Springer Texts in Statistics

Advisors:
George Casella   Stephen Fienberg   Ingram Olkin

Springer
New York
Berlin
Heidelberg
Barcelona
Hong Kong
London
Milan
Paris
Singapore
Tokyo

Springer Texts in Statistics

Alfred: Elements of Statistics for the Life and Social Sciences
Berger: An Introduction to Probability and Stochastic Processes
Bilodeau and Brenner: Theory of Multivariate Statistics
Blom: Probability and Statistics: Theory and Applications
Brockwell and Davis: An Introduction to Times Series and Forecasting
Chow and Teicher: Probability Theory: Independence, Interchangeability, Martingales, Third Edition
Christensen: Plane Answers to Complex Questions: The Theory of Linear Models, Second Edition
Christensen: Linear Models for Multivariate, Time Series, and Spatial Data
Christensen: Log-Linear Models and Logistic Regression, Second Edition
Creighton: A First Course in Probability Models and Statistical Inference
Dean and Voss: Design and Analysis of Experiments
du Toit, Steyn, and Stumpf: Graphical Exploratory Data Analysis
Durrett: Essentials of Stochastic Processes
Edwards: Introduction to Graphical Modelling, Second Edition
Finkelstein and Levin: Statistics for Lawyers
Flury: A First Course in Multivariate Statistics
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
Keyfitz: Applied Mathematical Demography, Second Edition
Kiefer: Introduction to Statistical Inference
Kokoska and Nevison: Statistical Tables and Formulae
Kulkarni: Modeling, Analysis, Design, and Control of Stochastic Systems
Lehmann: Elements of Large-Sample Theory
Lehmann: Testing Statistical Hypotheses, Second Edition
Lehmann and Casella: Theory of Point Estimation, Second Edition
Lindman: Analysis of Variance in Experimental Design
Lindsey: Applying Generalized Linear Models
Madansky: Prescriptions for Working Statisticians
McPherson: Applying and Interpreting Statistics: A Comprehensive Guide, Second Edition
Mueller: Basic Principles of Structural Equation Modeling: An Introduction to LISREL and EQS

(continued after index)

John O. Rawlings
Sastry G. Pantula
David A. Dickey

Applied Regression Analysis

A Research Tool

Second Edition

With 78 Figures

Springer

John O. Rawlings
Sastry G. Pantula
David A. Dickey
Department of Statistics
North Carolina State University
Raleigh, NC 27695
USA

Library of Congress Cataloging-in-Publication Data
Rawlings, John O., 1932–
Applied regression analysis: a research tool. — 2nd ed. / John O. Rawlings, Sastry G. Pentula, David A. Dickey.
p. cm. — (Springer texts in statistics)
Includes bibliographical references and indexes.
ISBN 0-387-98454-2 (hardcover: alk. paper)
1. regression analysis. I. Pentula, Sastry G. II. Dickey, David A. III. Title. IV. Series.
QA278.2.R38 1998
519.5′36—dc21 97-48858

Printed on acid-free paper.

9 8 7 6 5 4 3 2 1

ISBN 0-387-98454-2 Springer-Verlag New York Berlin Heidelberg SPIN 10660129

To

Our Families

PREFACE

This text is a new and improved edition of Rawlings (1988). It is the outgrowth of several years of teaching an applied regression course to graduate students in the sciences. Most of the students in these classes had taken a two-semester introduction to statistical methods that included experimental design and multiple regression at the level provided in texts such as Steel, Torrie, and Dickey (1997) and Snedecor and Cochran (1989). For most, the multiple regression had been presented in matrix notation.

The basic purpose of the course and this text is to develop an understanding of least squares and related statistical methods without becoming excessively mathematical. The emphasis is on regression concepts, rather than on mathematical proofs. Proofs are given only to develop facility with matrix algebra and comprehension of mathematical relationships. Good students, even though they may not have strong mathematical backgrounds, quickly grasp the essential concepts and appreciate the enhanced understanding. The learning process is reinforced with continuous use of numerical examples throughout the text and with several case studies. Some numerical and mathematical exercises are included to whet the appetite of graduate students.

