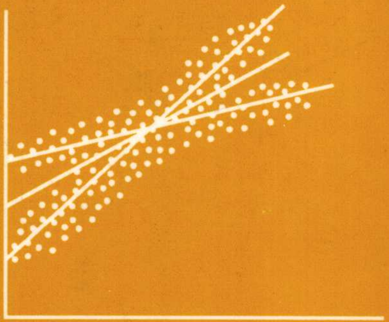
S T SPRINGER TEXTS IN STATISTICS

J.D. Jobson

Applied Multivariate Data Analysis

Volume I: Regression and Experimental Design



Including Floppy Disk

Springer-Verlag

J.D. Jobson

Applied Multivariate Data Analysis

Volume I: Regression and Experimental Design

With 93 Illustrations in 116 parts

With a diskette

Springer-Verlag
New York Berlin Heidelberg London Paris
Tokyo Hong Kong Barcelona Budapest

贈送

J.D. Jobson
Faculty of Business
University of Alberta
Edmonton, Alberta T6G 2R6
Canada

Mathematics Subject Classification: 62-07, 62J05, 62J10, 62K10, 62K15

Printed on acid-free paper.

Photocomposed copy prepared from the author's TEX file.
Printed and bound by R.R. Donnelley & Sons, Inc., Harrisonburg, Virginia.
Printed in the United States of America.

9 8 7 6 5 4 3 2 (corrected second printing)

ISBN 0-387-97660-4 Springer-Verlag New York Berlin Heidelberg
ISBN 3-540-97660-4 Springer-Verlag Berlin Heidelberg New York

Springer Texts in Statistics

Alfred	Elements of Statistics for the Life and Social Sciences
Berger	An Introduction to Probability and Stochastic Processes
Blom	Probability and Statistics: Theory and Applications
Chow and Teicher	Probability Theory: Independence, Interchangeability, Martingales, Second Edition
Christensen	Plane Answers to Complex Questions: The Theory of Linear Models
Christensen	Linear Models for Multivariate, Time Series, and Spatial Data
Christensen	Log-Linear Models
Creighton	A First Course in Probability Models and Statistical Inference
du Toit, Steyn and Stumpf	Graphical Exploratory Data Analysis
Finkelstein and Levin	Statistics for Lawyers
Jobson	Applied Multivariate Data Analysis, Volume I: Regression and Experimental Design
Jobson	Applied Multivariate Data Analysis, Volume II: Categorical and Multivariate Methods
Kalbfleisch	Probability and Statistical Inference, Volume I: Probability, Second Edition
Kalbfleisch	Probability and Statistical Inference, Volume II: Statistical Inference, Second Edition
Karr	Probability

Continued at end of book

Some Quotations From Early Statisticians

All sciences of observation follow the same course. One begins by observing a phenomenon, then studies all associated circumstances, and finally, if the results of observation *can be expressed numerically* [Quetelet's italics], estimates the intensity of the causes that have concurred in its formation. This course has been followed in studying purely material phenomena in physics and astronomy; it will likely also be the course followed in the study of phenomena dealing with moral behavior and the intelligence of man. Statistics begins with the gathering of numbers; these numbers, collected on a large scale with care and prudence, have revealed interesting facts and have led to the conjecture of laws ruling the moral and intellectual world, much like those that govern the material world. It is the whole of these laws that appears to me to constitute *social physics*, a science which, while still in its infancy, becomes incontestably more important each day and will eventually rank among those sciences most beneficial to man. (Quetelet, 1837)

The investigation of causal relations between economic phenomena presents many problems of peculiar difficulty, and offers many opportunities for fallacious conclusions. Since the statistician can seldom or never make experiments for himself, he has to accept the data of daily experience, and discuss as best he can the relations of a whole group of changes; he cannot, like the physicist, narrow down the issue to the effect of one variation at a time. The problems of statistics are in this sense far more complex than the problems of physics. (Yule, 1897)

Some people hate the very name of statistics, but I find them full of beauty and interest. Whenever they are not brutalized, but delicately handled by the higher methods, and are warily interpreted, their power of dealing with complicated phenomena is extraordinary. They are the only tools by which an opening can be cut through the formidable thicket of difficulties that bars the path of those who pursue the Science of man. (Galton, 1908)

To Leone

Preface

A Second Course in Statistics

The past decade has seen a tremendous increase in the use of statistical data analysis and in the availability of both computers and statistical software. Business and government professionals, as well as academic researchers, are now regularly employing techniques which go far beyond the standard two-semester, introductory course in statistics. While for this group of users short courses in various specialized topics are often available, there is a need to improve the statistics training of future users of statistics while they are still at colleges and universities. In addition, there is a need for a survey reference text for the many practitioners who cannot obtain specialized courses.

