike the preceding two equations, we are able to greatly relax the requirement that $H(x)$ be a known function or that $M(x)$ be restricted to being the difference between the empirical distribution function and the distribution function. By sequential conditioning upon $\mathcal{F}(x_i)$ where $\mathcal{F}(x_i)$ are increasing sequence of sets denoting observations on $M(x)$ and $H(x)$, for all values of x less than or equal to x_i , we can derive results of wide applicability. In particular, we can now take $M(x)$ and $H(x)$ to be stochastic processes. Some restrictions are still needed for $M(x)$, in particular that the incremental means and variances exist. We will suppose that

$$E\{M(x_i) - M(x_{i-1}) | \mathcal{F}(x_{i-1})\} = 0, \quad (\text{B.11})$$

in words, when given $\mathcal{F}(x_{i-1})$, the quantity $M(x_{i-1})$ is fixed and known and the expected size of the increment is zero. This is not a strong requirement and only supposes the existence of the mean. If the expected size of the increment is other than zero, then we can subtract this difference to recover the desired property. Furthermore, given $\mathcal{F}(x)$, the quantity $H(x)$ is fixed. The trick is then to exploit the device of double expectation whereby for events, \mathcal{A} and \mathcal{B} , it is always true that $E(\mathcal{A}) = EE(\mathcal{A}|\mathcal{B})$. In the context of this expression, $\mathcal{B} = \mathcal{F}(x_{i-1})$, leading to

$$E(U) = \sum H(x_{i-1}) E\{M(x_i) - M(x_{i-1}) | \mathcal{F}(x_{i-1})\} = 0, \quad (\text{B.12})$$

and under the assumption that the increments are uncorrelated we have the variance is the sum of the variance of each component to the sum. Thus

$$\text{Var}(U) = \sum E\{H^2(x_{i-1})[M(x_i) - M(x_{i-1})]^2 | \mathcal{F}(x_{i-1})\}. \quad (\text{B.13})$$

In order to keep the presentation uncluttered we use a single operator E in the above expressions, but there are some subtleties that ought not to go unremarked.

For instance, in Equation B.13, the inner expectation is taken with respect to repetitions over all possible outcomes in which the set $\mathcal{F}(x_{i-1})$ remains unchanged, whereas the outer expectation is taken with respect to all possible repetitions. In Equation B.12 the outer expectation, taken with respect to the distribution of all potential realizations of all the sets $\mathcal{F}(x_{i-1})$, is not written and is necessarily zero since all of the inner expectations are zero. The analogous device to double expectation for the variance is not so simple since $\text{Var}(Y) = E\text{Var}(Y|Z) + \text{Var}E(Y|Z)$. Applying this we have

$$\text{Var}\{M(x_i) - M(x_{i-1})\} = E\text{Var}\{M(x_i) - M(x_{i-1})|\mathcal{F}(x_{i-1})\} \quad (\text{B.14})$$

since $\text{Var}E\{M(x_i) - M(x_{i-1})|\mathcal{F}(x_{i-1})\}$ is equal to zero, this being the case because each term is itself equal to the constant zero. The first term also requires a little thought, the outer expectation indicated by E being taken with respect to the distribution of $\mathcal{F}(x_{i-1})$, i.e., all the conditional distributions $M(x)$ and $H(x)$ where $x \leq x_{i-1}$. The next key point arises through the sequential nesting. These outer expectations, taken with respect to the distribution of $\mathcal{F}(x_{i-1})$ are the same as those taken with respect to the distribution of any $\mathcal{F}(x)$ for which $x \geq x_{i-1}$. This is an immediate consequence of the fact that the lower-dimensional distribution results from integrating out all the additional terms in the higher-dimensional distribution. Thus, if x_{\max} is the greatest value of x for which observations are made then we can consider that all of these outer expectations are taken with respect to $\mathcal{F}(x_{\max})$. Each time that we condition upon $\mathcal{F}(x_{i-1})$ we will treat $H(x_{i-1})$ as a fixed constant and so it can be simply squared and moved outside the inner expectation. It is still governed by the outer expectation which we will take to be with respect to the distribution of $\mathcal{F}(x_{\max})$. Equation B.13 then follows.

Making a normal approximation for U , and from the theory of estimating equations, given any set of observations, that U depends monotonically on some parameter β , then it is very straightforward to set up hypothesis tests for $\beta = \beta_0$. Many situations, including that of proportional hazards regression, lead to estimating equations of the form of U . The above set-up, which is further developed below in a continuous form, i.e., after having “added in” the term $\lim_{\max \Delta_i \rightarrow 0}$, applies very broadly. We need the concept of a process, usually indexed by time t , the conditional means and variances of the increments, given the accumulated information up until time t .

We have restricted our attention here to the Riemann-Stieltjes definition of the integral. The broader Lebesgue definition allows the inclusion of subsets of t tolerating serious violations of our conditions such as conditional means and variances not existing. The conditioning sets can be also very much more involved. Only in a very small number of applications has this extra generality been exploited. Given that it makes the main ideas much more difficult to all but those familiar with measure theory, it seems preferable to avoid it altogether. As

already mentioned, in exchange for a reduction in generality by the imposition of some constraints, e.g., putting the time frame on the interval $(0,1)$, we are able to appeal to standard central limit theorems in order to obtain large sample behavior.

Counting processes

The above discussion started off with some consideration of the empirical cumulative distribution function $F_n(t)$ which is discussed in much more detail in Section C.5. Let's consider the function $N(t) = \{nF_n(t) : 0 \leq t \leq 1\}$. We can view this as a stochastic process, indexed by time t so that, given any t we can consider $N(t)$ to be a random variable taking values from 0 to n . We include here a restriction that we generally make which is that time has some upper limit, without loss of generality, we call this 1. This restriction can easily be avoided but it implies no practical constraint and is often convenient in practical applications. We can broaden the definition of $N(t)$ beyond that of $nF_n(t)$ and we have:

Definition B.1. A counting process $N = \{N(t) : 0 \leq t \leq 1\}$ is a stochastic process that can be thought of as counting the occurrences (as time t proceeds) of certain type of events. We suppose these events occur singly.

Very often $N(t)$ can be expressed as the sum of n individual counting processes, $N_i(t)$, each one counting no more than a single event. In this case $N_i(t)$ is a simple step function, taking the value zero at $t = 0$ and jumping to the value one at the time of an event. The realizations of $N(t)$ are integer-valued step functions with jumps of size +1 only. These functions are right continuous and $N(t)$ is the (random) number of events in the time interval $[0, t]$. We associate with the stochastic process $N(t)$ an intensity function $\alpha(t)$. The intensity function serves the purpose of standardizing the increments to have zero mean. In order to better grasp what is happening here, the reader might look back to Equation B.11 and the two sentences following that equation. The mean is not determined in advance but depends upon \mathcal{F}_{t-} where, in a continuous framework, \mathcal{F}_{t-} is to \mathcal{F}_t what $\mathcal{F}_{x_{i-1}}$ is to $\mathcal{F}(x_i)$. In technical terms:

Definition B.2. A filtration, \mathcal{F}_t , is an increasing right continuous family of sub sigma-algebras (see A3 for the meaning of sigma-algebra).

This definition may not be very transparent to those unfamiliar with the requirement of sigma additivity for probability spaces and there is no real need to expand on it here. The requirement is a theoretical one which imposes a mathematical restriction on the size, in an infinite sense, of the set of subsets of \mathcal{F}_t . The restriction guarantees that the probability we can associate with any infinite sum of disjoint sets is simply the sum of the probabilities associated with those sets composing the sum. For our purposes, the only key idea of importance is that \mathcal{F}_{t-} is a set containing all the accumulated information (hence “increasing”) on all processes contained in the past up until but not including

the time point t (hence “right continuous”). We write, $\alpha = \{\alpha(t) : 0 \leq t \leq 1\}$ where

$$\alpha(t)dt = \Pr\{N(t) \text{ jumps in } [t, t+dt) | \mathcal{F}_{t-}\} = E\{dN(t) | \mathcal{F}_{t-}\},$$

the equality being understood in an infinitesimal sense, i.e., the functional part of the left-hand side, $\alpha(t)$, is the limit of the right-hand side divided by $dt > 0$ as dt goes to zero. In the chapter on survival analysis we will see that the hazard function, $\lambda(t)$, expressible as the ratio of the density, $f(t)$, to the survivorship function, $S(t)$, i.e., $f(t)/S(t)$, can be expressed in fundamental terms by first letting $Y(t) = I(T \geq t)$. Under this interpretation, we can also write

$$\lambda(t)dt = \Pr\{N(t) \text{ jumps in } [t, t+dt) | Y(t) = 1\} = E\{dN(t) | Y(t) = 1\}.$$

It is instructive to compare the above definitions of $\alpha(t)$ and $\lambda(t)$. The first definition is the more general since, choosing the sets \mathcal{F}_t to be defined from the at-risk function $Y(t)$ when it takes the value one, enables the first definition to reduce to a definition equivalent to the second. The difference is an important one in that if we do not provide a value for $I(T \geq t)$ then this is a $(0, 1)$ random variable, and in consequence, $\alpha(t)$ is a $(0, \lambda(t))$ random variable. For this particular case we can express this idea succinctly via the formula

$$\alpha(t)dt = Y(t)\lambda(t)dt. \quad (\text{B.15})$$

Replacing $Y(t)$ by a more general “at risk” indicator variable will allow for great flexibility, including the ability to obtain a simple expression for the intensity in the presence of censoring as well as the ability to take on-board multistate problems where the transitions are not simply from alive to dead but from, say, state j to state k summarized via $\alpha_{jk}(t)dt = Y_{jk}(t)\lambda_{jk}(t)dt$ in which $Y_{jk}(t)$ is left continuous and therefore equal to the limit $Y_{jk}(t-\epsilon)$ as $\epsilon > 0$ goes to zero through positive values, an indicator variable taking the value one if the subject is in state j and available to make a transition to state k at time $t - \epsilon$ as $\epsilon \rightarrow 0$. The hazards $\lambda_{jk}(t)$ are known in advance, i.e., at $t = 0$ for all t , whereas the $\alpha_{jk}(t)$ are randomly viewed from time point s where $s < t$, with the subtle condition of left continuity which leads to the notion of “predictability” described below. The idea of sequential standardization, the repeated subtraction of the mean, that leans on the evaluation of intensities, can only work when the mean exists. This requires a further technical property, that of being “adapted.” We say

Definition B.3. A stochastic process $X(t)$ is said to be adapted to the filtration \mathcal{F}_t if $X(t)$ is a random variable with respect to \mathcal{F}_t .

Once again the definition is not particularly transparent to nonprobabilists and the reader need not be over-concerned since it will not be referred to here apart from in connection with the important concept of a predictable process. The basic idea is that the relevant quantities upon which we aim to use the

tools of probability modeling should all be contained in \mathcal{F}_t . If any probability statement we wish to construct concerning $X(t)$ cannot be made using the set \mathcal{F}_t but requires the set \mathcal{F}_{t+u} , where $u > 0$, then $X(t)$ is not adapted to \mathcal{F}_t . In our context just about all of the stochastic processes that are of interest to us are adapted and so this need not be a concern. A related property, of great importance, and which also will hold for all of those processes we focus attention on, is that of predictability. We have

Definition B.4. *A real-valued stochastic process, $H(t)$, that is left continuous and adapted to the filtration \mathcal{F}_t is called a predictable process.*

Since $H(t)$ is adapted to \mathcal{F}_t it is a random variable with respect to \mathcal{F}_t . Since the process is left continuous it is also adapted to \mathcal{F}_{t-} . Therefore, whenever we condition upon \mathcal{F}_{t-} , $H(t)$ is simply a fixed and known constant. This is the real sense of the term “predictable” and, in practice, the property is a very useful one. It is frequently encountered in the probabilistic context upon which a great number of tests are constructed. Counting processes can be defined in many different ways and such a formulation allows for a great deal of flexibility. Suppose for instance that we have events of type 1 and events of type 2, indicated by $N_1(t)$ and $N_2(t)$ respectively. Then $N(t) = N_1(t) + N_2(t)$ counts the occurrences of events of either type. For this counting process we have

$$\alpha(t)dt = P(N(t) \text{ jumps in } [t, t+dt] | \mathcal{F}_{t-}),$$

i.e., the same as $P(N_1(t) \text{ or } N_2(t) \text{ jump in } [t, t+dt] | \mathcal{F}_{t-})$ and, if as is reasonable in the great majority of applications, where, we assume to be negligible the probability of seeing events occurring simultaneously compared to seeing them occur singly, then

$$\alpha(t)dt = E\{dN_1(t) + dN_2(t) | \mathcal{F}_{t-}\} = \alpha_1(t) + \alpha_2(t).$$

This highlights a nice linearity property of intensities, not shared by probabilities themselves. For example, if we consider a group of n subjects and n individual counting processes $N_i(t)$, then the intensity function, $\alpha(t)$, for the occurrence of an event, regardless of individual, is simply $\sum \alpha_i(t)$. This result does not require independence of the processes, only that we can consider as negligible the intensities we might associate with simultaneous events.

Another counting process of great interest in survival applications concerns competing risks. Suppose there are two types of event but that they cannot both be observed. The most common example of this is right censoring where, once the censoring event has occurred, it is no longer possible to make observations on $N_i(t)$. This is discussed more fully in the following chapters and we limit ourselves here to the observation that $N_i(t)$ depends on more than one variable. In the absence of further assumptions, we are not able to determine the intensity function, but if we are prepared to assume that the censoring mechanism is

independent of the failure mechanism, i.e., that $\Pr(T_i > t | C_i > c) = \Pr(T_i > t)$, then a simple result is available.

Theorem B.3. *Let the counting process, $N_i(t)$, depend on two independent and positive random variables, T_i and C_i such that $N_i(t) = I\{T_i \leq t, T_i \leq C_i\}$. Let $X_i = \min(T_i, C_i)$, $Y_i(t) = I(X_i \geq t)$; then $N_i(t)$ has intensity process*

$$\alpha_i(t)dt = Y_i(t)\lambda_i(t)dt. \quad (\text{B.16})$$

The counting process, $N_i(t)$, is one of great interest to us since the response variable in most studies will be of such a form, i.e., an observation when the event of interest occurs but an observation that is only possible when the censoring variable is greater than the failure variable. Also, when we study a heterogeneous group, our principal focus in this book, the theorem still holds in a modified form. Thus, if we can assume that $\Pr(T_i > t | C_i > c, Z = z) = \Pr(T_i > t | Z = z)$, we then have:

Theorem B.4. *Let the counting processes, $N_i(t)$, depend on two independent and positive random variables, T_i and C_i , as well as Z such that*

$$N_i(t) = I\{T_i \leq t, T_i \leq C_i, Z = z\}. \quad (\text{B.17})$$

Then the intensity process for $N_i(t)$ can be written as $\alpha_i(t, z)dt = Y_i(t)\lambda_i(t, z)dt$.

The assumption needed for Theorem B.4, known as the conditional independence assumption, is weaker than that needed for B.3 in that the latter theorem contains the former as a special case. Note that the stochastic processes $Y_i(t)$ and $\alpha_i(t)$ are left continuous and adapted to \mathcal{F}_t . They are therefore predictable stochastic processes, which means that, given \mathcal{F}_{t-} , we treat $Y_i(t)$, $\alpha_i(t)$ and, assuming that $Z(t)$ is predictable, $\alpha_i(t, z)$ as fixed constants.

B.4 Inference for martingales and stochastic integrals

The reader might look over Section B.3 for the probability background behind martingales. A martingale $M = \{M(t) : t \geq 0\}$ is a stochastic process whose increment over an interval $(u, v]$, given the past up to and including time u , has expectation zero, i.e., $E\{M(v) - M(u) | \mathcal{F}_u\} = 0$ for all $0 \leq u < v < 1$. Equation B.11 provides the essential idea for the discrete time case. We can rewrite the above defining property of martingales by taking the time instants u and v to be

just before and just after the time instant t . Letting both v and u tend to t and u play the role of $t-$, we can write;

$$E\{dM(t)|\mathcal{F}_{t-}\} = 0. \quad (\text{B.18})$$

Note that this is no more than a formal way of stating that, whatever the history \mathcal{F}_t may be, given this history, expectations exist. If these expectations are not themselves equal to zero then we only need to subtract the nonzero means to achieve this end. A counting process $N_i(t)$ is not of itself a martingale but note, for $0 \leq u < v \leq 1$, that $E\{N_i(v)|\mathcal{F}_u\} > E\{N_i(u)|\mathcal{F}_u\}$ and, as above, by taking the time instants u and v to be just before and just after the time instant t , letting v and u tend to t and u play the role of $t-$, we have

$$E\{dN_i(t)|\mathcal{F}_{t-}\} > 0. \quad (\text{B.19})$$

A stochastic process $N_i(t)$ with the above property is known as a submartingale. Again, providing expectations are finite, it is only a matter of subtracting the sequentially calculated means in order to bring a submartingale under the martingale heading. This idea is made precise by the theorem of Doob-Meyer.

Doob-Meyer decomposition

For the submartingale $N_i(t)$, having associated intensity process $\alpha(t)$, we have from Equation B.15 that $E\{dN(t)|\mathcal{F}_{t-}\} = \alpha(t)dt$. If we write $dM(t) = dN(t) - \alpha(t)dt$ then $E\{dM(t)|\mathcal{F}_{t-}\} = 0$. Thus $M(t)$ is a martingale. For the counting processes of interest to us we will always be able to integrate $\alpha(t)$ and we define $A(t) = \int_0^t \alpha(s)ds$. We can write

$$N_i(t) = M_i(t) + A_i(t). \quad (\text{B.20})$$

Such a decomposition of a submartingale into the sum of a martingale and a predictable stochastic process, $A_i(t)$, is an example of a more general theorem for such decompositions known as the Doob-Meyer theorem. It can be applied to quite general submartingales, the precise conditions under which require measure-theoretic arguments. For the counting processes of interest to us in survival analysis the theorem always applies. The predictable process $A_i(t)$ is called the compensator of $N_i(t)$. In simple terms the compensator is used to make the means zero thereby producing the martingales that our theory needs.

The compensator $A_i(t)$

A counting process, $N_i(t)$, is simply a random variable indexed by t . This is the definition of a stochastic process so that $N_i(t)$ is, in particular, a stochastic process. For the majority of applications in survival analysis, $N_i(t)$ will count no further than one; at the outset, $N_i(0)$ takes the value zero and, subsequently, the

value one for all times greater than or equal to that at which the event of interest occurs. But, generally, $N_i(t)$ may assume many, or all, integer values. Note that any sum of counting processes can be immediately seen to be a counting process in its own right. An illustrative example could be the number of goals scored during a soccer season by some team. Here, the indexing variable t counts the minutes from the beginning of the season. The expectation of $N_i(t)$ (which must exist given the physical constraints of the example) may vary in a complex way with t , certainly non-decreasing and with long plateau when it is not possible for a goal to be scored, for instance when no game is being played. At time $t = 0$, it might make sense to look forward to any future time t and to consider the expectation of $N_i(t)$.

As the season unfolds, at each t , depending on how the team performs, we may exceed, possibly greatly, or fall short of, the initial expectation of $N_i(t)$. As the team's performance is progressively revealed to us, the original expectations are of diminishing interest and it is clearly more useful to consider those conditional expectations in which we take account of the accumulated history at time point t . Working this out as we go along, we determine $A_i(t)$ so that $N_i(t) - A_i(t)$, given all that has happened up to time t , has zero expectation. When $\alpha_i(s)$ is the intensity function for $N_i(s)$, then

$$A_i(t) = \int_0^t \alpha_i(s) ds$$

and this important result is presented in Theorem B.5 given immediately below.

Predictable variation process

Linear statistics of the form U described in Section D.1, following standardization, are, not surprisingly, well approximated by standard normal variates. We will see this below using results for estimating equations and results for sums of independent, although not necessarily identical, random variables. The martingale central limit theorem can also be used in this context, and for all of our applications, it is possible to apply it in a simple form avoiding measure-theoretic arguments. Such an approach would then coincide with standard results for sums of independent random variables. In order to standardize U we will require an estimate of the variance as well as the mean. Unlike, say, Brownian motion or the Ornstein-Uhlenbeck processes mentioned in the previous chapter, where at time t we have a very simple expression for the variance, the variance of U can be complex and will clearly depend on $H(x)$. One way of addressing this question is through the use of the predictable variation process. We know from the above that:

$$E\{dN(t)|\mathcal{F}_{t-}\} = \alpha(t)dt, \quad E\{dM(t)|\mathcal{F}_{t-}\} = 0.$$

Conditional upon \mathcal{F}_{t-} , we can view the random variable $dN(t)$ as a Bernoulli $(0,1)$ having mean $\alpha(t)dt$ and variance given by $\alpha(t)dt\{1-\alpha(t)dt\}$. In contrast, the random variable $dM(t)$, conditional on \mathcal{F}_{t-} , has mean zero and the same variance. This follows since, given \mathcal{F}_{t-} , $\alpha(t)$ is fixed and known. As usual, all the equations are in an infinitesimal sense, the equal sign indicating a limiting value as $dt \rightarrow 0$. In this sense $\alpha^2(t)(dt)^2$ is negligible when compared to $\alpha(t)dt$ since the ratio of the first to the second goes to zero as t goes to zero. Thus, the incremental variances are simply the same as the means, i.e., $\alpha(t)dt$. This, of course, ties in exactly with the theory for Poisson counting processes.

Definition B.5. *The predictable variation process of a martingale $M(t)$, denoted by $\langle M \rangle = \{\langle M \rangle(t) : t \geq 0\}$ is such that*

$$d\langle M \rangle(t) = E\{[dM(t)]^2 | \mathcal{F}_{t-}\} = \text{Var}\{dM(t) | \mathcal{F}_{t-}\}. \quad (\text{B.21})$$

The use of pointed brackets has become standard notation here and, indeed, the process is often referred to as the pointed brackets process. Note that $\langle M \rangle$ is clearly a stochastic process and that the process is predictable and non-decreasing. It can be thought of as the sum of conditional variances of the increments of M over small time intervals partitioning $[0, t]$, each conditional variance being taken given what has happened up to the beginning of the corresponding interval. We then have the following important result:

Theorem B.5. *Let $M_i(t) = N_i(t) - A_i(t)$ where $A_i(t) = \int_0^t \alpha_i(s)ds$. Then*

$$\langle M_i \rangle(t) = A_i(t). \quad (\text{B.22})$$

Corollary B.4. *Define, for all t and $i \neq j$, the predictable covariation process, $\langle M_i, M_j \rangle$, of two martingales, M_i and M_j , analogously to the above. Then*

$$\langle M_i, M_j \rangle(t) = 0. \quad (\text{B.23})$$

The corollary follows readily if, for $i \neq j$, the counting processes $N_i(t)$ and $N_j(t)$ can never jump simultaneously. In this case the product $dN_i(t)dN_j(t)$ is always equal to zero. Thus, the conditional covariance between $dN_i(t)$ and $dN_j(t)$ is $-\alpha_i(t)dt \cdot \alpha_j(t)dt$.

Stochastic integrals

The concept of a stochastic integral is very simple; essentially we take a Riemann-Stieltjes integral, from zero to time point t , of a function which, at the outset when $t = 0$ and looking forward, would be random. Examples of most immediate interest to us are: $N(t) = \int_0^t dN(s)$, $A(t) = \int_0^t dA(s)$ and $M(t) = \int_0^t dM(s)$. Of particular value are integrals of the form $\int_0^t H(s)dM(s)$ where $M(s)$ is a martingale and $H(s)$ a predictable function. By predictable we mean that if we know all the values of $H(s)$ for s less than t then we also know

$H(t)$, and this value is the same as the limit of $H(s)$ as $s \rightarrow t$ for values of s less than t .

The martingale transform theorem provides a tool for carrying out inference in the survival context. Many statistics arising in practice will be of a form U described in Appendix D on estimating equations. For these the following result will find immediate application:

Theorem B.6. *Let M be a martingale and H a predictable stochastic process. Then M^* is also a martingale where it is defined by:*

$$M^*(t) = \int_0^t H(s)dM(s). \quad (\text{B.24})$$

Corollary B.5. *The predictable variation process of the stochastic process $M^*(t)$ can be written*

$$\langle M^* \rangle(t) = \int_0^t H(s)^2 d\langle M \rangle(s) = \int_0^t H^2(s)dA(s). \quad (\text{B.25})$$

There is a considerable theory for stochastic integrals, much of it developed in the econometric and financial statistical literature. For our purposes and for all the tests that have been developed in the framework of counting processes for proportional hazards models, the above theorem and corollary are all that are needed. A very considerable array of test procedures come directly under this heading. Many modern approaches to survival analysis lean on the theory of stochastic integrals in order to carry out inference. The main approaches here are sometimes different, based instead on the main theorem of proportional hazards regression, Donsker's theorem, and known results concerning Brownian motion and functions of Brownian motion. Behind both approaches are the ideas relating to the limits of sums of conditionally independent increments. In some situations the resulting statistics that we use are identical.

