

## 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

Copyright © 2006 by John Wiley & Sons, Inc. All rights reserved.

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

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-4470, or on the web at [www.copyright.com](http://www.copyright.com). Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at <http://www.wiley.com/go/permission>.

**Limit of Liability/Disclaimer of Warranty:** While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic format. For information about Wiley products, visit our web site at [www.wiley.com](http://www.wiley.com).

***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

---

<b>Preface</b>	<b>xiii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 What Is Regression Analysis?	1
1.2 Publicly Available Data Sets	2
1.3 Selected Applications of Regression Analysis	3
1.3.1 Agricultural Sciences	3
1.3.2 Industrial and Labor Relations	3
1.3.3 History	4
1.3.4 Government	6
1.3.5 Environmental Sciences	6
1.4 Steps in Regression Analysis	7
1.4.1 Statement of the Problem	11
1.4.2 Selection of Potentially Relevant Variables	11
1.4.3 Data Collection	11
1.4.4 Model Specification	12
1.4.5 Method of Fitting	14
1.4.6 Model Fitting	14
1.4.7 Model Criticism and Selection	16
1.4.8 Objectives of Regression Analysis	16
1.5 Scope and Organization of the Book	17
Exercises	18
	<b>vii</b>

<b>2</b>	<b>Simple Linear Regression</b>	<b>21</b>
2.1	Introduction	21
2.2	Covariance and Correlation Coefficient	21
2.3	Example: Computer Repair Data	26
2.4	The Simple Linear Regression Model	28
2.5	Parameter Estimation	29
2.6	Tests of Hypotheses	32
2.7	Confidence Intervals	37
2.8	Predictions	37
2.9	Measuring the Quality of Fit	39
2.10	Regression Line Through the Origin	42
2.11	Trivial Regression Models	44
2.12	Bibliographic Notes	45
	Exercises	45
<b>3</b>	<b>Multiple Linear Regression</b>	<b>53</b>
3.1	Introduction	53
3.2	Description of the Data and Model	53
3.3	Example: Supervisor Performance Data	54
3.4	Parameter Estimation	57
3.5	Interpretations of Regression Coefficients	58
3.6	Properties of the Least Squares Estimators	60
3.7	Multiple Correlation Coefficient	61
3.8	Inference for Individual Regression Coefficients	62
3.9	Tests of Hypotheses in a Linear Model	64
3.9.1	Testing All Regression Coefficients Equal to Zero	66
3.9.2	Testing a Subset of Regression Coefficients Equal to Zero	69
3.9.3	Testing the Equality of Regression Coefficients	71
3.9.4	Estimating and Testing of Regression Parameters Under Constraints	73
3.10	Predictions	74
3.11	Summary	75
	Exercises	75
	Appendix: Multiple Regression in Matrix Notation	82
<b>4</b>	<b>Regression Diagnostics: Detection of Model Violations</b>	<b>85</b>
4.1	Introduction	85
4.2	The Standard Regression Assumptions	86
4.3	Various Types of Residuals	88
4.4	Graphical Methods	90
4.5	Graphs Before Fitting a Model	93

4.5.1	One-Dimensional Graphs	93
4.5.2	Two-Dimensional Graphs	93
4.5.3	Rotating Plots	96
4.5.4	Dynamic Graphs	96
4.6	Graphs After Fitting a Model	97
4.7	Checking Linearity and Normality Assumptions	97
4.8	Leverage, Influence, and Outliers	98
4.8.1	Outliers in the Response Variable	100
4.8.2	Outliers in the Predictors	100
4.8.3	Masking and Swamping Problems	100
4.9	Measures of Influence	103
4.9.1	Cook's Distance	103
4.9.2	Welsch and Kuh Measure	104
4.9.3	Hadi's Influence Measure	105
4.10	The Potential-Residual Plot	107
4.11	What to Do with the Outliers?	108
4.12	Role of Variables in a Regression Equation	109
4.12.1	Added-Variable Plot	109
4.12.2	Residual Plus Component Plot	110
4.13	Effects of an Additional Predictor	114
4.14	Robust Regression	115
	Exercises	115
<b>5</b>	<b>Qualitative Variables as Predictors</b>	<b>121</b>
5.1	Introduction	121
5.2	Salary Survey Data	122
5.3	Interaction Variables	125
5.4	Systems of Regression Equations	128
5.4.1	Models with Different Slopes and Different Intercepts	130
5.4.2	Models with Same Slope and Different Intercepts	137
5.4.3	Models with Same Intercept and Different Slopes	138
5.5	Other Applications of Indicator Variables	139
5.6	Seasonality	140
5.7	Stability of Regression Parameters Over Time	141
	Exercises	143
<b>6</b>	<b>Transformation of Variables</b>	<b>151</b>
6.1	Introduction	151
6.2	Transformations to Achieve Linearity	153
6.3	Bacteria Deaths Due to X-Ray Radiation	155
6.3.1	Inadequacy of a Linear Model	156
6.3.2	Logarithmic Transformation for Achieving Linearity	158



