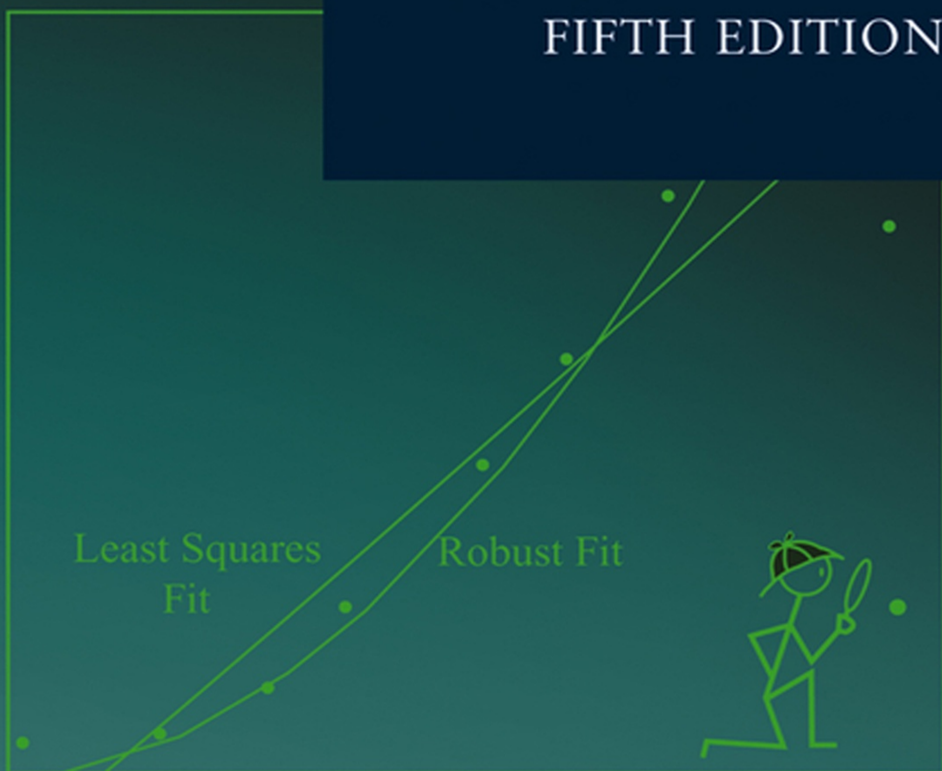
WILEY SERIES IN PROBABILITY AND STATISTICS

Regression Analysis *by* Example

FIFTH EDITION



Samprit Chatterjee and Ali S. Hadi

WILEY

WWW.
LINK AVAILABLE

Regession Analysis
By Example

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

Regression Analysis By Example

Fifth Edition

Samprit Chatterjee
Ali S. Hadi

WILEY

A JOHN WILEY & SONS, INC., PUBLICATION

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

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 formats. For more information about Wiley products, visit our web site at www.wiley.com.

Library of Congress Cataloging-in-Publication Data:

Chatterjee, Samprit, 1938–
Regression analysis by example / Samprit Chatterjee, New York University, Leonard N. Stern School of Business, New York, NY, Ali S. Hadi, American University in Cairo, Actuarial Science Program, Cairo, Egypt. — Fifth edition.
pages cm. — (Wiley series in probability and statistics)
Includes index.
ISBN 978-0-470-90584-5 (hardback)
1. Regression analysis. I. Hadi, Ali S. II. Title.
QA278.2.C5 2006
519.5'36—dc23 2012009795

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.

The memory of my parents – 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		**xiii**
1	**Introduction**		**1**
	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	4
		1.3.3 Government	6
		1.3.4 History	6
		1.3.5 Environmental Sciences	9
		1.3.6 Industrial Production	9
		1.3.7 The Space Shuttle Challenger	12
		1.3.8 Cost of Health Care	12
	1.4	Steps in Regression Analysis	13
		1.4.1 Statement of the Problem	13
		1.4.2 Selection of Potentially Relevant Variables	15
		1.4.3 Data Collection	15
		1.4.4 Model Specification	16
		1.4.5 Method of Fitting	19
		1.4.6 Model Fitting	19
		1.4.7 Model Criticism and Selection	19
		1.4.8 Objectives of Regression Analysis	20

1.5 Scope and Organization of the Book 21
Exercises 23

2 **Simple Linear Regression** **25**

2.1 Introduction 25
2.2 Covariance and Correlation Coefficient 25
2.3 Example: Computer Repair Data 30
2.4 The Simple Linear Regression Model 32
2.5 Parameter Estimation 33
2.6 Tests of Hypotheses 36
2.7 Confidence Intervals 41
2.8 Predictions 41
2.9 Measuring the Quality of Fit 43
2.10 Regression Line Through the Origin 46
2.11 Trivial Regression Models 48
2.12 Bibliographic Notes 49
Exercises 49