Central limit theorem for martingale processes

The hard part of the work in obtaining a large sample result for stochastic integrals is to find a convergence in probability result for the variation process $\langle M \rangle(t)$. If this limit exists then we write it as $A(t)$. The result can be summarized in a theorem which depends on two conditions.

Theorem B.7. *Suppose we have the following two conditions:*

1. *As n increases without bound, $\langle M \rangle(t)$ converges in probability to $A(t)$,*
2. *As n increases, the jumps in $M(t)$ tend to zero.*

Then, the martingale $M(t)$ converges to a Gaussian process with mean zero and variance $A(t)$.

Added conditions can make it easier to obtain the first one of these conditions, and as a result, there are a number of slightly different versions of these two criteria. The multivariate form has the same structure. In practical situations, we take $M(\infty)$ to be $\mathcal{N}(0, \sigma^2)$ where we estimate σ^2 by $\langle M \rangle(\infty)$.

Censoring and at-risk functions

The counting process structure does well on an intuitive level in dealing with the concept of censoring and the concept of being at risk. We will only observe the actual counts but we can imagine that the probability, more precisely the intensity when referring to infinitely small time periods, can change in complex ways through time. In particular, there may be time periods when, although a key event of interest may occur, we are unable to observe it because of some censoring phenomenon. As an example, allow $N(t)$ to count some event of interest and define $Y(t)$ to take the value zero when t lies in the semi-closed interval $(a, b]$, (when we are unable to observe the event of interest) and the value one otherwise. The counting process $N^*(t)$ where,

$$N^*(t) = \int_0^t Y(s)dN(s), \quad (\text{B.26})$$

counts observable events. If the censoring does not modify the compensator, $A(t)$, of $N(t)$, then $N(t)$ and $N^*(t)$ have the same compensator. The difference, $M^*(t) = N^*(t) - A(t)$ would typically differ from the martingale $M(t)$ but would nonetheless still be a martingale in its own right. In addition, it is easily anticipated how we might go about tackling the much more complex situation in which the censoring would not be independent of the failure mechanism. Here, the compensators for $N^*(t)$ and $N(t)$ do not coincide. For this more complex case, we would need some model, $A^*(t)$, for the compensator of $N^*(t)$ in order that $M^*(t) = N^*(t) - A^*(t)$ would be a martingale.

The most common and the simplest form of the at-risk indicator $Y(t)$ is one where it assumes the value one at $t = 0$, retaining this value until censored or failed, beyond which time point it assumes the value zero. When dealing with n individuals, and n counting processes, we can write $\bar{N}(t) = \sum_{i=1}^n N_i(t)$ and use the at-risk indicator to denote the risk set. If $Y_i(t)$ refers to individual i , then $\bar{Y}(t) = \sum_{i=1}^n Y_i(t)$ is the risk set at time t . The compensator for $N_i(t)$ is $\alpha_i(t) = Y_i(t)\lambda_i(t)$, where $\lambda_i(t)$ is the hazard for subject i , written simply as $\lambda(t)$ in the case of i.i.d. replications. Then, the compensator, $\bar{A}(t)$, for $\bar{N}(t)$ is:

$$\bar{A}(t) = \int_0^t \{\sum_{i=1}^n Y_i(s)\}\lambda(s)ds = \int_0^t \bar{Y}(s)\lambda(s)ds.$$

The intensity process for $\bar{N}(t)$ is then given by $\bar{Y}(t)\lambda(t)$. The multiplicative intensity model Aalen (1978) has as its cornerstone the product of the fully

observable quantity $\bar{Y}(t)$ and the hazard rate, $\lambda(t)$ which, typically, will involve unknown model parameters. In testing specific hypotheses we might fix some of these parameters at particular population values, most often the value zero.

Nonparametric statistics

The multiplicative intensity model just described and first recognized by Aalen (1978) allows a simple expression, and simple inference, for a large number of nonparametric statistics that have been used in survival analysis over the past half century. Martingales are immediate candidates for forming an estimating equation with which inference can be made on unknown parameters in the model. For our specific applications, these estimating equations will almost always present themselves in the form of a martingale. For example, the martingale structure can be used to underpin several nonparametric tests. Using a martingale as an estimating equation we can take $\bar{N}(t)$ as an estimator of its compensator $\bar{A}(t)$. Dividing by the risk set (assumed to be always greater than zero), we have:

$$\int_0^t \frac{d\bar{N}(s)}{\bar{Y}(s)} - \int_0^t \lambda(s)ds,$$

a quantity which is a martingale and has, in consequence, zero expectation. An estimate of the cumulative risk $\hat{A}(t)$ is given by $\hat{A}(t) = \int_0^t \bar{Y}(s)^{-1} d\bar{N}(s)$ which is the estimator of Nelson-Aalen. Simple nonparametric statistics for the comparison of two groups can be obtained immediately from this. If we consider some predictable weighting process $W(s)$ (for an explanation of what we mean by “predictable” see the beginning of this appendix), then define

$$K(t) = \int_0^t W(s) \left\{ \frac{d\bar{N}_1(s)}{\bar{Y}_1(s)} - \frac{d\bar{N}_2(s)}{\bar{Y}_2(s)} \right\} \quad (\text{B.27})$$

where a subscript 1 denotes subjects from group 1 and a 2 from group 2. The choice of the weighting function $W(s)$ can be made by the user and might be chosen when some particular alternative is in mind. The properties of different weights were investigated by Prentice (1978) and Harrington and Fleming (1982). We can readily claim that $K(\infty)$ converges in probability to $\mathcal{N}(0, \sigma^2)$. We estimate σ^2 by $\langle K \rangle(\infty)$ where

$$\langle K \rangle(t) = \int_0^t \left\{ \frac{W(s)}{\bar{Y}_1(s)} \right\}^2 d\bar{N}_1(s) + \int_0^t \left\{ \frac{W(s)}{\bar{Y}_2(s)} \right\}^2 d\bar{N}_2(s). \quad (\text{B.28})$$

The choice $W(s) = \bar{Y}_1(s)\bar{Y}_2(s)/[\bar{Y}_1(s) + \bar{Y}_2(s)]$ leads to the log-rank test statistic and would maintain good power under a proportional hazards alternative of a constant group difference as opposed to the null hypothesis of no group differences. The choice $W(s) = \bar{Y}_1(s)\bar{Y}_2(s)$ corresponds to the weighting suggested

by Gehan (1965) in his generalization of the Wilcoxon statistic. This test may offer improved power over the log-rank test in situations where the group difference declines with time. These weights and therefore the properties of the test are impacted by the censoring, and in order to obtain a test free of the impact of censoring, Prentice (1978) suggested the weights, $W(s) = \hat{S}_1(s)\hat{S}_2(s)$. These weights would also offer the potential for improved power when the regression effect declines with time. These weights are also considered in the chapter on test statistics.

Appendix C

Limit theorems

C.1 Empirical processes and central limit theorems

We outline the main theorems providing inference for sums of random variables. The theorem of de Moivre-Laplace is a well-known special case of the central limit theorem and helps provide the setting. Our main interest is in sums which can be considered to be composed of independent increments. The empirical distribution function $F_n(t)$ is readily seen to be a consistent estimator for $F(t)$ at all continuity points of $F(t)$. However, we can also view $F_n(t)$ as a constant number multiplying a sum of independent Bernoulli variates and this enables us to construct inference for $F(t)$ on the basis of $F_n(t)$. Such inference can then be extended to the more general context of estimating equations. Inference for counting processes and stochastic integrals is described since this is commonly used in this area and, additionally, shares a number of features with an approach based on empirical processes.

The importance of estimating equations is stressed, in particular equations based on the method of moments and equations derived from the likelihood. Resampling techniques can also be of great value for problems in inference. All of the statistics that arise in practical situations of interest can be seen quite easily to fall under the headings described here. These statistics will have known large sample distributions. We can then appeal immediately to known results from Brownian motion and other functions of Brownian motion. Using this approach to inference is reassuring since (1) the building blocks are elementary ones, well-known to those who have followed introductory courses on inference (this is not the case, for instance, for the martingale central limit theorem) and (2) we obtain, as special cases, statistics that are currently widely used, the most notable examples being the partial likelihood score test and weighted log-rank statistics. However, we will obtain many more statistics, all of which can be seen to sit in a

single solid framework and some of which, given a particular situation of interest, will suggest themselves as being potentially more suitable than others.

C.2 Limit theorems for sums of random variables

The majority of statistics of interest that arise in practical applications are directly or indirectly (e.g., after taking the logarithm to some base) expressible as sums of random variables. It is therefore of immense practical value that the distribution theory for such sums can, in a wide variety of cases, be approximated by normal distributions. Moreover, we can obtain some idea as to how well the approximation may be expected to behave. It is also possible to refine the approximation. In this section we review the main limit theorems applicable to sums of random variables.

Weak law of large numbers for non-correlated variables

A central theorem provides the conditions under which we obtain the convergence in probability of the empirical mean for centered variables.

Theorem C.1. *Let X_1, X_2, \dots, X_n be random variables having mean zero and uncorrelated. We assume that there exists a constant $C \in \mathbb{R}^+$ such that $V(X) = \text{Var}(X)$, $V(X_i) \leq C$, for $i = 1, \dots, n$. Then we have:*

$$\frac{1}{n} \sum_{i=1}^n X_i \xrightarrow[n \rightarrow \infty]{P} 0.$$

Theorem of De Moivre-Laplace

Let $N_n = \sum_{i=1}^n X_i$ be the number of successes in n independent Bernoulli trials X_i , each trial having probability of success equal to p . Then

$$\{N_n - np\} / \sqrt{np(1-p)} \rightarrow \mathcal{N}(0, 1)$$

where \rightarrow means convergence in distribution. This is the oldest result of a central limit type and is the most well-known special case of the more general result, just below, for sums of independent and identically distributed random variables.

Central limit theorem for i.i.d. variables

Let $X_i, i = 1, 2, \dots$ be independent random variables having the same distribution $F(\cdot)$. We assume that $\int u^2 dF(u) < \infty$. Let $\sigma^2 = \int u^2 dF(u) - \mu^2$ where $\mu = \int u dF(u)$. Let $\bar{x} = \int u dF_n(u)$ where $F_n(t) = n^{-1} \sum_{i=1}^n I(T_i \leq t)$. Then the central limit theorem states that

$$\sigma \sqrt{n} (\bar{x} - \mu) \rightarrow \mathcal{N}(0, 1).$$

Less formally we state that \bar{x} converges to a normal distribution with mean μ and variance σ^2/n . This result is extremely useful and also quite general. For example, applying the mean value theorem, then for $g(\bar{x})$, where $g(x)$ is a differentiable function of x , we can see, using the same kind of informal statement, that $g(\bar{x})$ converges to a normal distribution with mean $g(\mu)$ and variance $\{g'(\mu)\}^2\sigma^2/n$.

Central limit theorem for independent variables

For nonidentically distributed random variables the problem is very much more involved. This is because of the large number of potential situations that need be considered. The most succinct solution appeared as the condition described below. Let $X_i, i = 1, 2, \dots$ be independent random variables having distributions $F_i(\cdot)$. Let $\sigma_i^2 = \int u^2 dF_i(u) < \infty$ and $\mu_i = \int u dF_i(u)$. Let $B_n^2 = \sum \sigma_i^2$ and define \int_ϵ to be an integral over the real line such that $|t - \mu_i| > \epsilon B_n$. Introduce the following:

Condition C.1. For each $\epsilon > 0$, $\sum B_n^{-2} \int_\epsilon (t - \mu_i)^2 \rightarrow 0$, as $n \rightarrow \infty$. If this condition is satisfied then

$$nB_n^{-1}(\bar{x} - n^{-1} \sum \mu_i) \rightarrow \mathcal{N}(0, 1).$$

This condition is known as the Lindeberg condition. The statement is an “only if” statement. Less formally we say that \bar{x} converges to a normal distribution with mean $\sum \mu_i/n$ and variance B_n^2/n^2 . The condition is simply a way of formulating or expressing mathematically the need that the sum be composed of independent “relevant” contributions. If a single term or group of terms dominate the sum such that the remaining contributions are in some sense negligible, then our intuition may tell us it would not be reasonable to anticipate the central limit theorem to generally apply. This could happen in various ways, in particular if there is “too much” information in the tails of the distributions, i.e., the tails are too heavy or σ_i^2 diminishes with increasing i at too fast a rate. It follows from the Lindeberg condition that

$$B_n^{-2} \sigma_n^2 \rightarrow 0, \quad B_n \rightarrow \infty, \text{ as } n \rightarrow \infty.$$

It can be fairly easily shown that the condition below implies the Lindeberg condition and provides a more ready way of evaluating whether or not asymptotic normality obtains.

Condition C.2. The Lindeberg condition holds if, for $k > 2$,

$$B_n^{-k} \sum \kappa_k \rightarrow 0 \quad \text{as } n \rightarrow \infty.$$

Central limit theorem for dependent variables

Let $X_i, i = 1, 2, \dots$ be a sequence of random variables having distributions $F_i(\cdot)$. Let $\sigma_i^2 = \int u^2 dF_i(u) < \infty$ and $\mu_i = \int u dF_i(u)$. As before, let $B_n^2 = \sum \sigma_i^2$. Then, under certain conditions:

$$nB_n^{-1} \left(\bar{x} - n^{-1} \sum \mu_i \right) \rightarrow \mathcal{N}(0, 1).$$

As we might guess, the conditions in this case are much more involved and we need to use array notation in order to express the cross dependencies that are generated. If we take an extreme case we see immediately why the dependencies have to be carefully considered for, suppose $X_i = \alpha_{i-1} X_{i-1}$ where the α_i are nonzero deterministic coefficients such that $\sum_1^n \alpha_i \rightarrow 1$, then clearly X_n converges in distribution to X_1 which can be any chosen distribution. In rough terms, there needs to be enough independence between the variables for the result to hold. Describing what is meant by “enough” is important in certain contexts, time series analysis being an example, but, since it is not needed in this work, we do not spend any time on it here. A special case of nonidentical distributions, of value in survival analysis, is the following.

Central limit theorem for weighted sums of i.i.d. variables

Let $X_i, i = 1, 2, \dots$ be independent random variables having the same distribution $F(\cdot)$. Let $\sigma^2 = \int u^2 dF(u) < \infty$ and $\mu = \int u dF(u)$. Let $a_i, i = 1, \dots, n$, be constants, $S_n = n^{-1/2} \sum_{i=1}^n a_i (X_i - \mu)$ and $\sigma_S^2(n) = \sigma^2 \sum_{i=1}^n a_i^2 / n$, then $S_n / \sigma_S(n) \rightarrow \mathcal{N}(0, 1)$ where $\sigma_S(n) = \{\sigma_S^2(n)\}^{1/2}$, whenever the following condition holds;

Condition C.3. *The coefficients a_i are constants and are such that*

$$\frac{\max |a_i|}{\sqrt{\sum_{j=1}^n a_j^2}} \rightarrow 0.$$

Many statistics arising in nonparametric theory come under this heading, e.g., linear sums of ranks. The condition is a particularly straightforward one to verify and leads us to conclude large sample normality for the great majority of the commonly used rank statistics in nonparametric theory. A related condition, which is sometimes of more immediate applicability, can be derived as a consequence of the above large sample result together with an application of Slutsky’s theorem. Suppose, as before, that $X_i, i = 1, 2, \dots$ are independent random variables having the same distribution $F(\cdot)$, that $\sigma^2 = \int u^2 dF(u) < \infty$, $\mu = \int u dF(u)$ and that $a_i, i = 1, \dots, n$, are constants. Again, letting $S_n = n^{-1/2} \sum_{i=1}^n a_i (X_i - \mu)$ and $\sigma_S^2(n) = \sigma^2 \sum_{i=1}^n a_i^2 / n$, then $S_n \rightarrow \mathcal{N}(0, \sigma^2 a^2)$ where:

Condition C.4. *The mean of the constant coefficients a_i converges and*

$$\frac{1}{n} \sum_{j=1}^n a_j^2 \rightarrow \alpha^2, \quad 0 < \alpha^2 < \infty.$$

The condition is useful in that it will allow us to both conclude normality for the linear combination S_n , and at the same time, provide us with a variance for the linear combination. Weighted log-rank statistics and score statistics under non-proportional hazards models can be put under this heading. The weights in that case are not fixed in advance but, since the weights are typically bounded and converge to given quantities, it is relatively straightforward to put in the extra steps to obtain large sample normality in those cases too.

C.3 Functional central limit theorem

If we limit our attention to sums of random variables, each of which is indexed by a value t lying between 0 and 1 (we lose no generality in practice by fixing an upper limit 1 rather than infinity), we can obtain many useful results. The important idea here is that of the order among the random variables indexed by t , since t will be a real number between 0 and 1. As always the sums of interest will be finite, sums of quantities evaluated at some finite number of time points on the interval $(0,1)$, and, we will appeal to known results concerning the limiting continuous distributions, as the interval is “filled out,” as a means to approximate the exact, but necessarily complicated, finite sample distributions. For this reason it is helpful to begin reasoning in terms of sums, indexed by a finite number of points, and consider what such sums look like as the number of points increases without limit.

Sums of i.i.d. variables on interval $(0,1)$

Imagine a process starting at the origin and making successive displacements $X_i, i = 1, \dots, n$, where the X_i are all independent. For every k , where $1 \leq k \leq n$, the total distance traveled from the origin can be represented by $U_k = \sum_{i \leq k} X_i$ (random walk). The simplest way of looking at such a process is to consider the interval $(0,1)$ divided into n equal nonoverlapping intervals each of size $1/n$. This can only be achieved in one way. We make observations X_i , and therefore U_i , at the points $t = i/n, i = 1, \dots, n$. The increments X_i are independent. We have $E(X_i) = 0$ and $\text{Var}(X_i) = \sigma^2 < \infty$. In consequence we see that $E(U_k) = 0$, that $E(U_k^2) = k\sigma^2$ and, in view of the central limit theorem, that $E(U_k^\ell) \rightarrow 0, \forall \ell \text{ odd}$. We make the process continuous by linearly interpolating between the points at which $U_i, i = 1, \dots, n$ is defined. Note that there are much more general developments of the limiting process than we obtain here (Brownian motion) and that continuity can be demonstrated as a property of the limiting process.

However, it seems easier to construct the process already having continuity as a property for finite situations. This avoids technical difficulties and, perhaps more importantly, helps illustrate why and how, in practice, we can construct processes that will look like Brownian motion. Indeed, not only will these processes look like Brownian motion, but their probabilistic behavior, of practical interest to us, can be accurately approximated by the known properties of Brownian motion. Finally, just as in the standardizations of the preceding sections, we need to standardize the variance of our process. This we do by considering the sum

$$U_k^* = (\sigma\sqrt{n})^{-1} U_k = (\sigma\sqrt{n})^{-1} \sum_{i \leq k} X_i,$$

from which, letting $t = k/n$, we readily obtain the mean and the variance of U_k^* as

$$E(U_k^*) = (\sigma\sqrt{n})^{-1} \sum_{i \leq k} E(X_i) = 0; \quad \text{Var}(U_k^*) = (\sigma\sqrt{n})^{-2} k \sigma^2 = t. \quad (\text{C.1})$$

Although this and the following section are particularly simple, the reader should make sure that he or she has a very solid understanding as to what is taking place. It underscores all the main ideas behind the methods of inference that are used. An example of such a process in which $\sigma^2 = 1$ and $n = 30$ is shown in Figure C.1 As for $\text{Var}(U_k^*)$ we see in the same way that;

Theorem C.2. *For $k < m$, $\text{Cov}(U_k^*, U_m^*) = t$ where $t = k/n$.*

The important thing to note is that the increments are independent, implying convergence to a Gaussian process. All we then need is the covariance process. Figure C.1 and Figure C.2 represent approximations to Brownian motion in view of discreteness and the linear interpolation. The figures indicate two realizations from the above summed processes, and the reader is encouraged to carry out his or her own such simulations, an easy exercise, and yet invaluable in terms of building good intuition. An inspection of any small part of the curve (take, for example, the curve between 0.30 and 0.31 where the curve is based on less than 100 points), might easily be a continuous straight line, nothing at all like the limiting process, Brownian motion. But imagine, as very often is the case in applications, that we are only concerned about some simple aspect of the process, for instance, the greatest absolute distance traveled from the origin for the transformed process, tied down at $t = 1$. With as few as 30 observations our intuition would correctly lead us to believe that the distribution of this maximum will be accurately approximated by the same distribution, evaluated under the assumption of a Brownian bridge. Of course, such a statement can be made more technically precise via use, for example, of the law of the iterated logarithm or the use of Berry-Esseen bounds.

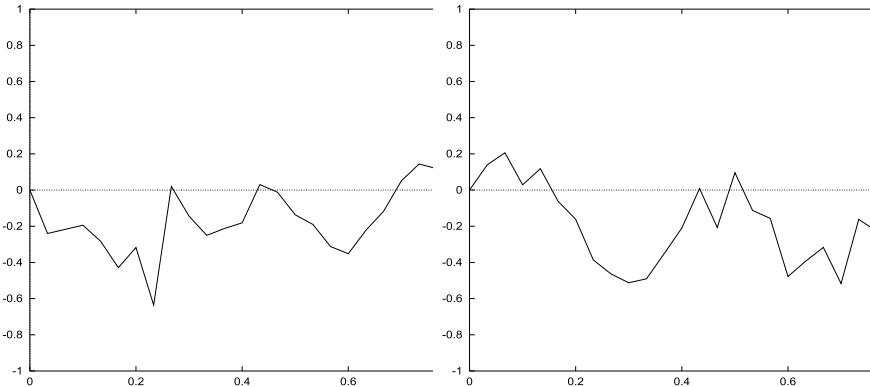


Figure C.1: Two independent simulations of sums of 30 points on interval $(0,1)$.

Sums of independent variables on $(0,1)$

The simple set-up of the previous section, for which the large sample theory, described above, is well established, can be readily extended to the non i.i.d. case. Let's begin by relaxing the assumption that the variances of the X_i do not depend on i . Suppose that as before $E(X_i) = 0$ and let $\text{Var}(X_i) = \sigma_i^2 < \infty$. Then clearly the process

$$U_k^* = (\sqrt{n})^{-1} \sum_{i \leq k} \sigma_i^{-1} X_i$$

will look like the process defined above. In particular, straightforward manipulation as above shows that $E(U_k^*) = 0$ and $\text{Cov}(U_k^*, U_m^*) = t$ where $k < m$ and $t = k/n$. We allow k to increase at the same rate, i.e., $k = nt$ where $0 < t < 1$. As $n \rightarrow \infty$ the number of possible values of t , $t \in (0,1)$ also increases without limit to the set of all rationals on this interval. We can also suppose that as $k, n \rightarrow \infty$; $k < n$, such that $k/n = t$ then σ_t^2 converges almost everywhere to some function $\sigma^2(t)$. We then allow n to increase without bound.

The functional central limit theorem states that the above process goes to a limit. The limiting process is defined on the real interval. Choosing any set of points $\{t_1, \dots, t_k\}$, $(0 < t_i < 1, i = 1, \dots, k)$ then the process $U_{t_1}^*, U_{t_2}^*, \dots, U_{t_k}^*$ converges in distribution to the multivariate normal. As indicated above the covariance only depends on the distance between points so that $\text{Cov}\{U_s^*, U_t^*\} = s; s < t$. The basic idea is that the increments, making up the sum $U^*(t)$, get smaller and smaller as n increases. The increments have expectation equal to zero unless there is drift. Also, the way in which the increments become smaller with n is precisely of order \sqrt{n} . The variance therefore increases linearly with time out in the process. In practical applications, it is only necessary that the increments be independent and that these increments have a finite variance. It

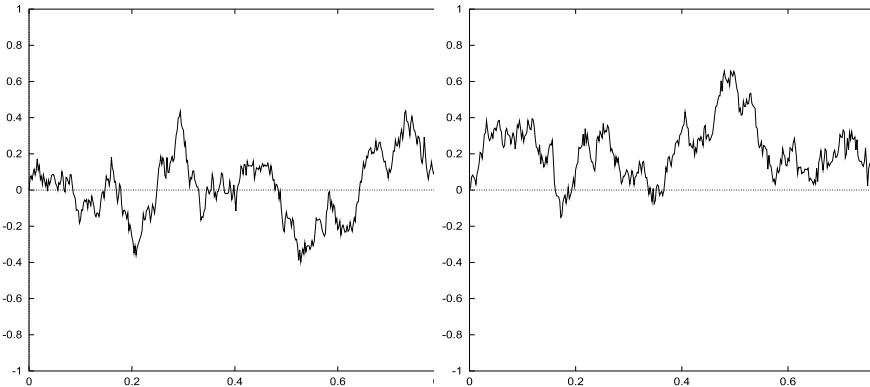


Figure C.2: Two independent simulations of sums of 500 points on interval $(0,1)$.

is then straightforward to carry out a time transformation to obtain the limiting process as an immediate consequence of the functional central limit theorem. The functional central limit theorem differs very little in essence from the usual central limit theorem, from which it derives. The key additional idea is that of sequential standardization. It is all very simple but, as we shall see, very powerful.

