

Use R!

Series Editors:

Robert Gentleman Kurt Hornik Giovanni Parmigiani

More information about this series at <http://www.springer.com/series/6991>

Use R!

Wickham: ggplot2 (2nd ed. 2016)

Luke: A User's Guide to Network Analysis in R

Monogan: Political Analysis Using R

Cano/M. Moguerza/Prieto Corcoba: Quality Control with R

Schwarzer/Carpenter/Rücker: Meta-Analysis with R

Gondro: Primer to Analysis of Genomic Data Using R

Chapman/Feit: R for Marketing Research and Analytics

Willekens: Multistate Analysis of Life Histories with R

Cortez: Modern Optimization with R

Kolaczyk/Csardi: Statistical Analysis of Network Data with R

Swenson/Nathan: Functional and Phylogenetic Ecology in R

Nolan/Temple Lang: XML and Web Technologies for Data Sciences with R

Nagarajan/Scutari/Lèbre: Bayesian Networks in R

van den Boogaart/Tolosana-Delgado: Analyzing Compositional Data with R

Bivand/Pebesma/Gómez-Rubio: Applied Spatial Data

Analysis with R (2nd ed. 2013)

Eddelbuettel: Seamless R and C++ Integration with Rcpp

Knoblauch/Maloney: Modeling Psychophysical Data in R

Lin/Shkedy/Yekutieli/Amaratunga/Bijnens: Modeling Dose-Response Microarray

Data in Early Drug Development

Experiments Using R

Cano/M. Moguerza/Redchuk: Six Sigma with R

Soetaert/Cash/Mazzia: Solving Differential Equations in R

Dirk F. Moore

Applied Survival Analysis Using R

 Springer

Dirk F. Moore
Department of Biostatistics
Rutgers School of Public Health
Piscataway, NJ, USA

ISSN 2197-5736

Use R!

ISBN 978-3-319-31243-9

DOI 10.1007/978-3-319-31245-3

ISSN 2197-5744 (electronic)

ISBN 978-3-319-31245-3 (eBook)

Library of Congress Control Number: 2016940055

© Springer International Publishing Switzerland 2016

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, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG Switzerland

To Lynne, Molly, and Emily

Preface

This book serves as an introductory guide for students and analysts who need to work with survival time data. The minimum prerequisites are basic applied courses in linear regression and categorical data analysis. Students who also have taken a master's level course in statistical theory will be well prepared to work through this book, since frequent reference is made to maximum likelihood theory. Students lacking this training may still be able to understand most of the material, provided they have an understanding of the basic concepts of differential and integral calculus. Specifically, students should understand the concept of the limit, and they should know what derivatives and integrals are and be able to evaluate them in some basic cases.

The material for this book has come from two sources. The first source is an introductory class in survival analysis for graduate students in epidemiology and biostatistics at the Rutgers School of Public Health. Biostatistics students, as one would expect, have a much firmer grasp of more mathematical aspects of statistics than do epidemiology students. Still, I have found that those epidemiology students with strong quantitative backgrounds have been able to understand some mathematical statistical procedures such as score and likelihood ratio tests, provided that they are not expected to symbolically differentiate or integrate complex formulas. In this book I have, when possible, used the numerical capabilities of the R system to substitute for symbolic manipulation. The second source of material is derived from collaborations with physicians and epidemiologists at the Rutgers Cancer Institute of New Jersey and at the Rutgers Robert Wood Johnson Medical School. A number of the data sets in this text are derived from these collaborations. Also, the experience of training statistical analysts to work on these data sets provided additional inspiration for the book.

The first chapter introduces the concepts of survival times and how right censoring occurs and describes several of the datasets that will be used throughout the book. Chapter 2 presents fundamentals of survival theory. This includes hazard, probability density, survival functions, and how they are related. The hazard function is illustrated using both life table data and using some common parametric distributions. The chapter ends with a brief introduction to properties of maximum

likelihood estimates using the exponential distribution as an illustration. Chapter 3 discusses the Kaplan-Meier estimate of the survival function and several related concepts such as the median survival and its confidence interval. Also discussed in this chapter are smoothing of the hazard function and how to accommodate left truncation into the Kaplan-Meier estimate.

Chapter 4 discusses the log-rank test for comparing survival distributions and also some modified linear rank tests. Stratified tests are also discussed, along with an example where stratification can reverse the apparent direction of a treatment effect in a survival example of Simpson's paradox. In Chapter 5, we present the Cox proportional hazards model and partial likelihood function in the context of comparing two groups of survival data. There we illustrate the Wald, score, and likelihood ratio tests in this basic context. Left-truncated survival data and the partial likelihood are also discussed.