3 **Multiple Linear Regression** **57**

3.1 Introduction 57
3.2 Description of the Data and Model 57
3.3 Example: Supervisor Performance Data 58
3.4 Parameter Estimation 59
3.5 Interpretations of Regression Coefficients 62
3.6 Centering and Scaling 64
3.6.1 Centering and Scaling in Intercept Models 65
3.6.2 Scaling in No-Intercept Models 66
3.7 Properties of the Least Squares Estimators 67
3.8 Multiple Correlation Coefficient 68
3.9 Inference for Individual Regression Coefficients 69
3.10 Tests of Hypotheses in a Linear Model 71
3.10.1 Testing All Regression Coefficients Equal to Zero 73
3.10.2 Testing a Subset of Regression Coefficients Equal to Zero 75
3.10.3 Testing the Equality of Regression Coefficients 78
3.10.4 Estimating and Testing of Regression Parameters Under Constraints 79
3.11 Predictions 81
3.12 Summary 82
Exercises 82
Appendix: Multiple Regression in Matrix Notation 89

4 **Regression Diagnostics: Detection of Model Violations** 93

4.1 Introduction 93
4.2 The Standard Regression Assumptions 94
4.3 Various Types of Residuals 96
4.4 Graphical Methods 98
4.5 Graphs Before Fitting a Model 101
4.5.1 One-Dimensional Graphs 101
4.5.2 Two-Dimensional Graphs 101
4.5.3 Rotating Plots 104
4.5.4 Dynamic Graphs 104
4.6 Graphs After Fitting a Model 105
4.7 Checking Linearity and Normality Assumptions 105
4.8 Leverage, Influence, and Outliers 106
4.8.1 Outliers in the Response Variable 108
4.8.2 Outliers in the Predictors 108
4.8.3 Masking and Swamping Problems 108
4.9 Measures of Influence 111
4.9.1 Cook's Distance 111
4.9.2 Welsch and Kuh Measure 112
4.9.3 Hadi's Influence Measure 113
4.10 The Potential-Residual Plot 115
4.11 What to Do with the Outliers? 116
4.12 Role of Variables in a Regression Equation 117
4.12.1 Added-Variable Plot 117
4.12.2 Residual Plus Component Plot 118
4.13 Effects of an Additional Predictor 121
4.14 Robust Regression 123
Exercises 123

5 **Qualitative Variables as Predictors** 129

5.1 Introduction 129
5.2 Salary Survey Data 130
5.3 Interaction Variables 133
5.4 Systems of Regression Equations 137
5.4.1 Models with Different Slopes and Different Intercepts 138
5.4.2 Models with Same Slope and Different Intercepts 145
5.4.3 Models with Same Intercept and Different Slopes 146
5.5 Other Applications of Indicator Variables 147
5.6 Seasonality 148
5.7 Stability of Regression Parameters Over Time 150
Exercises 154

6 **Transformation of Variables** 163

6.1 Introduction 163
6.2 Transformations to Achieve Linearity 165
6.3 Bacteria Deaths Due to X-Ray Radiation 167
6.3.1 Inadequacy of a Linear Model 168
6.3.2 Logarithmic Transformation for Achieving Linearity 170
6.4 Transformations to Stabilize Variance 171
6.5 Detection of Heteroscedastic Errors 176
6.6 Removal of Heteroscedasticity 178
6.7 Weighted Least Squares 179
6.8 Logarithmic Transformation of Data 180
6.9 Power Transformation 181
6.10 Summary 185
Exercises 186

7 **Weighted Least Squares** 191

7.1 Introduction 191
7.2 Heteroscedastic Models 192
7.2.1 Supervisors Data 192
7.2.2 College Expense Data 194
7.3 Two-Stage Estimation 195
7.4 Education Expenditure Data 197
7.5 Fitting a Dose-Response Relationship Curve 206
Exercises 208

8 **The Problem of Correlated Errors** 209

8.1 Introduction: Autocorrelation 209
8.2 Consumer Expenditure and Money Stock 210
8.3 Durbin-Watson Statistic 212
8.4 Removal of Autocorrelation by Transformation 214
8.5 Iterative Estimation with Autocorrelated Errors 216
8.6 Autocorrelation and Missing Variables 217
8.7 Analysis of Housing Starts 218
8.8 Limitations of the Durbin-Watson Statistic 222
8.9 Indicator Variables to Remove Seasonality 223
8.10 Regressing Two Time Series 226
Exercises 228