C.4 Brownian motion as limit process

Perhaps the most well-known application of the above is to the empirical distribution function. A great deal can be said about the empirical distribution function by appealing to the large sample results that can be anticipated from Donsker's theorem. Brownian motion can be seen to be the limit process obtained by applying the functional central limit theorem. We make wide use of these results in this book. Donsker's theorem focuses on the case of linear interpolation of a stochastic process having independent increments with the same distribution and, specifically, having the same variance.

Theorem C.3. (Donsker, 1951). *Let $(\xi_n)_{n \in \mathbb{N}}$ be a sequence of i.i.d. random variables on the probability space (Ω, \mathcal{F}, P) such that $E(\xi_n) = 0$ et $V(\xi_n) = \sigma^2$ for all $n \in \mathbb{N}$, then*

$$X_n \xrightarrow[n \rightarrow \infty]{\mathcal{L}} \mathcal{W}(t),$$

where, for all $t \in [0, 1]$,

$$X_n(t) = \frac{1}{\sigma \sqrt{n}} \sum_{i=1}^{\lfloor nt \rfloor} \xi_i + (nt - \lfloor nt \rfloor) \frac{1}{\sigma \sqrt{n}} \xi_{\lfloor nt \rfloor + 1}.$$

We can relax the assumption of the identical distribution and limit our attention to zero mean random variables not having the same distribution, and specif-

ically having different variances, we can still apply a functional central limit theorem described by Helland (1982). The main idea here is to view the sequences of random variables as martingale differences with respect to a particular family of σ -algebras. Unlike Donsker we do not use linear interpolation. In practice of course linear interpolation does not add on any essential restriction and, important properties such as continuity, follow very easily.

Theorem C.4. (Helland, 1982). *Let $\xi_{j,n}$ be a random variable defined on the probability space (Ω, \mathcal{F}, P) , for $j = 1, \dots, n$. Let $\mathcal{F}_{j,n}$ be a σ -algebra such that $\xi_{j,n}$ is $\mathcal{F}_{j,n}$ -measurable and $\mathcal{F}_{j-1,n} \subset \mathcal{F}_{j,n} \subset \mathcal{F}$ for all $j = 2, \dots, n$. Let r_n be a function defined on \mathbb{R}^+ such that $r_n(t)$ is a stopping time with respect to $\mathcal{F}_{j,n}, j = 1, \dots, n$. We suppose that the paths r_n are right continuous and increasing with $r_n(0) = 0$. Note that,*

$$X_n(t) = \sum_{j=1}^{r_n(t)} \xi_{j,n}. \quad (\text{C.2})$$

Let f be a positive measurable function such that $\int_0^t f^2(s)ds < \infty, \forall t > 0$. When the following conditions are verified:

a. $\xi_{j,n}$ is a difference of martingales, i.e.,

$$E(\xi_{j,n} | \mathcal{F}_{j-1,n}) = 0, \quad j = 1, \dots, n,$$

$$\text{b. } \sum_{j=1}^{r_n(t)} E(\xi_{j,n}^2 | \mathcal{F}_{j-1,n}) \xrightarrow[n \rightarrow \infty]{P} \int_0^t f^2(s)ds,$$

$$\text{c. } \sum_{j=1}^{r_n(t)} E(\xi_{j,n}^2 I(|\xi_{j,n}| > \varepsilon) | \mathcal{F}_{j-1,n}) \xrightarrow[n \rightarrow \infty]{P} 0, \quad \forall \varepsilon > 0,$$

Then,

$$X_n \xrightarrow[n \rightarrow \infty]{\mathcal{L}} f\mathcal{W},$$

where $f\mathcal{W}(t) = \int_0^t f(s)d\mathcal{W}(s)$ for $t \in \mathbb{R}^+$.

When f is constant and equal to one, the process X_n converges in distribution toward standard Brownian motion. Condition (c) follows from the condition of Lyapunov, denoted (c') such that

$$\exists \delta > 0, \quad \sum_{j=1}^{r_n(t)} E(\xi_{j,n}^{2+\delta} | \mathcal{F}_{j-1,n}) \xrightarrow[n \rightarrow \infty]{P} 0. \quad (\text{C.3})$$

These somewhat elementary results provide the framework in which we can readily anticipate the large sample behavior of the processes of interest to us. Such behavior we outline specifically at the relevant place via key theorems.

Theorem C.5. (Helland, 1982). Let $n, m \in \mathbb{N}^*$. Let $\xi_{j,n}^{(l)}$ be a random variable defined on the probability space (Ω, \mathcal{F}, P) , for $j = 1, \dots, n$ and $l = 1, \dots, m$. We assume that the sets $\{\xi_{j,n}^{(l)}, j = 1, \dots, n, n = 1, 2, \dots\}$ are tables of martingale differences with respect to the increasing sequence of σ -algebras $(\mathcal{F}_{j,n})_{j=1,2,\dots,n}$, for $l = 1, \dots, m$. Let r_n be a function defined on \mathbb{R}^+ such that $r_n(t)$ is a stopping time with respect to $\mathcal{F}_{j,n}, j = 1, \dots, n$. We suppose also that the paths r_n are right continuous and increasing with $r_n(0) = 0$. We have:

$$X_n(t) = \left(X_n^{(1)}(t), \dots, X_n^{(m)}(t) \right), \quad X_n^{(l)}(t) = \sum_{j=1}^{r_n(t)} \xi_{j,n}^{(l)}, \quad l = 1, \dots, m. \quad (\text{C.4})$$

Let f_1, \dots, f_m, m be positive measurable functions such that $\int_0^t f_l^2(s)ds < \infty, \forall t > 0$, for $l = 1, \dots, m$. When the following conditions hold the, for all $i, l = 1, \dots, m$:

- a. $\sum_{j=1}^{r_n(t)} E \left(\xi_{j,n}^{(l)} \xi_{j,n}^{(i)} \middle| \mathcal{F}_{j-1,n} \right) \xrightarrow[n \rightarrow \infty]{P} 0$, for $t > 0$ and $l \neq i$.
- b. $\sum_{j=1}^{r_n(t)} E \left(\left(\xi_{j,n}^{(l)} \right)^2 \middle| \mathcal{F}_{j-1,n} \right) \xrightarrow[n \rightarrow \infty]{P} \int_0^t f_l^2(s)ds,$
- c. $\sum_{j=1}^{r_n(t)} E \left(\left(\xi_{j,n}^{(l)} \right)^2 I(|\xi_{j,n}^{(l)}| > \varepsilon) \middle| \mathcal{F}_{j-1,n} \right) \xrightarrow[n \rightarrow \infty]{P} 0$, $\forall \varepsilon > 0$,

Then, $X_n \xrightarrow[n \rightarrow \infty]{\mathcal{L}} (\int f_1 d\mathcal{W}_1, \dots, \int f_m d\mathcal{W}_m)$ with respect to the product topology of Skorokhod where $\int_0^t f_l d\mathcal{W}_l = \int_0^t f_l(s) d\mathcal{W}_l(s)$ for $t \in \mathbb{R}^+, l = 1, \dots, m$ and $\mathcal{W}_1, \dots, \mathcal{W}_m$ are m independent Brownian motions.

C.5 Empirical distribution function

The above results can be directly applied to the sample empirical distribution function $F_n(t)$, defined for a sample of size n (uncensored) to be the number of observations less than or equal to t divided by n , i.e., $F_n(t) = n^{-1} \sum_{i=1}^n I(T_i \leq t)$. For each t , and we may assume $F(t)$ to be a continuous function of t , we would hope that $F_n(t)$ converges to $F(t)$ in probability. This is easy to see but, in fact, we have stronger results, starting with the Glivenko-Cantelli theorem whereby:

$$D_n = \sup_{0 \leq t \leq \infty} |F_n(t) - F(t)| = \sup_{0 \leq t \leq \infty} |S_n(t) - S(t)|$$

converges to zero with probability one and where $S_n(t) = 1 - F_n(t)$. This is analogous to the law of large numbers, and although important, is not all that informative. A central limit theorem can tell us much more and this obtains

by noticing how $F_n(t)$ will simulate a process relating to the Brownian bridge. To see this it suffices to note that $nF_n(t)$, for each value of t , is a sum of independent Bernoulli variables. Therefore, for each t as $n \rightarrow \infty$, we have that $\sqrt{n}\{F_n(t) - F(t)\}$ converges to normal with mean zero and variance $F(t)\{1 - F(t)\}$. We have marginal normality. However, we can claim conditional normality in the same way, since, for each t and s ($s < t$), $nF_n(t)$ given $nF_n(s)$, is also a sum of independent Bernoulli variables. Take k_1 and k_2 to be integers ($1 < k_1 < k_2 < n$) such that k_1/n is the nearest rational smaller than or equal to s (i.e., $k_1 = \max j; j \in \{1, \dots, n\}, j \leq ns$) and k_2/n is the nearest rational smaller than or equal to t . Thus, k_1/n converges with probability one to s , k_2/n to t , and $k_2 - k_1$ increases without bound at the same rate as n . We then have:

Theorem C.6. $\sqrt{n}\{F_n(t) - F(t)\}$ is a Gaussian process with mean zero and covariance given by:

$$\text{Cov}[\sqrt{n}\{F_n(s)\}, \sqrt{n}\{F_n(t)\}] = F(s)\{1 - F(t)\}. \quad (\text{C.5})$$

It follows immediately that, for T uniform, the process $\sqrt{n}\{F_n(t) - t\}$ ($0 \leq t \leq 1$), converges in distribution to the Brownian bridge. But note that, whatever the distribution of T , as long as it has a continuous distribution function, monotonic increasing transformations on T leave the distribution of $\sqrt{n}\{F_n(t) - F(t)\}$ unaltered. This means that we can use the Brownian bridge for inference quite generally. In particular, consider results of the Brownian bridge, such as the distribution of the supremum over the interval $(0, 1)$, that do not involve any particular value of t (and thereby $F(t)$). These results can be applied without modification to the process $\sqrt{n}\{F_n(t) - F(t)\}$ whether or not $F(t)$ is uniform. Among other useful results concerning the empirical distribution function we have:

Law of iterated logarithm

This law tells us something about the extreme deviations of the process. The following theorem (Serfling, 2009) provides the rate with n at which the largest absolute discrepancy between $F_n(t)$ and $F(t)$ is tending to zero.

Theorem C.7. With probability one,

$$\lim_{n \rightarrow \infty} \frac{\sqrt{n}D_n}{(2\log\log n)^{\frac{1}{2}}} = \sup\{F(t)(1 - F(t))\}^{\frac{1}{2}} \quad (\text{C.6})$$

For the most common case in which $F(t)$ is continuous, we know that $\sup F(t)(1 - F(t))$ is equal to 0.5. As an illustration, for 50 subjects, we find that D_n is around 0.12. For 50 i.i.d. observations coming from some known or some hypothesized distribution, if the hypothesis is correct then we expect to see the greatest discrepancy between the empirical and the hypothesized distribution to be close to 0.12. Values far removed from that might then be indicative of either a rare event or that the assumed distribution is not correct. Other quantities,

indicating how close to $F(t)$ we can anticipate $F_n(t)$ to be, are of interest, one in particular being;

$$C_n = n \int_0^\infty \{F_n(t) - F(t)\}^2 f(t) dt.$$

As for D_n the asymptotic distribution of C_n does not depend upon $F(t)$. For this case the law of the iterated logarithm is expressed as follows:

Theorem C.8. *With probability one,*

$$\lim_{n \rightarrow \infty} \frac{C_n}{(2 \log \log n)^{\frac{1}{2}}} = \frac{1}{\pi^2}. \quad (\text{C.7})$$

For D_n , inference can be based on the maximum of a Brownian bridge. In the case of C_n inference is less straightforward and is based on the following lemma;

Lemma C.1. *Letting $\eta = \sum_{j=1}^{\infty} \chi_j^2 (\pi j)^{-2}$ where the χ_j^2 are independent chi-square variates then*

$$\lim_{n \rightarrow \infty} P(C_n \leq c) = P(\eta \leq c). \quad (\text{C.8})$$

The results for both D_n and C_n are large sample ones but can nonetheless provide guidance when dealing with actual finite samples. Under assumed models it is usually possible to calculate the theoretical distribution of some quantity which can also be observed. We are then able to contrast the two and test the plausibility of given hypotheses.

Appendix D

Inferential tools

D.1 Theory of estimating equations

Most researchers, together with a large section of the general public, even if uncertain as to what the study of statistics entails, will be familiar with the concept, if not the expression itself, of the type $\bar{T} = n^{-1} \sum_{i=1}^n T_i$. The statistician may formulate this in somewhat more abstract terms, stating that; $\bar{T} = n^{-1} \sum_{i=1}^n T_i$ is a solution to the linear estimating equation for the parameter μ , the population mean of the random variable T , in terms of the n i.i.d. replicates of T . The estimating equation is simply $\mu - n^{-1} \sum_{i=1}^n T_i = 0$. This basic idea is very useful in view of the potential for immediate generalization.

The most useful approach to analyzing data is to postulate plausible models that may approximate some unknown, most likely very complex, mechanism generating the observations. These models involve unknown parameters and we use the observations, in conjunction with an estimating equation, to replace the unknown parameters by estimates. Deriving "good" estimating equations is a sizeable topic whose surface we only need to scratch here. We appeal to some general principles, the most common of which are very briefly recalled below, and note that, unfortunately, the nice simple form for the estimating equation for μ just above is more an exception than the rule. Estimating equations are mostly nonlinear and need to be solved by numerical algorithms. Nonetheless, an understanding of the linear case is more helpful than it may at first appear since solutions to the nonlinear case are achieved by local linearization (called also Newton-Raphson approximation) in the neighborhood of the solution. A fundamental result in the theory of estimation is described in the following theorem. Firstly, we define two important functions, $L(\theta)$ and $I(\theta)$ of the parameter θ by

$$L(\theta) = f(t_1, t_2, \dots, t_n; \theta); \quad I(\theta) = -\partial^2 \log L(\theta) / \partial \theta^2. \quad (\text{D.1})$$

We refer to $L(\theta)$ as the observed likelihood, or simply just the likelihood (note that, for $n = 1$, the expected log-likelihood is the negative of the entropy, also called the information). When the observations T_i , $i = 1, \dots, n$, are independent and identically distributed then we can write $L(\theta) = \prod_{i=1}^n f(t_i; \theta)$ and $\log L(\theta) = \sum_{i=1}^n \log f(t_i; \theta)$. We refer to $I(\theta)$ as the information in the sample. Unfortunately the negative of the entropy is also called the information (the two are of course related, both quantifying precision in some sense). The risks of confusion are small given that the contexts are usually distinct. The function $I(\theta)$ is random because it depends on the data and reaches a maximum in the neighborhood of θ_0 since this is where the slope of the log-likelihood is changing the most quickly.

Theorem D.1. *For a statistic T we can write the following;*

$$\text{Var}(T) \geq \{\partial E(T)/\partial\theta\}^2/E\{I(\theta)\}.$$

This inequality is called the Cramer-Rao inequality (Cox and Hinkley, 1979). When T is an unbiased estimate of θ then $\partial E(T)/\partial\theta = 1$ and $\text{Var}(T) \geq 1/E\{I(\theta)\}$. The quantity $1/E\{I(\theta)\}$ is called the Cramer-Rao bound. Taking the variance as a measure of preciseness then, given unbiasedness, we prefer the estimator T that has the smallest variance. The Cramer-Rao bound provides the best that we can do in this sense, and below, we see that the maximum likelihood estimator achieves this for large samples, i.e., the variance of the maximum likelihood estimator becomes progressively closer to the bound as sample size increases without limit.

Basic equations

For a scalar parameter θ_0 we will take some function $U(\theta)$ that depends on the observations as well as θ . We then use $U(\theta)$ to obtain an estimate $\hat{\theta}$ of θ_0 via an estimating equation of the form

$$U(\hat{\theta}) = 0. \tag{D.2}$$

This is too general to be of use and so we limit the class of possible choices of $U(\cdot)$. We require the first two moments of U to exist in which case, without loss of generality we can say

$$E U(\theta_0) = 0, \quad \text{Var } U(\theta_0) = \sigma^2. \tag{D.3}$$

Two widely used methods for constructing $U(\theta)$ are described below. It is quite common that U be expressible as a sum of independent and identically distributed contributions, U_i , each having a finite second moment. An immediate application of the central limit theorem then provides the large sample normality for $U(\theta_0)$. For independent but nonidentically distributed U_i , it is still usually not

difficult to verify the Lindeberg condition and apply the central limit theorem for independent sums. Finally, in order for inference for U to carry over to $\hat{\theta}$, some further weak restrictions on U will be all we need. These require that U be monotone and continuous in θ and differentiable in some neighborhood of θ_0 . This is less restrictive than it sounds. In practice it means that we can simply apply the mean value theorem (A.2) whereby:

Corollary D.1. *For any $\epsilon > 0$, when $\hat{\theta}$ lies in an interval $(\theta_0 - \epsilon, \theta_0 + \epsilon)$ within which $U(\theta)$ is continuously differentiable, then there exists a real number $\xi \in (\theta_0 - \epsilon, \theta_0 + \epsilon)$ such that*

$$U(\hat{\theta}) = U(\theta_0) - (\hat{\theta} - \theta_0)I(\xi).$$

This expression is useful for the following reasons. A likelihood for θ will, with increasing sample size, look more and more normal. As a consequence, $I(\xi)$ will look more and more like a constant, depending only on sample size, and not ξ itself. This is useful since ξ is unknown. We approximate $I(\xi)$ by $I(\hat{\theta})$. We can then express $\hat{\theta}$ in terms of approximate constants and $U(\hat{\theta})$ whose distribution we can approximate by a normal distribution.

Finding equations

The guiding principle is always the same, that of replacing unknown parameters by values that minimize the distance between empirical (observed) quantities and their theoretical (model-based) equivalents. The large range of potential choices stem from two central observations: (1) there can be many different definitions of distance (indeed, the concept of distance is typically made wider than the usual mathematical one which stipulates that the distance between a and b must be the same as that between b and a) and (2) there may be a number of competing empirical and theoretical quantities to consider. To make this more concrete, consider a particular situation in which the mean is modeled by some parameter θ such that $E_\theta(T)$ is monotone in θ . Let's say that the true mean $E(T)$ corresponds to the value $\theta = \theta_0$. Then the mean squared error, variance about a hypothesized $E_\theta(T)$, let's say $\sigma^2(\theta)$, can be written as

$$\sigma^2(\theta) = E\{T - E_\theta(T)\}^2 = E\{T - E_\theta(T)\}^2 + \{E_\theta(T) - E_{\theta_0}(T)\}^2.$$

The value of θ that minimizes this expression is clearly θ_0 . An estimating equation derives from minimizing the empirical equivalent of $\sigma^2(\theta)$. Minimum chi-squared estimates have a similar motivation. The idea is to bring, in some sense, via our choice of parameter value, the hypothesized model as close as possible to the data. Were we to index the distribution F by this parameter, calling this say $F_\theta(t)$, we could re-express the definition for $D_n(\theta)$ given earlier as

$$D_n(\theta) = \sup_{0 \leq t \leq \infty} |F_n(t) - F_\theta(t)|.$$

Minimizing $D_n(\theta)$ with respect to θ will often provide a good, although not necessarily very tractable, estimating equation. The same will apply to C_n and related expressions such as the Anderson-Darling statistic. We will see later that the so-called partial likelihood estimate for the proportional hazards model can be viewed as an estimate arising from an empirical process. It can also be seen as a method of moments estimate and closely relates to the maximum likelihood estimate. Indeed, these latter two methods of obtaining estimating equations are those most commonly used, and in particular, the ones given the closest attention in this work. It is quite common for different techniques, and even contending approaches from within the same technique, to lead to different estimators. It is not always easy to argue in favor of one over the others.

Other principles can sometimes provide guidance in practice, the principle of efficiency holding a strong place in this regard. The idea of efficiency is to minimize the sample size required to achieve any given precision, or equivalently, to find estimators having the smallest variance. However, since we are almost always in situations where our models are only approximately correct, and on occasion, even quite far off, it is more useful to focus attention on other qualities of an estimator. How can it be interpreted when the data are generated by a mechanism much wider than that assumed by the model? How useful is it to us in our endeavor to build predictive models, even when the model is, at least to some extent, incorrectly specified. This is the reality of modeling data and efficiency, as an issue for us to be concerned with, does not take us very far. On the other hand, estimators that have demonstrably poor efficiency, when model assumptions are correct, are unlikely to redeem themselves in a broader context and so it would be a mistake to dismiss efficiency considerations altogether even though they are rather limited.

Method of moments

This very simple method derives immediately as an application of the Helly-Bray theorem (Theorem A.3). The idea is to equate population moments to empirical ones obtained from the observed sample. Given that $\mu = \int x dF(x)$, the above example is a special case since, we can write $\bar{\mu} = \bar{x} = \int x dF_n(x)$. Properties of the estimate can be deduced from the well-known properties of $F_n(x)$ as an estimate of $F(x)$ (see Section C.5). For the broad exponential class of distributions, the method of moments estimator, based on the first moment, coincides with the maximum likelihood estimator recalled below. In the survival context we will see that the so-called partial likelihood estimator can be viewed as a method of moments estimator. The main difficulty with method of moments estimators is that they are not uniquely defined for any given problem. For example, suppose we wish to aim to estimate the rate parameter λ from a series of observations, assumed to have been generated by a Poisson distribution. We can use either the empirical mean or the empirical variance as an estimate for λ . Typically they will

not be the same. Indeed we can construct an infinite class of potential estimators as linear combinations of the two.

Maximum likelihood estimation

A minimum chi-square estimator for θ derives by minimizing an expression of variance. It can also appear equally natural to minimize an estimate of the entropy as a function of θ , i.e., maximize the observed information as a function of θ . We write the information as $V(\theta)$ where

$$V(\theta) = E \log f(T; \theta).$$

Given the observations, T_1, \dots, T_n , we replace the unknown function $V(\theta)$ by $\bar{V}(\theta) = n^{-1} \sum_{i=1}^n \log f(T_i; \theta)$. The maximization is easily accomplished when the parameter or parameter vector θ can only assume some finite number of discrete values. It is then sufficient to examine all the cases and select θ such that $\bar{V}(\theta)$ is maximized. For all the models under consideration here we can assume that $V(\theta)$ and $\bar{V}(\theta)$ are continuous smooth functions of θ . By smooth we mean that the first two derivatives, at least, exist for all values of θ . This is not at all a restrictive assumption and models that do not have such differentiability properties can nearly always be replaced by models that do via useful reparameterization. For instance, there are cases where a model, defined for all positive real θ , may break down at $\theta = 0$, the entropy not being differentiable at that point, whereas under the reparameterization $\theta = \exp(\alpha)$ for α defined over the whole real line, the problem disappears.

Two fundamental theorems and three corollaries enable us to appreciate the great utility of the maximum likelihood approach. All that we need are “suitable regularity conditions.” We return to these immediately below. Assuming these conditions (a valid assumption for all the models in this book), we have a number of important results concerning $V(\theta)$ and consistent estimates of $V(\theta)$.

Theorem D.2. *Viewed as a function of θ , $V(\theta)$ satisfies*

$$\left\{ \frac{\partial V(\theta)}{\partial \theta} \right\}_{\theta=\theta_0} = E \left\{ \frac{\partial \log f(T; \theta)}{\partial \theta} \right\}_{\theta=\theta_0} = 0. \quad (\text{D.4})$$

Note the switching of the operations, integration and differentiation, in the above equations. In many texts describing the likelihood method it is common to only focus on the second part of the equation. It helps understanding to also keep the first part of the equation in mind since this will enable us to establish the solid link between the information measure and likelihood. Having divided by sample size we should view the log-likelihood as an empirical estimate of $V(\theta)$. The law of large numbers alone would provide us with a convergence result but we can do better, in terms of fitting in with elementary results for estimating

equations, by assuming some smoothness in $V(\theta)$ and a consistent estimator, $\bar{V}(\theta)$, as functions of θ . More precisely:

Corollary D.2. *Let $\bar{V}(\theta)$ be a consistent estimate of $V(\theta)$. Then $\{\partial\bar{V}(\theta)/\partial\theta\}_{\theta=\theta_0}$ converges, with probability one, to zero.*

The expression “suitable regularity conditions” is not very transparent, and for those lacking a good grounding in analysis, or simply a bit rusty, it might be less than illuminating. We dwell on it for a moment since it appears frequently in the literature. We require continuity of the function $V(\theta)$, at least for general situations, in order to be able to claim that as $V(\theta)$ approaches $V(\theta_0)$ then θ approaches θ_0 . This is nearly enough, although not quite. Differentiability is a stronger requirement since we can see that a function that is differentiable at some point must also be continuous at that same point. The converse is not so and can be seen immediately in a simple example, $y = |x|$, a function that is continuous everywhere but not differentiable at the origin. We need the differentiability condition in order to obtain the estimating equation and just a tiny bit more, the tiny bit more not being easily described but amounting to authorizing the switching of the processes of integration and differentiation. Such switching has to take place in order to be able to demonstrate the validity of the above theorem, and the one just below. All of this is summarized by the expression “suitable regularity conditions” and the reader need not worry about them since they will hold in all the practical cases of interest to us. The main result follows as a further corollary:

Corollary D.3. *Suppose that $U(\alpha) = \{\partial\bar{V}(\theta)/\partial\theta\}_{\theta=\alpha}$ and that, for sample size n , $\hat{\theta}_n$ is the solution to the equation $U(\theta) = 0$. Then, if $\bar{V}(\theta)$ is consistent for $V(\theta)$, $\hat{\theta}_n$ converges with probability one to θ_0 .*

This is almost obvious, and certainly very intuitive, but it provides a solid foundation to likelihood theory. The result is a strong and useful one, requiring only the so-called regularity conditions referred to above and that $\bar{V}(\theta)$ be consistent for $V(\theta)$. Independence of the observations or that the observations arise from random sampling is not appealed to or needed. However, when we do have independent and identically distributed observations then, not only do our operations become very straightforward, we also can appeal readily to central limit theorems, initially for the left-hand side of the estimating equation, and then, by extension, to a smooth function of the estimating equation. The smooth function of interest is of, course, $\hat{\theta}_n$ itself.

