A Handbook of Statistical Analyses Using R

Brian S. Everitt and Torsten Hothorn

Preface

This book is intended as a guide to data analysis with the R system for statistical computing. R is an environment incorporating an implementation of the S programming language, which is powerful, flexible and has excellent graphical facilities (R Development Core Team, 2005). In the Handbook we aim to give relatively brief and straightforward descriptions of how to conduct a range of statistical analyses using R. Each chapter deals with the analysis appropriate for one or several data sets. A brief account of the relevant statistical background is included in each chapter along with appropriate references, but our prime focus is on how to use R and how to interpret results. We hope the book will provide students and researchers in many disciplines with a self-contained means of using R to analyse their data. R is an open-source project developed by dozens of volunteers for more than ten years now and is available from the Internet under the General Public Licence. R has become the *lingua franca* of statistical computing. Increasingly, implementations of new statistical methodology first appear as R add-on packages. In some communities, such as in bioinformatics, R already is the primary workhorse for statistical analyses. Because the sources of the R system are open and available to everyone without restrictions and because of its powerful language and graphical capabilities, R has started to become the main computing engine for reproducible statistical research (Leisch, 2002a,b, 2003, Leisch and Rossini, 2003, Gentleman, 2005). For a reproducible piece of research, the original observations, all data preprocessing steps, the statistical analysis as well as the scientific report form a unity and all need to be available for inspection, reproduction and modification by the readers. Reproducibility is a natural requirement for textbooks such as the 'Handbook of Statistical Analyses Using R' and therefore this book is fully reproducible using an R version greater or equal to 2.5.0. All analyses and results, including figures and tables, can be reproduced by the reader without having to retype a single line of R code. The data sets presented in this book are collected in a dedicated add-on package called *HSAUR* accompanying this book. The package can be installed from the Comprehensive R Archive Network (CRAN) via

```
R> install.packages("HSAUR")
```

and its functionality is attached by

```
R> library("HSAUR")
```

The relevant parts of each chapter are available as a *vignette*, basically a document including both the R sources and the rendered output of every

analysis contained in the book. For example, the first chapter can be inspected by

```
R> vignette("Ch_introduction_to_R", package = "HSAUR")
```

and the R sources are available for reproducing our analyses by

```
R> edit(vignette("Ch_introduction_to_R", package = "HSAUR"))
```

An overview on all chapter vignettes included in the package can be obtained from

```
R> vignette(package = "HSAUR")
```

We welcome comments on the R package *HSAUR*, and where we think these add to or improve our analysis of a data set we will incorporate them into the package and, hopefully at a later stage, into a revised or second edition of the book. Plots and tables of results obtained from R are all labelled as 'Figures' in the text. For the graphical material, the corresponding figure also contains the 'essence' of the R code used to produce the figure, although this code may differ a little from that given in the *HSAUR* package, since the latter may include some features, for example thicker line widths, designed to make a basic plot more suitable for publication. We would like to thank the R Development Core Team for the R system, and authors of contributed add-on packages, particularly Uwe Ligges and Vince Carey for helpful advice on *scatterplot3d* and *gee*. Kurt Hornik, Ludwig A. Hothorn, Fritz Leisch and Rafael Weißbach provided good advice with some statistical and technical problems. We are also very grateful to Achim Zeileis for reading the entire manuscript, pointing out inconsistencies or even bugs and for making many suggestions which have led to improvements. Lastly we would like to thank the CRC Press staff, in particular Rob Calver, for their support during the preparation of the book. Any errors in the book are, of course, the joint responsibility of the two authors.

Brian S. Everitt and Torsten Hothorn
London and Erlangen, December 2005

Bibliography

Gentleman, R. (2005), "Reproducible research: A bioinformatics case study," *Statistical Applications in Genetics and Molecular Biology*, 4, URL http://www.bepress.com/sagmb/vol4/iss1/art2, article 2.

Leisch, F. (2002a), "Sweave: Dynamic generation of statistical reports using literate data analysis," in *Compstat 2002 — Proceedings in Computational Statistics*, eds. W. Härdle and B. Rönz, Physica Verlag, Heidelberg, pp. 575–580, ISBN 3-7908-1517-9.

Leisch, F. (2002b), "Sweave, Part I: Mixing R and LaTeX," *R News*, 2, 28–31, URL http://CRAN.R-project.org/doc/Rnews/.

Leisch, F. (2003), "Sweave, Part II: Package vignettes," *R News*, 3, 21–24, URL http://CRAN.R-project.org/doc/Rnews/.

Leisch, F. and Rossini, A. J. (2003), "Reproducible statistical research," *Chance*, 16, 46–50.

R Development Core Team (2005), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, URL http://www.R-project.org, ISBN 3-900051-07-0.