9 **Analysis of Collinear Data** 233

9.1 Introduction 233
9.2 Effects of Collinearity on Inference 234
9.3 Effects of Collinearity on Forecasting 240

9.4 Detection of Collinearity 245
 9.4.1 Simple Signs of Collinearity 245
 9.4.2 Variance Inflation Factors 248
 9.4.3 The Condition Indices 251
 Exercises 255

10 Working With Collinear Data 259

10.1 Introduction 259
10.2 Principal Components 259
10.3 Computations Using Principal Components 263
10.4 Imposing Constraints 265
10.5 Searching for Linear Functions of the β's 268
10.6 Biased Estimation of Regression Coefficients 271
10.7 Principal Components Regression 272
10.8 Reduction of Collinearity in the Estimation Data 274
10.9 Constraints on the Regression Coefficients 276
10.10 Principal Components Regression: A Caution 277
10.11 Ridge Regression 279
10.12 Estimation by the Ridge Method 281
10.13 Ridge Regression: Some Remarks 286
10.14 Summary 287
10.15 Bibliographic Notes 287
 Exercises 288
 Appendix 10.A: Principal Components 292
 Appendix 10.B: Ridge Regression 294
 Appendix 10.C: Surrogate Ridge Regression 296

11 Variable Selection Procedures 299

11.1 Introduction 299
11.2 Formulation of the Problem 300
11.3 Consequences of Variables Deletion 300
11.4 Uses of Regression Equations 302
 11.4.1 Description and Model Building 302
 11.4.2 Estimation and Prediction 302
 11.4.3 Control 302
11.5 Criteria for Evaluating Equations 303
 11.5.1 Residual Mean Square 303
 11.5.2 Mallows C_p 304
 11.5.3 Information Criteria 305
11.6 Collinearity and Variable Selection 306
11.7 Evaluating All Possible Equations 306
11.8 Variable Selection Procedures 307
 11.8.1 Forward Selection Procedure 307

11.8.2 Backward Elimination Procedure 308
11.8.3 Stepwise Method 308
11.9 General Remarks on Variable Selection Methods 309
11.10 A Study of Supervisor Performance 310
11.11 Variable Selection with Collinear Data 314
11.12 The Homicide Data 314
11.13 Variable Selection Using Ridge Regression 317
11.14 Selection of Variables in an Air Pollution Study 318
11.15 A Possible Strategy for Fitting Regression Models 326
11.16 Bibliographic Notes 328
Exercises 328
Appendix: Effects of Incorrect Model Specifications 331

12 Logistic Regression 335

12.1 Introduction 335
12.2 Modeling Qualitative Data 336
12.3 The Logit Model 336
12.4 Example: Estimating Probability of Bankruptcies 338
12.5 Logistic Regression Diagnostics 341
12.6 Determination of Variables to Retain 342
12.7 Judging the Fit of a Logistic Regression 345
12.8 The Multinomial Logit Model 347
12.8.1 Multinomial Logistic Regression 347
12.8.2 Example: Determining Chemical Diabetes 348
12.8.3 Ordinal Logistic Regression 352
12.8.4 Example: Determining Chemical Diabetes Revisited 353
12.9 Classification Problem: Another Approach 354
Exercises 355

13 Further Topics 359

13.1 Introduction 359
13.2 Generalized Linear Model 359
13.3 Poisson Regression Model 360
13.4 Introduction of New Drugs 361
13.5 Robust Regression 363
13.6 Fitting a Quadratic Model 364
13.7 Distribution of PCB in U.S. Bays 366
Exercises 370

Appendix A: Statistical Tables 371

References 381

Index 389

PREFACE

It is with great pleasure that we introduce the fifth edition of *Regression Analysis by Example* first published in 1977. The statistical community has been most supportive, and we have benefitted greatly from their suggestions in improving the text.

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 toward 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 Website: http://www.aucegypt.edu/faculty/hadi/RABE5. This Website contains, among other things, all the data sets that are included in this book and more.

Major changes have been made in streamlining the text, removing ambiguities, and correcting errors pointed out by readers and others detected by the authors. New examples of data sets have been added in Chapter 1. The material on centering and scaling has been moved from Chapter 9 to Section 3.6. Chapters 9 and 10 have been rearranged, so the updated material flows more smoothly. The Appendix to Chapter 10 now includes a brief description of surrogate ridge regression, a recently introduced topic in the literature. New references have also been added. We have rewritten some of the exercises, and increased the number of exercises at the end of the chapters. We feel that the exercises reinforce the understanding of the material in the preceding chapters.