Corollary D.4. *Suppose that T_1, \dots, T_n are independent identically distributed random variables having density $f(t; \theta_0)$. Then $\hat{\theta}_n$ converges with probability one to θ_0 where $\hat{\theta}_n$ is such that $U(\hat{\theta}_n) = 0$, where $U(\theta) = \sum_{i=1}^n U_i(\theta)$ and where $U_i(\alpha) = \{\partial \log f(T_i; \theta)/\partial\theta\}_{\theta=\alpha}$.*

We can say much more about $\hat{\theta}_n$. A central limit theorem result applies immediately to $U(\theta_0)$, and via Slutsky’s theorem, we can then also claim large sample

normality for $U(\hat{\theta}_n)$. By expressing $\hat{\theta}_n$ as a smooth (not necessarily explicit) function of U we can also then claim large sample normality for $\hat{\theta}_n$. The fact that $U(\hat{\theta}_n)$ (having subtracted off the mean and divided by its standard deviation) will converge in distribution to a standard normal and that a smooth function of this, notably θ_n , will do the same, does not mean that their behavior can be considered to be equivalent. The result is a large sample one, i.e., as n tends to infinity, a concept that is not so easy to grasp, and for finite samples, behavior will differ. Since U is a linear sum we may expect the large sample approximation to be more accurate, more quickly, than for θ_n itself. Inference is often more accurate if we work directly with $U(\hat{\theta}_n)$, exploiting the monotonicity of $U(\cdot)$, and inverting intervals for $U(\hat{\theta}_n)$ into intervals for $\hat{\theta}_n$. In either case we need some expression for the variance and this can be obtained from the second important theorem:

Theorem D.3. *Viewed as a function of θ , $V(\theta)$ satisfies:*

$$\left(\frac{\partial V(\theta)}{\partial \theta} \right)_{\theta=\theta_0}^2 = \left(-\frac{\partial^2 V(\theta)}{\partial \theta^2} \right)_{\theta=\theta_0} = E \left(-\frac{\partial^2 \log f(T; \theta)}{\partial \theta^2} \right)_{\theta=\theta_0}. \quad (\text{D.5})$$

As in the previous theorem, note the switching of the operations of integration (expectation) and differentiation. Since $E U(\theta_0) = E \{\partial \log f(T; \theta) / \partial \theta\}_{\theta=\theta_0} = 0$, then, from the above theorem, $\text{Var} U(\theta_0) = E \{\partial^2 \log f(T; \theta) / \partial \theta^2\}_{\theta=\theta_0}$. In practical applications we approximate the variance expression by replacing θ_0 by its maximum likelihood estimate. It is also interesting to note that the above inequality will usually break down when the model is incorrectly specified and that, in some sense, the further away is the assumed model from that which actually generates the data, then the greater the discrepancy between the two quantities will be. This idea can be exploited to construct goodness-of-fit tests or to construct more robust estimators.

Accurate inference for linear estimating equations

For the case described above where T_1, \dots, T_n are independent identically distributed random variables having density $f(t; \theta_0)$, we write the estimating equation in terms of $U(T_1, \dots, T_n : \theta_0)$ having solution $\hat{\theta}_n$ where $U(T_1, \dots, T_n : \hat{\theta}_n) = 0$. We already know that $\hat{\theta}_n$ converges with probability one to θ_0 . In the linear case we write: $U(\theta) = \sum_{i=1}^n U_i(\theta) = \sum_{i=1}^n U_{(i)}(\theta)$ where the parenthesized subscript denotes order statistics (Appendix A.9) and where $U_i(\alpha) = \{\partial \log f(T_i; \theta) / \partial \theta\}_{\theta=\alpha}$. The large sample theory applies but consider also the estimate $\hat{\theta}_n^w$, obtained from the weighted estimating equation, $\sum_{i=1}^n W_i U_{(i)}(\hat{\theta}_n^w) = 0$, where for all i , $W_i > 0$. In general, the large sample properties of $\hat{\theta}_n^w$ and $\hat{\theta}_n$ are the same but can differ for finite samples. Naturally, some conditions are needed, specifically that $\text{Cov}(W_i, U_i) = 0$ and with some restrictions on the relative sizes of the W_i . Standardization will solve most problems, usually

achieved via the constraint, $\sum_{i=1}^n W_i = 1$. We might wish to make the condition, $\text{Cov}(W_i, U_i) = 0$, a large sample rather than a finite sample result and that would also work providing the rate of convergence is higher than \sqrt{n} . Suppose that $X_i, i = 1, \dots, n$, are standard exponential variates and we define W_i by: $W_i = X_i / \{\sum_{j=1}^n X_j\}$.

Rubin (1981) and, more recently, Xu and Meeker (2020) studied inference based on such weighted estimating equations. Inference is fully efficient and particularly easy to carry out. We can view inference as being conditional upon the data, i.e., the observations $T_1 \dots T_n$. Rather than consider theoretical replications of $U_1 \dots U_n$, we treat these as being fixed. In particular, we can write down the empirical distribution for the ordered $U_{(1)} \dots U_{(n)}$. If we then make an appeal to the probability integral transform (Appendix A.6), we see that the W_i , the gaps between the cumulative observations, i.e., the empirical distribution function, are exactly distributed as described above. This is a consequence of the fact that the gaps between uniform order statistics are distributed as: $W_i = X_i / \{\sum_{j=1}^n X_j\}$ (Appendix A.9). Large sample inference will agree to a very high degree with inference based on normal approximations, and for finite samples, we can often obtain more precise and more robust results.

Three further advantages of such an approach to inference for linear estimating equations are worthy of mention. The first is that we can either do any calculations analytically, given the particularly simple form of the exponential distribution, or via simulated replications which, again, are very simple to set up. For small samples where the bootstrap may fail as a result of unboundedness of the estimator itself, use of the continuously weighted estimating equation will avoid such difficulties. The bootstrap for survival data is described below and it is not difficult to see how the non-zero chance of obtaining samples with infinite parameter estimates can cause problems. The wild bootstrap is an alternative way around this problem but we do not investigate that here.

Finally, when a parameter is defined in terms of large sample results for an estimating equation, we are able to interpret what is being estimated when our model assumptions are not correct. But, if our estimating equations are linear, we can do more and we can use these weighted linear estimating equations to make valid inferences even when the data are generated by a model that is different from that assumed. This is of great value in cases where we assume some narrow model, e.g., proportional hazards, but the mechanism generating the observations is a broader one.

D.2 Efficiency in estimation and in tests

Relative efficiency for consistent estimators

Intuitively we would prefer the estimator $\hat{\theta}$ to $\tilde{\theta}$ of the parameter θ if, on average, it was the closer of the two. The word “closer” gives us the clue that this is not

going to be straightforward since there is a very large number of possibilities for distance that we could choose to work with. Suppose we choose to work with Euclidean distance. Assuming θ to be fixed, a reasonable guide could be provided by $E(\hat{\theta} - \theta)^2$, the mean squared error. This would depend on both the bias and the variance of the distribution of $\hat{\theta}$. In practice we will mostly work with estimators that are consistent or asymptotically unbiased so that our focus can then be limited to variances of the distributions for $\hat{\theta}$ and $\tilde{\theta}$. Recalling the expressions for the likelihood, $L(\theta)$, and information $I(\theta)$ of the parameter θ by

$$L(\theta) = f(t_1, t_2, \dots, t_n; \theta); \quad I(\theta) = -\partial^2 \log L(\theta) / \partial \theta^2,$$

for a statistic T we can write the following:

$$\text{Var}(T) \geq \{\partial E(T)/\partial \theta\}^2 / E\{I(\theta)\}.$$

which is the Cramer-Rao inequality (Cox and Hinkley, 1979). When T is an unbiased estimate of θ then $\partial E(T)/\partial \theta = 1$ and $\text{Var}(T) \geq 1/E\{I(\theta)\}$. The quantity $1/E\{I(\theta)\}$ is called the Cramer-Rao bound and it provides the best that we can do. This bound divided by the variance of any alternative unbiased estimator measures the relative efficiency of the estimator. This would depend on sample size n and, when this ratio achieves a limit as n increases without bound then we call this the asymptotic relative efficiency. Since the ratio of the variance of the maximum likelihood estimator and this bound tends to one with increasing n we can see that the maximum likelihood estimator is fully efficient.

Relative efficiency for consistent tests

Against local alternatives, indeed against any alternatives, the power of a consistent test tends to 100% as sample size increases without bound. In order to weigh up the relative advantages of one consistent test over another, in terms of power, it is common to investigate relative efficiency at some different, plausible, sample sizes. Typically simulations are used, and assuming some smoothness with sample size, we are able to obtain a good sense of the advantages or disadvantages of different consistent tests. Asymptotic efficiency is, again, a limiting behavior although somewhat more complex than for estimation in that not only the parameter stipulated by the test, say $\theta = \theta_0$ comes into play but also the size, α and power, β , need be taken into account. For two consistent tests, T_1 and T_2 , and for given fixed α and β , we require n_1 and n_2 subjects. The relative efficiency, $e(T_1, T_2)$ is defined to be

$$e(T_1, T_2) = n_1/n_2.$$

In this expression, if we know which is the more powerful test, then this would correspond to n_1 so that this ratio lets us know how much extra work is needed, in terms of sample size, if we prefer to use test T_2 . If, for α and β fixed, for

a sequence of local alternatives to the null hypothesis, $H_0 : \theta = \theta_0$, where θ approaches θ_0 at rate \sqrt{n} , the ratio n_1/n_2 approaches a limit, then we call this limit, $e(T_1, T_2)$ the Pitman relative efficiency of the tests. Two other measures of asymptotic relative efficiency are commonly used, the Hodges-Lehmann and the Bahadur measures. For the Hodges-Lehmann measure, we fix β and θ and consider the limit as $\alpha \rightarrow 0$. For the Bahadur measure, we fix α and θ and consider the limit as $\beta \rightarrow 1$. Note further (Van Eeden, 1963) that, under broad conditions, for Pitman asymptotic relative efficiency, $e(T_1, T_2) = \rho^2(T_1, T_2)$ where $\rho^2(T_1, T_2)$ is the limiting correlation between T_1 and T_2 .

D.3 Inference using resampling techniques

Bootstrap resampling

The purpose of bootstrap resampling is twofold: (1) to obtain more accurate inference, in particular more accurate confidence intervals, than is available via the usual normal approximation, and (2) to facilitate inference for parameter estimators in complex situations. A broad discussion including several challenging applications is provided by Politis (1998). Here we will describe the basic ideas in so far as they are used for most problems arising in survival analysis. Consider the empirical distribution function $F_n(t)$ as an estimate for the unknown distribution function $F(t)$. The observations are T_1, T_2, \dots, T_n . A parameter of interest, such as the mean, the median, some percentile, let's say θ , depends only on F . This dependence can be made more explicit by writing $\theta = \theta(F)$. The core idea of the bootstrap can be summarized via the simple expression $\tilde{\theta} = \theta(F_n)$ as an estimator for $\theta(F)$.

Taking infinitely many i.i.d. samples, each of size n , from F would provide us with the exact sampling properties of any estimator $\tilde{\theta} = \theta(F_n)$. If, instead of taking infinitely many samples, we were to take a very large number, say B , of samples, each sample again of size n from F , then this would provide us with accurate approximations to the sampling properties of $\tilde{\theta}$, the errors of the approximations diminishing to zero as B becomes infinitely large. Since F is not known we are unable to carry out such a prescription. However, we do have available our best possible estimator of $F(t)$, the empirical distribution function $F_n(t)$. The bootstrap idea is to sample from $F_n(t)$, which is known and available, instead of from $F(t)$ which, apart from theoretical investigations, is typically unknown and unavailable.

Empirical bootstrap distribution

The conceptual viewpoint of the bootstrap is to condition on the observed T_1, \dots, T_n and its associated empirical cumulative distribution function $F_n(t)$, thereafter treating these quantities as though they were a population of interest, rather than a sample. From this “population” we can draw samples with replacement, each

sample having size n . We repeat this whole process B times where B is a large number, typically in the thousands. Each sample is viewed as an i.i.d. sample from $F_n(t)$. The i th resample of size n can be written $T_{1i}^*, T_{2i}^*, \dots, T_{ni}^*$ and has empirical distribution $F_n^{*i}(t)$. For any parameter of interest θ , the mean, median coefficient of variation for example, it is helpful to remind ourselves of the several quantities of interest, $\theta(F)$, $\theta(F_n)$, $\theta(F_n^{*i})$ and $F_B(\theta)$, the significance of each of these quantities needing a little explanation. First, $\theta(F)$ is simply the population quantity of interest. Second, $\theta(F_n)$ is this same quantity defined with respect to the empirical distribution of the data T_1, \dots, T_n . Third, $\theta(F_n^{*i})$ is again the same quantity defined with respect to the i th empirical distribution of the resamples $T_{1i}^*, T_{2i}^*, \dots, T_{ni}^*$. Finally, $F_B(\theta)$ is the bootstrap distribution of $\theta(F_n^{*i})$, i.e., the empirical distribution of $\theta(F_n^{*i})$ ($i = 1, \dots, B$).

To keep track of our asymptotic thinking we might note that, as $B \rightarrow \infty$, $\int u dF_B(u)$ converges in probability to $\theta(F_n)$ and, as $n \rightarrow \infty$, $\theta(F_n)$ converges in probability to $\theta(F)$. Thus, there is an important conceptual distinction between F_B and the other distribution functions. These latter concern the distribution of the original observations or resamples of these observations. F_B itself deals with the distribution of $\theta(F_n^{*i})$ ($i = 1, \dots, n$) and therefore, when our focus of interest changes from one parameter to another, from say θ_1 to θ_2 the function F_B will be generally quite different. This is not the case for F , F_n , and F_n^{*i} which are not affected by the particular parameter we are considering. Empirical quantities with respect to the bootstrap distribution, F_B are evaluated in a way entirely analogous to those evaluated with respect to F_n . For example,

$$\text{Var}\{\theta(F_n)\} = \sigma_B^2 = \int u^2 dF_B(u) - \left(\int u dF_B(u) \right)^2, \quad (\text{D.6})$$

where it is understood that the variance operator, $\text{Var}()$ is with respect to the distribution $F_B(t)$. Of greater interest in practice is the fact that $\text{Var}\{\theta(F_n)\}$, where $\text{Var}()$ is with respect to the distribution $F_B(t)$, can be used as an estimator of $\text{Var}\{\theta(F_n)\}$, where $\text{Var}(\cdot)$ is with respect to the distribution $F(t)$.

Bootstrap confidence intervals

For the normal distribution the standardized percentiles z_α are defined from $\Phi(z_\alpha) = \alpha$. Supposing that $\text{Var}\theta(F_n) = \sigma^2$, the variance operator being taken with respect to F , the scaled percentiles Q_α are then given by $Q_\alpha = \sigma z_\alpha$, leading to a normal approximation for a confidence interval for θ as

$$I_{1-\alpha}(\theta) = \{\theta(F_n) - Q_{1-\alpha/2}, \theta(F_n) + Q_{\alpha/2}\} \quad (\text{D.7})$$

which obtains from a rearrangement of the expression $\Pr[\sigma z_{\alpha/2} < \theta(F_n) - \theta < \sigma z_{1-\alpha/2}] = 1 - \alpha$. Since σ^2 is not generally available we would usually work with σ_B^2 . Instead of using the normal approximation it is possible to define Q_α

differently, directly from the observed bootstrap distribution F_B . We can define Q_α via the equation $F_B(Q_\alpha) = \alpha$. In view of the finiteness of B (and also n) this equation may not have an exact solution and we will, in practice, take the nearest point from $F_B^{-1}(\alpha)$ as Q_α . The values of $Q_{\alpha/2}$ and $Q_{1-\alpha/2}$ are then inserted into equation (D.7). Such intervals are referred to as bootstrap “root” intervals. The more common approach to constructing bootstrap confidence intervals is, however, slightly different and has something of a Bayesian or fiducial inference flavor to it. We simply tick off the percentiles, $Q_{\alpha/2}$ and $Q_{1-\alpha/2}$ and view the distribution F_B as our best estimate of a distribution for θ . The intervals are then written as

$$I_{1-\alpha}(\theta) = \{\theta(F_n) + Q_{\alpha/2}, \theta(F_n) + Q_{1-\alpha/2}\}. \quad (\text{D.8})$$

These intervals are called percentile bootstrap confidence intervals. Whenever the distribution F_B is symmetric then the root intervals and the percentile intervals coincide since $Q_{\alpha/2} + Q_{1-\alpha/2} = 0$. In particular, they coincide if we make a normal approximation.

Accuracy of bootstrap confidence intervals

Using theoretical tools for investigating statistical distributions, the Edgeworth expansion in particular, it can be shown that the accuracy of the three types of interval described above is the same. The argument in favor of the bootstrap is then not compelling, apart from the fact that they can be constructed in cases where variance estimates may not be readily available. However, it is possible to improve on the accuracy of both the root and the percentile intervals. One simple, albeit slightly laborious, way to accomplish this is to consider studentized methods. By these we mean, in essence, that the variance in the “population” F_n is not considered fixed and known for each subsample but is estimated several, precisely B , times across the B bootstrap samples.

To get an intuitive feel for this it helps to recall the simple standard set-up we work with in the comparison of two estimated means, \bar{X}_1 and \bar{X}_2 where, when the variance σ^2 is known we use as test statistic, $(\bar{X}_1 - \bar{X}_2)/\sigma$, referring the result to standard normal tables. When σ^2 is unknown and replaced by its usual empirical estimate s^2 then, for moderate to large samples, we do the same thing. However, for small samples, in order to improve on the accuracy of inference, we appeal to the known distribution of $(\bar{X}_1 - \bar{X}_2)/s$ when the original observations follow a normal distribution. This distribution was worked out by Student and is his so-called Student’s t distribution. For the i th bootstrap sample our estimate of the variance is

$$\text{Var}\{\theta(F_n^{*i})\} = \sigma_{*i}^2 = \int u^2 dF_n^{*i}(u) - \left(\int u dF_n^{*i}(u) \right)^2 \quad (\text{D.9})$$

and we then consider the standardized distribution of the quantity, $\theta(F_n^{*i})/\sigma_{*i}$. The essence of the studentized approach, having thus standardized, is to use the bootstrap sampling force to focus on the higher-order questions, those concerning bias and skewness in particular. Having, in some sense, spared our bootstrap resources from being dilapidated to an extent via estimation of the mean and variance, we can make real gains in accuracy when applying the resulting distributional results to the statistic of interest. For most day-to-day situations this is probably the best approach to take. It is computationally intensive (no longer a serious objection) but very simple conceptually. An alternative approach, with the same ultimate end in mind, is not to standardize the statistic but to make adjustments to the derived bootstrap distribution, the adjustments taking into account any bias and skewness. These lead to the so-called bias corrected, accelerated intervals (often written as BC_a intervals).

D.4 Conditional, marginal, and partial likelihood

A reasonably accurate description of a mathematical function, including multi-dimensional functions, would be that of a recipe. Throw in the ingredients (arguments) and out comes the result. The result is unique as long as we use the same ingredients. The likelihood comes under this heading and we refer to it as the likelihood function. It is a function of a model's parameters (arguments). The data involved in the likelihood expression are taken to be fixed. As is so often the case, we will have conditioned on (taken as fixed) some aspect of the observations and then considered the likelihood as a function of the parameters. The precise nature of the functional output will depend upon the probability measure we associate with the observations. For any given model, and set of observations, it is not common to give much thought to the type of likelihood on which to base inference. We tend to talk about the "likelihood of the data," as though it were clear and unambiguous.

While this is often the case we might remind ourselves that there may be many, even an infinite number of potential likelihoods we are at liberty to choose. Take, for example, the i.i.d. observations, $X_1 \dots X_n$. We may decide to neglect all contributions with an odd index reducing our sample size by 50%. Now, none of the large sample properties that underlie our motivation are in any way compromised. They work no less well with only half of the sample as with the full sample, at least in as far as asymptotic properties are concerned. Finite samples are, of course, another story and even though large sample properties are unaffected, none of us would have any difficulty in choosing whether to work with all of the observations or only half of them. The point is an important one even so, and in many cases, it may not be at all clear, given several competing likelihoods, which one will provide the more precise results. Deciding that some feature of the observations ought to be treated as though they were fixed and known will lead to a number of possibilities coming under the heading of conditional likelihood.

Empirical likelihood in which certain parameters are assumed fixed at some value will also come under this heading. We next describe an interesting approach to inference in which those aspects of the observations that impact the precision and not the location of the parameter of interest could be assumed fixed at the values observed in the data.

Conditional likelihood

A conditional likelihood, that we may like to view as a conditional density, simply fixes some parameters at certain values. An obvious example is that of a parameter that is a function of some variance, a quantity which indicates the precision in an estimate of other parameters but tells us nothing about where such parameters may be located. It has been often argued that it is appropriate to condition on such quantities. The variance parameter is then taken as fixed and known and the remaining likelihood is a conditional likelihood.

Cox (1958) gives an example of two instruments measuring the same quantity but having very different precision. His argument, which is entirely persuasive, says that we should use the conditional likelihood and not the full likelihood that involves the probabilities of having chosen one or the other instrument. The fact that we may have chosen the less precise instrument at a given rate of the time is neither here nor there. All that matters are the observations and the particular instruments from which they arise. Another compelling example arises in binomial sampling. It is not at all uncommon that we would not know in advance the exact number n of repetitions of the experiment. However, we rarely would think of working with a full likelihood in which the distribution of n is explicitly expressed. Typically, we would simply use the actual value of n that we observed. This amounts to working with a conditional likelihood.

Fisher (1934) derived an exact expression for the distribution of the maximum likelihood estimate for a location or scale parameter conditional on observed spread in the data. Fisher's expression is particularly simple and corresponds to the likelihood function itself standardized so that the integral with respect to the parameter over the whole of the parameter space is equal to one. It is quite an extraordinary result in its generality and the fact that it is exact. Mostly we are happy to use large sample approximations based on central limit theorems for the score statistic and Taylor series approximations applied to a development around the true value of an unknown parameter. Here we have an exact result regardless of how small our sample is as long as we are prepared to accept the argument that it is appropriate to condition on the observed spread, the so-called "configuration" of the sample. For a model in which the parameter of interest is a location parameter, the configuration of the sample is simply the set of distances between the observations and the empirical mean. For a location-scale family this set consists of these same quantities standardized by the square root of the variance.

Fisher's results were extended to the very broad class of models coming under the heading of curved exponential family models (Efron et al., 1978). This extension was carried out for more general situations by conditioning on the observed information in the sample (a quantity analogous to the information contained in the set of standardized residuals). Although no longer an exact result the result turns out to be very accurate and has been extensively studied by Barndorff-Nielsen and Hall (1988). For the proportional hazards model we can use these results to obtain $g(\beta)$ as the conditional distribution of the maximum likelihood estimator via the expression

$$g(u) = \prod_i \pi(u, X_i) / \int_u \left\{ \prod_i \pi(u, X_i) \right\} du.$$

In Figure D.1 we illustrate the function $g(u)$, i.e., the estimated conditional distribution of the maximum likelihood estimator for the proportional hazards model given the two sample data from the Freireich study. It is interesting to compare the findings with those available via more standard procedures. Rather than base inference on the mode of this distribution (i.e., the maximum partial likelihood estimate) and the second derivative of the log-likelihood we can consider

$$\tilde{\beta} = \int_u ug(u) du, \quad v(\beta) = \int_u u^2 g(u) du - \left(\int_u ug(u) du \right)^2,$$

and base tests, point, and interval estimation on these. Estimators of this kind have been studied extensively by Pitman (1948), and generally, have smaller mean squared error than estimators based on the mode. Note also, in a way analogous to the calculation of bootstrap percentile intervals, that we can simply consider the curve $g(u)$ and tick off areas of size $\alpha/2$ on the upper and lower halves of the function to obtain a $100(1-\alpha)\%$ confidence interval for β . Again, analogous to bootstrap percentile intervals, these intervals can pick up asymmetries and provide more accurate confidence intervals for small samples than those based on the symmetric large sample normal approximation. For the Freireich data, agreement between the approaches is good and that is to be expected since, in this case, the normal curve is clearly able to give a good approximation to the likelihood function.

