Proposal for the Second Edition of Event History Analysis with R

Göran Broström

July 7, 2017

Contents

1 Introduction

I have been asked by Chapman & Hall to write a second edition of my book *Event History Analysis with R* (Broström, 2012). I have accepted, and this is a proposal for the second edition. It includes both updates as a consequence

of the development of the companion **R** (R Core Team, 2017) package `eha` (Broström, 2017) and new material. A general description of news is given in Section 2 and suggested updates are given in Section 3.

Finally, in Section 4, a tentative layout of Chapters and Sections in the new edition is given.

2 News

2.1 Reproducible research and RStudio

The importance of *reproducible research* is recognized today (Gandrud, 2015; Stodden et al., 2014). The topic will be introduced, tools for it recommended, and they will be used throughout the book. This includes using RStudio (RStudio Team, 2016) and Rmarkdown (Allaire et al., 2017).

2.2 Register-based data methods

I am working (well, professor emeritus) at CEDAR, Umeå University, Umeå, Sweden. CEDAR stands for *Centre for Demographic and Ageing Research* and it is a centre for both research and data base building and keeping. One of the largest data bases at CEDAR is the *Linnæus data base*, consisting of yearly individual information about the population of Sweden between the years 1986 and 2013. Sweden has about 9 million inhabitants today, so this data base is huge.

Several research projects utilizing the *Linnæus* data base are applying event history analysis one way or the other. They regularly encounter two typical problems:

1. Data are discrete in nature (observation of state one day per year), so performing *Cox regression* is not straightforward.

2. The huge amount of data makes standard survival analysis programs choke.

I intend to show how to handle these aspects with register-based data. The R package eha (Broström, 2017) will be developed accordingly.

2.3 Reducing huge data sets by statistical sufficiency

The sufficiency principle can sometimes be used to reduce the size of a data set to a fixed-dimension data set (a table). This will speed up compilation considerably given that certain conditions are met. This is something to implement in the eha package in the first place, but a description of how it works will be included in the book as well.

2.4 The weird bootstrap

The weird bootstrap (Andersen et al., 1993, p. 323) is implemented in the package eha, but not explained in the first edition of the book. It will now be introduced.

2.5 Presentation

Presentation refers to how to present the results of a statistical investigation, for instance of a Cox regression. My view is that graphical presentations should be promoted and tables kept to a minimum, so I will investigate how certain kinds of table content can be translated into a graph.

2.5.1 The Hauck-Donner effect, p-values, and contrasts

Traditional ways of presenting the results a Cox or AFT regression (or any regression) are neglecting the multiple testing problem, which is especially serious in connection with contrasts for categorical covariates. I will show the correct way(s) of doing things.

See http://capa.ddb.umu.se/ds/contrasts.html for a hint of what I have in mind here.

2.5.2 Tables

The issue is what information should go into a table: I am especially critical to the way p-values are presented in standard layouts. One interpretation of a p-value is that it is just another statistic, computable from data. That is fine, but problem arises when the p-value is interpreted as a probability and statements about "significance" are made.

2.6 Cox regression

This is more about a changing focus of the eha package: The non-parametric approach (Cox regression) is well covered by the survival package (Therneau, 2017), and the corresponding parts of the eha package will be toned down. Instead, things that eha can do but survival cannot will be emphasized.

However, this is more about changes in the package eha, and it will not change the presentation in the book significantly.

This will also be reflected in the second edition of the book.

2.7 Parametric survival models

2.7.1 The Gompertz baseline distribution

The Gompertz distribution is traditionally used in modeling human adult mortality, by good reasons. However, a "common knowledge" that *the Gom-*

pertz distribution does not lend itself to be used in AFT modeling (Kleinbaum and Klein, 2005, p. 285) is *not true*. On the contrary, the Gompertz distribution can be used both in PH and AFT modeling. This will be explained in the book, and the benefits from this will be pointed out. It includes the possibility to perform mediation analysis in a simple way because the AFT model is log-linear.

2.7.2 The Piecewise Constant Hazard model

This parametric model is close to a non-parametric approach (let the sizes of the pieces tend to zero ...) and is an excellent replacement for Cox regression with massive data sets: With categorical covariates data can be reduced to fixed-dimensional tables by sufficiency and the equivalence with Poisson regression can be utilized for very fast and accurate estimation. This will be implemented ih the eha package and described in the book. It should be noted that this is already possible to do with base R and some programming, ant that is what will be implemented and described.