Chapter 6 presents methods for model selection and extends and illustrates the proportional hazards model in situations where there are multiple possible predictor covariates. Chapter 7 presents diagnostic residual plots that are useful for assessing model assumptions. Chapter 8 discusses how to adapt the survival models discussed earlier to allow for time-dependent covariates.

The next few chapters discuss some important special situations. Chapter 9 discusses multiple outcomes, which can occur as clustered survival times or in a competing risks framework, where only the first of multiple outcomes can be observed. Chapter 10 discusses parametric survival models, and Chapter 11 covers the critically important design question of how to determine the power and sample size of a proposed study that has a survival outcome. Finally, Chapter 12 presents some additional topics, including the piecewise exponential distribution, methods for handling interval censoring, and the lasso method for handling survival data with large numbers of predictors. Many of the data sets discussed in the text are available in the accompanying R package "asaur" (for "Applied Survival Analysis Using R"), while others are in other packages. All are freely available for download from the Central R Archive Network at cran.r-project.org. The R-code discussed in the book is available for download at <http://www.springer.com/us/book/9783319312439>

A key feature of this book is the integration of the R statistical system with the survival analysis material. Not only do we show the reader how to use R functions to fit survival models and how to interpret the results, but we also use R to illustrate how survival quantities are computed. Typically we use small examples to illustrate in detail how one constructs survival tests, partial likelihood models, and diagnostics and then proceed to more complicated examples. Most of the survival functions will require that the "survival" library be attached using the "library(survival)" statement. The "survival" package is included by default; other packages referred to in the text must be explicitly downloaded and installed. The appendix includes both some basics of the R language and special features relevant to the survival calculations used elsewhere in the book. Users not already familiar with the R system should refer to one of the many online resources for more detailed information.

I would like to thank Rebecca Moss for permission to use the “pancreatic” data and Michael Steinberg for permission to use the “pharmacoSmoking” data. Both of these data sets are used repeatedly throughout the text. I would also like to thank Grace Lu-Yao, Weichung Joe Shih, and Yong Lin for years-long collaborations on using the SEER-Medicare data for studying the survival trajectories of prostate cancer patients. These collaborations led to the development of the “prostateSurvival” data set discussed in this text in Chapter 9. I thank the Division of Cancer Epidemiology and Genetics of the US National Cancer Institute for providing the “asheknazi” data. I also thank Wan Yee Lau for making the “hepatoCellular” data publically available in the online Dryad data repository and for allowing me to include it in the “asaur” R package.

Piscataway, NJ, USA
October 2015

Dirk F. Moore

Contents

1	Introduction	1
1.1	What Is Survival Analysis?	1
1.2	What You Need to Know to Use This Book	2
1.3	Survival Data and Censoring	2
1.4	Some Examples of Survival Data Sets	6
1.5	Additional Notes	9
2	Basic Principles of Survival Analysis	11
2.1	The Hazard and Survival Functions	11
2.2	Other Representations of a Survival Distribution	13
2.3	Mean and Median Survival Time	14
2.4	Parametric Survival Distributions	15
2.5	Computing the Survival Function from the Hazard Function	19
2.6	A Brief Introduction to Maximum Likelihood Estimation	20
2.7	Additional Notes	23
3	Nonparametric Survival Curve Estimation	25
3.1	Nonparametric Estimation of the Survival Function	25
3.2	Finding the Median Survival and a Confidence Interval for the Median	30
3.3	Median Follow-Up Time	32
3.4	Obtaining a Smoothed Hazard and Survival Function Estimate ...	32
3.5	Left Truncation	36
3.6	Additional Notes	41
4	Nonparametric Comparison of Survival Distributions	43
4.1	Comparing Two Groups of Survival Times	43
4.2	Stratified Tests	49
4.3	Additional Note	52