For higher dimensions the approach can be extended almost immediately, at least conceptually. A potential difficulty arises in the following way: suppose we are interested in β_1 and we have a second parameter, β_2 , in the model. Logically we would just integrate the two-dimensional density with respect to β_2 , leaving us with a single marginal density for β_1 . This is straightforward since we have all that we need to do this, although of course there will usually be no analytical solution and it will be necessary to appeal to numerical techniques. The difficulty occurs since we could often parameterize a model differently, say incorporating β_2^2 instead of β_2 alone. The marginal distribution for β_1 , after having integrated

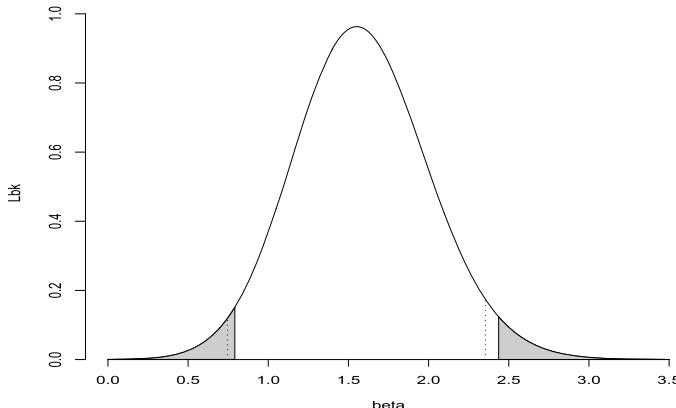


Figure D.1: The standardized (i.e., area integrates to one) partial likelihood function for the proportional hazards model based on the Freireich data.

out β_2 , will not generally be the same as before. Nonetheless, for the proportional hazards model at least, the parameterization is quite natural and it would suffice to work with the models as they are usually expressed.

Partial likelihood

Again, starting from a full likelihood, or full density, we can focus interest on some subset, “integrating out” those parameters of indirect interest. Integrating the full density (likelihood) with respect to those parameters of secondary interest produces a marginal likelihood. Cox (1975) develops a particular expression for the full likelihood in which an important component term is called the partial likelihood. Careful choices lead us back to conditional likelihood and marginal likelihood and Cox then describes these as special cases. For a given problem, Cox provides some guidelines for finding partial likelihoods: (1) the omitted factors should have distributions depending in an essential way on nuisance parameters and should contain no information about the parameters of interest and (2) incidental parameters, in particular the nuisance parameters, should not occur in the partial likelihood.

Lemma D.1. *For the proportional hazards model with constant effects:*

$$L(\beta) = \prod_{i=1}^n \left\{ \frac{\exp(\beta Z_i)}{\sum_{j=1}^n Y_j(X_i) \exp(\beta Z_j)} \right\}^{\delta_i} \quad (\text{D.10})$$

satisfies the two guidelines provided by Cox.

For this particular case then the term “partial likelihood” applies. We use the term more generally for likelihoods of this form even though we may not be

able to verify the extent to which the guidelines apply. Unfortunately, at least when compared to usual likelihood, marginal and conditional likelihood, or profile likelihood (maximizing out rather than integrating out nuisance parameters), partial likelihood is a very difficult concept. For general situations, it is not at all clear how to proceed. Furthermore, however obtained, such partial likelihoods are unlikely to be unique. Unlike the more commonly used likelihoods, great mathematical skill is required and even well steeled statistical modelers, able to obtain likelihood estimates in complex applied settings, will not generally be able to do the same should they wish to proceed on the basis of partial likelihood. For counting processes Andersen et al. (1993), it requires some six pages (pages 103–109), in order to formulate an appropriate partial likelihood.

However, none of this impedes our development since the usual reason for seeking an expression for the likelihood is to be able to take its logarithm, differentiate it with respect to the unknown parameters, and equating this to zero, to enable the construction of an estimating equation. Here we already have, via the main theorem (Section 7.5), or stochastic integral considerations, appropriate estimating equations. Thus, and in light of the above mentioned difficulties, we do not put emphasis on partial likelihood as a concept or as a general statistical technique. Nonetheless, for the main models we consider here, the partial likelihood, when calculated, can be seen to coincide with other kinds of likelihood, derived in different ways, and that the estimating equation, arising from use of the partial likelihood, can be seen to be a reasonable estimating equation in its own right, however obtained. The above expression was first presented in Cox's original paper where it was described as a conditional likelihood.

Cox's conditional likelihood

At any time point t there is a maximum of n subjects under study. We can index the j th subject as having hazard rate $\lambda_j(X_i)$ at time point X_i . In other words we let $Y_j(t)$ take the value 1 if an individual is available to make a transition, zero otherwise. Then, for the j th subject we have an intensity function $\alpha_j(t) = Y_j(t)\lambda_j(t)$. The n patients can be viewed as a system. At time point t the system either remains the same (no failure), changes in a total of n possible ways, in which a single patient fails or the system may change in a more complicated way in which there can be more than a single failure.

If we are prepared to assume that at any given point there can be no more than a single failure, then the system can change in a maximum of n ways. This defines a simple stochastic process. Notice that, conditional upon there being a transition, then a straightforward application of Bayes formula enables us to deduce the probability that the transition is of type j . This is simply

$$\frac{\alpha_j(X_i)}{\sum_{\ell=1}^n \alpha_\ell(X_i)} = \frac{Y_j(X_i)\lambda_j(X_i)}{\sum_{\ell=1}^n Y_\ell(X_i)\lambda_\ell(X_i)} = \frac{\exp(\beta Z_j)}{\sum_{\ell=1}^n Y_\ell(X_i)\exp(\beta Z_\ell)}.$$

For subjects having either failed or being lost to follow-up before time t we can still carry out the sum over all n subjects in our evaluation at time t . This is because of the indicator variable $Y_j(t)$ that takes the value zero for all such subjects, so that their transition probabilities become zero. The same idea can be expressed via the concept of risk sets, i.e., those subjects alive and available to make the transition under study. However, whenever possible, it is preferable to make use of the indicator variables $Y_j(t)$, thereby keeping the sums over n .

Multiplying all the above terms over the observed failure times produces $L(\beta)$. In his 1972 paper Cox described this as a conditional likelihood and suggested it be treated as a regular likelihood for the purposes of inference. In their contribution to the discussion of Cox's paper, Kalbfleisch and Prentice point out that the above likelihood does not have the usual probabilistic interpretation. If we take times to be fixed then $\exp(\beta Z_j)/\sum_\ell Y_\ell(X_i) \exp(\beta Z_\ell)$ is the probability of the subject indexed by j failing at X_i and that all other subjects, regardless of order, occur after X_i . Cox's deeper study (Cox, 1975) into $L(\beta)$ led to the recognition of $L(\beta)$ as a partial likelihood and not a conditional likelihood in the usual sense.

The flavor of $L(\beta)$ is nonetheless that of a conditional quantity, even if the conditioning is done sequentially and not all at once. Cox's discovery of $L(\beta)$, leading to a host of subsequent applications (time-dependent effects, time-dependent covariates, random effects), represents one of the most important statistical advances of the twentieth century. Although it took years of subsequent research in order to identify the quantity introduced by Cox as the relevant quantity with which to carry out inference, and although it was argued that Cox's likelihood was not a conditional likelihood in the usual sense (where all of the conditioning is done at once), his likelihood was all the same the right quantity to work with.

Marginal likelihood of ranks

Kalbfleisch and Prentice (1973) pointed out, that for fixed covariates Z the partial likelihood coincides with the likelihood, or probability, of the rank vector occurring as observed, under the model. We then have an important result;

Theorem D.4. *Under an independent censoring mechanism the probability of observing the particular order of the failures is given by $L(\beta)$.*

Alternatively we can express the likelihood as a function of the regression vector β and the underlying failure rate $\lambda_0(t)$. Writing this down we have:

$$L(\beta, \lambda_0(t)) = \prod_{i=1}^n \left[\lambda_0(t) e^{\beta Z_i} S_0(X_i)^{\exp(\beta Z_i)} \right]^{\delta_i} \left[S_0(X_i)^{\exp(\beta Z_i)} \right]^{1-\delta_i}.$$

From this we can break the likelihood into two components. We then have:

Theorem D.5. *$L(\lambda_0, \beta)$ can be written as the product $L_\lambda(\lambda_0, \beta)L(\beta)$.*

This argument is also sometimes given to motivate the idea of partial likelihood, stating that the full likelihood can be decomposed into a product of two terms, one of which contains the nuisance parameters $\lambda_0(t)$ inextricably mixed in with the parameter of interest β and a term that only depends upon β . This second term is then called the partial likelihood. Once again, however, any such decomposition is unlikely to be unique and it is not clear how to express in precise mathematical terms just what we mean by “inextricably mixed in” since this is more of an intuitive notion suggesting that we do not know how to separate the parameters. Not knowing how to separate the parameters do not mean that no procedure exists that might be able to separate them. And, if we were to sharpen the definition by stating, for example, that within some large class there exists no transformation or re-parameterization that would separate out the parameters, then we would be left with the difficulty of verifying this in practice, a task that would not be feasible.

Appendix E

Simulating data under the non-proportional hazards model

We look at two methods described by Chauvel (2014) that will allow us to simulate samples $\{T_i, Z_i, i = 1, \dots, n\}$ under the non-proportional hazards model. The first allows us to simulate data when we have a piecewise constant $\beta(t)$, corresponding to a change-point model, using the distribution of T given Z . The second is based on the distribution of Z given T , and allows us to generate data for any $\beta(t)$ which is non-constant over time. Censoring is independent of this; we thus simulate the times $\{C_1, \dots, C_n\}$ independently of the data.

E.1 Method 1—Change-point models

Consider the non-proportional hazards model with a piecewise constant regression coefficient given by: for any $t \in [0, \mathcal{T}]$:

$$\beta(t) = \sum_{i=1}^k \beta_i \mathbf{1}_{t_i \leq t < t_{i+1}}, \quad t_1 = 0, \quad t_{k+1} = +\infty,$$

where the values β_1, \dots, β_k are constants. To begin with, we suppose that the $k - 1$ break-points t_2, t_3, \dots, t_k are known and we show how to simulate data under the model with parameter $\beta(t)$. After that, we show how to define the change-points corresponding to the places where the value of $\beta(t)$ changes.

Simulating the data

To simplify the presentation, we will work with a constant (over time) one-dimensional covariate $Z \in \mathbb{R}$ and suppose that the base risk is constant: $\lambda_0(t) = 1$ for all $t \in [0, \mathcal{T}]$. Generalizing this to several (potentially time-dependent) covariates is straightforward. Data is thus simulated under the model:

$$\lambda(t | Z) = \lambda_0(t) \exp\{\beta(t)Z\} = \exp\left(Z \sum_{i=1}^k \beta_i \mathbf{1}_{t_i \leq t < t_{i+1}}\right), \quad t \in [0, T]. \quad (\text{E.1})$$

Calculating $S(\cdot | Z)$

Let $t \in [0, T]$ and calculate the value of the conditional survival $S(t | Z)$. If t is between $t_1 = 0$ and t_2 , the survival is

$$\begin{aligned} S(t | Z) &= \exp\left(-\int_0^t \lambda(u | Z) du\right) = \exp\left(-\int_0^t \exp(\beta(u)Z) du\right) \\ &= \exp\left(-\int_{t_1=0}^t \exp(\beta_1 Z) du\right) = \exp(-t \exp(\beta_1 Z)). \end{aligned}$$

Then, if t is between t_j and t_{j+1} , $j \in \{2, \dots, k\}$, the survival of T given Z calculated at t is

$$\begin{aligned} S(t | Z) &= \exp\left(-\int_0^t \lambda(u | Z) du\right) = \exp\left(-\int_0^t \exp(\beta(u)Z) du\right) \\ &= \exp\left(-\sum_{i=1}^{j-1} \int_{t_i}^{t_{i+1}} \exp(\beta_i Z) du - \int_{t_j}^t \exp(\beta_j Z) du\right) \\ &= \exp\left(-\sum_{i=1}^{j-1} (t_{i+1} - t_i) \exp(\beta_i Z) - (t - t_j) \exp(\beta_j Z)\right). \end{aligned}$$

For simulating a random variable whose survival function, given the covariates Z , is $S(\cdot | Z)$, we use the fact that, if $U \sim \mathcal{U}[0, 1]$, then $F^{-1}(U)$ has the cumulative distribution function F . All that is needed then is to inverse the cumulative distribution function $F(\cdot | Z) = 1 - S(\cdot | Z)$.

The inverse of $F(\cdot | Z)$

Let $\gamma \in [0, 1]$ be such that

$$F(t_j | Z) \leq \gamma < F(t_{j+1} | Z), \quad j \in \{1, \dots, k\}.$$

We are looking for the t such that $\gamma = F(t | Z)$. If $0 \leq \gamma < F(t_2 | Z)$:

$$\gamma = F(t | Z) = 1 - \exp(-t \exp(\beta_1 Z)),$$

which is equivalent to $t = -\log(1 - \gamma) \exp(-\beta_1 Z)$. If γ is such that $F(t_j | Z) \leq \gamma < F(t_{j+1} | Z)$, then for $j \in \{2, \dots, k\}$ we have

$$\gamma = F(t | Z) = 1 - \exp \left(- \sum_{i=1}^{j-1} (t_{i+1} - t_i) \exp(\beta_i Z) - (t - t_j) \exp(\beta_j Z) \right),$$

which is the same as

$$\begin{aligned} -\log(1 - \gamma) &= \sum_{i=1}^{j-1} (t_{i+1} - t_i) \exp(\beta_i Z) + (t - t_j) \exp(\beta_j Z) \\ &= \sum_{i=2}^j t_i \{\exp(\beta_{i-1} Z) - \exp(\beta_i Z)\} + t \exp(\beta_j Z). \end{aligned}$$

In conclusion:

$$t = \exp(-\beta_j Z) \left(-\log(1 - \gamma) - \sum_{i=2}^j t_i \{\exp(\beta_{i-1} Z) - \exp(\beta_i Z)\} \right).$$

Thus, overall, for any $\gamma \in [0, 1]$, we have:

$$\begin{aligned} F^{-1}(\gamma | Z) &= \sum_{j=2}^k \left(e^{-\beta_j Z} \left(-\log(1 - \gamma) - \sum_{i=2}^j t_i (e^{\beta_{i-1} Z} - e^{\beta_i Z}) \right) \right) \mathbf{1}_{F(t_j | Z) \leq \gamma < F(t_{j+1} | Z)} \\ &\quad - \log(1 - \gamma) e^{-\beta_1 Z} \mathbf{1}_{\gamma \leq F(t_2 | Z)}. \end{aligned}$$

Starting with a variable Z , a coefficients vector $(\beta_1, \dots, \beta_k)$, and a vector with the break-points (t_2, \dots, t_k) , we simulate U from the uniform distribution $\mathcal{U}[0, 1]$; then, $F^{-1}(U | Z)$ is a random variable with cumulative distribution function $F(\cdot | Z)$. This variable is the survival time T and is coherent with the non-proportional hazards model.

The break-point times need to be selected in such a way that we do not run out of individuals. In effect, the times need to be calibrated so that individuals are present and of sufficient number in each interval $[t_j, t_{j+1}]$. The choice of times is made as a function of the coefficients β_1, \dots, β_k that have been selected for the study.

Choosing change-point locations

Break-point times t_2, \dots, t_{k-1} are chosen numerically in such a way that there is a comparable number of individuals in each time interval. Consider the following example: we wish to simulate survival data with model (E.1) and three successive values of β :

$$\beta(t) = \sum_{i=1}^3 \beta_i \mathbf{1}_{t_i \leq t < t_{i+1}}, \quad t_1 = 0, \quad t_4 = +\infty,$$

and $\beta_1 = 2$, $\beta_2 = 1$ and $\beta_3 = 0$. Before being able to simulate the data using the method presented in Section E, we need to choose the values t_2 and t_3 .

Step 1: We simulate survival data for 2000 individuals according to model (E.1) with $\beta(t) = \beta_1 = 2$ for all t . We then calculate the corresponding Kaplan-Meier estimator. Next, we select a time such that 1/3 of individuals have died before that time. This value corresponds to t_2 . In our example, we get $t_2 = 0.1$.

Step 2: We simulate a new dataset with 2000 individuals according to model (E.1) with $\beta(t) = \beta_1 \mathbf{1}_{t \leq t_2} + \beta_2 \mathbf{1}_{t \geq t_2} = 2\mathbf{1}_{t \leq t_2} + \mathbf{1}_{t \geq t_2}$. We then calculate the corresponding Kaplan-Meier estimator. The time t_3 corresponds to when 2/3 of the individuals have died. In our example, $t_3 = 0.4$. We simulated a dataset with the break-points determined as above. Figure E.1 shows the Kaplan-Meier estimator conditional on the covariates $\hat{S}(\cdot | Z)$ as a function of time, as well as its transform $\log(-\log(\hat{S}(\cdot | Z)))$. After applying this transform, each line on the plot represents the log of the instantaneous hazard for one group. We see that there is a different distance between the lines in each of the three time periods (between 0 and $t_2 = 0.1$, t_2 and $t_3 = 0.4$, and after t_3). We choose these time intervals so that the same number of deaths occurs in each. It is possible instead to have unequal numbers, as long as we keep in mind that we need individuals in each time period; otherwise, the corresponding effect will not be represented in the dataset.

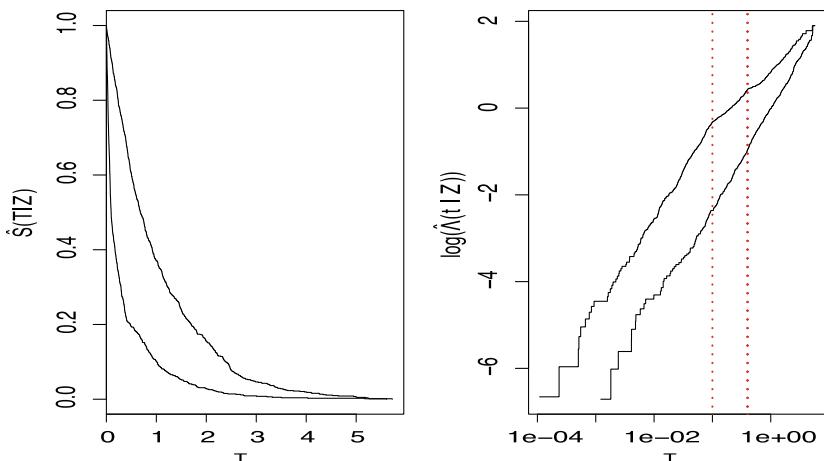


Figure E.1: The Kaplan-Meier estimator conditional on the covariate $\hat{S}(\cdot | Z)$ and its transform $\log(-\log(\hat{S}(\cdot | Z)))$ as a function of time.

E.2 Method 2—Non-proportional hazards models

In order to generate data under the more general non-proportional hazards models, we use the conditional distribution of Z given T . We work with a single covariate Z that does not change with time; extensions to multiple covariates are straightforward with the help of the prognostic index.

We simulate the vectors for times of death (T_1, \dots, T_n), censoring (C_1, \dots, C_n), and covariate vectors (Z_1, \dots, Z_n) independently, via their marginal distributions. With the help of the probabilities $\{\pi_i(\beta(t), t), i = 1, \dots, n\}$, we connect up the covariates and the ranked times of death in the following way:

1. We order the times of death: $T_{(1)} \leq T_{(2)} \leq \dots \leq T_{(n)}$.
2. We set $\mathcal{R} = \{1, \dots, n\}$, which corresponds to the current set of at-risk individuals.
3. For i from 1 to n :
 - We calculate $\pi_j(\beta(T_{(i)}), T_{(i)})$ for each individual $j \in \mathcal{R}$, setting $Y_l(t) = 1$ if $l \in \mathcal{R}$ and $Y_l(t) = 0$ otherwise.
 - We randomly select an individual j^* in \mathcal{R} , where individual j in \mathcal{R} is chosen with probability $\pi_j(\beta(T_{(i)}), T_{(i)})$.
 - We link Z_{j^*} with $T_{(i)}$.
 - We remove j^* from \mathcal{R} .
4. We set $X_i = \min(T_i, C_i)$ and $\delta_i = \mathbf{1}_{T_i \leq C_i}$ for $i = 1, \dots, n$.

Hence, in our dataset $\{X_i, \delta_i, Z_i; i = 1, \dots, n\}$, times of death are simulated under the non-proportional hazards model with parameter $\beta(t)$, and censoring is independent of times of death. Abrahamowicz et al. (1996) have proposed a similar method in which the sampling probabilities for Z are not the same when there is censoring as when there is not.

Further exercises and proofs

The appendices are referred to throughout the text and provide the background to many important results. They are of interest in their own right and, for this reason, we include here proofs of some key results as well as classwork and exercises. These are for the benefit of instructors and students who wish to dig a little more deeply into this background. A course in survival analysis might well include certain aspects covered in these appendices and this would depend on the type of course being given as well as the flavor that the instructor wishes for it to assume.

RANDOM VARIABLES, DISTRIBUTIONS, ORDER STATISTICS

1. Use a simple sketch to informally demonstrate the mean value theorem.
2. Newton-Raphson iteration provides sequentially updated estimates to the solution to the equation $f(x_0) = 0$. At the n th step, we write $x_{n+1} = x_n - f(x_n)/f'(x_n)$ and claim that x_n converges (in the analytical sense) to x_0 . Use the mean value theorem, and again, a simple sketch to show this. Intuitively, which conditions will lead to convergence and which ones can lead to failure of the algorithm.
3. Let $g(x)$ take the value 0 for $-\infty < x \leq 0$; $1/2$ for $0 < x \leq 1$; 1 for $1 < x \leq 2$; and 0 otherwise. Let $f(x) = x^2 + 2$. Evaluate the Riemann-Stieltjes integral of $f(x)$ with respect to $g(x)$ over the real line.
4. Note that $\sum_{i=1}^n i = n(n+1)/2$. Describe a function such that a Riemann-Stieltjes integral of it is equal to $n(n+1)/2$. Viewing integration as an area under a curve, conclude that this integral converges to n^2 as n becomes large.
5. Suppose that in the Helly-Bray theorem for $\int h(x)dF_n(x)$, the function $h(x)$ is unbounded. Break the integral into components over the real line. For regions where $h(x)$ is bounded the theorem holds. For the other regions obtain conditions that would lead to the result holding generally.

6. Prove the probability integral transformation by finding the moment-generating function of the random variable $Y = F(X)$ where X has the continuous cumulative distribution function $F(x)$ and a moment-generating function that exists.
7. If X is a continuous random variable with probability density function $f(x) = 2(1-x), 0 < x < 1$, find that transformation $Y = \psi(X)$ such that the random variable Y has the uniform distribution over $(0,2)$.
8. The order statistics for a random sample of size n from a discrete distribution are defined as in the continuous case except that now we have $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$. Suppose a random sample of size 5 is taken with replacement from the discrete distribution $f(x) = 1/6$ for $x = 1, 2, \dots, 6$. Find the probability mass function of $X_{(1)}$, the smallest order statistic.
9. Ten points are chosen randomly and independently on the interval $(0,1)$. Find (a) the probability that the point nearest 1 exceeds 0.8, (b) the number c such that the probability is 0.4 that the point nearest zero will exceed c .
10. Find the expected value of the largest order statistic in a random sample of size 3 from (a) the exponential distribution $f(x) = \exp(-x)$ for $x > 0$, (b) the standard normal distribution.
11. Find the probability that the range of a random sample of size n from the population $f(x) = 2e^{-2x}$ for $x \geq 0$ does not exceed the value 4.
12. Approximate the mean and variance of (a) the median of a sample of size 13 from a normal distribution with mean 2 and variance 9, (b) the fifth-order statistic of a random sample of size 15 from the standard exponential distribution.
13. Simulate 100 observations from a uniform distribution. Do the same for an exponential, Weibull, and log-logistic distribution with different parameters. Next, generate normal and log-normal variates by summing a small number of uniform variates. Obtain histograms. Do the same for 5000 observations.
14. Obtain the histogram of 100 Weibull observations. Obtain the histogram of the logarithms of these observations. Compare this with the histogram obtained by the empirical transformation to normality.
15. Suppose that T_1, \dots, T_n are n exponential variates with parameter λ . Show that, under repeated sampling, the smallest of these also has an exponential distribution. Is the same true for the largest observation? Suppose we are only given the value of the smallest of n observations from an exponential distribution with parameter λ . How can this observation be used to estimate λ .

