The R Book

Michael J. Crawley

Imperial College London at Silwood Park, UK

John Wiley & Sons, Ltd
## Contents

*Preface* vii

1 Getting Started 1
2 Essentials of the R Language 9
3 Data Input 97
4 Dataframes 107
5 Graphics 135
6 Tables 183
7 Mathematics 195
8 Classical Tests 279
9 Statistical Modelling 323
10 Regression 387
11 Analysis of Variance 449
12 Analysis of Covariance 489
13 Generalized Linear Models 511
14 Count Data 527
15 Count Data in Tables 549
16 Proportion Data 569
17 Binary Response Variables 593
18 Generalized Additive Models 611
19 Mixed-Effects Models 627
20 Non-linear Regression 661
21 Tree Models 685
22 Time Series Analysis 701
23 Multivariate Statistics 731
24 Spatial Statistics 749
25 Survival Analysis 787
26 Simulation Models 811
27 Changing the Look of Graphics 827

*References and Further Reading* 873

*Index* 877
Preface

R is a high-level language and an environment for data analysis and graphics. The design of R was heavily influenced by two existing languages: Becker, Chambers and Wilks’ S and Sussman’s Scheme. The resulting language is very similar in appearance to S, but the underlying implementation and semantics are derived from Scheme. This book is intended as an introduction to the riches of the R environment, aimed at beginners and intermediate users in disciplines ranging from science to economics and from medicine to engineering. I hope that the book can be read as a text as well as dipped into as a reference manual. The early chapters assume absolutely no background in statistics or computing, but the later chapters assume that the material in the earlier chapters has been studied. The book covers data handling, graphics, mathematical functions, and a wide range of statistical techniques all the way from elementary classical tests, through regression and analysis of variance and generalized linear modelling, up to more specialized topics such as spatial statistics, multivariate methods, tree models, mixed-effects models and time series analysis. The idea is to introduce users to the assumptions that lie behind the tests, fostering a critical approach to statistical modelling, but involving little or no statistical theory and assuming no background in mathematics or statistics.

Why should you switch to using R when you have mastered a perfectly adequate statistical package already? At one level, there is no point in switching. If you only carry out a very limited range of statistical tests, and you don’t intend to do more (or different) in the future, then fine. The main reason for switching to R is to take advantage of its unrivalled coverage and the availability of new, cutting edge applications in fields such as generalized mixed-effects modelling and generalized additive models. The next reason for learning R is that you want to be able to understand the literature. More and more people are reporting their results in the context of R, and it is important to know what they are talking about. Third, look around your discipline to see who else is using R: many of the top people will have switched to R already. A large proportion of the world’s leading statisticians use R, and this should tell you something (many, indeed, contribute to R, as you can see below). Another reason for changing to R is the quality of back-up and support available. There is a superb network of dedicated R wizards out there on the web, eager to answer your questions. If you intend to invest sufficient effort to become good at statistical computing, then the structure of R and the ease with which you can write your own functions are major attractions. Last, and certainly not least, the product is free. This is some of the finest integrated software in the world, and yet it is yours for absolutely nothing.

Although much of the text will equally apply to S-PLUS, there are some substantial differences, so in order not to confuse things I concentrate on describing R. I have made no attempt to show where S-PLUS is different from R, but if you have to work in S-PLUS, then try it and see if it works.
Acknowledgements

S is an elegant, widely accepted, and enduring software system with outstanding conceptual integrity, thanks to the insight, taste, and effort of John Chambers. In 1998, the Association for Computing Machinery (ACM) presented him with its Software System Award, for ‘the S system, which has forever altered the way people analyze, visualize, and manipulate data’. R was inspired by the S environment that was developed by John Chambers, and which had substantial input from Douglas Bates, Rick Becker, Bill Cleveland, Trevor Hastie, Daryl Pregibon and Allan Wilks.

R was initially written by Ross Ihaka and Robert Gentleman at the Department of Statistics of the University of Auckland in New Zealand. Subsequently, a large group of individuals contributed to R by sending code and bug reports. John Chambers graciously contributed advice and encouragement in the early days of R, and later became a member of the core team. The current R is the result of a collaborative effort with contributions from all over the world.

Since mid-1997 there has been a core group known as the ‘R Core Team’ who can modify the R source code archive. The group currently consists of Doug Bates, John Chambers, Peter Dalgaard, Robert Gentleman, Kurt Hornik, Stefano Iacus, Ross Ihaka, Friedrich Leisch, Thomas Lumley, Martin Maechler, Duncan Murdoch, Paul Murrell, Martyn Plummer, Brian Ripley, Duncan Temple Lang, Luke Tierney, and Simon Urbanek.

R would not be what it is today without the invaluable help of the following people, who contributed by donating code, bug fixes and documentation: Valerio Aimale, Thomas Baier, Roger Bivand, Ben Bolker, David Brahm, Göran Broström, Patrick Burns, Vince Carey, Saikat DebRoy, Brian D’Urso, Lyndon Drake, Dirk Eddelbuettel, Claus Ekström, John Fox, Paul Gilbert, Frank E. Harrell Jr, Torsten Hothorn, Robert King, Kjetil Kjernsmo, Roger Koenker, Philippe Lambert, Jan de Leeuw, Uwe Ligges, Jim Lindsey, Patrick Lindsey, Catherine Loader, Gordon Maclean, John Maindonald, David Meyer, Eiji Nakama, Jens Oehlschaegel, Steve Onclcy, Richard O’Keefe, Hubert Palme, Roger D. Peng, Jose C. Pinheiro, Tony Plate, Anthony Rossini, Jonathan Rougier, Deepayan Sarkar, Guenther Sawitzki, Marc Schwartz, Detlef Steuer, Bill Simpson, Gordon Smyth, Adrian Trapletti, Terry Therneau, Rolf Turner, Bill Venables, Gregory R. Warnes, Andreas Weingessel, Morten Welinder, James Wettenhall, Simon Wood and Achim Zeileis. I have drawn heavily on the R help pages in writing this book, and I am extremely grateful to all the R contributors who wrote the help files.

Special thanks are due to the generations of graduate students on the annual GLIM course at Silwood. It was their feedback that enabled me to understand those aspects of R that are most difficult for beginners, and highlighted the concepts that require the most detailed explanation. Please tell me about the errors and omissions you find, and send suggestions for changes and additions to m.crawley@imperial.ac.uk.

M.J. Crawley
Ascot
September 2006