The first four chapters of the book provide a review of simple regression in algebraic notation (Chapter 1), an introduction to key matrix operations and the geometry of vectors (Chapter 2), and a review of ordinary least squares in matrix notation (Chapters 3 and 4). Chapter 4 also provides a foundation for the testing of hypotheses and the properties of sums of squares used in analysis of variance. Chapter 5 is a case study giving a complete multiple regression analysis using the methods reviewed in the first four chapters. Then Chapter 6 gives a brief geometric interpretation of least squares illustrating the relationships among the data vectors, the link between the analysis of variance and the lengths of the vectors, and the role of degrees of freedom. Chapter 7 discusses the methods and criteria for determining which independent variables should be included in the models. The next two chapters include special classes of multiple regression models. Chapter 8 introduces polynomial and trigonometric regression models. This chapter also discusses response curve models that are linear in the parameters. Class variables and the analysis of variance of designed experiments (models of less than full rank) are introduced in Chapter 9.

Chapters 10 through 14 address some of the problems that might be encountered in regression. A general introduction to the various kinds of problems is given in Chapter 10. This is followed by discussions of regression diagnostic techniques (Chapter 11), and scaling or transforming variables to rectify some of the problems (Chapter 12). Analysis of the correlational structure of the data and biased regression are discussed as techniques for dealing with the collinearity problem common in observational data (Chapter 13). Chapter 14 is a case study illustrating the analysis of data in the presence of collinearity.

Models that are nonlinear in the parameters are presented in Chapter 15. Chapter 16 is another case study using polynomial response models, nonlinear modeling, transformations to linearize, and analysis of residuals. Chapter 17 addresses the analysis of unbalanced data. Chapter 18 (new to this edition) introduces linear models that have more than one random effect. The ordinary least squares approach to such models is given. This is followed by the definition of the variance–covariance matrix for such models and a brief introduction to mixed effects and random coefficient models. The use of iterative maximum likelihood estimation of both the variance components and the fixed effects is discussed. The final chapter, Chapter 19, is a case study of the analysis of unbalanced data.

We are grateful for the assistance of many in the development of this book. Of particular importance have been the dedicated editing of the earlier edition by Gwen Briggs, daughter of John Rawlings, and her many suggestions for improvement. It is uncertain when the book would have been finished without her support. A special thanks goes to our former student, Virginia Lesser, for her many contributions in reading parts of the manuscript, in data analysis, and in the enlistment of many data sets from her graduate student friends in the biological sciences. We are indebted to our friends, both faculty and students, at North Carolina State University for bringing us many interesting consulting problems over the years that have stimulated the teaching of this material. We are particularly indebted to those (acknowledged in the text) who have generously allowed the use of their data. In this regard, Rick Linthurst warrants special mention for his stimulating discussions as well as the use of his data. We acknowledge the encouragement and valuable discussions of colleagues in the Department of Statistics at NCSU, and we thank Matthew Sommerville for checking answers to the exercises. We wish to thank Sharon Sullivan and Dawn Haines for their help with LaTeX. Finally, we want to express appreciation for the critical reviews and many suggestions provided for the first edition by the Wadsworth Brooks/Cole reviewers: Mark Conaway, University of Iowa; Franklin Graybill, Colorado State University; Jason Hsu, Ohio State University; Kenneth Koehler, Iowa State University; B. Lindsay, The Pennsylvania State University; Michael Meridith, Cornell University; M. B. Rajarshi, University of Poona (India); Muni Srivastava, University of Toronto; and Patricia Wahl, University of Washington; and for the second edition by the Springer-Verlag reviewers.

Acknowledgment is given for the use of material in the appendix tables. Appendix Table A.7 is reproduced in part from Tables 4 and 6 of Durbin and Watson (1951) with permission of the Biometrika Trustees. Appendix Table A.8 is reproduced with permission from Shapiro and Francia (1972), *Journal of the American Statistical Association*. The remaining appendix tables have been computer generated by one of the authors. We gratefully acknowledge permission of other authors and publishers for use of material from their publications as noted in the text.

