Statistics for Biology and Health

John O'Quigley

Proportional Hazards Regression

Springer

Statistics for Biology and Health

John O'Quigley

Proportional Hazards Regression

Springer

Statistics for Biology and Health

Statistics for Biology and Health

Bacchieri/Cioppa: Fundamentals of Clinical Research
Borchers/Buckland/Zucchini: Estimating Animal Abundance: Closed Populations
Burzykowski/Molenberghs/Buyse: The Evaluation of Surrogate Endpoints
Duchateau/Janssen: The Frailty Model
Everitt/Rabe-Hesketh: Analyzing Medical Data Using S-PLUS
Ewens/Grant: Statistical Methods in Bioinformatics: An Introduction, 2nd ed.
Gentleman/Carey/Huber/Irizarry/Dudoit: Bioinformatics and Computational Biology Solutions Using R and Bioconductor
Hougaard: Analysis of Multivariate Survival Data
Keyfitz/Caswell: Applied Mathematical Demography, 3rd ed.
Klein/Moeschberger: Survival Analysis: Techniques for Censored and Truncated Data, 2nd ed.
Kleinbaum/Klein: Survival Analysis: A Self-Learning Text, 2nd ed.
Kleinbaum/Klein: Logistic Regression: A Self-Learning Text, 2nd ed.
Lange: Mathematical and Statistical Methods for Genetic Analysis, 2nd ed.
Manton/Singer/Suzman: Forecasting the Health of Elderly Populations
Martinussen/Scheike: Dynamic Regression Models for Survival Data
Moyé: Multiple Analyses in Clinical Trials: Fundamentals for Investigators
Nielsen: Statistical Methods in Molecular Evolution
O'Quigley: Proportional Hazards Regression
Parmigiani/Garrett/Irizarry/Zeger: The Analysis of Gene Expression Data: Methods and Software
Proschan/LanWittes: Statistical Monitoring of Clinical Trials: A Unified Approach
Siegmund/Yakir: The Statistics of Gene Mapping
Simon/Korn/McShane/Radmacher/Wright/Zhao: Design and Analysis of DNA Microarray Investigations
Sorensen/Gianola: Likelihood, Bayesian, and MCMC Methods in Quantitative Genetics
Stallard/Manton/Cohen: Forecasting Product Liability Claims: Epidemiology and Modeling in the Manville Asbestos Case
Sun: The Statistical Analysis of Interval-censored Failure Time Data
Therneau/Grambsch: Modeling Survival Data: Extending the Cox Model
Ting: Dose Finding in Drug Development
Vittinghoff/Glidden/Shiboski/McCulloch: Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models
Wu/Ma/Casella: Statistical Genetics of Quantitative Traits: Linkage, Maps, and QTL
Zhang/Singer: Recursive Partitioning in the Health Sciences
Zuur/Ieno/Smith: Analysing Ecological Data

John O'Quigley

Proportional Hazards Regression

Springer

John O'Quigley
Institut Curie
26 rue d'Ulm
75005 Paris, France

ISBN 978-0-387-25148-6 e-ISBN 978-0-387-68639-4
DOI: 10.1007/978-0-387-68639-4

Library of Congress Control Number: 2008920048

Printed on acid-free paper.

9 8 7 6 5 4 3 2 1

springer.com

To Máire Proinsias

Preface

Proportional hazards models and their extensions (models with time-dependent covariates, models with time dependent regression coefficients, 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 statistician the ability to tackle the thorniest of questions which may arise when dealing with survival data.

The main goal of this text 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 (prediction, goodness of fit, model construction, inference and interpretation in the presence of misspecified models) from the standpoint of the proportional hazards and the non-proportional hazards models.

This standpoint is essentially mathematical in that the aim is to put all of the inferential questions on a firm conceptual footing. However, unlike the current widely accepted approach based almost entirely on counting processes, stochastic integrals and the martingale central limit theorem for multivariate counting processes, we mostly work with much more classic 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 particular interest, the functional central limit theorem establishing the Brownian motion limit for standardized univariate sums. Delicate measure theoretic arguments borrowed from mathematical analysis can then be wholly avoided. Admittedly, some tricky situations can still be more readily resolved by making an appeal to the martingale central limit theorem and it may also be true that the use of martingale techniques for multivariate counting processes affords greater generality. Nonetheless, in the author's view, a very high percentage of practical problems can be tackled by making an appeal to the arguably less general but more standard and well known central limit theorems for univariate sums.

Mathematicians always strive for the greatest generality and, in the author's view - at least as far as survival analysis is concerned - this undertaking has not been without some unfortunate drawbacks. The measure theoretic underpinning of the counting processes and martingale approach is quite opaque. The subject is very difficult and while outstanding efforts have been made across the globe in leading statistics and biostatistics departments to explain the essential ideas behind the material, few would claim that students, other than the small minority already well steeled in mathematical analysis, ever really fully grasp just what is going on. This is a situation that we need be concerned about. The author, having taught such courses in a number of institutions, speculates that even for the most successful students entering research careers and publishing articles in our leading journals it is not easy for them to do other than reiterate well rehearsed near inscrutable arguments. Their work, reviewed by their peers - alumni survivors of similar courses - may clear the publishing hurdle and achieve technical excellence but somehow, along the way, creativity is stifled.

