

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

Editorial Board

George Casella
Biometrics Unit
Cornell University
Ithaca, NY 14853-7801
USA

Stephen Fienberg
Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213-3890
USA

Ingram Olkin
Department of Statistics
Stanford University
Stanford, CA 94305
USA

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

Apart from any fair dealing for the purposes of research or private study, or criticism or review, as permitted under the Copyright, Designs and Patents Act 1988, this publication may only be reproduced, stored or transmitted, in any form or by any means, with the prior permission in writing of the publishers, or in the case of reprographic reproduction in accordance with the terms of licences issued by the Copyright Licensing Agency. Enquiries concerning reproduction outside those terms should be sent to the publishers.

ISBN 1-85233-882-2

Springer Science+Business Media
springeronline.com

© Springer-Verlag London Limited 2005

The use of registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant laws and regulations and therefore free for general use.

The publisher makes no representation, express or implied, with regard to the accuracy of the information contained in this book and cannot accept any legal responsibility or liability for any errors or omissions that may be made.

Whilst we have made considerable efforts to contact all holders of copyright material contained in this book, we have failed to locate some of them. Should holders wish to contact the Publisher, we will be happy to come to some arrangement with them.

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

Preface	vii
1 Multivariate Data and Multivariate Analysis	1
1.1 Introduction	1
1.2 Types of Data	1
1.3 Summary Statistics for Multivariate Data	4
1.3.1 Means	4
1.3.2 Variances	5
1.3.3 Covariances	5
1.3.4 Correlations	6
1.3.5 Distances	7
1.4 The Multivariate Normal Distribution	9
1.5 The Aims of Multivariate Analysis	13
1.6 Summary	15
2 Looking at Multivariate Data	16
2.1 Introduction	16
2.2 Scatterplots and Beyond	17
2.2.1 The Convex Hull of Bivariate Data	22
2.2.2 The Chiplot	23
2.2.3 The Bivariate Boxplot	25
2.3 Estimating Bivariate Densities	29
2.4 Representing Other Variables on a Scatterplot	32
2.5 The Scatterplot Matrix	33
2.6 Three-Dimensional Plots	35
2.7 Conditioning Plots and Trellis Graphics	37
2.8 Summary	40
Exercises	40
3 Principal Components Analysis	41
3.1 Introduction	41
3.2 Algebraic Basics of Principal Components	42

3.2.1	Rescaling Principal Components	45
3.2.2	Choosing the Number of Components	46
3.2.3	Calculating Principal Component Scores	47
3.2.4	Principal Components of Bivariate Data with Correlation Coefficient r	48
3.3	An Example of Principal Components Analysis: Air Pollution in U.S. Cities	49
3.4	Summary	61
	Exercises	62
4	Exploratory Factor Analysis	65
4.1	Introduction	65
4.2	The Factor Analysis Model	65
4.2.1	Principal Factor Analysis	68
4.2.2	Maximum Likelihood Factor Analysis	69
4.3	Estimating the Numbers of Factors	69
4.4	A Simple Example of Factor Analysis	70
4.5	Factor Rotation	71
4.6	Estimating Factor Scores	76
4.7	Two Examples of Exploratory Factor Analysis	77
4.7.1	Expectations of Life	77
4.7.2	Drug Usage by American College Students	82
4.8	Comparison of Factor Analysis and Principal Components Analysis	85
4.9	Confirmatory Factor Analysis	88
4.10	Summary	88
	Exercises	89
5	Multidimensional Scaling and Correspondence Analysis	91
5.1	Introduction	91
5.2	Multidimensional Scaling (MDS)	93
5.2.1	Examples of Classical Multidimensional Scaling	96
5.3	Correspondence Analysis	104
5.3.1	Smoking and Motherhood	109
5.3.2	Hodgkin's Disease	111
5.4	Summary	112
	Exercises	112
6	Cluster Analysis	115
6.1	Introduction	115
6.2	Agglomerative Hierarchical Clustering	115
6.2.1	Measuring Intercluster Dissimilarity	118
6.3	K -Means Clustering	122
6.4	Model-Based Clustering	128
6.5	Summary	134
	Exercises	135

7 Grouped Multivariate Data: Multivariate Analysis of Variance and Discriminant Function Analysis	137
7.1 Introduction	137
7.2 Two Groups: Hotellings T^2 Test and Fisher's Linear Discriminant Function Analysis	137
7.2.1 Hotellings T^2 Test	137
7.2.2 Fisher's Linear Discriminant Function	142
7.2.3 Assessing the Performance of a Discriminant Function	146
7.3 More Than Two Groups: Multivariate Analysis of Variance (MANOVA) and Classification Functions	147
7.3.1 Multivariate Analysis of Variance	147
7.3.2 Classification Functions and Canonical Variates	149
7.4 Summary	155
Exercises	156
8 Multiple Regression and Canonical Correlation	157
8.1 Introduction	157
8.2 Multiple Regression	157
8.3 Canonical Correlations	160
8.4 Summary	167
Exercises	167
9 Analysis of Repeated Measures Data	171
9.1 Introduction	171
9.2 Linear Mixed Effects Models for Repeated Measures Data	174
9.3 Dropouts in Longitudinal Data	190
9.4 Summary	198
Exercises	198
Appendix: An Aide Memoir for R and S-PLUS®	200
1. Elementary commands	200
2. Vectors	201
3. Matrices	204
4. Logical Expressions	205
5. List Objects	207
6. Data Frames	209
References	211
Index	217