Regression Analysis by Example

WILEY SERIES IN PROBABILITY AND STATISTICS

Established by WALTER A. SHEWHART and SAMUEL S. WILKS

Editors: *David J. Balding, Noel A. C. Cressie, Nicholas I. Fisher, Iain M. Johnstone, J. B. Kadane, Geert Molenberghs, Louise M. Ryan, David W. Scott, Adrian F. M. Smith, Jozef L. Teugels*
Editors Emeriti: *Vic Barnett, J. Stuart Hunter, David G. Kendall*

A complete list of the titles in this series appears at the end of this volume.

Regression Analysis by Example

Fourth Edition

SAMPRIT CHATTERJEE
Department of Health Policy
Mount Sinai School of Medicine
New York, NY

ALI S. HADI
Department of Mathematics
The American University in Cairo
Cairo, Egypt

A JOHN WILEY & SONS, INC., PUBLICATION

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Library of Congress Cataloging-in-Publication Data:

Chatterjee, Samprit, 1938–
Regression analysis by example. — 4th ed. / Samprit Chatterjee, Ali S. Hadi.
p. cm.
Includes bibliographical references and index.
ISBN-13 978-0-471-74696-6 (cloth : acid-free paper)
ISBN-10 0-471-74696-7 (cloth : acid-free paper)
1. Regression analysis. I. Title.

QA278.2.C5 2006
519.5'36—dc22 2006044595

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

Dedicated to:

Allegra, Martha, and Rima – S. C.

My mother and the memory of my father – A. S. H.

It's a gift to be simple . . .

Old Shaker hymn

True knowledge is knowledge of why things are as they are, and not merely what they are.

Isaiah Berlin

CONTENTS

PREFACE

Regression analysis has become one of the most widely used statistical tools for analyzing multifactor data. It is appealing because it provides a conceptually simple method for investigating functional relationships among variables. The standard approach in regression analysis is to take data, fit a model, and then evaluate the fit using statistics such as t, F,and R^2. Our approach is much broader. We view regression analysis as a set of data analytic techniques that examine the interrelationships among a given set of variables. The emphasis is not on formal statistical tests and probability calculations. We argue for an informal analysis directed towards uncovering patterns in the data.

We utilize most standard and some not so standard summary statistics on the basis of their intuitive appeal. We rely heavily on graphical representations of the data, and employ many variations of plots of regression residuals. We are not overly concerned with precise probability evaluations. Graphical methods for exploring residuals can suggest model deficiencies or point to troublesome observations. Upon further investigation into their origin, the troublesome observations often turn out to be more informative than the well-behaved observations. We notice often that more information is obtained from a quick examination of a plot of residuals than from a formal test of statistical significance of some limited null-hypothesis. In short, the presentation in the chapters of this book is guided by the principles and concepts of exploratory data analysis.

Our presentation of the various concepts and techniques of regression analysis relies on carefully developed examples. In each example, we have isolated one

or two techniques and discussed them in some detail. The data were chosen to highlight the techniques being presented. Although when analyzing a given set of data it is usually necessary to employ many techniques, we have tried to choose the various data sets so that it would not be necessary to discuss the same technique more than once. Our hope is that after working through the book, the reader will be ready and able to analyze his/her data methodically, thoroughly, and confidently.

The emphasis in this book is on the analysis of data rather than on formulas, tests of hypotheses, or confidence intervals. Therefore no attempt has been made to derive the techniques. Techniques are described, the required assumptions are given, and finally, the success of the technique in the particular example is assessed. Although derivations of the techniques are not included, we have tried to refer the reader in each case to sources in which such discussion is available. Our hope is that some of these sources will be followed up by the reader who wants a more thorough grounding in theory.

We have taken for granted the availability of a computer and a statistical package. Recently there has been a qualitative change in the analysis of linear models, from model fitting to model building, from overall tests to clinical examinations of data, from macroscopic to the microscopic analysis. To do this kind of analysis a computer is essential and we have assumed its availability. Almost all of the analyses we use are now available in software packages. We are particularly heartened by the arrival of the package **R**, available on the Internet under the General Public License (GPL). The package has excellent computing and graphical features. It is also free!

The material presented is intended for anyone who is involved in analyzing data. The book should be helpful to those who have some knowledge of the basic concepts of statistics. In the university, it could be used as a text for a course on regression analysis for students whose specialization is not statistics, but, who nevertheless, use regression analysis quite extensively in their work. For students whose major emphasis is statistics, and who take a course on regression analysis from a book at the level of Rao (1973), Seber (1977), or Sen and Srivastava (1990), this book can be used to balance and complement the theoretical aspects of the subject with practical applications. Outside the university, this book can be profitably used by those people whose present approach to analyzing multifactor data consists of looking at standard computer output (t, F, R^2, standard errors, etc.), but who want to go beyond these summaries for a more thorough analysis.

The book has a Web site: http://www.ilr.cornell.edu/~hadi/RABE4. This Web site contains, among other things, all the data sets that are included in this book and more.

Several new topics have been introduced in this edition. The discussion in Section 2.10 about the regression line through the origin has been considerably expanded. In the chapter on variable selection (Chapter 11), we introduce information measures and illustrate their use. The information criteria help in variable selection by

balancing the conflicting requirements of accuracy and complexity. It is a useful tool for arriving at parsimonious models.

The chapter on logistic regression (Chapter 12) has been considerably expanded. This reflects the increased use of the logit models in statistical analysis. In addition to binary logistic regression, we have now included a discussion of multinomial logistic regression. This extends the application of logistic regression to more diverse situations. The categories in some multinomial are ordered, for example in attitude surveys. We also discuss the application of the logistic model to ordered response variable.

A new chapter titled Further Topics (Chapter 13) has been added to this edition. This chapter is intended to be an introduction to a more advanced study of regression analysis. The topics discussed are generalized linear models (GLM) and robust regression. We introduce the concept of GLM and discuss how the linear regression and logistic regression models can be regarded as special cases from a large family of linear models. This provides a unifying view of linear models. We discuss Poisson regression in the context of GLM, and its use for modeling count data.

We have attempted to write a book for a group of readers with diverse backgrounds. We have also tried to put emphasis on the art of data analysis rather than on the development of statistical theory.

We are fortunate to have had assistance and encouragement from several friends, colleagues, and associates. Some of our colleagues at New York University and Cornell University have used portions of the material in their courses and have shared with us their comments and comments of their students. Special thanks are due to our friend and former colleague Jeffrey Simonoff (New York University) for comments, suggestions, and general help. The students in our classes on regression analysis have all contributed by asking penetrating questions and demanding meaningful and understandable answers. Our special thanks go to Nedret Billor (Cukurova University, Turkey) and Sahar El-Sheneity (Cornell University) for their very careful reading of an earlier edition of this book. We also thank Amy Hendrickson for preparing the Latex style files and for responding to our Latex questions, and Dean Gonzalez for help with the production of some of the figures.

SAMPRIT CHATTERJEE
ALI S. HADI

Brooksville, Maine
Cairo, Egypt