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

CHAPTER 1

INTRODUCTION

1.1 WHAT IS REGRESSION ANALYSIS?

Regression analysis is a conceptually simple method for investigating functional relationships among variables. A real estate appraiser may wish to relate the sale price of a home from selected physical characteristics of the building and taxes (local, school, county) paid on the building. We may wish to examine whether cigarette consumption is related to various socioeconomic and demographic variables such as age, education, income, and price of cigarettes. The relationship is expressed in the form of an equation or a model connecting the *response* or *dependent* variable and one or more *explanatory* or *predictor* variables. In the cigarette consumption example, the response variable is cigarette consumption (measured by the number of packs of cigarette sold in a given state on a per capita basis during a given year) and the explanatory or predictor variables are the various socioeconomic and demographic variables. In the real estate appraisal example, the response variable is the price of a home and the explanatory or predictor variables are the characteristics of the building and taxes paid on the building.

We denote the response variable by Y and the set of predictor variables by $X_1, X_2, \cdots, X_p$, where p denotes the number of predictor variables. The true relationship between Y and $X_1, X_2, \cdots, X_p$ can be approximated by the regression

Regression Analysis by Example, Fifth Edition. By Samprit Chatterjee and Ali S. Hadi model

$$Y = f(X_1, X_2, \cdots, X_p) + \varepsilon, \tag{1.1}$$

where ε is assumed to be a random error representing the discrepancy in the approximation. It accounts for the failure of the model to fit the data exactly. The function $f(X_1, X_2, \cdots, X_p)$ describes the relationship between Y and $X_1, X_2, \cdots$, X_p. An example is the linear regression model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon, \tag{1.2}$$

where $\beta_0, \beta_1, \cdots, \beta_p$, called the *regression parameters* or *coefficients*, are unknown constants to be determined (estimated) from the data. We follow the commonly used notational convention of denoting unknown parameters by Greek letters.

The predictor or explanatory variables are also called by other names such as *independent* variables, *covariates*, *regressors*, *factors*, and *carriers*. The name independent variable, though commonly used, is the least preferred, because in practice the predictor variables are rarely independent of each other.

1.2 PUBLICLY AVAILABLE DATA SETS

Regression analysis has numerous areas of applications. A partial list would include economics, finance, business, law, meteorology, medicine, biology, chemistry, engineering, physics, education, sports, history, sociology, and psychology. A few examples of such applications are given in Section 1.3. Regression analysis is learned most effectively by analyzing data that are of direct interest to the reader. We invite the readers to think about questions (in their own areas of work, research, or interest) that can be addressed using regression analysis. Readers should collect the relevant data and then apply the regression analysis techniques presented in this book to their own data. To help the reader locate real-life data, this section provides some sources and links to a wealth of data sets that are available for public use.

A number of data sets are available in books and on the Internet. The book by Hand et al. (1994) contains data sets from many fields. These data sets are small in size and are suitable for use as exercises. The book by Chatterjee, Handcock, and Simonoff (1995) provides numerous data sets from diverse fields. The data are included in a diskette that comes with the book and can also be found at the Website.[1]

Data sets are also available on the Internet at many other sites. Some of the Websites given below allow the direct copying and pasting into the statistical package of choice, while others require downloading the data file and then importing them into a statistical package. Some of these sites also contain further links to yet other data sets or statistics-related Websites.

The Data and Story Library (DASL, pronounced "dazzle") is one of the most interesting sites that contains a number of data sets accompanied by the "story" or

[1] http://www.stern.nyu.edu/~jsimonof/Casebook

background associated with each data set. DASL is an online library[2] of data files and stories that illustrate the use of basic statistical methods. The data sets cover a wide variety of topics. DASL comes with a powerful search engine to locate the story or data file of interest.

Another Website, which also contains data sets arranged by the method used in the analysis, is the Electronic Dataset Service.[3] The site also contains many links to other data sources on the Internet.

Finally, this book has a Website[4], which contains, among other things, all the data sets that are included in this book and more. These and other data sets can be found at the book's Website.

1.3 SELECTED APPLICATIONS OF REGRESSION ANALYSIS

Regression analysis is one of the most widely used statistical tools because it provides simple methods for establishing a functional relationship among variables. It has extensive applications in many subject areas. The cigarette consumption and the real estate appraisal, mentioned above, are but two examples. In this section, we give a few additional examples demonstrating the wide applicability of regression analysis in real-life situations. Some of the data sets described here will be used later in the book to illustrate regression techniques or in the exercises at the end of various chapters.