5	Regression Analysis Using the Proportional Hazards Model	55
5.1	Covariates and Nonparametric Survival Models.....	55
5.2	Comparing Two Survival Distributions Using a Partial Likelihood Function	56
5.3	Partial Likelihood Hypothesis Tests.....	59
5.3.1	The Wald Test.....	60
5.3.2	The Score Test	60
5.3.3	The Likelihood Ratio Test.....	60
5.4	The Partial Likelihood with Multiple Covariates	63
5.5	Estimating the Baseline Survival Function.....	64
5.6	Handling of Tied Survival Times	65
5.7	Left Truncation	69
5.8	Additional Notes	71
6	Model Selection and Interpretation	73
6.1	Covariate Adjustment	73
6.2	Categorical and Continuous Covariates	74
6.3	Hypothesis Testing for Nested Models.....	78
6.4	The Akaike Information Criterion for Comparing Non-nested Models.....	81
6.5	Including Smooth Estimates of Continuous Covariates in a Survival Model	84
6.6	Additional Note	86
7	Model Diagnostics	87
7.1	Assessing Goodness of Fit Using Residuals	87
7.1.1	Martingale and Deviance Residuals	87
7.1.2	Case Deletion Residuals.....	92
7.2	Checking the Proportion Hazards Assumption	94
7.2.1	Log Cumulative Hazard Plots	94
7.2.2	Schoenfeld Residuals.....	96
7.3	Additional Note	100
8	Time Dependent Covariates	101
8.1	Introduction.....	101
8.2	Predictable Time Dependent Variables	106
8.2.1	Using the Time Transfer Function	107
8.2.2	Time Dependent Variables That Increase Linearly with Time	109
8.3	Additional Note	110
9	Multiple Survival Outcomes and Competing Risks	113
9.1	Clustered Survival Times and Frailty Models.....	113
9.1.1	Marginal Survival Models.....	115
9.1.2	Frailty Survival Models	116
9.1.3	Accounting for Family-Based Clusters in the “ashkenazi” Data	117

- 9.1.4 Accounting for Within-Person Pairing of Eye
Observations in the Diabetes Data 120
- 9.2 Cause-Specific Hazards 121
 - 9.2.1 Kaplan-Meier Estimation with Competing Risks 121
 - 9.2.2 Cause-Specific Hazards and Cumulative
Incidence Functions 123
 - 9.2.3 Cumulative Incidence Functions for Prostate
Cancer Data 126
 - 9.2.4 Regression Methods for Cause-Specific Hazards 127
 - 9.2.5 Comparing the Effects of Covariates on
Different Causes of Death 130
- 9.3 Additional Notes 134
- 10 Parametric Models 137**
 - 10.1 Introduction 137
 - 10.2 The Exponential Distribution 138
 - 10.3 The Weibull Model 138
 - 10.3.1 Assessing the Weibull Distribution as a Model
for Survival Data in a Single Sample 138
 - 10.3.2 Maximum Likelihood Estimation of Weibull
Parameters for a Single Group of Survival Data 141
 - 10.3.3 Profile Weibull Likelihood 142
 - 10.3.4 Selecting a Weibull Distribution to Model
Survival Data 143
 - 10.3.5 Comparing Two Weibull Distributions Using
the Accelerated Failure Time and Proportional
Hazards Models 146
 - 10.3.6 A Regression Approach to the Weibull Model 148
 - 10.3.7 Using the Weibull Distribution to Model
Survival Data with Multiple Covariates 149
 - 10.3.8 Model Selection and Residual Analysis with
Weibull Survival Data 151
 - 10.4 Other Parametric Survival Distributions 153
 - 10.5 Additional Note 154
- 11 Sample Size Determination for Survival Studies 157**
 - 11.1 Power and Sample Size for a Single Arm Study 157
 - 11.2 Determining the Probability of Death in a Clinical Trial 161
 - 11.3 Sample Size for Comparing Two Exponential Survival
Distributions 163
 - 11.4 Sample Size for Comparing Two Survival Distributions
Using the Log-Rank Test 165
 - 11.5 Determining the Probability of Death
from a Non-parametric Survival Curve Estimate 166
 - 11.6 Example: Calculating the Required Number of Patients
for a Randomized Study of Advanced Gastric Cancer Patients 169

- 11.7 Example: Calculating the Required Number of Patients
for a Randomized Study of Patients with Metastatic
Colorectal Cancer 170
- 11.8 Using Simulations to Estimate Power 171
- 11.9 Additional Notes 174
- 12 Additional Topics 177**
 - 12.1 Using Piecewise Constant Hazards to Model Survival Data 177
 - 12.2 Interval Censoring 187
 - 12.3 The Lasso Method for Selecting Predictive Biomarkers 192
- Erratum E1**
- A A Basic Guide to Using R for Survival Analysis 201**
 - A.1 The R System 201
 - A.1.1 A First R Session 202
 - A.1.2 Scatterplots and Fitting Linear Regression Models 204
 - A.1.3 Accommodating Non-linear Relationships 207
 - A.1.4 Data Frames and the Search Path for Variable Names 209
 - A.1.5 Defining Variables Within a Data Frame 211
 - A.1.6 Importing and Exporting Data Frames 211
 - A.2 Working with Dates in R 212
 - A.2.1 Dates and Leap Years 213
 - A.2.2 Using the “as.date” Function 213
 - A.3 Presenting Coefficient Estimates Using Forest Plots 215
 - A.4 Extracting the Log Partial Likelihood and Coefficient
Estimates from a coxph Object 217
 - References 218
- Index 223**
- R Package Index 225**