6.4	Transformations to Stabilize Variance	159
6.5	Detection of Heteroscedastic Errors	164
6.6	Removal of Heteroscedasticity	166
6.7	Weighted Least Squares	167
6.8	Logarithmic Transformation of Data	168
6.9	Power Transformation	169
6.10	Summary	173
	Exercises	174
<b>7</b>	<b>Weighted Least Squares</b>	<b>179</b>
7.1	Introduction	179
7.2	Heteroscedastic Models	180
7.2.1	Supervisors Data	180
7.2.2	College Expense Data	182
7.3	Two-Stage Estimation	183
7.4	Education Expenditure Data	185
7.5	Fitting a Dose-Response Relationship Curve	194
	Exercises	196
<b>8</b>	<b>The Problem of Correlated Errors</b>	<b>197</b>
8.1	Introduction: Autocorrelation	197
8.2	Consumer Expenditure and Money Stock	198
8.3	Durbin-Watson Statistic	200
8.4	Removal of Autocorrelation by Transformation	202
8.5	Iterative Estimation With Autocorrelated Errors	204
8.6	Autocorrelation and Missing Variables	205
8.7	Analysis of Housing Starts	206
8.8	Limitations of Durbin-Watson Statistic	210
8.9	Indicator Variables to Remove Seasonality	211
8.10	Regressing Two Time Series	214
	Exercises	216
<b>9</b>	<b>Analysis of Collinear Data</b>	<b>221</b>
9.1	Introduction	221
9.2	Effects on Inference	222
9.3	Effects on Forecasting	228
9.4	Detection of Multicollinearity	233
9.5	Centering and Scaling	239
9.5.1	Centering and Scaling in Intercept Models	240
9.5.2	Scaling in No-Intercept Models	241
9.6	Principal Components Approach	243
9.7	Imposing Constraints	246

9.8	Searching for Linear Functions of the $\beta$ 's	248
9.9	Computations Using Principal Components	252
9.10	Bibliographic Notes	254
	Exercises	254
	Appendix: Principal Components	255
<b>10</b>	<b>Biased Estimation of Regression Coefficients</b>	<b>259</b>
10.1	Introduction	259
10.2	Principal Components Regression	260
10.3	Removing Dependence Among the Predictors	262
10.4	Constraints on the Regression Coefficients	264
10.5	Principal Components Regression: A Caution	265
10.6	Ridge Regression	268
10.7	Estimation by the Ridge Method	269
10.8	Ridge Regression: Some Remarks	272
10.9	Summary	275
	Exercises	275
	Appendix: Ridge Regression	277
<b>11</b>	<b>Variable Selection Procedures</b>	<b>281</b>
11.1	Introduction	281
11.2	Formulation of the Problem	282
11.3	Consequences of Variables Deletion	282
11.4	Uses of Regression Equations	284
	11.4.1 Description and Model Building	284
	11.4.2 Estimation and Prediction	284
	11.4.3 Control	284
11.5	Criteria for Evaluating Equations	285
	11.5.1 Residual Mean Square	285
	11.5.2 Mallows $C_p$	286
	11.5.3 Information Criteria: Akaike and Other Modified Forms	287
11.6	Multicollinearity and Variable Selection	288
11.7	Evaluating All Possible Equations	288
11.8	Variable Selection Procedures	289
	11.8.1 Forward Selection Procedure	289
	11.8.2 Backward Elimination Procedure	290
	11.8.3 Stepwise Method	290
11.9	General Remarks on Variable Selection Methods	291
11.10	A Study of Supervisor Performance	292
11.11	Variable Selection With Collinear Data	296
11.12	The Homicide Data	296

11.13	Variable Selection Using Ridge Regression	299
11.14	Selection of Variables in an Air Pollution Study	300
11.15	A Possible Strategy for Fitting Regression Models	307
11.16	Bibliographic Notes	308
	Exercises	308
	Appendix: Effects of Incorrect Model Specifications	313
<b>12</b>	<b>Logistic Regression</b>	<b>317</b>
12.1	Introduction	317
12.2	Modeling Qualitative Data	318
12.3	The Logit Model	318
12.4	Example: Estimating Probability of Bankruptcies	320
12.5	Logistic Regression Diagnostics	323
12.6	Determination of Variables to Retain	324
12.7	Judging the Fit of a Logistic Regression	327
12.8	The Multinomial Logit Model	329
	12.8.1 Multinomial Logistic Regression	329
	12.8.2 Example: Determining Chemical Diabetes	330
	12.8.3 Ordered Response Category: Ordinal Logistic Regression	334
	12.8.4 Example: Determining Chemical Diabetes Revisited	335
12.9	Classification Problem: Another Approach	336
	Exercises	337
<b>13</b>	<b>Further Topics</b>	<b>341</b>
13.1	Introduction	341
13.2	Generalized Linear Model	341
13.3	Poisson Regression Model	342
13.4	Introduction of New Drugs	343
13.5	Robust Regression	345
13.6	Fitting a Quadratic Model	346
13.7	Distribution of PCB in U.S. Bays	348
	Exercises	352
	<b>Appendix A: Statistical Tables</b>	<b>353</b>
	<b>References</b>	<b>363</b>
	<b>Index</b>	<b>371</b>

# 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.

SAMPRIIT CHATTERJEE  
ALI S. HADI

*Brooksville, Maine*  
*Cairo, Egypt*