1.3.1 Agricultural Sciences

The Dairy Herd Improvement Cooperative (DHI) in upstate New York collects and analyzes data on milk production. One question of interest here is how to develop a suitable model to predict current milk production from a set of measured variables. The response variable (current milk production in pounds) and the predictor variables are given in Table 1.1. Samples are taken once a month during milking. The period that a cow gives milk is called lactation. Number of lactations is the number of times a cow has calved or given milk. The recommended management practice is to have the cow produce milk for about 305 days and then allow a 60-day rest period before beginning the next lactation. The data set, consisting of 199 observations, was compiled from the DHI milk production records. The Milk Production data can be found at the book's Website.

---

[2] http://lib.stat.cmu.edu/DASL

[3] http://www-unix.oit.umass.edu/~statdata

[4] http://www.aucegypt.edu/faculty/hadi/RABE5

Table 1.1 Variables in Milk Production Data

Variable	Definition
Current	Current month milk production in pounds
Previous	Previous month milk production in pounds
Fat	Percent of fat in milk
Protein	Percent of protein in milk
Days	Number of days since present lactation
Lactation	Number of lactations
I79	Indicator variable (0 if Days $\leq$ 79 and 1 if Days $>$ 79)

Table 1.2 Variables in Right-To-Work Laws Data

Variable	Definition
COL	Cost of living for a four-person family
PD	Population density (person per square mile)
URate	State unionization rate in 1978
Pop	Population in 1975
Taxes	Property taxes in 1972
Income	Per capita income in 1974
RTWL	Indicator variable (1 if there are right-to-work laws in the state and 0 otherwise)

1.3.2 Industrial and Labor Relations

In 1947, the United States Congress passed the Taft-Hartley Amendments to the Wagner Act. The original Wagner Act had permitted the unions to use a *Closed Shop Contract*[5] unless prohibited by state law. The Taft-Hartley Amendments made the use of Closed Shop Contract illegal and gave individual states the right to prohibit union shops[6] as well. These right-to-work laws have caused a wave of concern throughout the labor movement. A question of interest here is: What are the effects of these laws on the cost of living for a four-person family living on an intermediate budget in the United States? To answer this question a data set consisting of 38 geographic locations has been assembled from various sources. The variables used are defined in Table 1.2. The Right-To-Work Laws data are given in Table 1.3 and can also be found at the book's Website.

[5] Under a Closed Shop Contract provision, all employees must be union members at the time of hire and must remain members as a condition of employment.

[6] Under a Union Shop clause, employees are not required to be union members at the time of hire, but must become a member within two months, thus allowing the employer complete discretion in hiring decisions.

Table 1.3 The Right-To-Work Laws Data

City	COL	PD	URate	Pop	Taxes	Income	RTWL
Atlanta	169	414	13.6	1790128	5128	2961	1
Austin	143	239	11	396891	4303	1711	1
Bakersfield	339	43	23.7	349874	4166	2122	0
Baltimore	173	951	21	2147850	5001	4654	0
Baton Rouge	99	255	16	411725	3965	1620	1
Boston	363	1257	24.4	3914071	4928	5634	0
Buffalo	253	834	39.2	1326848	4471	7213	0
Champaign-Urbana	117	162	31.5	162304	4813	5535	0
Cedar Rapids	294	229	18.2	164145	4839	7224	1
Chicago	291	1886	31.5	7015251	5408	6113	0
Cincinnati	170	643	29.5	1381196	4637	4806	0
Cleveland	239	1295	29.5	1966725	5138	6432	0
Dallas	174	302	11	2527224	4923	2363	1
Dayton	183	489	29.5	835708	4787	5606	0
Denver	227	304	15.2	1413318	5386	5982	0
Detriot	255	1130	34.6	4424382	5246	6275	0
Green Bay	249	323	27.8	169467	4289	8214	0
Hartford	326	696	21.9	1062565	5134	6235	0
Houston	194	337	11	2286247	5084	1278	1
Indianapolis	251	371	29.3	1138753	4837	5699	0
Kansas City	201	386	30	1290110	5052	4868	0
Lancaster, PA	124	362	34.2	342797	4377	5205	0
Los Angeles	340	1717	23.7	6986898	5281	1349	0
Milwaukee	328	968	27.8	1409363	5176	7635	0
Minneapolis, St. Paul	265	433	24.4	2010841	5206	8392	0
Nashville	120	183	17.7	748493	4454	3578	1
New York	323	6908	39.2	9561089	5260	4862	0
Orlando	117	230	11.7	582664	4613	782	1
Philadelphia	182	1353	34.2	4807001	4877	5144	0
Pittsburgh	169	762	34.2	2322224	4677	5987	0
Portland	267	201	23.1	228417	4123	7511	0
St. Louis	184	480	30	2366542	4721	4809	0
San Diego	256	372	23.7	1584583	4837	1458	0
San Francisco	381	1266	23.7	3140306	5940	3015	0
Seattle	195	333	33.1	1406746	5416	4424	0
Washington	205	1073	21	3021801	6404	4224	0
Wichita	206	157	12.8	384920	4796	4620	1
Raleigh-Durham	126	302	6.5	468512	4614	3393	1

