Preface

R is a language and environment for data analysis and graphics. It may be considered an implementation of S, an award-winning language initially developed at Bell Laboratories since the late 1970s. The R project was initiated by Robert Gentleman and Ross Ihaka at the University of Auckland, New Zealand, in the early 1990s, and has been developed by an international team since mid-1997.

Historically, econometricians have favored other computing environments, some of which have fallen by the wayside, and also a variety of packages with canned routines. We believe that R has great potential in econometrics, both for research and for teaching. There are at least three reasons for this: (1) R is mostly platform independent and runs on Microsoft Windows, the Mac family of operating systems, and various flavors of Unix/Linux, and also on some more exotic platforms. (2) R is free software that can be downloaded and installed at no cost from a family of mirror sites around the globe, the Comprehensive R Archive Network (CRAN); hence students can easily install it on their own machines. (3) R is open-source software, so that the full source code is available and can be inspected to understand what it really does, learn from it, and modify and extend it. We also like to think that platform independence and the open-source philosophy make R an ideal environment for reproducible econometric research.

This book provides an introduction to econometric computing with R; it is not an econometrics textbook. Preferably readers have taken an introductory econometrics course before but not necessarily one that makes heavy use of matrices. However, we do assume that readers are somewhat familiar with matrix notation, specifically matrix representations of regression models. Thus, we hope the book might be suitable as a "second book" for a course with sufficient emphasis on applications and practical issues at the intermediate or beginning graduate level. It is hoped that it will also be useful to professional economists and econometricians who wish to learn R. We cover linear regression models for cross-section and time series data as well as the common nonlinear models of microeconometrics, such as logit, probit, and tobit

models, as well as regression models for count data. In addition, we provide a chapter on programming, including simulations, optimization, and an introduction to `Sweave()`—an environment that allows integration of text and code in a single document, thereby greatly facilitating reproducible research. (In fact, the entire book was written using `Sweave()` technology.)

We feel that students should be introduced to challenging data sets as early as possible. We therefore use a number of data sets from the data archives of leading applied econometrics journals such as the *Journal of Applied Econometrics* and the *Journal of Business & Economic Statistics*. Some of these have been used in recent textbooks, among them Baltagi (2002), Davidson and MacKinnon (2004), Greene (2003), Stock and Watson (2007), and Verbeek (2004). In addition, we provide all further data sets from Baltagi (2002), Franses (1998), Greene (2003), and Stock and Watson (2007), as well as selected data sets from additional sources, in an R package called **AER** that accompanies this book. It is available from the CRAN servers at `http://CRAN.R-project.org/` and also contains all the code used in the following chapters. These data sets are suitable for illustrating a wide variety of topics, among them wage equations, growth regressions, dynamic regressions and time series models, hedonic regressions, the demand for health care, or labor force participation, to mention a few.

In our view, applied econometrics suffers from an underuse of graphics—one of the strengths of the R system for statistical computing *and graphics*. Therefore, we decided to make liberal use of graphical displays throughout, some of which are perhaps not well known.

The publisher asked for a compact treatment; however, the fact that R has been mainly developed by statisticians forces us to briefly discuss a number of statistical concepts that are not widely used among econometricians, for historical reasons, including factors and generalized linear models, the latter in connection with microeconometrics. We also provide a chapter on R basics (notably data structures, graphics, and basic aspects of programming) to keep the book self-contained.

The production of the book

The entire book was typeset by the authors using LaTeX and R's `Sweave()` tools. Specifically, the final manuscript was compiled using R version 2.7.0, **AER** version 0.9-0, and the most current version (as of 2008-05-28) of all other CRAN packages that **AER** depends on (or suggests). The first author started under Microsoft Windows XP Pro, but thanks to a case of theft he switched to Mac OS X along the way. The second author used Debian GNU/Linux throughout. Thus, we can confidently assert that the book is fully reproducible, for the version given above, on the most important (single-user) platforms.

Settings and appearance

R is mainly run at its default settings; however, we found it convenient to employ a few minor modifications invoked by

```
R> options(prompt="R> ", digits=4, show.signif.stars=FALSE)
```

This replaces the standard R prompt > by the more evocative R>. For compactness, `digits = 4` reduces the number of digits shown when printing numbers from the default of 7. Note that this does not reduce the precision with which these numbers are internally processed and stored. In addition, R by default displays one to three stars to indicate the significance of p values in model summaries at conventional levels. This is disabled by setting `show.signif.stars = FALSE`.

Typographical conventions

We use a `typewriter` font for all code; additionally, function names are followed by parentheses, as in `plot()`, and class names (a concept that is explained in Chapters 1 and 2) are displayed as in "`lm`". Furthermore, boldface is used for package names, as in **AER**.

Acknowledgments

This book would not exist without R itself, and thus we thank the R Development Core Team for their continuing efforts to provide an outstanding piece of open-source software, as well as all the R users and developers supporting these efforts. In particular, we are indebted to all those R package authors whose packages we employ in the course of this book.

Several anonymous reviewers provided valuable feedback on earlier drafts. In addition, we are grateful to Rob J. Hyndman, Roger Koenker, and Jeffrey S. Racine for particularly detailed comments and helpful discussions. On the technical side, we are indebted to Torsten Hothorn and Uwe Ligges for advice on and infrastructure for automated production of the book. Regarding the accompanying package **AER**, we are grateful to Badi H. Baltagi, Philip Hans Franses, William H. Greene, James H. Stock, and Mark W. Watson for permitting us to include all the data sets from their textbooks (namely Baltagi 2002; Franses 1998; Greene 2003; Stock and Watson 2007). We also thank Inga Diedenhofen and Markus Hertrich for preparing some of these data in R format. Finally, we thank John Kimmel, our editor at Springer, for his patience and encouragement in guiding the preparation and production of this book. Needless to say, we are responsible for the remaining shortcomings.

May, 2008   Christian Kleiber, Basel
                                                    Achim Zeileis, Wien