2.7.3 Mediation analysis

Ways of performing mediation analysis in parametric survival models are given, especially in the accelerated failure time framework.

3 Updates

3.1 Writing the book

The first edition was written using Sweave (Leisch, 2002) and LaTeX (Lamport, 1994), but I consider switching to either knitr (Xie, 2015) and LaTeX or bookdown (Xie, 2016). For now, I will go with *bookdown*.

3.1.1 Graphics

In the first edition of the book (Broström, 2012), and in the current version of the eha package (Broström, 2017), the basic graphics system is used. I will investigate possible gains given by switching to ggplot2, lattice, or a mixture.

3.2 Updating wrt the R package eha

Since the first edition was published in 2012, the eha package has developed, and that will be picked up in the book.

3.3 Correct possible errors and mistakes

Readers and myself have spotted some errors and cryptic formulations over the years. Will of course be dealt with.

3.4 Improve examples and descriptions

The data used in the examples of the first edition will be checked. Old examples may be replaced by new.

4 Contents of the Second edition

Chapter 1: Event history and survival data

Remains essentially as is.

Chapter 2: Single sample data

Essentially as is, with the addition that Subsection 2.5. Doing it in **R** will introduce *RStudio*.

Chapter 3: Reproducible research and RStudio

New chapter: The important concept of *reproducible research* is introduced, and it is shown how *RStudio* can be assisting in applying it. *Rmarkdown* is introduced. These tools are used throughout the book.

Chapter 4: Cox regression

The old Chapter 3.

Chapter 5: Poisson regression

The old Chapter 4.

Chapter 6: More on Cox regression

The old Chapter 5.

Chapter 7: Parametric models

The old Chapter 6. The *Gompertz* baseline distribution will get an own subsection, as will the *Piecewise Constant Hazard* model.

Chapter 8: Register-based survival data methods

New Chapter.

Chapter 9: Multivariate survival models

The old Chapter 7.

Chapter 10: Causality and mediation

Replaces and expands the old Chapter 9.

Chapter 11: Competing risks models

The old Chapter 8.

Appendices

Essentially as is, but much have aged and will be rewritten.

References

Allaire, J., Cheng, J., Xie, Y., McPherson, J., Chang, W., Allen, J., Wickham, H., Atkins, A., Hyndman, R., and Arslan, R. (2017). *rmarkdown: Dynamic Documents for R*. R package version 1.6.

Andersen, P., Borgan, Ø., Gill, R., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, Berlin.

Broström, G. (2017). *eha: Event History Analysis*. R package version 2.4-5.

Broström, G. (2012). *Event History Analysis with R*. Chapman and Hall/CRC, Boca Raton, Florida. ISBN 978-1439831649.

Gandrud, C. (2015). *Reproducible Research with R and RStudio: Second Edition*. Chapman and Hall/CRC, Boca Raton, Florida. ISBN-13 987-1-4987-1537-9.

Kleinbaum, D. and Klein, M. (2005). *Survival Analysis: A Self-Learning Text*. Springer, Second edition.

Lamport, L. (1994). *LaTeX: A Document Preparation System (2nd Edition)*. Addison-Wesley, Reading, Mass.

Leisch, F. (2002). Sweave: Dynamic generation of statistical reports using literate data analysis. In Haerdle, W. and Roenz, B., editors, *Proceedings in Computational Statistics*, pages 575—-580. Physica Verlag. ISBN 3-7908-1517-9.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

RStudio Team (2016). *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA.

Stodden, V., Leisch, F., and Peng, R. D. (2014). *Implementing Reproducible Research*. Chapman and Hall/CRC, Boca Raton, Florida. ISBN 978-1-4665-6159-5.

Therneau, T. (2017). *survival: A package for survival analysis in S*. R package version 2.41-3.

Xie, Y. (2015). *Dynamic Documents with R and knitr*. Chapman and Hall/CRC, Boca Raton, Florida, 2nd edition. ISBN 978-1498716963.

Xie, Y. (2016). *bookdown: Authoring Books and Technical Documents with R Markdown*. Chapman and Hall/CRC, Boca Raton, Florida. ISBN 978-1138700109.