# Multivariate statistical analysis in geography

A primer on the general linear model

R. J. Johnston

# Multivariate statistical analysis in geography

A primer on the general linear model

R. J. Johnston

Longman
London and New York

**Longman Group Limited** London

*Associated companies, branches and representatives throughout the world*

*Published in the United States of America by Longman Inc., New York*



*First published 1978*

ISBN 0 582 48677 7

---

**Library of Congress Cataloging in Publication Data**

Johnston, Ronald John
Multivariate statistical analysis in geography

Bibliography: p.
Includes index.
1. Geography—Statistical methods. 2 Multivariate analysis. I. Title.
G70.3.J65 910'.01'82 77-22424
ISBN 0-582-48677-7

---

Printed in Great Britain by
Richard Clay (The Chaucer Press) Ltd, Bungay, Suffolk

# Preface

The use of statistics in geographical research has increased rapidly during the last two or three decades. Most students being trained in geography at tertiary educational institutions receive some instruction in statistical methods; all are given things to read, particularly research reports in journals, in which statistical procedures form a basic part of the reasoning. And yet relatively few geography students have a strong background in school mathematics, and even fewer continue with mathematical studies in parallel with their geography. How, then, can they obtain sufficient knowledge of statistical methods in order to appreciate the arguments which they must read and digest?

Two main approaches to this problem have been suggested and applied. The first is generally known as the 'cookbook', which outlines the computational procedures involved in using a statistical method and, by way of examples, illustrates the method's application within the general contexts of hypothesis-testing and statistical inference. The focus is on 'how' and 'what for'. The other approach is based much more firmly on 'why', arguing that in order to use a method it is necessary to understand how it works. Some detailed mathematical and statistical grounding is required for this.

The major texts written specifically for geographers to date have been concerned with relatively simple and straightforward statistical techniques, and they have largely been based on the 'cookbook' approach. Notable among these are the texts by Gregory (3rd ed, 1973) and by Hammond and McCullagh (1974). Two other volumes, by King (1969) and by Yeates (1974), have essayed a stronger 'why' component, but seem to have fallen between the two approaches: in addition, both aim to cover a very wide range of material in a relatively small space, a feature of the later chapters of a recent text by Smith (1975), also.

The present book follows introductory texts, such as Gregory's and Hammond and McCullagh's, both of which have twin purposes of introducing students to the fundamentals of statistical analysis and of outlining in some detail the computational procedures by which relatively simple analyses may be conducted. But many of the research reports which students must read use much more complex techniques, because the world being studied is more complex than is allowed for by the simple procedures covered by introductory texts. These more complex techniques are based on the simple ones, but students are given no introduction to them which crosses what is basically the gap between bivariate – or single cause – and

multivariate – multiple cause-relationships. The present volume attempts to build a bridge over that gap with respect to one main family of techniques, the general linear model.

This is not a 'cookbook', for two reasons. First, it is impractical to attempt the use of most of the methods discussed here by hand calculations, or even with calculators of the type many people can now afford. Computers are needed, and without them many of the procedures, especially those introduced in the later chapters, cannot be employed. Secondly, the book is not aimed to help those who would make considerable use of the procedures in their own research: for that purpose, much more intensive mathematical and statistical background work is necessary, and students would need to graduate from this volume to the more detailed statistical texts relating to particular procedures. The target here is appreciation. If students must read about multiple regressions and factor analyses, dummy variables and linear discriminant functions, then they must have a general knowledge of what those procedures do. Just as appreciation of a good meal does not need understanding of the recipe, so appreciation of a research paper by someone who wishes to use its findings does not require a mathematical understanding of the calculations undertaken. The general student needs to know what in broad terms the adopted procedure does and how its results should be interpreted.

There are dangers in the approach adopted here. The wide availability of computer programs invites the 'number-crunching' of ill-assembled data sets through partly-comprehended statistical routines, and those intending research using the methods outlined here are advised that further reading is necessary before they start their analyses. But for those who only wish to know what it was that X did, whether it was relevant, and what his results mean, have a more basic requirement, which is a general appreciation of the method. It cannot be obtained without equations and formulae, but it does not demand much mathematics, beyond simple arithmetic, geometry and algebra; matrix algebra, however, is a necessity for those who would travel further. Words, diagrams, and simple examples are used here, in the hope that careful, perhaps guided, reading and study of this text will at least enable the mass of geography students to know enough about what they read to appreciate the contribution it makes to the discipline: hence the sub-title.

Statistical analysis is not the only research method in geography; it is the one which is far superior to any other when the aim is to make precise and unambiguous statements about relationships and patterns in sets of numbers. The statements which the researcher wants to make, and the numbers themselves, are based in geographical theory and the purposes of geographical work. Chapter 1 briefly discusses this background to the use of statistics, and reviews some basic concepts relevant to successful reading of the rest of the book.

As indicated by the title and sub-title, discussion here is restricted to the various aspects of the general linear model, and particularly to multivariate analyses. Most students will have encountered the bivariate aspects

– simple regression and correlation; one-way analysis of variance – prior to using this book, but as understanding of these is basic to later discussion of multivariate procedures, they are covered in some detail in Chapter 2, to provide the necessary foundation. From them, Chapters 3 and 4 develop the multivariate extensions, showing how multi-causal hypotheses can be tested with independent variables that are either interval/ratio measured, or nominally measured, or both.