Table 1.4 Variables in Study of Domestic Immigration

Variable	Definition
State	State name
NDIR	Net domestic immigration rate over the period 1990–1994
Unemp	Unemployment rate in the civilian labor force in 1994
Wage	Average hourly earnings of production workers in manufacturing in 1994
Crime	Violent crime rate per 100,000 people in 1993
Income	Median household income in 1994
Metrop	Percentage of state population living in metropolitan areas in 1992
Poor	Percentage of population who fall below the poverty level in 1994
Taxes	Total state and local taxes per capita in 1993
Educ	Percentage of population 25 years or older who have a high school degree or higher in 1990
BusFail	The number of business failures divided by the population of the state in 1993
Temp	Average of the 12 monthly average temperatures (in degrees Fahrenheit) for the state in 1993
Region	Region in which the state is located (northeast, south, midwest, west)

1.3.3 Government

Information about domestic immigration (the movement of people from one state or area of a country to another) is important to state and local governments. It is of interest to build a model that predicts domestic immigration or to answer the question of why do people leave one place to go to another? There are many factors that influence domestic immigration, such as weather conditions, crime, taxes, and unemployment rates. A data set for the 48 contiguous states has been created. Alaska and Hawaii are excluded from the analysis because the environments of these states are significantly different from the other 48, and their locations present certain barriers to immigration. The response variable here is net domestic immigration, which represents the net movement of people into and out of a state over the period 1990–1994 divided by the population of the state. Eleven predictor variables thought to influence domestic immigration are defined in Table 1.4. The data are given in Tables 1.5 and 1.6, and can also be found at the book's Website.

1.3.4 History

A question of historical interest is how to estimate the age of historical objects based on some age-related characteristics of the objects. For example, the variables

Table 1.5 First Six Variables of Domestic Immigration Data

State	NDIR	Unemp	Wage	Crime	Income	Metrop
Alabama	17.47	6.0	10.75	780	27196	67.4
Arizona	49.60	6.4	11.17	715	31293	84.7
Arkansas	23.62	5.3	9.65	593	25565	44.7
California	−37.21	8.6	12.44	1078	35331	96.7
Colorado	53.17	4.2	12.27	567	37833	81.8
Connecticut	−38.41	5.6	13.53	456	41097	95.7
Delaware	22.43	4.9	13.90	686	35873	82.7
Florida	39.73	6.6	9.97	1206	29294	93.0
Georgia	39.24	5.2	10.35	723	31467	67.7
Idaho	71.41	5.6	11.88	282	31536	30.0
Illinois	−20.87	5.7	12.26	960	35081	84.0
Indiana	9.04	4.9	13.56	489	27858	71.6
Iowa	0.00	3.7	12.47	326	33079	43.8
Kansas	−1.25	5.3	12.14	469	28322	54.6
Kentucky	13.44	5.4	11.82	463	26595	48.5
Louisiana	−13.94	8.0	13.13	1062	25676	75.0
Maine	−9.770	7.4	11.68	126	30316	35.7
Maryland	−1.55	5.1	13.15	998	39198	92.8
Massachusetts	−30.46	6.0	12.59	805	40500	96.2
Michigan	−13.19	5.9	16.13	792	35284	82.7
Minnesota	9.46	4.0	12.60	327	33644	69.3
Mississippi	5.33	6.6	9.40	434	25400	34.6
Missouri	6.97	4.9	11.78	744	30190	68.3
Montana	41.50	5.1	12.50	178	27631	24.0
Nebraska	−0.62	2.9	10.94	339	31794	50.6
Nevada	128.52	6.2	11.83	875	35871	84.8
New Hampshire	−8.72	4.6	11.73	138	35245	59.4
New Jersey	−24.90	6.8	13.38	627	42280	100.0
New Mexico	29.05	6.3	10.14	930	26905	56.0
New York	−45.46	6.9	12.19	1074	31899	91.7
North Carolina	29.46	4.4	10.19	679	30114	66.3
North Dakota	−26.47	3.9	10.19	82	28278	41.6
Ohio	−3.27	5.5	14.38	504	31855	81.3
Oklahoma	7.37	5.8	11.41	635	26991	60.1
Oregon	49.63	5.4	12.31	503	31456	70.0
Pennsylvania	−4.30	6.2	12.49	418	32066	84.8
Rhode Island	−35.32	7.1	10.35	402	31928	93.6
South Carolina	11.88	6.3	9.99	1023	29846	69.8
South Dakota	13.71	3.3	9.19	208	29733	32.6
Tennessee	32.11	4.8	10.51	766	28639	67.7
Texas	13.00	6.4	11.14	762	30775	83.9
Utah	31.25	3.7	11.26	301	35716	77.5
Vermont	3.94	4.7	11.54	114	35802	27.0
Virginia	6.94	4.9	11.25	372	37647	77.5
Washington	44.66	6.4	14.42	515	33533	83.0
West Virginia	10.75	8.9	12.60	208	23564	41.8
Wisconsin	11.73	4.7	12.41	264	35388	68.1
Wyoming	11.95	5.3	11.81	286	33140	29.7