16. Suppose that X_i $i = 1, \dots, n$ are independent exponential variates with parameter λ . Determine, via simple calculation, the variance of $\min(X_1, \dots, X_n)$.
17. Having some knowledge of the survival distribution governing observations we are planning to study, how might we determine an interval of time to obtain with high probability a given number of failures? How should we proceed in the presence of censoring?
18. Derive the Bienaymé-Chebyshev inequality. Describe the advantages and drawbacks of using this inequality to construct confidence intervals in a general setting.
19. Recall that the information contained in the density $g(t)$ with respect to $h(t)$ is defined by $V(g, h) = E \log g(T) = \int \log g(t)h(t)dt$ and the entropy is just $-V(g, h)$. Suppose that the entropy depends on a parameter θ and is written $-V_\theta(f, f)$. Consider $-V_\alpha(f, f)$ as a function of α . Show that this function is maximized when $\alpha = \theta$.
20. For random variables, T and Z with finite second moments, use the device of double expectation to show that: $\text{Var}(T) = \text{Var}E(T|Z) + E\text{Var}(T|Z)$. Thus, the total variance, $\text{Var}(T)$, breaks down into two parts. Why is this breakdown interpreted as one component corresponding to “signal” and one component corresponding to “noise?”
21. Suppose that θ_n converges in probability to θ and that the variance of θ_n is given by $\psi(\theta)/n$. Using Equation A.17, find a transformation of θ_n for which, at least approximately, the variance does not depend on θ .
22. Consider a stochastic process $X(t)$ on the interval $(2, 7)$ with the following properties: (a) $X(0) = 2$, (b) $X(t), t \in (2, 7)$ has increments such that (c), for each $t \in (2, 7)$ the distribution of $X(t)$ is Weibull with mean $2 + \lambda t^\gamma$. Can these increments be independent and stationary? Can the process be described using the known results of Brownian motion?
23. For Brownian motion, explain why the conditional distribution of $X(s)$ given $X(t)$ ($t > s$) is normal with $E\{X(s)|X(t) = w\} = ws/t$ and $\text{Var}\{X(s)|X(t) = w\} = s(t-s)/t$. Deduce the mean and the covariance process for the Brownian bridge.
24. The Ornstein-Uhlenbeck process can be thought of as transformed Brownian motion in which the variance has been standardized. Explain why this is the case.

25. Reread the subsection headed “Time-transformed Brownian motion” (Section B.2) and conclude that the only essential characteristic underwriting the construction of Brownian motion is that of independent increments.
26. Find the value of $t \in (0,1)$ for which the variance of a Brownian bridge is maximized.
27. Suppose that under H_0 , $X(t)$ is Brownian motion. Under H_1 , $X(t)$ is Brownian motion with drift, having drift parameter 2 as long as $X(t) < 1$ and drift parameter minus 2 otherwise. Describe likely paths for reflected Brownian motion under both H_0 and H_1 . As a class exercise simulate ten paths under both hypotheses. Comment on the resulting figures.

APPROXIMATIONS AND LARGE SAMPLE THEORY

1. Suppose that $\sum_{i=1}^{\infty} a_i = 1$. Also suppose that there exist random variable X_i such that $X_i = a_i X_1$. Show that $\sum X_i$, standardized by mean and variance, would not generally tend to a normal distribution. Comment.
2. Describe in simple terms what is quantified by the Lindeburg condition.
3. For a weighted sum of independent identically distributed random variables, suppose that each variable is standard uniform and the weights a_i are defined as $a_i = 10^i$. Will a central limit result hold, and if not, why not.
4. As a class project, construct graphs based on summing (1) 20 uniform random variates, (2) 100 uniform random variates, and (3) 10000 uniform random variates. Standardize the x axis to lie between 0 and 1 and the y axis such that the mean is equal to zero and the variance increases linearly with x . Replicate each graph ten times. What conclusions can be drawn from the figures, in particular the influence of the number of variates summed?
5. Repeat the above class exercise, replacing the uniform distribution by (1) the log-logistic distribution with different means and variances, (2) the exponential distribution, and (3) the normal distribution. Again replicate each graph ten times. Comment on your findings.
6. Consider two hypotheses for the sequence of observations: X_1, \dots, X_n in which $\mu_i = E(X_i)$; $H_1 : \mu_i = 0, \forall i$, against $H_1 : \mu_i = b^2 i$ for some $b \neq 0$. Construct different tests based on Brownian motion that would enable us to test H_0 versus H_1 . Discuss the relative merits of the different tests.
7. Simulate 20 values from a standard exponential distribution and evaluate the greatest absolute distance between $F_n(t)$ and $F(t)$. Repeat this 1000 times, store

the 1000 values of D_{20} , and then plot a histogram of these values. Add to the histogram the distribution of D_{20} using the Brownian bridge approximation.

8. For a sample size of 220, provide an approximate calculation of how large you would anticipate the greatest discrepancy between $F_n(t)$ and $F(t)$ to be.

9. Let T_1, \dots, T_n be i.i.d. observations from an exponential distribution with mean θ . Obtain an estimating equation for θ in terms of $E(T_i)$. Obtain an estimating equation for θ in terms of $\text{Var}(T_i)$. Which of the two equations, if any, would be the most preferable. Give reasons.

10. Consider the following bivariate situation. We have a random variable, T , which is continuous and a binary variable Z , which take the values 0 or 1. Given that $T = 0$, $\Pr(Z = 0) = 0.5$. Given that $Z = 0$, the distribution of T is exponential with mean equal to 1. Given that $Z = 1$ the distribution of T is exponential with mean equal to $\exp(\beta)$. Given n pairs of observations (T_i, Z_i) our purpose is to estimate the parameter β . Considering the values of Z as being fixed, obtain an estimating equation based on the observations T_i ($i = 1, \dots, n$). Secondly, we could view the random aspect of the experiment differently and now take the observations T_1 to T_n as being fixed. Derive the conditional distribution of Z given $T = T_i$. Use this to obtain a different estimating equation. Discuss the relative merits and disadvantages of the two sets of estimating equations.

11. In the next chapter we discuss the idea of right censoring where, for certain of the observations T_i , the exact value is not known. All that we can say for sure is that it is greater than some censoring time. How might the discussion of the previous exercise on the two types of estimating equations arising from reversing the conditioning variable have a bearing on this.

12. Consider the pair of independent random variables (Y_i, X_i) , $i = 1, \dots, n$. The null hypothesis is that $Y_i = \phi(X_i) + \epsilon_i$ where ϵ_i is an error term independent of the pair (Y_i, X_i) and where $\phi(u)$ is a nondecreasing function of u . Describe how you would carry out tests based on D_n and C_n of Section C.5.

13. By investigating different classes of functions ϕ , describe the relative advantages and disadvantages of tests based upon C_n rather than D_n .

OUTLINE OF PROOFS

Proof of Theorem A.2, Theorem A.7 and Corollary A.9. The importance of the first of these two theorems is difficult to overstate. All the useful large sample results, for instance, hinge ultimately on the theorem. An elegant proof of the theorem, together with well thought out illustrations and some examples, is given

in Shenk (1979). For Theorem A.7 let T have a continuous and invertible distribution function $F(t)$ and let $U = F(T)$. The inverse function is denoted F^{-1} so that $F^{-1}\{F(t)\} = t$. Then

$$\begin{aligned} P(U < u) &= P\{F(T) < u\} = P\{T < F^{-1}(u)\} \\ &= P\{T < F^{-1}F(t)\} = P\{T < t\} = F(t) = u. \end{aligned}$$

Thus U has the distribution of a standard uniform. The proof of the corollary leans upon some straightforward manipulation of elementary events. An outline is provided in David (1994). \square

Proof of Theorem A.9 and Corollary A.4. The theorem states that, letting $F(x) = P(X \leq x)$ and $F_r(x) = P(X_{(r)} \leq x)$ then;

$$F_r(x) = \sum_{i=r}^n \binom{n}{i} F^i(x)[1 - F(x)]^{n-i}.$$

Recall that the X_i from the parent population with distribution $F(x)$ are i.i.d. The event that $X_{(r)} \leq x$ is the event that at least r of the X_i are less than or equal to x . This is then the sum of the binomial probabilities summed over all values of i greater than or equal to r . The first part of the corollary is clear upon inspection. For the second part note that

$$\sum_{i=0}^n \binom{n}{i} F^i(x)[1 - F(x)]^{n-i} = 1, \quad \binom{n}{0} F^0(x)[1 - F(x)]^n = [1 - F(x)]^n.$$

\square

Proof of Proposition A.3. Let $n \in \mathbb{N}^*$ and the functions $f = (f^1, \dots, f^p) \in (C[0, 1], \mathbb{R}^p)$ and $g_n = (g_n^1, \dots, g_n^p) \in (D[0, 1], \mathbb{R}^p)$. We have the relation,

$$\begin{aligned} \sup_{0 \leq t \leq 1} \|f(t) - g_n(t)\| &= \max_{i=1, \dots, p} d(f^i, g_n^i) \geq \frac{1}{p} \sum_{i=1}^p d(f^i, g_n^i) \\ &\geq \frac{1}{p} \sum_{i=1}^p \delta(f^i, g_n^i) = \frac{1}{p} \delta_p(f, g_n), \end{aligned}$$

where the final inequality is obtained by applying Equation (A.4). The conclusion follows. \square

Proof of Theorem B.1 and Corollaries B.1 and B.2. We have that:

$$\begin{aligned} \Pr\{X(t+s) > x | X(s) = x_s, X(u), 0 \leq u < s\} \\ &= \Pr\{X(t+s) - X(s) > x - x_s | X(s) = x_s, X(u), 0 \leq u < s\} \\ &= \Pr\{X(t+s) - X(s) > x - x_s\} = \Pr\{X(t+s) > x | X(s) = x_s\}. \end{aligned}$$

Corollary B.1 follows since:

$$\begin{aligned} f_{s|t}(x|w) &= f_s(x)f_{t-s}(w-x)/f_t(w) = \text{const} \exp\{-x^2/2s - (w-x)^2/2(t-s)\} \\ &= \text{const} \exp\{-t(x-ws/t)^2/2s(t-s)\}. \end{aligned}$$

For Corollary B.2 note that the process $V(t)$ is a Gaussian process in which:

$$E\{V(t)\} = 0, \quad \text{Cov}\{V(t), V(s)\} = E\{\exp(-\alpha t/2) \exp(-\alpha s/2) X(e^{\alpha t}) X(e^{\alpha s})\}$$

This can be written: $\exp(-\alpha(t+s)/2) \times E\{X(e^{\alpha t}) X(e^{\alpha s})\}$ which in turn is then equal to $\exp(-\alpha(t+s)/2) \exp \alpha t$ and this is just $\exp(-\alpha(t-s)/2)$. \square

Proof of Theorem B.2 and Corollary B.3. Note that:

$$\begin{aligned} \text{Cov}(X(s), X(t)|X(1)=0) &= E(X(s), X(t)|X(1)=0) \\ &= E\{E(X(s), X(t)|X(t), X(1)=0|X(1)=0)\}. \\ &= E\{X(t)E(X(s)|X(t))|X(1)=0\} \\ &= E\{X(t)(s/t)X(t)|X(1)=0\} \\ &= (s/t)E\{X^2(t)|X(1)=0\} = (s/t)t(1-t) = s(1-t). \end{aligned}$$

Corollary B.3 can be deduced from this theorem via the following simple steps: For $s < t$,

$$\text{Cov}\{W^0(s), W^0(t)\} = E\{W(s) - sW(1)\}\{W(t) - tW(1)\}$$

and

$$\begin{aligned} E\{W(s)W(t) - sW(1)W(t) + stW(1)W(1) - tW(1)W(s)\} \\ = s - st + st - st = s(1-t). \end{aligned}$$

\square

Proof of Lemma B.1.

$$\text{Cov}\{X(s), X(t)\} = (s+1)(t+1) \times \text{Cov}(Z(t/(t+1)), Z(s/(s+1))).$$

Next, from the definition of the Brownian bridge, $Z(t) = W(t) - tW(1)$ so that, expanding the above expression we obtain:

$$\begin{aligned} \text{Cov}\{W(t/(t+1)), W(s/(s+1))\} - t/(t+1) \times \text{Cov}\{W(1), W(s/(s+1))\} - \\ s/(s+1) \text{Cov}\{W(1), W(t/(t+1))\} + st/((s+1)(t+1)) \times \text{Cov}\{W(1), W(1)\} \\ = s(t+1) - st - st + st = s. \end{aligned}$$

\square

Proof of Lemma B.2. $\text{Cov}\{Z(s), Z(t)\} = E\{Z(s)Z(t)\}$.

We write this as $E \int_0^s X(y) dy \int_0^t X(u) du$. Bringing together the two integral

operators we have:

$$\begin{aligned} E \int_0^s \int_0^t X(y)X(u)dydu &= \int_0^s \int_0^t EX(y)X(u)dydu \\ &= \int_0^s \int_0^t \min(y,u)dydu = \int_0^s \int_0^u ydy + \int_u^t udydu = s^2(t/2 - s/6). \end{aligned}$$

□

Proof of Theorem C.1. Assume that $\varepsilon > 0$. The variables X_1, X_2, \dots, X_n are uncorrelated so that $V(\sum_{i=1}^n X_i) = \sum_{i=1}^n V(X_i)$. A simple application of the Chebyshev-Bienaymé inequality implies that:

$$P\left(\left|\frac{1}{n} \sum_{i=1}^n X_i\right| > \varepsilon\right) \leq \frac{1}{\varepsilon^2} V\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n^2 \varepsilon^2} \sum_{i=1}^n V(X_i) \leq \frac{C}{n \varepsilon^2}.$$

As a result, $\lim_{n \rightarrow \infty} P\left(\left|\frac{1}{n} \sum_{i=1}^n X_i\right| > \varepsilon\right) = 0$ which is what we needed to show. □

Proof of Theorem C.2. Suppose that $k < m$,

$$\begin{aligned} \text{Cov}(S_k^*, S_m^*) &= E(S_k^* S_m^*) - E(S_k^*)E(S_m^*) \\ &= (\sigma \sqrt{n})^{-2} E\{S_k S_m\} \\ &= (\sigma \sqrt{n})^{-2} E\{S_k (S_k + \sum_{k < i \leq m} X_i)\}. \end{aligned}$$

Developing the bracket gives,

$$\begin{aligned} &(\sigma \sqrt{n})^{-2} \{E(S_k^2) + E(S_k)E(\sum_{k < i \leq m} X_i)\} \\ &= (\sigma \sqrt{n})^{-2} E(S_k^2) = (\sigma \sqrt{n})^{-2} k \sigma^2 = k/n = t. \end{aligned}$$

□

Proof of Theorem C.6 and B.6. For the first of these two theorems, note that the marginal and conditional normality of any sequence of $\sqrt{n}\{F_n(t_i) - F(t_i)\}$, ($0 < t_1 < \dots < t_m$), for some m , indicates the multivariate normality of $\sqrt{n}\{F_n(t_i) - F(t_i)\}$, ($0 < t_1 < \dots < t_m$). Thus $\sqrt{n}\{F_n(t) - F(t)\}$ is a Gaussian process. It has mean zero and variance $F(t)\{1 - F(t)\}$. To obtain the covariance, note first that the indicator variables $I(T_i \leq t)$ and $I(T_j \leq s)$ are independent for all t and s and $i \neq j$. It only remains to evaluate, for $s < t$

$$\begin{aligned}\text{Cov}\{I(T_i \leq t), I(T_i \leq s)\} &= E\{I(T_i \leq t)I(T_i \leq s)\} - E\{I(T_i \leq t)\}E\{I(T_i \leq s)\} \\ &= F(s) - F(t)F(s) = F(s)\{1 - F(t)\}.\end{aligned}$$

For the second of these two theorems, note that:

$$E\{dM'(t)|\mathcal{F}_{t-}\} = E\{H(t)dM(t)|\mathcal{F}_{t-}\} = H(t)E\{dM(t)|\mathcal{F}_{t-}\} = 0.$$

Furthermore,

$$\text{Var}\{dM'(t)|\mathcal{F}_{t-}\} = \text{Var}\{H(t)dM(t)|\mathcal{F}_{t-}\} = H(t)^2\text{Var}\{dM(t)|\mathcal{F}_{t-}\} = H(t)^2d\langle M \rangle(t).$$

Similarly,

$$\left\langle \int HdM, \int KdM' \right\rangle(t) = \int_0^t H(s)K(s)d\langle M, M' \rangle(s).$$

□

Proof of Theorem D.4. Suppose k individuals are observed to fail at times t_1, \dots, t_k and have corresponding explanatory variables z_1 to z_k . Assume times are ordered. Then the probability, conditional upon the observed failure times, of obtaining a particular ordering is:

$$\begin{aligned}\Pr(r = (1), (2), \dots, (n)) &= \Pr(t_1 < t_2 < t_3) = \int_0^\infty \int_{t_1}^\infty \dots \int_{t_{n-1}}^\infty f(t_1|z_1) \\ &\dots f(t_n|z_n) dt_n \dots dt_1 = \int_0^\infty \int_{t_1}^\infty \dots \int_{t_{n-1}}^\infty \prod_{i=1}^n \{h_0(t_i)e^{\beta z_i} \exp[-H_0(t_i)e^{\beta z_i}]\} \\ &= \prod_{i=1}^n \{\exp(\beta Z_i) / \sum_{j=1}^n Y_j(X_i) \exp(\beta Z_j)\}^{\delta_i}\end{aligned}$$

□

Bibliography

- O. Aalen, A model for nonparametric regression analysis of counting processes, in *Mathematical Statistics and Probability Theory. Lecture Notes in Statistics* 2 (Springer, New York, 1980), pp. 1–25
- O. Aalen, O. Borgan, H. Gjessing, *Survival and Event History Analysis: A Process Point of View* (Springer, New-York, 2008)
- O.O. Aalen, Nonparametric inference for a family of counting processes. *Ann. Stat.* **6**, 701–726 (1978)
- O.O. Aalen, A linear regression model for the analysis of life times. *Stat. Med.* **8**, 907–925 (1989)
- M. Abrahamowicz, T. MacKenzie, J. Esdaile, Time-dependent hazard ratio: modeling and hypothesis testing with application in lupus nephritis. *J. Am. Stat. Assoc.* **91**, 1432–1439 (1996)
- M. Abramowitz, I.A. Stegun, Handbook of mathematical functions with formulas, graphs, and mathematical table, in *US Department of Commerce*. National Bureau of Standards Applied Mathematics series 55 (1965)
- Acute Leukemia Group B, E.J. Freireich, E. Gehan, E. Frei, L.R. Schroeder, I.J. Wolman, R. Anbari, E.O. Burgert, S.D. Mills, D. Pinkel, O.S. Selawry, J.H. Moon, B.R. Gendel, C.L. Spurr, R. Storrs, F. Haurani, B. Hoogstraten, S. Lee, The effect of 6-mercaptopurine on the duration of steroid-induced remissions in acute leukemia: a model for evaluation of other potentially useful therapy. *Blood* **21**(6), 699–716 (1963)
- D.G. Altman, P.K. Andersen, A note on the uncertainty of a survival probability estimated from Cox's regression model. *Biometrika* **73**(3), 722–724 (1986)
- S.Z. Abildstrom, P.K. Andersen, S. Rosthøj, Competing risks as a multi-state model. *Stat. Methods Med. Res.* **11**(2), 203–215 (2002)
- P. Andersen, Testing goodness of fit of Cox's regression and life model. *Biometrics* **38**, 67–77 (1982)

- P.K. Andersen, R.D. Gill, Cox's regression model for counting processes: a large sample study. *Ann. Stat.* **10**, 1100–1120 (1982)
- P.K. Andersen, O. Borgan, R. Gill, N. Keiding, *Statistical Models Based on Counting Processes* (Springer, New York, 1993)
- P.K. Andersen, E. Christensen, L. Fauerholdt, P. Schlichting, Evaluating prognoses based on the proportional hazards model. *Scand. J. Stat.* **10**(2), 141–144 (1983)
- J.A. Anderson, A. Senthilselvan, A two-step regression model for hazard function. *J. R. Stat. Soc. Ser. C* **31**, 44–51 (1982)
- I. Annesi, T. Moreau, J. Lellouch, Efficiency of the logistic regression and cox proportional hazards models in longitudinal studies. *Stat. Med.* **8**(12), 1515–1521 (1989)
- E. Arjas, A graphical method for assessing goodness of fit in Cox's proportional hazards model. *J. Am. Stat. Assoc.* **83**, 204–212 (1988)
- P.C. Austin, J.P. Fine, Accounting for competing risks in randomized controlled trials: a review and recommendations for improvement. *Stat. Med.* **36**(8), 1203–1209 (2017)
- G. Bakoyannis, G. Touloumi, Practical methods for competing risks data: a review. *Stat. Methods Med. Res.* **21**(3), 257–272 (2012)
- N. Balakrishnan, N. Johnson, Samuel kotz. Continuous univariate distributions, vol. 1 (1994)
- I. Baltazar-Aban, E. Peña, Properties of hazard-based residuals and implications in model diagnostics. *J. Am. Stat. Assoc.* **90**, 185–219 (1995)
- W.E. Barlow, R.L. Prentice, Residuals for relative risk regression. *Biometrika* **75**(1), 65–74 (1988)
- O.E. Barndorff-Nielsen, P. Hall, On the level-error after Bartlett adjustment of the likelihood ratio statistic. *Biometrika* **75**(2), 374–378 (1988)
- O.E. Barndorff-Nielsen, D.R. Cox, *Inference and Asymptotics*, tome 13 (Chapman & Hall London, 1994)
- S. Bennett, Analysis of survival data by the proportional odds model. *Stat. Med.* **2**, 273–277 (1983a)
- S. Bennett, Log-logistic regression models for survival data. *Appl. Stat.* **32**, 165–171 (1983b)

- R. Beran, Nonparametric regression with randomly censored survival data. Technical Report, University of California, Berkeley (1981)
- R.N. Bhattacharya, E.C. Waymire, *Stochastic Processes with Applications* (Wiley, New York, 1990)
- P. Billingsley, *Convergence of Probability Measures*, 2nd edn. (Wiley, New York, 1999)
- D.A. Binder, Fitting cox's proportional hazards models from survey data. *Biometrika* **79**(1), 139–147 (1992)
- O. Borgan, K. Liestol, A note on confidence intervals and bands for the survival function based on transformations. *Scand. J. Stat.* 35–41 (1990)
- L. Breiman, *Classification and Regression Trees* (Routledge, 2017)
- N. Breslow, A generalized Kruskal-Wallis test for comparing K samples subject to unequal patterns of censorship. *Biometrika* **57**, 579–594 (1970)
- N. Breslow, Discussion of the paper by D. R. Cox. *J. R. Stat. Soc. B* **34**, 216–217 (1972)
- N. Breslow, The proportional hazards model: applications in epidemiology. *Commun. Stat. Theory Methods* **7**(4), 315–332 (1978)
- N. Breslow, J. Crowley, A large sample study of the life table and product limit estimates under random censorship. *Ann. Stat.* **437–453** (1974)
- N. Breslow, L. Elder, L. Berger, A two sample censored-data rank test for acceleration. *Biometrics* **40**, 1042–1069 (1984)
- N.E. Breslow, Statistics in epidemiology: the case-control study. *J. Am. Stat. Assoc.* **91**(433), 14–28 (1996)
- M.S. Brose, T.R. Rebbeck, K.A. Calzone, J.E. Stopfer, K.L. Nathanson, B.L. Weber, Cancer risk estimates for brca1 mutation carriers identified in a risk evaluation program. *J. Natl. Cancer Inst.* **94**(18), 1365–1372 (2002)
- C. Brown, On the use of indicator variables for studying the time-dependence of parameters in a response-time model. *Biometrics* **31**, 863–872 (1975)
- Z. Cai, Y. Sun, Local linear estimation for time-dependent coefficients in Cox's regression models. *Scand. J. Stat.* **30**, 93–111 (2003)
- H.-S.D. Cain, C. Kevin, R.J. Little, B. Nan, M. Yosef, J.R. Taffe, M.R. Elliott, Bias due to left truncation and left censoring in longitudinal studies of developmental and disease processes. *Am. J. Epidemiol.* **173**(9), 1078–1084 (2011)

