Springer Texts in Statistics

Advisors:
George Casella Stephen Fienberg Ingram Olkin

Brian S. Everitt

An R and S-PLUS® Companion to Multivariate Analysis

With 59 Figures

Springer

Brian Sidney Everitt, BSc, MSc
Emeritus Professor, King's College, London, UK

British Library Cataloguing in Publication Data
Everitt, Brian
An R and S-PLUS® companion to multivariate analysis.
(Springer texts in statistics)
1. S-PLUS (Computer file) 2. Multivariate analysis-Computer programs. 3. Multivariate analysis-Data processing
I. Title
519.5′35′0285

ISBN 1852338822

Library of Congress Cataloging-in-Publication Data
Everitt, Brian.
An R and S-PLUS® companion to multivariate analysis/Brian S. Everitt.
p. cm.—(Springer texts in statistics)
Includes bibliographical references and index.
ISBN 1-85233-882-2 (alk. paper)
1. Multivariate analysis. 2. S-Plus. 3. R (Computer program language) I. Title. II. Series.

QA278.E926 2004
519.5′35—dc22 2004054963

ISBN 1-85233-882-2
Springer Science+Business Media
springeronline.com

Printed in the United States of America
Typeset by Techset Composition Limited
12/3830-543210 Printed on acid-free paper SPIN 10969380

To my dear daughters, Joanna and Rachel

Preface

The majority of data sets collected by researchers in all disciplines are multivariate. In a few cases it may be sensible to isolate each variable and study it separately, but in most cases all the variables need to be examined simultaneously in order to fully grasp the structure and key features of the data. For this purpose, one or another method of multivariate analysis might be most helpful, and it is with such methods that this book is largely concerned.

Multivariate analysis includes methods both for describing and exploring such data and for making formal inferences about them. The aim of all the techniques is, in a general sense, to display or extract the signal in the data in the presence of noise, and to find out what the data show us in the midst of their apparent chaos.

The computations involved in applying most multivariate techniques are considerable, and their routine use requires a suitable software package. In addition, most analyses of multivariate data should involve the construction of appropriate graphs and diagrams and this will also need to be carried out by the same package. R and S-PLUS® are statistical computing environments, incorporating implementations of the S programming language. Both are powerful, flexible, and, in addition, have excellent graphical facilities. It is for these reasons that they appear in this book. R is available free through the Internet under the General Public License; see R Development Core Team (2004), R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, or visit their website www.R-project.org. S-PLUS is a registered trademark of Insightful Corporation, www.insightful.com. It is distributed in the United Kingdom by

Insightful Limited
5th Floor
Network House
Basing View
Basingstoke
Hampshire
RG21 4HG

Tel: +44 (0) 1256 339800

Fax: +44 (0) 1256 339839
info.uk@insightful.com

and in the United States by

Insightful Corporation
1700 Westlake Avenue North
Suite 500
Seattle, WA 98109-3044

Tel: (206) 283-8802
Fax: (206) 283-8691
info@insightful.com

We assume that readers have had some experience using either R or S-PLUS, although they are not assumed to be experts. If, however, they require to learn more about either program, we recommend Dalgaard (2002) for R and Krause and Olson (2002) for S-PLUS. An appendix very briefly describes some of the main features of the packages, but is intended primarily as nothing more than an *aide memoire*. One of the most powerful features of both R and S-PLUS (particularly the former) is the increasing number of functions being written and made available by the user community. In R, for example, CRAN (Comprehensive R Archive Network) collects libraries of functions for a vast variety of applications. Details of the libraries that can be used within R can be found by typing in `help.start()`. Additional libraries can be accessed by clicking on **Packages** followed by **Load package** and then selecting from the list presented.

In this book we concentrate on what might be termed the "core" multivariate methodology, although mention will be made of recent developments where these are considered relevant and useful. Some basic theory is given for each technique described but not the complete theoretical details; this theory is separated out into "displays." Suitable R and S-PLUS code (which is often identical) is given for each application. All data sets and code used in the book can be found at http://biostatistics.iop.kcl.ac.uk/publications/everitt/. In addition, this site contains the code for a number of functions written by the author and used at a number of places in the book. These can no doubt be greatly improved! After the data files have been downloaded by the reader, they can be read using the `source` function

R: `name<-source("path")$value`

For example,

```
huswif<-source("c:\\allwork\\rsplus\\chap1huswif.dat")$value
```

S-PLUS: `name<-source("path")`

For example,

```
huswif<-source("c:\\allwork\\rsplus\\chap1huswif.dat")
```

Since the output from S-PLUS and R is not their most compelling or attractive feature, such output has often been edited in the text and the results then displayed in a different form from this output to make them more readable; on a few occasions, however, the exact output itself is given. In one or two places the "click-and-point" features of the S-PLUS GUI are illustrated.

This book is aimed at students in applied statistics courses at both the undergraduate and postgraduate levels. It is also hoped that many applied statisticians dealing with multivariate data will find something of interest.

Since this book contains the word "companion" in the title, prospective readers may legitimately ask "companion to what?" The answer is, to a multivariate analysis textbook that covers the theory of each method in more detail but does not incorporate the use of any specific software. Some examples are Mardia, Kent, and Bibby (1979), Everitt and Dunn (2002), and Johnson and Wichern (2003).

I am very grateful to Dr. Torsten Hothorn for his advice about using R and for pointing out errors in my initial code. Any errors that remain, of course, are entirely due to me.

Finally I would like to thank my secretary, Harriet Meteyard, who, as always, provided both expertise and support during the writing of this book.

London, UK Brian S. Everitt

Contents