Note to the Reader

Most research is aimed at quantifing relationships among variables that either measure the end result of some process or are likely to affect the process. The process in question may be any biological, chemical, or physical process of interest to the scientist. The quantification of the process may be as simple as determining the degree of association between two variables or as complicated as estimating the many parameters of a very detailed nonlinear mathematical model of the system.

Regardless of the degree of sophistication of the model, the most commonly used statistical method for estimating the parameters of interest is the method of **least squares**. The criterion applied in least squares estimation is simple and has great intuitive appeal. The researcher chooses the model that is believed to be most appropriate for the project at hand. The parameters for the model are then estimated such that the predictions from the model and the observed data are in as good agreement as possible as measured by the **least squares criterion**, minimization of the sum of squared differences between the predicted and the observed points.

Least squares estimation is a powerful research tool. Few assumptions are required and the estimators obtained have several desirable properties. Inference from research data to the true behavior of a process, however, can be a difficult and dangerous step due to unrecognized inadequacies in the data, misspecification of the model, or inappropriate inferences of causality. As with any research tool it is important that the least squares method be thoroughly understood in order to eliminate as much misuse or misinterpretation of the results as possible. There is a distinct difference between understanding and pure memorization. Memorization can make a good technician, but it takes understanding to produce a master. A discussion of the **geometric interpretation** of least squares is given to enhance your understanding. You may find your first exposure to the geometry of least squares somewhat traumatic but the visual perception of least squares is worth the effort. We encourage you to tackle the topic in the spirit in which it is included.

The general topic of least squares has been broadened to include statistical techniques associated with **model development and testing**. The backbone of least squares is the classical multiple regression analysis using the linear model to relate several independent variables to a response or dependent variable. Initially, this classical model is assumed to be appropriate. Then methods for detecting inadequacies in this model and possible remedies are discussed.

The connection between the analysis of variance for designed experiments and multiple regression is developed to build the foundation for the analysis of **unbalanced data**. (This also emphasizes the generality of the least squares method.) Interpretation of unbalanced data is difficult. It is important that the application of least squares to the analysis of such data be understood if the results from computer programs designed for the analysis of unbalanced data are to be used correctly.

The objective of a research project determines the amount of effort to be devoted to the development of realistic models. If the intent is one of prediction only, the degree to which the model might be considered realistic is immaterial. The only requirement is that the predictions be adequately precise in the region of interest. On the other hand, realism is of primary importance if the goal is a thorough understanding of the system. The simple linear additive model can seldom be regarded as a realistic model. It is at best an approximation of the true model. Almost without exception, models developed from the basic principles of a process will be nonlinear in the parameters. The least squares estimation principle is still applicable but the mathematical methods become much more difficult. You are introduced to **nonlinear least squares regression methods** and some of the more common nonlinear models.

Least squares estimation is controlled by the correlational structure observed among the independent and dependent variables in the data set. Observational data, data collected by observing the state of nature according to some sampling plan, will frequently cause special problems for least squares estimation because of strong correlations or, more generally, near-linear dependencies among the independent variables. The seriousness of the problems will depend on the use to be made of the analyses. Understanding the correlational structure of the data is most helpful in interpreting regression results and deciding what inferences might be made. Principal component analysis is introduced as an aid in characterizing the correlational structure of the data. A graphical procedure, Gabriel's biplot, is introduced to help visualize the correlational structure. Principal component analysis also serves as an introduction to **biased regression methods**. Biased regression methods are designed to alleviate the deleterious effects of near-linear dependencies (among the independent variables) on ordinary least squares estimation.

Least squares estimation is a powerful research tool and, with modern low cost computers, is readily available. This ease of access, however, also facilitates misuse. Proper use of least squares requires an understanding of the basic method and assumptions on which it is built, and an awareness of the possible problems and their remedies. In some cases, alternative methods to least squares estimation might be more appropriate. It is the intent of this text to convey the basic understanding that will allow you to use least squares as an effective research tool.

The data sets used in this text are available on the internet at
http://www.stat.ncsu.edu/publications/rawlings/applied_least_squares
or through a link at the Springer-Verlag page. The "readme" file explains the contents of each data set.

Raleigh, North Carolina
March 4, 1998

John O. Rawlings
Sastry G. Pantula
David A. Dickey

CONTENTS