- B.P. Carlin, J.S. Hodges, Hierarchical proportional hazards regression models for highly stratified data. *Biometrics* **55**(4), 1162–1170 (1999)
- C.L. Chang, *Introduction to Stochastic Processes in Biostatistics* (Wiley, New York, 1968)
- C. Chauvel, *Empirical Processes for Inference in the Non-Proportional Hazards Model*. Thèse de doctorat, PhD thesis, Université Pierre et Marie Curie-Paris 6, Paris (2014)
- C. Chauvel, J. O'Quigley, Tests for comparing estimated survival functions. *Biometrika* **101**(3), 535–552 (2014)
- C. Chauvel, J. O'Quigley, Survival model construction guided by fit and predictive strength. *Biometrics* **73**(2), 483–494 (2017)
- Q. Chen, R.C. May, J.G. Ibrahim, H. Chu, S.R. Cole, Joint modeling of longitudinal and survival data with missing and left-censored time-varying covariates. *Stat. Med.* **33**(26), 4560–4576 (2014)
- S. Cheng, L. Wei, Z. Ying, Analysis of transformation models with censored data. *Biometrika* **82**, 835–845 (1995)
- B. Choodari-Oskooei, P. Royston, M.K.B. Parmar, A simulation study of predictive ability measures in a survival model I: explained variation measures. *Stat. Med.* **31**(23), 2627–2643 (2012)
- A. Ciampi, J. Lawless, S. McKinney, K. Singhal, Regression and recursive partition strategies in the analysis of medical survival data. *J. Clin. Epidemiol.* **41**(8), 737–748 (1988)
- A. Ciampi, Z. Lou, Q. Lin, A. Negassa, Recursive partition and amalgamation with the exponential family: theory and applications. *Appl. Stoch. Models Data Anal.* **7**(2), 121–137 (1991)
- W.G. Cochran, Some methods for strengthening the common χ^2 tests. *Biometrics* **10**(4), 417–451 (1954)
- Fibrinogen Studies Collaboration, Measures to assess the prognostic ability of the stratified cox proportional hazards model. *Stat. Med.* **28**(3), 389–411 (2009)
- J. Cologne, W.-L. Hsu, R.D. Abbott, W. Ohishi, E.J. Grant, S. Fujiwara, H.M. Cullings, Proportional hazards regression in epidemiologic follow-up studies: an intuitive consideration of primary time scale. *Epidemiology* 565–573 (2012)
- D.R. Cox, *Planning of Experiments* (1958)
- D.R. Cox, Regression models and life-tables (with discussion). *J. R. Stat. Soc. Ser. B* **34**(2), 187–220 (1972)

- D.R. Cox, Partial likelihood. *Biometrika* **63**, 269–276 (1975)
- D.R. Cox, D.V. Hinkley, *Theoretical Statistics* (Chapman and Hall/CRC, 1979)
- D.R. Cox, E.J. Snell, A general definition of residuals. *J. R. Stat. Soc. Ser. B* **30**(2), 248–275 (1968)
- H. Cramér, *Random Variables and Probability Distributions*, tome 36 (Cambridge University Press, 2004)
- M. Crowder, Identifiability crises in competing risks. *Int. Stat. Rev.* 379–391 (1994)
- J.J. Crowley, B.E. Storer, Comment on 'A reanalysis of the Stanford Heart Transplant Data', by M. Aitkin, N. Laird and B. Francis. *J. Am. Stat. Assoc.* **78**, 277–281 (1983)
- M.J. Crowther, K.R. Abrams, P.C. Lambert, Joint modeling of longitudinal and survival data. *Stata J.* **13**(1), 165–184 (2013)
- S. CsÖrgÖ, L. Horvath, On the koziol-green model for random censorship. *Biometrika* **68**(2), 391–401 (1981)
- D.M. Dabrowska, K.A. Doksum, Partial likelihood in transformation models with censored data. *Scand. J. Stat.* **18**, 1–23 (1988)
- H. Daniels, Saddlepoint approximations for estimating equations. *Biometrika* **70**(1), 89–96 (1983)
- H.E. Daniels, Saddlepoint approximations in statistics. *Ann. Math. Stat.* 631–650 (1954)
- H.E. Daniels, Exact saddlepoint approximations. *Biometrika* **67**(1), 59–63 (1980)
- H.E. Daniels, Tail probability approximations. *Int. Stat. Rev./Revue Internationale de Statistique* 37–48 (1987)
- H. David, Concomitants of extreme order statistics, in *Extreme Value Theory and Applications* (Springer, 1994), pp. 211–224
- R.B. Davies, Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* **64**(2), 247–254 (1977)
- R.B. Davies, Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* **74**(1), 33–43 (1987)
- Z.-Q. Dignam, J. James, M. Kocherginsky, The use and interpretation of competing risks regression models. *Clin. Cancer Res.* **18**(8), 2301–2308 (2012)

- M. Donsker, An invariance principle for certain probability limit theorems. *Mem. Am. Math. Soc.* **6**, 1–10 (1951)
- N.R. Draper, The Box-Wetz criterion versus R2. *J. R. Stat. Soc. Ser. A (General)* **147**(1), 100–103 (1984)
- N.R. Draper, Corrections: the Box-Wetz criterion versus R2. *J. R. Stat. Soc. Ser. A (General)* **148**(4), 357–357 (1985)
- L. Duchateau, P. Janssen, *The Frailty Model* (Springer Science & Business Media, 2007)
- B. Efron, Censored data and the bootstrap. *J. Am. Stat. Assoc.* **76**(374), 312–319 (1981a)
- B. Efron, Nonparametric estimates of standard error: the jackknife, the bootstrap and other methods. *Biometrika* **68**(3), 589–599 (1981b)
- B. Efron, Nonparametric standard errors and confidence intervals. *Can. J. Stat.* **9**(2), 139–158 (1981c)
- B. Efron, Better bootstrap confidence intervals. *J. Am. Stat. Assoc.* **82**(397), 171–185 (1987)
- B. Efron, D.V. Hinkley, Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher information. *Biometrika* **65**, 457–483 (1978)
- B. Efron, C. Stein, The jackknife estimate of variance. *Ann. Stat.* 586–596 (1981)
- B. Efron et al., The geometry of exponential families. *Ann. Stat.* **6**(2), 362–376 (1978)
- S.S. Ellenberg, J.M. Hamilton, Surrogate endpoints in clinical trials: cancer. *Stat. Med.* **8**(4), 405–413 (1989)
- K.H. Eng, M.R. Kosorok, A sample size formula for the supremum log-rank statistic. *Biometrics* **61**, 86–91 (2005)
- M. Evans, N. Hastings, B. Peacock, *Statistical Distributions* (2001)
- G. Fang, W. Liu, L. Wang, A machine learning approach to select features important to stroke prognosis. *Comput. Biol. Chem.* 107316 (2020)
- P. Feigl, M. Zelen, Estimation of exponential survival probabilities with concomitant information. *Biometrics* **21**(4), 826–838 (1965)
- J.P. Fine, R.J. Gray, A proportional hazards model for the subdistribution of a competing risk. *J. Am. Stat. Assoc.* **94**(446), 496–509 (1999)

- J.P. Fine, H. Jiang, R. Chappell, On semi-competing risks data. *Biometrika* **88**(4), 907–919 (2001)
- R.A. Fisher, Two new properties of mathematical likelihood. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, vol. 144(852), pp. 285–307 (1934)
- E. Fix, J. Neyman, A simple stochastic model of recovery, relapse, death and loss of patients. *Hum. Biol.* **23**(3), 205–241 (1951)
- P. Flandre, J. O'Quigley, A two-stage procedure for survival studies with surrogate endpoints. *Biometrics* 969–976 (1995)
- P. Flandre, J. O'Quigley, Comparing Kaplan–Meier curves with delayed treatment effects: applications in immunotherapy trials. *J. R. Stat. Soc. Ser. C (Appl. Stat.)* (2019)
- P. Flandre, J. O'Quigley, The short-term and long-term hazard ratio model: parameterization inconsistency. *Am. Stat.* 1–7 (2020)
- P. Flandre, Y. Saidi, Estimating the proportion of treatment effect explained by a surrogate marker by DY Lin, TR Fleming and V. De Gruttola. *Stat. Med.* **16**, 1515–1527 (1997). *Stat. Med.* **18**(1), 107–109 (1999)
- T.R. Fleming, D.P. Harrington, *Counting Processes and Survival Analysis* (Wiley, New York, 1991)
- T.R. Fleming, D.P. Harrington, *Counting Processes and Survival Analysis*, 2nd edn. (Wiley, New York, 2005)
- T.R. Fleming, D.P. Harrington, Evaluation of censored survival data test procedures based on single and multiple statistics, in *Topics in Applied Statistics* (Marcel Dekker, New York, 1984), pp. 97–123
- T.R. Fleming, D.P. Harrington, M. O'Sullivan, Supremum versions of the log-rank and generalized Wilcoxon statistics. *J. Am. Stat. Assoc.* **82**, 312–320 (1987)
- T.R. Fleming, J.R. O'Fallon, P.C. O'Brien, D.P. Harrington, Modified Kolmogorov-Smirnov test procedures with application to arbitrarily right-censored data. *Biometrics* **36**, 607–625 (1980)
- L.S. Freedman, B.I. Graubard, A. Schatzkin, Statistical validation of intermediate endpoints for chronic diseases. *Stat. Med.* **11**(2), 167–178 (1992)
- B. Freidlin, E.L. Korn, Testing treatment effects in the presence of competing risks. *Stat. Med.* **24**(11), 1703–1712 (2005)

- J.E. Freund, A bivariate extension of the exponential distribution. *J. Am. Stat. Assoc.* **56**(296), 971–977 (1961)
- A. Gaddah, R. Braekers, Weak convergence for the conditional distribution function in a koziol-green model under dependent censoring. *J. Stat. Plann. Infer.* **139**(3), 930–943 (2009)
- U. Gather, J. Pawlitschko, Estimating the survival function under a generalized koziol-green model with partially informative censoring. *Metrika* **48**(3), 189–207 (1998)
- E. Gehan, A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika* **52**, 203–223 (1965)
- R.B. Geskus, Cause-specific cumulative incidence estimation and the fine and gray model under both left truncation and right censoring. *Biometrics* **67**(1), 39–49 (2011)
- D. Ghosh, Semiparametric inferences for association with semi-competing risks data. *Stat. Med.* **25**(12), 2059–2070 (2006)
- R. Gill, Censoring and stochastic integrals. *Mathematical Center Tract* **124** (1980)
- R. Gill, M. Schumacher, A simple test of the proportional hazards assumption. *Biometrika* **74**(2), 289–300 (1987)
- S.M. Gore, S.J. Pocock, G.R. Kerr, Regression models and non-proportional hazards in the analysis of breast cancer survival. *J. R. Stat. Soc. Ser. C (Appl. Stat.)* **33**(2), 176–195 (1984)
- P.M. Grambsch, T.M. Therneau, Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika* **81**(3), 515–526 (1994)
- R. Gray, Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *J. Am. Stat. Assoc.* **87**, 942–951 (1992)
- M. Greenwood, *A Report on the Natural Duration of Cancer. Reports of Public Health and Related Subjects*, vol. 33 (HMSO, London, 1926)
- R.G. Gutierrez, Parametric frailty and shared frailty survival models. *Stata J.* **2**(1), 22–44 (2002)
- M. Gönen, G. Heller, Concordance probability and discriminatory power in proportional hazards regression. *Biometrika* **92**, 965–970 (2005)
- P. Haara, A note on the asymptotic behaviour of the empirical score in Cox's regression model for counting processes, in *Proceedings of the 1st World Congress of the Bernoulli Society* (VNU Science Press, Tashkent, Soviet Union, 1987), pp. 139–142

- B. Haller, G. Schmidt, K. Ulm, Applying competing risks regression models: an overview. *Lifetime Data Anal.* **19**(1), 33–58 (2013)
- D.D. Hanagal, *Modeling Survival Data Using Frailty Models* (Chapman and Hall, CRC, 2011)
- S. Haneuse, K.H. Lee, Semi-competing risks data analysis: accounting for death as a competing risk when the outcome of interest is nonterminal. *Circ. Cardiovasc. Qual. Outcomes* **9**(3), 322–331 (2016)
- T.E. Hanson , A. Jara, L. Zhao, A Bayesian semi-parametric temporally stratified proportional hazards model with spatial frailties. *Bayesian Anal.* **7**, 147–188 (2012)
- D.P. Harrington, T.R. Fleming, A class of rank test procedures for censored survival data. *Biometrika* **69**(3), 553–566 (1982)
- T. Hastie, R. Tibshirani, Exploring the nature of covariate effects in the proportional hazards model. *Biometrics* **46**, 1005–1016 (1990)
- M. Healy, The use of R2 as a measure of goodness of fit. *J. R. Stat. Soc. Ser. A (General)* **147**(4), 608–609 (1984)
- I. Helland, Central limit theorems for martingales with discrete or continuous time. *Scand. J. Stat.* **9**, 79–94 (1982)
- R. Henderson, P. Diggle, A. Dobson, Joint modelling of longitudinal measurements and event time data. *Biostatistics* **1**(4), 465–480 (2000)
- J. Herson, The use of surrogate endpoints in clinical trials (an introduction to a series of four papers). *Stat. Med.* **8**(4), 403–404 (1989)
- A. Hillis, D. Seigel, Surrogate endpoints in clinical trials: ophthalmologic disorders. *Stat. Med.* **8**(4), 427–430 (1989)
- N.L. Hjort, On inference in parametric survival data models. *Int. Stat. Rev./Revue Internationale de Statistique* 355–387 (1992)
- F.S. Hodi, V. Chiarion-Silenti, R. Gonzalez, J.-J. Grob, P. Rutkowski, C.L. Cowey, C.D. Lao, D. Schadendorf, J. Wagstaff, R. Dummer et al., Nivolumab plus ipilimumab or nivolumab alone versus ipilimumab alone in advanced melanoma (checkmate 067): 4-year outcomes of a multicentre, randomised, phase 3 trial. *Lancet Oncol.* **19**(11), 1480–1492 (2018)
- F. Hsieh, Y.-K. Tseng, J.-L. Wang, Joint modeling of survival and longitudinal data: likelihood approach revisited. *Biometrics* **62**(4), 1037–1043 (2006)

- J.-J. Hsieh, Y.-T. Huang, Regression analysis based on conditional likelihood approach under semi-competing risks data. *Lifetime Data Anal.* **18**(3), 302–320 (2012)
- X. Huang, L. Liu, A joint frailty model for survival and gap times between recurrent events. *Biometrics* **63**(2), 389–397 (2007)
- R. Huang, R. Xu, P.S. Dulai, Sensitivity analysis of treatment effect to unmeasured confounding in observational studies with survival and competing risks outcomes. *Stat. Med.* 1–15 (2020)
- J. Hyde, Testing survival under right censoring and left truncation. *Biometrika* **64**(2), 225–230 (1977)
- H. Jiang, J.P. Fine, R. Chappell, Semiparametric analysis of survival data with left truncation and dependent right censoring. *Biometrics* **61**(2), 567–575 (2005)
- M.P. Jones, J. Crowley, A general class of nonparametric tests for survival analysis. *Biometrics* **45**, 157–170 (1989)
- M.P. Jones, J. Crowley, Asymptotic properties of a general class of nonparametric tests for survival analysis. *Ann. Stat.* **18**, 1203–1220 (1990)
- J.D. Kalbfleisch, R.L. Prentice, Marginal likelihood based on Cox's regression and life model. *Biometrika* **60**, 267–278 (1973)
- J.D. Kalbfleisch, R.L. Prentice, *The Statistical Analysis of Failure Time Data* (Wiley, 2002)
- E.L. Kaplan, P. Meier, Nonparametric estimation from incomplete observations. *J. Am. Stat. Assoc.* **53**(282), 457–481 (1958)
- R. Kay, Proportional hazard regression models and the analysis of censored survival data. *J. R. Stat. Soc. Ser. C (Appl. Stat.)* **26**(3), 227–237 (1977)
- N. Keiding, Historical controls and modern survival analysis. *Lifetime Data Anal.* **1**(1), 19–25 (1995)
- N. Keiding, Event history analysis and inference from observational epidemiology. *Stat. Med.* **18**(17–18), 2353–2363 (1999)
- N. Keiding, T. Bayer, S. Watt-Boolsen, Confirmatory analysis of survival data using left truncation of the life times of primary survivors. *Stat. Med.* **6**(8), 939–944 (1987)
- N. Keiding, R.D. Gill, Random truncation models and Markov processes. *Ann. Stat.* 582–602 (1990)

- M.G. Kendall, A. Stuart, J.K. Ord, S.F. Arnold, Kendall's advanced theory of statistics (1987)
- J.T. Kent, J. O'Quigley, Measures of dependence for censored survival data. *Biometrika* **75**(3), 525–534 (1988)
- H.T. Kim, Cumulative incidence in competing risks data and competing risks regression analysis. *Clin. Cancer Res.* **13**(2), 559–565 (2007)
- K. Kim, A.A. Tsiatis, Study duration for clinical trials with survival response and early stopping rule. *Biometrics* 81–92 (1990)
- J.P. Klein, M.L. Moeschberger, *Survival Analysis Techniques for Censored and Truncated Data* (Springer, 2003)
- J.P. Klein, S.-C. Lee, M. Moeschberger, A partially parametric estimator of survival in the presence of randomly censored data. *Biometrics* 795–811 (1990)
- A.N. Kolmogorov, *Foundations of the Theory of Probability: Second English Edition* (2018)
- M.R. Kosorok, C.Y. Lin, The versatility of function-indexed weighted log-rank statistics. *J. Am. Stat. Assoc.* **94**, 320–332 (1999)
- J.A. Koziol, S.B. Green, A Cramer-von Mises statistic for randomly censored data. *Biometrika* **63**(3), 465–474 (1976)
- J.A. Koziol, J.-Y. Zhang, C.A. Casiano, X.-X. Peng, F.-D. Shi, A.C. Feng, E.K. Chan, E.M. Tan, Recursive partitioning as an approach to selection of immune markers for tumor diagnosis. *Clin. Cancer Res.* **9**(14), 5120–5126 (2003)
- T.O. Kvaalseth, Cautionary note about R2. *Am. Stat.* **39**(4), 279–285 (1985)
- S. Lagakos, D. Schoenfeld, Properties of proportional-hazards score tests under misspecified regression models. *Biometrics* 1037–1048 (1984)
- S. Lagakos, The graphical evaluation of explanatory variables in proportional hazards regression models. *Biometrika* **68**, 93–98 (1981)
- S. Lagakos, The loss in efficiency from misspecifying covariates in proportional hazards regression models. *Biometrika* **75**(1), 156–160 (1988)
- S.W. Lagakos, A stochastic model for censored-survival data in the presence of an auxiliary variable. *Biometrics* 551–559 (1976)
- S.W. Lagakos, Using auxiliary variables for improved estimates of survival time. *Biometrics* 399–404 (1977)
- S.W. Lagakos, L.L. Kim, J.M. Robins, Adjusting for early treatment termination in comparative clinical trials. *Stat. Med.* **9**, 1417–1424 (1990)

- S.W. Lagakos, C.J. Sommer, M. Zelen, Semi-Markov models for partially censored data. *Biometrika* **65**(2), 311–317 (1978)
- M.G. Larson, G.E. Dinse, A mixture model for the regression analysis of competing risks data. *J. R. Stat. Soc. Ser. C (Appl. Stat.)* **34**(3), 201–211 (1985)
- B. Lau, S.R. Cole, S.J. Gange, Competing risk regression models for epidemiologic data. *Am. J. Epidemiol.* **170**(2), 244–256 (2009)
- M. LeBlanc, J. Crowley, Relative risk trees for censored survival data. *Biometrics* 411–425 (1992)
- J. Lee, Some versatile tests based on the simultaneous use of weighted log-rank statistics. *Biometrics* **52**, 721–725 (1996)
- K.H. Lee, S. Haneuse, D. Schrag, F. Dominici, Bayesian semi-parametric analysis of semi-competing risks data: investigating hospital readmission after a pancreatic cancer diagnosis. *J. R. Stat. Soc. Ser. C, Appl. Stat.* **64**(2), 253 (2015)
- S.-H. Lee, Maximum of the weighted Kaplan-Meier tests for the two-sample censored data. *J. Stat. Comput. Simul.* **81**, 1017–1026 (2011)
- E.L. Lehmann et al., The power of rank tests. *Ann. Math. Stat.* **24**(1), 23–43 (1953)
- E. Lenglart, Relation de domination entre deux processus. *Annales de l'Institut Henri Poincaré* **13**, 171–179 (1977)
- S. Leurgans, Three classes of censored data rank tests: strengths and weaknesses under censoring. *Biometrika* **70**, 651–658 (1983)
- S. Leurgans, Asymptotic behavior of two-sample rank tests in the presence of random censoring. *Ann. Stat.* **12**, 572–589 (1984)
- Y. Li, L. Ryan, Modeling spatial survival data using semiparametric frailty models. *Biometrics* **58**(2), 287–297 (2002)
- D. Lin, Goodness-of-fit analysis for the Cox regression model based on a class of parameter estimators. *J. Am. Stat. Assoc.* **86**, 153–180 (1991)
- D. Lin, Cox regression analysis of multivariate failure time data: the marginal approach. *Stat. Med.* **13**(21), 2233–2247 (1994)
- D. Lin, J. Robins, L. Wei, Comparing two failure time distributions in the presence of dependent censoring. *Biometrika* **83**, 381–393 (1996)
- D. Lin, L. Wei, Z. Ying, Checking the Cox model with cumulative sums of martingale based residuals. *Biometrika* **80**, 557–572 (1993)

- D.Y. Lin, Z. Ying, Semiparametric analysis of the additive risk model. *Biometrika* **81**(1), 61–71 (1994)
- D.Y. Lin, Transformation models, in *Handook of Survival Analysis* (Chapman & Hall/CRC, 2013)
- D.Y. Lin, L.-J. Wei, The robust inference for the cox proportional hazards model. *J. Am. Stat. Assoc.* **84**(408), 1074–1078 (1989)
- D.Y. Lin, Z. Ying, Semiparametric analysis of general additive-multiplicative intensity for counting processes. *Ann. Stat.* **23**, 1712–1734 (1995)
- A. Linden, P.R. Yarnold, Modeling time-to-event (survival) data using classification tree analysis. *J. Eval. Clin. Pract.* **23**(6), 1299–1308 (2017)
- C.L. Link, Confidence intervals for the survival function using Cox's proportional-hazard model with covariates. *Biometrics* 601–609 (1984)
- L. Liu, R.A. Wolfe, X. Huang, Shared frailty models for recurrent events and a terminal event. *Biometrics* **60**(3), 747–756 (2004)
- W.-Y. Loh, Classification and regression trees. *Wiley Interdisc. Rev. Data Min. Knowl. Disc.* **1**(1), 14–23 (2011)
- W.-Y. Loh, Fifty years of classification and regression trees. *Int. Stat. Rev.* **82**(3), 329–348 (2014)
- M. Lunn, D. McNeil, Applying Cox regression to competing risks. *Biometrics* 524–532 (1995)
- H.M. Malani, A modification of the redistribution to the right algorithm using disease markers. *Biometrika* **82**(3), 515–526 (1995)
- N. Mantel, Chi-square tests with one degree of freedom; extensions of the Mantel-Haenszel procedure. *J. Am. Stat. Assoc.* **58**(303), 690–700 (1963)
- N. Mantel, Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother. Rep.* **50**, 163–70 (1966)
- N. Mantel, W. Haenszel, Statistical aspects of the analysis of data from retrospective studies of disease. *J. Natl. Cancer Inst.* **22**(4), 719–748 (1959)
- N. Mantel, D.M. Stablein, The crossing hazard function problem. *The Statistician* **37**, 59–64 (1988)
- T. Martinussen, L. Peng, Alternatives to the Cox model, in *Handook of Survival Analysis*, ed. by J.P. Klein, H.C. van Houwelingen, J.G. Ibrahim, T.H. Scheike (Chapman & Hall/CRC, 2013)

- T. Martinussen, T.H. Scheike, *Dynamic Regression Models for Survival Data* (Springer, 2006)
- I.W. McKeague, P.D. Sasieni, A partly parametric additive risk model. *Biometrika* **81**, 501–514 (1994)
- I.W. McKeague, K.J. Utikal, Inference for a nonlinear counting process regression model. *Ann. Stat.* **18**, 1172–1187 (1990)
- M.L. Moeschberger, J.P. Klein, A comparison of several methods of estimating the survival function when there is extreme right censoring. *Biometrics* **41**(1), 253–259 (1985)
- M.L. Moeschberger, K.P. Tordoff, N. Kochhar, A review of statistical analyses for competing risks. *Handb. Stat.* **27**, 321–341 (2007)
- S.H. Moolgavkar, E.T. Chang, H.N. Watson, E.C. Lau, An assessment of the cox proportional hazards regression model for epidemiologic studies. *Risk Anal.* **38**(4), 777–794 (2018)
- T. Moreau, J. O'Quigley, M. Mesbah, A global goodness-of-fit statistic for the proportional hazards model. *J. R. Stat. Soc. Ser. C* **34**(3), 212–218 (1985)
- S. Murphy, Testing for a time-dependent coefficient in Cox's regression model. *Scand. J. Stat.* **20**, 35–50 (1993)
- S.A. Murphy, P.K. Sen, Time-dependent coefficients in a Cox-type regression model. *Stoch. Process. Appl.* **39**, 153–180 (1991)
- S. Murray, A.A. Tsiatis, Sequential methods for comparing years of life saved in the two sample censored data problem. *Biometrics* **55** (1999)
- S. Murray, A.A. Tsiatis, Using auxiliary time-dependent covariates to recover information in nonparametric testing with censored data. *Lifetime Data Anal.* **7** (2001)
- S. Murray, A.A. Tsiatis, Nonparametric survival estimation using prognostic longitudinal covariates. *Biometrics* 137–151 (1996)
- M.H. Myers, B.F. Hankey, N. Mantel, A logistic-exponential model for use with response-time data involving regressor variables. *Biometrics* 257–269 (1973)
- L. Natarajan, J. O'Quigley, Predictive capability of stratified proportional hazards models. *J. Appl. Stat.* **29**(8), 1153–1163 (2002)
- W. Nelson, Hazard plotting for incomplete failure data. *J. Qual. Technol.* **1**(1), 27–52 (1969)