Table 1.6 Last Six Variables of Domestic Immigration Data

State	Poor	Taxes	Educ	BusFail	Temp	Region
Alabama	16.4	1553	66.9	0.20	62.77	South
Arizona	15.9	2122	78.7	0.51	61.09	West
Arkansas	15.3	1590	66.3	0.08	59.57	South
California	17.9	2396	76.2	0.63	59.25	West
Colorado	9.0	2092	84.4	0.42	43.43	West
Connecticut	10.8	3334	79.2	0.33	48.63	Northeast
Delaware	8.3	2336	77.5	0.19	54.58	South
Florida	14.9	2048	74.4	0.36	70.64	South
Georgia	14.0	1999	70.9	0.33	63.54	South
Idaho	12.0	1916	79.7	0.31	42.35	West
Illinois	12.4	2332	76.2	0.18	50.98	Midwest
Indiana	13.7	1919	75.6	0.19	50.88	Midwest
Iowa	10.7	2200	80.1	0.18	45.83	Midwest
Kansas	14.9	2126	81.3	0.42	52.03	Midwest
Kentucky	18.5	1816	64.6	0.22	55.36	South
Louisiana	25.7	1685	68.3	0.15	65.91	South
Maine	9.4	2281	78.8	0.31	40.23	Northeast
Maryland	10.7	2565	78.4	0.31	54.04	South
Massachusetts	9.7	2664	80.0	0.45	47.35	Northeast
Michigan	14.1	2371	76.8	0.27	43.68	Midwest
Minnesota	11.7	2673	82.4	0.20	39.30	Midwest
Mississippi	19.9	1535	64.3	0.12	63.18	South
Missouri	15.6	1721	73.9	0.23	53.41	Midwest
Montana	11.5	1853	81.0	0.20	40.40	West
Nebraska	8.8	2128	81.8	0.25	46.01	Midwest
Nevada	11.1	2289	78.8	0.39	48.23	West
New Hampshire	7.7	2305	82.2	0.54	43.53	Northeast
New Jersey	9.2	3051	76.7	0.36	52.72	Northeast
New Mexico	21.1	2131	75.1	0.27	53.37	Midwest
New York	17.0	3655	74.8	0.38	44.85	Northeast
North Carolina	14.2	1975	70.0	0.17	59.36	South
North Dakota	10.4	1986	76.7	0.23	38.53	Midwest
Ohio	14.1	2059	75.7	0.19	50.87	Midwest
Oklahoma	16.7	1777	74.6	0.44	58.36	South
Oregon	11.8	2169	81.5	0.31	46.55	West
Pennsylvania	12.5	2260	74.7	0.26	49.01	Northeast
Rhode Island	10.3	2405	72.0	0.35	49.99	Northeast
South Carolina	13.8	1736	68.3	0.11	62.53	South
South Dakota	14.5	1668	77.1	0.24	42.89	Midwest
Tennessee	14.6	1684	67.1	0.23	57.75	South
Texas	19.1	1932	72.1	0.39	64.40	South
Utah	8.0	1806	85.1	0.18	46.32	West
Vermont	7.6	2379	80.8	0.30	42.46	Northeast
Virginia	10.7	2073	75.2	0.27	55.55	South
Washington	11.7	2433	83.8	0.38	46.93	Midwest
West Virginia	18.6	1752	66.0	0.17	52.25	South
Wisconsin	9.0	2524	78.6	0.24	42.20	Midwest
Wyoming	9.3	2295	83.0	0.19	43.68	West

Table 1.7 Variables in Egyptian Skulls Data

Variable	Definition
Year	Approximate year of skull formation (negative = B.C.; positive = A.D.)
MB	Maximum breadth of skull
BH	Basibregmatic height of skull
BL	Basialveolar length of skull
NH	Nasal Height of skull

in Table 1.7 can be used to estimate the age of Egyptian skulls. Here the response variable is Year and the other four variables are possible predictors. There are 150 observations in this data set. The original source of the data is Thomson and Randall-Maciver (1905), but they can be found in Hand et al. (1994), pp. 299–301. An analysis of the data can be found in Manly (1986). The Egyptian Skulls data can be found at the book's Website.