With the exception of the statistics major, most university students do not have sufficient time in their programs to enroll in a variety of specialized one-semester courses, such as data analysis, linear models, experimental design, multivariate methods, contingency tables, logistic regression, etc. There is a need for a second survey course that covers a wide variety of these techniques in an integrated fashion. It is also important that this second course combine an overview of theory with an opportunity to practice, including the use of statistical software and the interpretation of results obtained from 'real' data.

Topics

This two-volume survey is designed to provide a second two-semester course in statistics. The first volume outlines univariate data analysis and provides an extensive overview of regression models. The first volume also surveys the methods of analysis of variance and experimental design including their relationship to the regression model. The second volume begins with a survey of techniques for analyzing multidimensional contingency tables and then outlines the traditional topics of multivariate methods. The second volume also includes a discussion of logistic regression, cluster analysis, multidimensional scaling and correspondence analysis, which are not always included in surveys of multivariate methods. In each volume an appendix is provided to review the basic concepts of linear and matrix algebra. The appendix also includes a series of exercises in linear algebra for student practice.

Mathematics Background

The text assumes a background equivalent to one semester of each of linear algebra and calculus, as well as the standard two-semester introductory course in statistics. Calculus is almost never used in the text other than in the theoretical questions at the end of each chapter. The one semester of calculus is an indication of the ideal mathematics comfort level. The linear algebra background is needed primarily to understand the presentation of formulae. Competence with linear algebra however is required to complete many of the theoretical questions at the end of each chapter. These background prerequisites would seem to be a practical compromise given the wide variety of potential users.

Examples and Exercises
In addition to an overview of theory, the text also includes a large number of examples based on actual research data. Not only are numerical results given for the examples but also interpretations for the results are also discussed. The text also provides data analysis exercises and theoretical questions for student practice. The data analysis exercises are based on real data which is also provided with the text. The student is therefore able to improve by "working out" on the favorite local software. The theoretical questions can be used to raise the theoretical level of the course or can be omitted without any loss of the applied aspects of the course. The theoretical questions provide useful training for those who plan to take additional courses in statistics.

Use as a Text
The two volumes can be used independently for two separate courses. Volume I can be used for a course in regression and design, while Volume II can be used for a course in categorical and multivariate methods. A quick review of multiple regression and analysis of variance may be required if the second volume is to be used without the first. If the entire text is to be used in two semesters some material in each chapter can be omitted. A number of sections can be left for the student to read or for the student's future reference. Large portions of most chapters and/or entire topics can be omitted without affecting the understanding of other topics discussed later in the text. A course in applied multivariate data analysis for graduate students in a particular field of specialization can be derived from the text by concentrating on a particular selection of topics.

This two-volume survey should be useful for a second course in statistics for most college juniors or seniors. Also, for the undergraduate statistics major, this text provides a useful second course which can be combined with other specialized courses in time series, stochastic processes, sampling theory, nonparametric statistics and mathematical statistics. Because the text includes the topics normally found in traditional second courses, such as regression analysis or multivariate methods, this course provides a broader substitute which also includes other topics such as data analysis, multidimensional contingency tables, logistic regression, correspondence analysis and multidimensional scaling. The set of theoretical questions in the book can provide useful practice for statistics majors who have already been exposed to mathematical statistics.

For graduate students in business and the social and biological sciences, this survey of applied multivariate data analysis is a useful first year graduate course which could then be followed by other more specialized courses, such as econometrics, structural equation models, time series analysis or stochastic processes. By obtaining this background early in the graduate program the student is then well prepared to read the research literature in the chosen discipline and at a later stage to analyse research data. This course is also useful if taken concurrently with a course in the research methodology of the chosen discipline. I have found the first year of the Ph.D. program to be the ideal time for this course, since later in their programs Ph.D. students are too often preoccupied with their own area of specialization and research tasks.