- M.A. Newton, A.E. Raftery, Approximate Bayesian inference with the weighted likelihood bootstrap. *J. R. Stat. Soc. Ser. B (Methodological)* **56**(1), 3–26 (1994)
- D. Oakes, An approximate likelihood procedure for censored data. *Biometrics* 177–182 (1986)
- P.C. O'Brien, A nonparametric test for association with censored data. *Biometrics* 243–250 (1978)
- P.C. O'Brien, T.R. Fleming, A paired Prentice-Wilcoxon test for censored paired data. *Biometrics* 169–180 (1987)
- J. O'Quigley, Regression models and survival prediction. *Statistician* **31**, 106–116 (1982)
- J. O'Quigley, Confidence intervals for the survival function in the presence of covariates. *Biometrics* **42**(1), 219–220 (1986)
- J. O'Quigley, On a two-sided test for crossing hazards. *J. R. Stat. Soc. Ser. D (The Statistician)* **43**(4), 563–569 (1994)
- J. O'Quigley, Khmaladze-type graphical evaluation of the proportional hazards assumption. *Biometrika* **90**, 577–584 (2003)
- J. O'Quigley, *Proportional Hazards Regression* (Springer, New York, 2008)
- J. O'Quigley, Faulty brca1, brca2 genes: how poor is the prognosis? *Anna. Epidemiol.* **27**(10), 672–676 (2017)
- J. O'Quigley, P. Flandre, Predictive capability of proportional hazards regression. *Proc. Natl. Acad. Sci.* **91**(6), 2310–2314 (1994)
- J. O'Quigley, P. Flandre, Quantification of the prentice criteria for surrogate endpoints. *Biometrics* **62**(1), 297–300 (2006)
- J. O'Quigley, T. Moreau, Testing the proportional hazards regression model against some general alternatives. *Revue d'épidémiologie et de santé publique* **32**(3–4), 199–205 (1984)
- J. O'Quigley, L. Natarajan, Erosion of regression effect in a survival study. *Biometrics* **60**(2), 344–351 (2004)
- J. O'Quigley, F. Pessione, Score tests for homogeneity of regression effect in the proportional hazards model. *Biometrics* **45**, 135–144 (1989)
- J. O'Quigley, F. Pessione, The problem of a covariate-time qualitative interaction in a survival study. *Biometrics* **47**, 101–15 (1991)

- J. O'Quigley, R.L. Prentice, Nonparametric tests of association between survival time and continuously measured covariates: the logit-rank and associated procedures. *Biometrics* **117**–127 (1991)
- J. O'Quigley, J. Stare, Proportional hazards models with frailties and random effects. *Stat. Med.* **21**(21), 3219–3233 (2002)
- J. O'Quigley, R. Xu, Explained variation in proportional hazards regression, in *Handbook of Statistics in Clinical Oncology*, ed. by J. Crowley, A. Hoering (Marcel Dekker, New York, 2001), pp. 397–410
- J. O'Quigley, R. Xu, Goodness of fit in survival analysis. *Wiley StatsRef: Statistics Reference Online* (2014)
- J. O'Quigley, R. Xu, J. Stare, Explained randomness in proportional hazards models. *Stat. Med.* **24**(3), 479–489 (2005)
- M. Peckova, T.R. Fleming, Adaptive test for testing the difference in survival distributions. *Lifetime Data Anal.* **9**, 223–238 (2003)
- M.J. Pencina, R.B. D'Agostino, R.S. Vasan, Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat. Med.* **27**, 157–172 (2008)
- M.S. Pepe, T.R. Fleming, Weighted Kaplan-Meier statistics: a class of distance tests for censored survival data. *Biometrics* **45**, 497–507 (1989)
- M.S. Pepe, J. Fan, Z. Feng, T. Gerdts, J. Hilden, The net reclassification index (nri): a misleading measure of prediction improvement even with independent test data sets. *Stat. Biosci.* **7**(2), 282–295 (2015)
- M.P. Perme, J. Stare, J. Estève, On estimation in relative survival. *Biometrics* **68**(1), 113–120 (2012)
- R. Peto, J. Peto, Asymptotically efficient rank invariant test procedures. *J. R. Stat. Soc. Ser. A* **135**, 185–207 (1972)
- E. Pitman, Nonparametric statistical inference. *Lecture Notes* (Institute of Statistics, University of North Carolina, 1948)
- D.N. Politis, Computer-intensive methods in statistical analysis. *IEEE Signal Process. Mag.* **15**(1), 39–55 (1998)
- R.L. Prentice, Linear rank tests with right censored data. *Biometrika* **65**, 167–179 (1978)
- R.L. Prentice, Surrogate endpoints in clinical trials: definition and operational criteria. *Stat. Med.* **8**(4), 431–440 (1989)

- R.L. Prentice, J.D. Kalbfleisch, Hazard rate models with covariates. *Biometrics* **25**–39 (1979)
- H. Putter, M. Fiocco, R.B. Geskus, Tutorial in biostatistics: competing risks and multi-state models. *Stat. Med.* **26**(11), 2389–2430 (2007)
- E.R. Brown, J.G. Ibrahim, A Bayesian semiparametric joint hierarchical model for longitudinal and survival data. *Biometrics* **59**(2), 221–228 (2003)
- R.H. Randles, D.A. Wolfe, Introduction to the theory of nonparametric statistics. *Introduction to the Theory of Nonparametric Statistics*, by Randles, Ronald H.; Wolfe, Douglas A. Wiley, New York, c1979. Wiley series in probability and mathematical statistics (1979)
- C.R. Rao, C.R. Rao, M. Statistiker, C.R. Rao, C.R. Rao, *Linear Statistical Inference and Its Applications*, tome 2 (Wiley, New York, 1973)
- S. Rashid, J. O'Quigley, A. Axon, E. Cooper, Plasma protein profiles and prognosis in gastric cancer. *Br. J. Cancer* **45**(3), 390 (1982)
- R. Rebollo, Sur les applications de la théorie des martingales à l'étude statistique d'une famille de processus ponctuels. Dans *Journées de Statistique des Processus Stochastiques* (Springer, 1978), pp. 27–70
- R. Rebollo, Central limit theorems for local martingales. *Z. Wahrsch. Verw. Gebiete* **51**, 269–286 (1980)
- D. Rizopoulos, *Joint Models for Longitudinal and Time-to-Event Data: With Applications in R*. (Chapman and Hall/CRC, 2012)
- D. Rizopoulos, G. Molenberghs, E.M.E.H. Lesaffre, Dynamic predictions with time-dependent covariates in survival analysis using joint modeling and landmarking. *Biometrical J.* **59**(6), 1261–1276 (2017)
- J. Robinson, Saddlepoint approximations for permutation tests and confidence intervals. *J. R. Stat. Soc. Ser. B* **44**, 91–101 (1982)
- V.K. Rohatgi, A.K.M.E. Saleh, *An Introduction to Probability and Statistics* (Wiley, 2015)
- P.S. Rosenberg, W.F. Anderson, Proportional hazards models and age-period-cohort analysis of cancer rates. *Stat. Med.* **29**(11), 1228–1238 (2010)
- S. Ross, *Stochastic Processes*, 2nd edn. (Wiley, New York, 1996)
- D.B. Rubin, The Bayesian bootstrap. *Ann. Stat.* 130–134 (1981)
- P. Sasieni, On the expected number of cancer deaths during follow-up of an initially cancer-free cohort. *Epidemiology* **14**(1), 108–110 (2003)

- P. Sasieni, A. Winnett, Martingale difference residuals as a diagnostic tool for the Cox model. *Biometrika* **90**, 899–912 (2003)
- J.M. Satagopan, L. Ben-Porat, M. Berwick, M. Robson, D. Kutler, A.D. Auerbach, A note on competing risks in survival data analysis. *Br. J. Cancer* **91**(7), 1229–1235 (2004)
- T.H. Scheike, T. Martinussen, On estimation and tests of time-varying effects in the proportional hazards model. *Scand. J. Stat.* **31**, 51–62 (2004)
- T.H. Scheike, M. Zhang, An additive-multiplicative Cox-Aalen regression model. *Scand. J. Stat.* **29**, 79–92 (2002)
- T.H. Scheike, M.-J. Zhang, Flexible competing risks regression modeling and goodness-of-fit. *Lifetime Data Anal.* **14**(4), 464 (2008)
- D. Schoenfeld, Chi-squared goodness-of-fit tests for the proportional hazards regression model. *Biometrika* **67**, 145–153 (1980)
- D. Schoenfeld, The asymptotic properties of nonparametric tests for comparing survival distributions. *Biometrika* **68**(1), 316–319 (1981)
- D. Schoenfeld, Partial residuals for the proportional hazards regression model. *Biometrika* **69**, 239–241 (1982)
- M. Schumacher, Two-sample tests of Cramer-von Mises and Kolmogorov-Smirnov-type for randomly censored data. *Int. Stat. Rev.* **62**, 263–281 (1984)
- M. Schumacher, E. Graf, T. Gerdts, How to assess prognostic models for survival data: a case study in oncology. *Methods Inf. Med.* **42**(05), 564–571 (2003)
- A. Scott, C. Wild, Transformations and R 2. Am. Stat. **45**(2), 127–129 (1991)
- S. Self, An adaptive weighted log-rank test with application to cancer prevention and screening trials. *Biometrics* **47**, 975–86 (1991)
- R.J. Serfling, *Approximation Theorems of Mathematical Statistics*, tome 162 (Wiley, 2009)
- A.A. Seyerle, C.L. Avery, Genetic epidemiology: the potential benefits and challenges of using genetic information to improve human health. N. C. Med. J. **74**(6), 505–508 (2013)
- P.-S. Shen, An inverse-probability-weighted approach to estimation of the bivariate survival function under left-truncation and right-censoring. J. Stat. Plann. Infer. **136**(12), 4365–4384 (2006)
- Y. Shen, J. Cai, Maximum of the weighted Kaplan-Meier tests with application to cancer prevention and screening trials. *Biometrics* **57**, 837–843 (2001)

- A. Shenk, *Calculus and Analytic Geometry* (Santa Monica, 1979)
- E.V. Slud, L.V. Rubinstein, Dependent competing risks and summary survival curves. *Biometrika* **70**(3), 643–649 (1983)
- X. Song, C.Y. Wang, Semiparametric approaches for joint modeling of longitudinal and survival data with time-varying coefficients. *Biometrics* **64**(2), 557–566 (2008)
- D. Stablein, W. Carter, J. Novax, Analysis of survival data with non-proportional hazard functions. *Control. Clin. Trials* **2**, 149–159 (1981)
- C.A. Struthers, J.D. Kalbfleisch, Misspecified proportional hazards model. *Biometrika* **73**, 363–369 (1986)
- W. Stute, Strong consistency under the koziol-green model. *Stat. Prob. Lett.* **14**(4), 313–320 (1992)
- W. Stute, The central limit theorem under random censorship. *Ann. Stat.* **23**, 422–439 (1995)
- R.E. Tarone, On the distribution of the maximum of the log-rank statistic and the modified Wilcoxon statistic. *Biometrics* **37**, 79–85 (1981)
- R.E. Tarone, J.H. Ware, On distribution-free tests for equality for survival distributions. *Biometrika* **64**, 156–160 (1977)
- T.M. Therneau, P.M. Grambsch, *Modeling Survival Data: Extending the Cox Model* (Springer, New York, 2000)
- T. Therneau, P. Grambsch, T. Fleming, Martingale-based residuals for survival models. *Biometrika* **77**(1), 147–160 (1990)
- P.K. Trivedi, D.M. Zimmer, *Copula Modeling: An Introduction for Practitioners* (Now Publishers Inc., 2007)
- W.-Y. Tsai, N.P. Jewell, M.-C. Wang, A note on the product-limit estimator under right censoring and left truncation. *Biometrika* **74**(4), 883–886 (1987)
- A. Tsiatis, A nonidentifiability aspect of the problem of competing risks. *Proc. Natl. Acad. Sci.* **72**(1), 20–22 (1975)
- A. Tsiatis, Competing risks. *Encycl. Biostat.* **2** (2005)
- A.A. Tsiatis, Group sequential methods for survival analysis with staggered entry. *Lect. Notes Monogr. Ser.* **2**, 257–268 (1982)
- A.A. Tsiatis, M. Davidian, Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica* 809–834 (2004)

- C. Van Eeden, The relation between pitman's asymptotic relative efficiency of two tests and the correlation coefficient between their test statistics. *Ann. Math. Stat.* **34**(4), 1442–1451 (1963)
- J.W. Vaupel, K.G. Manton, E. Stallard, The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography* **16**(3), 439–454 (1979)
- D.J. Venzon, S.H. Moolgavkar, Origin-invariant relative risk functions for case-control and survival studies. *Biometrika* **75**(2), 325–333 (1988)
- L. Wei, Testing goodness-of-fit for proportional hazards model with censored observations. *J. Am. Stat. Assoc.* **79**, 649–652 (1984)
- L. Wei, The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis. *Stat. Med.* **11**, 1871–1879 (1992)
- A. Wienke, *Frailty Models in Survival Analysis* (CRC Press, 2010)
- F. Wilcoxon, Individual comparisons by ranking methods. *Biometrics Bull.* **1**, 80–83 (1945)
- J.B. Willett, J.D. Singer, Another cautionary note about R 2: its use in weighted least-squares regression analysis. *Am. Stat.* **42**(3), 236–238 (1988)
- J. Wittes, E. Lakatos, J. Probstfield, Surrogate endpoints in clinical trials: cardiovascular diseases. *Stat. Med.* **8**(4), 415–425 (1989)
- L. Wu, P.B. Gilbert, Flexible weighted log-rank tests optimal for detecting early and/or late survival differences. *Biometrics* **58**, 997–1004 (2002)
- L. Xu, C. Gotwalt, Y. Hong, C.B. King, W.Q. Meeker, Applications of the fractional-random-weight bootstrap. *Am. Stat.* 1–21 (2020)
- J. Xu, J.D. Kalbfleisch, B. Tai, Statistical analysis of illness-death processes and semicompeting risks data. *Biometrics* **66**(3), 716–725 (2010)
- R. Xu, *Inference for the proportional hazards model*. Thèse de doctorat, University of California, San Diego (1996)
- R. Xu, S. Adak, Survival analysis with time-varying regression effects using a tree-based approach. *Biometrics* **58**(2), 305–315 (2002)
- R. Xu, J. O'Quigley, Proportional hazards estimate of the conditional survival function. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **62**(4), 667–680 (2000)
- R. Xu, J. O'Quigley, Estimating average regression effect under non-proportional hazards. *Biostatistics* **1**(4), 423–439 (2000)

- X. Xue, M.Y. Kim, M.M. Gaudet, Y. Park, M. Heo, A.R. Hollenbeck, H.D. Strickler, M.J. Gunter, A comparison of the polytomous logistic regression and joint cox proportional hazards models for evaluating multiple disease subtypes in prospective cohort studies. *Cancer Epidemiol. Prev. Biomark.* **22**(2), 275–285 (2013)
- S. Yang, R. Prentice, Semiparametric analysis of short term and long term relative risks with two sample survival data. *Biometrika* **92**, 1–17 (2005)
- S. Yang, R. Prentice, Improved logrank-type tests for survival data using adaptive weights. *Biometrics* **66**, 30–38 (2010)
- M. Yu, J.M.G. Taylor, H.M. Sandler, Individual prediction in prostate cancer studies using a joint longitudinal survival-cure model. *J. Am. Stat. Assoc.* **103**(481), 178–187 (2008)
- M.J. Zhang, X. Zhang, J. Fine, A proportional hazards regression model for the subdistribution with right-censored and left-truncated competing risks data. *Stat. Med.* **30**(16), 1933–1951 (2011)
- B. Zhou, A. Latouche, V. Rocha, J. Fine, Competing risks regression for stratified data. *Biometrics* **67**(2), 661–670 (2011)
- M. Zhou, G. Li, Empirical likelihood analysis of the Buckley-James estimator. *J. Multivar. Anal.* **99**(4), 649–664 (2008)
- C. Zippin, P. Armitage, Use of concomitant variables and incomplete survival information in the estimation of an exponential survival parameter. *Biometrics* 665–672 (1966)
- D.M. Zucker, E. Lakatos, Weighted log rank type statistics for comparing survival curves when there is a time lag in the effectiveness of treatment. *Biometrika* **77**, 853–864 (1990)

Index

A

Approximations, 371

Cornish-Fisher approximation, 177,
372

Saddlepoint approximation, 177, 374

Arcsine test, 322

At-risk indicator, 19, 25

Repeated events, 26

Average value of $\beta(T)$, 165

Axioms of probability, 24, 354

B

Brownian bridge

Covariance process, 381

Brownian motion, 408

Maximum of process, 383

Probability results, 385

Process with drift, 384

Properties of drift process, 384

Brownian motion based tests

Brownian bridge, 312

C

Censoring, 34

Competing risks, 38

Conditionally independent censoring, 37, 202

Dependent censoring, 207

Finite censoring support, 37

Independent censoring, 37

Informative censoring, 36, 204

Koziol-Green model, 36

Left censoring; truncation, 8

Type I censoring, 35

Type II censoring, 35

Type III censoring, 36

Central limit theorems

Dependent variables, 404

Functional central limit theorem,
408

I.I.D. variables, 402

Independent variables, 403

Linear sums of ranks, 404

Stochastic integrals, 396

Sums of random variables, 405

Change-point models, 286

Characteristic function, 365

Clinical trials, 2

Coding of groups, 87

Coding variables, 90

Competing risks, 38, 391

Conditional logistic regression, 104

Conditional survivorship function, 22

Convergence, 355

Convergence in distribution, 355

Functions, 356

Counting processes, 385, 389, 391

Adapted process, 392

Filtrations, 389

Predictability, 391

Covariate distribution

Continuous, 91

Moments, 179

Cox regression model, 78

- H**
- Historical background, 86
 - Cramer-Rao inequality, 414
 - Cumulative generating function, 365
 - Cumulative hazard transform, 362
- D**
- DeMoivre-Laplace approximation, 366
 - Densities, 358
 - Density of a sum, 359
 - Distributions, 358
 - Difference of random variables, 359
 - Sums of random variables, 359
- E**
- Empirical distribution function, 353
 - Empirical estimates
 - Model verification, 69
 - Epidemiology, 97
 - Estimating equations, 141, 413
 - Censoring, 170
 - Distribution of solution, 178
 - Large sample properties, 169
 - Marginal survival, 51
 - Maximum likelihood, 50, 417
 - Method of moments, 416
 - Minimum chi-square, 415
 - Misspecified models, 164
 - Moments, 156
 - Newton-Raphson iteration, 51
 - Non-proportional hazards, 169
 - Proportional hazards, 153
 - Regularity conditions, 418
 - Relative risk models, 159
 - Residuals, 174
 - Semi-parametric, 151
 - Small samples, 176
 - Stratified models, 158
 - Expectation, 364
- F**
- Finance and insurance, 5
 - Frailty model, 9
 - Freireich data, 53, 147
- G**
- Glivenko-Cantelli theorem, 57
 - Goodness of fit, 278
 - Goodness-of-fit tests, 69
 - Graphical methods, 263
 - Greenwood's formula, 211
- H**
- Hazard and related functions, 21
 - Hazard function, 390
 - Helly-Bray theorem, 152, 353, 416
 - High dimensional sparse data, 203
 - Hypothesis tests, 301
 - Area above curve, 320
 - Area under curve, 315
 - Combined tests, 306
 - Concave alternatives, 328
 - Convex alternatives, 329
 - Delayed effect, 330
 - Diminishing effect, 331
 - Distance traveled, 310
 - Integrated log-rank, 318
 - Kolmogorov type tests, 312
 - Log-rank test, 303
 - Maximum statistics, 307
 - Non-responders, 332
 - Reflected Brownian motion, 313
 - Restrictive adaptive, 323
 - Supremum over cutpoints, 335
 - Weighted Kaplan-Meier, 308
 - Weighted log-rank, 304
- I**
- Individual risks, 116
 - Integrals
 - Lebesgue, 388
 - Riemann-Stieltjes, 388
 - Intensity functions, 23
 - Compartment models, 23
- J**
- Jensen's inequality, 364
 - Joint survival-covariate model, 12

K

- Kaplan-Meier estimate, 23, 56, 61, 62
Continuous version, 62
Greenwood's formula, 63
Mean and Median, 67
Precision, 63
Redistribution to the right, 66
Transformations, 63
Variance, 63, 67

L

- Large sample theory, 226
Law of iterated logarithm, 411
Learning and classification, 9
Likelihood
 Conditional, 426
 Cox's conditional likelihood, 429
 Exponential model, 145
 Marginal, 430
 Nonparametric exponential, 149
 Parametric models, 143
 Partial, 10, 428
Linear models
 Additive, 136
 Transformation models, 137
Log-minus-log transformation, 33
Logistic regression, 102

M

- Mantel-Haenszel estimate, 101
Marginal survival, 49
 Kaplan-Meier estimator, 49
Martingales, 385, 392
 Compensator, 393
 Doob-Meyer decomposition, 393
 Predictable variation process, 394
 Stochastic integrals, 386, 392, 395
Mean residual lifetime, 22
Mean value theorem, 352
Model building, 294
Multistate models, 25
Multivariate normal distribution, 360

N

- Nelson-Aalen estimate, 68
Non-proportional hazards, 77, 119, 161
Normal distribution, 358
 Mill's ratio, 361

O

- Observed information, 52
Order statistics, 366
 Distribution of difference, 367
 Expected values, 370
 Joint distribution, 367
 Markov property, 368
 Maximum of sample, 366
 Minimum of sample, 366
 Normal parent, 370

P

- Parametric goodness-of-fit tests, 150
Permutation test, 209
Predictive indices, 12
 Interpretation, 14
Probability integral transform, 181, 361
Probability that $T_i > T_j$, 193
Prognostic biomarkers, 199
Proportional hazards, 78
 Applications in epidemiology, 98
 Average effect, 172
 Changepoint models, 131
 Cox model, 79
 Explained variation, 270
 Models with intercept, 129
 Partial, 120
 Predictive ability, 270
 Random effects, frailties, 124
 Stratified models, 121
 Time-dependent effects, 131

R

- Random variables, 354
Registry data, 107
Regression effect process, 215
 Concave effects, 288
 Confidence bands, 267

- Sample size, 220, 235
 Several covariates, 231
 Time scale, 219
 Regression models, 75, 97, 119
 Relative efficiency, 334
 Relative survival, 107, 194
 Resampling methods, 422
 Accuracy, 424
 Bootstrap confidence intervals, 423
 Bootstrap, 422
 Empirical bootstrap, 422
 Studentized bootstrap, 424
 Riemann integral, 352
 Riemann-Stieltjes integral, 353
 Rolle's theorem, 352
- S**
- Saddlepoint approximation, 366
 Stochastic processes, 377
 Brownian bridge, 379, 380
 Brownian motion, 378
 Gaussian processes, 378
 Independent increments, 377
 Integrated Brownian motion, 382
 Ornstein-Uhlenbeck process, 380
 Reflected Brownian motion, 383
 Stationarity, 377
 Time-transformed Brownian motion, 380
 Sums of random variables, 402
 Weighted sums, 404
 Central limit theorem for dependent variables, 404
 Central limit theorem for i.i.d. variables, 402
 Central limit theorem for independent variables, 403
 De Moivre-Laplace, 402
 Empirical distribution function, 410
 Functional central limit theorem, 405
 Lindeberg condition, 403
 Sums on (0,1), 405
- Surrogate endpoints, 205
 Survival function
 Covariate information, 192
 Empirical estimate, 56, 58
 Estimation, 52, 197
 estimation, 145
 Exponential model, 52
 Given $Z \in H$, 195
 Kaplan-Meier estimate, 58
 Piecewise exponential, 54
 Survival models, 28
 Exponential, 28
 Extreme value, 33
 Freund model, 207
 Gompertz, 33
 Lagakos model, 207
 Log-normal, 34
 Nonparametric exponential, 70
 Parametric proportional hazards, 34
 Piecewise exponential, 30
 Proportional hazards exponential, 29
 Proportional hazards piecewise exponential, 31
 Proportional hazards Weibull, 32
 Random effects model, 201
 Stratified model, 201
 Weibull, 31
- T**
- Transformed covariate models, 160
 Two stage designs, 206
- X**
- Xu-O'Quigley estimator, 205
- Y**
- Yang and Prentice model, 138