# 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/Csàrdi:* 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/Gòmez-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 ISSN 2197-5744 (electronic)
Use R!
ISBN 978-3-319-31243-9 ISBN 978-3-319-31245-3 (eBook)
DOI 10.1007/978-3-319-31245-3

Library of Congress Control Number: 2016940055



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 Dirk F. Moore
October 2015

# Contents