Author's Motivation and Use of Text

The author's motivation for writing this text was to provide a two-semester overview of applied multivariate data analysis for beginning Ph.D. students in the Faculty of Business at the University of Alberta. The quantitative background assumed for the business Ph.D. student using this text is equivalent to what is required in most undergraduate business programs in North America — one semester of each of linear algebra and calculus and a two-semester introduction to statistics. Many entering Ph.D. students have more mathematics background but do not usually have more statistics background. A selection of topics from the text has also been used for an elective course in applied multivariate data analysis for second year MBA students. For the MBA elective course much less emphasis is placed on the underlying theory.

Because of the many different fields of interest within business Ph.D. programs — Accounting, Finance, Marketing, Organization Analysis and Industrial Relations — the topical needs, interests and level of mathematical sophistication of the graduate students differ greatly. Some will pursue a strong statistics minor, while others will take very little statistics training beyond this course.

In my Ph.D. class the wide variety of needs are handled simultaneously by assigning portfolios of theoretical questions to the statistics minor student, while the less theoretically oriented students are assigned a paper. The paper topic may involve a discussion of the application of one or more of the statistical techniques to a particular field or an overview of techniques not discussed in the text. A small number of classes are devoted exclusively to the discussion of the theory questions. For the theory classes only the 'theory folk' need attend. All students are required to complete data analysis exercises and to provide written discussions of the results. For the data analysis exercises great emphasis is placed on the quality of the interpretation of the results. Graduate students often have greater difficulty with the interpretation of results than with the understanding of the principles.

Quotations

The quotations by Quetelet (1837) and Yule (1897) were obtained from pages 193 and 348 respectively of *The History of Statistics: The Measurement of Uncertainty Before: 1900*, by Stephen Stigler, published by Harvard University Press, Cambridge, Mass., 1986.

The quotation by Galton (1908) was obtained from *An Introduction to Mathematical Statistics and its Applications*, Second Edition, by Richard J. Larcen and Morris L. Marx, published by Prentice–Hall, 1986.

Acknowledgments

The production of this text has benefited greatly from the input and assistance of many individuals. The Faculty of Business at the University of Alberta has born most of the cost of production. Steve Beveridge and John Brown were helpful in making funds available. John Brown and John Waterhouse also provided much encouragement during the development and early implementation of this text in our Ph.D. program.

The bulk of the typing has been done by two very able typists, Anna Fujita and Shelley Hey. Both individuals have been tireless in providing error-free typing through the uncountable number of drafts. The numerous graphs and figures could not have been carried out without the capable assistance of Berry Hsu and Anna Fujita.

The examples and data analysis exercises have been generated from data sets provided to me by my colleagues and former students. Colleagues who have allowed me to use their research data or have suggested examples include Alice and Masao Nakamura, Rodney Schneck, Ken Lemke, Chris Vaughn, John Waterhouse, Chris Janssen, Bernard Yeung and Jordan Louviére. Graduate students who have gifted me their data include Nancy Button, Nancy Keown, Pamela Norton, Diane Ewanishan, Clarke Carson, Caroline Pinkert-Rust, Frank Kobe and Cam Morrell. I am also grateful to G.C. McDonald for the use of the air pollution data and to SPSS for the bank employee salary data.

Subash Bagui, Sheila Mozejko, Alice Nakamura and Tom Johnson read parts of the manuscript and were extremely helpful in improving the overall presentation. Many graduate students have provided editorial assistance by pointing out errors in various drafts of the text. These students have also been very patient in tolerating the inconveniences resulting from the many errors. Three graduate students who provided editorial assistance by checking grammar, and table and figure references were Caroline Pinkert-Rust, Ellen Nygaard and Mary Allen. Any errors that remain are solely my responsibility.

I am also grateful to the University of Alberta who provided my undergraduate training in mathematics, 1959–1963; to Iowa State University who provided my graduate training in statistics, 1968–1971; and to the University of North Carolina who provided a statistics theory update during my study leave in 1980–81. I am extremely fortunate to have been

exposed to so many great teachers.

Last and most importantly I am indebted to my wife, Leone, who cheerfully attended to most of my domestic chores while I was preoccupied with this task. At times I think she was the only one who believed that the project would ever be completed. Perhaps that was wishful thinking, as the book was often spread out all over the house. A vote of thanks is also due my two daughters, Leslie and Heather, who cheerfully accepted DAD's excessive devotion to "the book."

J.D. Jobson

Contents of Volume II

Contents