Of the seven substantive chapters the first three deal with hypothesis-testing when causal and effect variables have been specified. In the other four, the process is reversed somewhat, with attention on those procedures which create new variables and test more complicated hypotheses with them. Chapters 5 and 6 deal with this topic for data measured on interval and ratio scales, with the former treating the procedures – principal components analysis and factor analysis – which are used to create new variables and the latter concentrating on methods, such as canonical correlation analysis, which test for relationships between hybrid variables. In Chapters 7 and 8 the focus turns to independent variables which are measured on nominal scales: classification of observations into categories is discussed in Chapter 7; testing of hypotheses with and about such categories by linear discriminant analysis is treated in Chapter 8. Perhaps paradoxically, these later chapters are about the more complex procedures, but they are also the shorter ones, reflecting the detailed basis given earlier.

Throughout the book, the various methods are illustrated with geographical examples. The assumption is always that there are no problems in the wholesale adoption of the methods to the research aims and data of the geographer. This is not necessarily so, and in the final chapter some of the problems raised by geographical data are discussed.

There is continuing debate within the geographical discipline concerning the methods discussed here. This has two strands. The first contends that statistical analyses of the types discussed here are invalid for much geographical work, certainly in human geography. This is a philosophical, sometimes ideological issue, whose discussion lies outside the present context. The assumptions here are (1) that the methods discussed are valuable in certain circumstances, and (2) that because the methods are widely used and reported in the geographical literature at present, it is a necessary part of a geographer's training to obtain some general comprehension of what is being done and why. The second strand is not a debate about the value of statistical techniques, but about the validity of using those discussed here in geographical research. As with the other strand, there is validity in parts of the argument against the techniques but the general case is far from proven and the previous argument still holds; the techniques are being widely used and geographers should be aware of them, thus enabling an informed assessment of their value.

The tricky issues are avoided here, therefore. The philosophy of this book is that a family of techniques that is widely used in geographical research should be appreciated – if not entirely understood – by all geographers. The book has been written in an attempt to provide such an

appreciation, with as few constraints on required numeracy and mathematical skills as possible.

# Acknowledgements

This book is the result of attempts to teach aspects of multivariate statistical analysis in three separate continents, largely to students with little mathematical background, let alone statistical. The methods developed during more than a decade of such teaching, based on verbal and pictorial representation of algebraic arguments, are now presented here for a wider audience, in the hope that they are useful to those who, like me, have little of the background in depth necessary to an appreciation of what is being done in so much geographical research today.

In developing the teaching method outlined here, I am grateful to those classes of students who have persevered with my attempts to help them understand the methods covered, and whose comments helped me to improve the presentation. Two 'students' in particular deserve special mention, my colleagues Stan Gregory and Alan Hay, with whom I have participated in a course involving this material. Stan Gregory encouraged me to write this book, and both he and Alan Hay read the full manuscript and made many valuable comments on it, for which I am extremely grateful. Arthur Hunt, too, read and commented on the draft and Keith Beavon provided a very meticulous critique of the whole manuscript in its penultimate form, which was extremely rewarding for me. To all of them, my thanks, whilst of course absolving them from any responsibility for the contents.

Technical preparation of a volume of this nature is a major task, and I am deeply indebted to several persons for their assistance. Joan Dunn performed marvels in transforming the manuscript into a typescript; Stephen Frampton provided an excellent set of diagrams; and the staff at Longman have translated both into an excellent book. My wife Rita meticulously checked everything in the script, a task which nobody should be asked to do. To all, my sincere thanks.

This book is dedicated to my wife, as a measly token of thanks for her fantastic support, in so many ways, throughout my career. Since no thanks could be sufficient, I trust this small mention at least records the depth of my gratitude.

*R. J. Johnston*

*May, 1977.*

We are grateful to the following for permission to reproduce copyright material:

Cambridge University Press for an extract from Table 4.23 in *The Urban Mosaic* by Timms; Geographical Analysis for Table 2 in 'On Structuring a Migration Model' by J. B. Riddell from *Geographical Analysis*, Vol. 2; the author for his Figure No. 8 by Peter Goheen from *Victorian Toronto*; Journal of Ecology for Figure 1 by Norris and Barkham from *Journal of Ecology*, 58; the author for his Figure No. 28 by R. A. Murdie from *Factorial Ecology of Metropolitan Toronto*; New Zealand Geographer for Figure 1 by W. A. V. Clark from 'The Use of Residuals for Regression' in *New Zealand Geographer*, 23, 1967; the authors for a table from 'Canonical Correlation in Geographical Analysis' by D. M. Ray and P. R. Lohnes in *Geographia Polonica*, XXV, 1973; Regional Science Association for Figure 5A by P. Haggett 'Trend-Surface Mapping in the Interregional Comparison of Intra-Regional Structures' from Haggett *Papers RSA*, Vol. 20; Regional Studies Association for Table 5 by Willis from *Regional Studies*, 6, 1972; Economic Geography for an extract from Table III by J. Morrison, M. W. Scripter and R. H. T. Smith 'Basic Measures of Manufacturing in the U.S.A. 1958' from *Economic Geography*, 44, 1968; University of Chicago Press for Figure 2 by O. D. Duncan 'Path Analysis – Sociological Examples' from *American Journal of Sociology*, 72, 1966.

# Contents

# List of figures

# List of tables