In brief, there is a real danger of the subject of survival analysis sustaining itself from within and reluctant to absorb input from without. The pressure to focus so much attention on the resolution of mathematical subtleties by non-mathematicians has led us away from those areas where we have traditionally done well ... abstract modeling of medical, biological, physical and social phenomena. The technical demands of those in the area of survival analysis are such that it is becoming difficult for them to construct or challenge models via considerations other than those pertaining to the correct application of

an abstruse theory. Those not in the area, a large and diverse pool of potential contributors, will typically throw up their hands and say that they are not sufficiently comfortable with survival analysis to make criticism of substance. A somewhat ambitious goal, or hope, for this text is to help change this state of affairs.

Work on this book began during the author's thesis. I would like to acknowledge the input of Dr Salah Rashid, a visiting surgeon to the University of Leeds Medical School, for asking a lot of awkward questions to a, then, very inexperienced statistician. Among these questions were 'how much of the variation is explained by the predictors', 'why would you assume that the strength of effect remains the same through time' and 'what is the relative importance of biological measurements to clinical measurements.' I believe that I can now attempt an answer to some of these questions although I fear, taking rather longer than expected to answer the clinician's concern - in this case some twenty odd years - that my good friend Dr Rashid may have moved on to other questions. Much of my thesis was based on collaborative work with Dr Rashid and his comments and questions then, and for decades to follow, provided an invaluable source of food for thought. I share the debt we all owe to Professor Sir David Cox for his great vision and scientific imagination, making all of this work possible, but also a personal debt to Sir David for having so very kindly agreed to be the external examinor on my own Ph.d thesis and for having patiently explained issues which, alone, I was unable to resolve.

My career at the Institut National de la Santé et de la Recherche Médicale in France was made possible thanks to the unfailing support of Professor Daniel Schwartz, one of the founders of the modern theory of clinical trials and I thank him warmly for that as well as for numerous discussions on parametric survival models, especially as they relate to problems in human fertility. A number of Professor Schwartz's colleagues, Joseph Lellouch, Denis Hémon, Alfred Spira in particular, were of great assistance to me in gaining understanding of the role played by survival analysis in quantitative epidemiology. However, my good fortune did not end there and I would like to offer my warm appreciation for the support, help and advice offered by Ross Prentice in inviting me to work at the Fred Hutchinson Cancer Research Center, Seattle during the late eighties, an opportunity which brought me into contact with a remarkable number of major contributors to the area of survival analysis. Among these I would like to express my gratitude to

Ross himself alongside Norman Breslow, John Crowley, Tom Fleming, Suresh Moolgavkar, Margaret Pepe and Steve Self, all of whom showed great generosity and forbearance in discussing general concepts along with their own ideas on different aspects of survival analysis.

Competing with Seattle as the world's leader in survival analysis is the Department of Biostatistics in the Harvard School of Public Health. I was given the opportunity of spending several months there in 1999 and would like to thank Nan Laird, the then department chair, for that. Many visits and collaborations followed and, although these were not in the area of survival analysis, I took advantage of the proximity to talk to those who have left quite a mark in the field. In particular I would like to offer my thanks to Victor DeGruttola, Dave Harrington, Michael Hughes, Steve Lagakos, David Schoenfeld, L.J. Wei, Marvin Zelen, all of whom, from very demanding schedules, gave time to the exchange of ideas.

I have been very fortunate in coming into contact with quite a number of the most creative researchers in this area, so many of whom have shown such scholarly patience in explaining their own views to a keen, always enthusiastic but often slow listener. I hesitate to list them since I will surely miss out some names and then, of course, if I were to credit all of the writings which have greatly helped me, this preface would take a good third of the whole book. But let me include a special mention for Janez Stare and Ronghui Xu who will recognize within these pages much which stems from our extensive collaborations. And, as an extension to this special mention, let me thank those colleagues whose kindness, as well as exceptional talent, helped provide part of a hard-to-define support structure without which this enduring task would most likely have been abandoned many years ago. I have in mind Jacques Bénichou, Claude Chastang (whose colorful view of statistics as well as life in general is so sorely missed), Michel Chavance, Philippe Flandre, Catherine Hill, Joe Ibrahim, Richard Kay, John Kent, Susanne May, Thierry Moreau, Loki Natarajan, Fabienne Pessione, Maja Pohar, Catherine Quantin, Peter Sassieni, Michael Schemper, Martin Schumacher, Lesley Struthers and Joe Whittaker. The many students who followed my course on Survival Analysis in the Department of Mathematics at the University of California at San Diego, the course upon which the skeleton of this book ended up being based, are sincerely thanked for their enthusiasm and obstinate questioning.

Finally, although the very word statistics, let alone proportional hazards regression, would leave them quite at a loss, this work owes its greatest debt to those closest to me - my nearest and dearest - for a contribution which involved untold patience and tender indulgence. My warmest gratitude goes to them.

Paris, February 2007.

Contents