1.3.5 Environmental Sciences

In a 1976 study exploring the relationship between water quality and land use, Haith (1976) obtained the measurements (shown in Table 1.8) on 20 river basins in New York State. A question of interest here is how the land use around a river basin contributes to the water pollution as measured by the mean nitrogen concentration (mg/liter). The data are shown in Table 1.9 and can also be found at the book's Website.

1.3.6 Industrial Production

Nambe Mills in Santa Fe, New Mexico, makes a line of tableware made from sand casting a special alloy of metals. After casting, the pieces go through a series of shaping, grinding, buffing, and polishing steps. Data was collected for 59 items produced by the company. The relation between the polishing time and the product diameters and the product types (Bowl, Casserole, Dish, Tray, and Plate) are used to estimate the polishing time for new products which are designed or suggested for design and manufacture. The data are given in Table 1.10. The variables representing product types are coded as binary variables (1 corresponds to the type and 0 otherwise). Diam is the diameter of the item (in inches), polishing time is measured in minutes, and price in dollars. The polishing time is the major item in the cost of the product. The production decision will be based on the estimated time of polishing. The data is obtained from the DASL library, can be found there, and also at the book's Website.

Table 1.8 Variables in Study of Water Pollution in New York Rivers

Variable	Definition
Y	Mean nitrogen concentration (mg/liter) based on samples taken at regular intervals during the spring, summer, and fall months
X_1	Agriculture: percentage of land area currently in agricultural use
X_2	Forest: percentage of forest land
X_3	Residential: percentage of land area in residential use
X_4	Commercial/Industrial: percentage of land area in either commercial or industrial use

Table 1.9 New York Rivers Data

Row	River	Y	X_1	X_2	X_3	X_4
1	Olean	1.10	26	63	1.2	0.29
2	Cassadaga	1.01	29	57	0.7	0.09
3	Oatka	1.90	54	26	1.8	0.58
4	Neversink	1.00	2	84	1.9	1.98
5	Hackensack	1.99	3	27	29.4	3.11
6	Wappinger	1.42	19	61	3.4	0.56
7	Fishkill	2.04	16	60	5.6	1.11
8	Honeoye	1.65	40	43	1.3	0.24
9	Susquehanna	1.01	28	62	1.1	0.15
10	Chenango	1.21	26	60	0.9	0.23
11	Tioughnioga	1.33	26	53	0.9	0.18
12	West Canada	0.75	15	75	0.7	0.16
13	East Canada	0.73	6	84	0.5	0.12
14	Saranac	0.80	3	81	0.8	0.35
15	Ausable	0.76	2	89	0.7	0.35
16	Black	0.87	6	82	0.5	0.15
17	Schoharie	0.80	22	70	0.9	0.22
18	Raquette	0.87	4	75	0.4	0.18
19	Oswegatchie	0.66	21	56	0.5	0.13
20	Cohocton	1.25	40	49	1.1	0.13

Table 1.10 Industrial Production

Row	Bowl	Casserole	Dish	Tray	Plate	Diam	Time	Price
1	0	1	0	0	0	10.7	47.65	144.0
2	0	1	0	0	0	14.0	63.13	215.0
3	0	1	0	0	0	9.0	58.76	105.0
4	1	0	0	0	0	8.0	34.88	69.0
5	0	0	1	0	0	10.0	55.53	134.0
6	0	1	0	0	0	10.5	43.14	129.0
7	0	0	0	1	0	16.0	54.86	155.0
8	0	0	0	1	0	15.0	44.14	99.0
9	0	0	1	0	0	6.5	17.46	38.5
10	0	0	1	0	0	5.0	21.04	36.5
11	0	0	0	1	0	25.0	109.38	260.0
12	1	0	0	0	0	10.4	17.67	54.0
13	1	0	0	0	0	7.4	16.41	39.0
14	1	0	0	0	0	5.4	12.02	29.5
15	0	1	0	0	0	15.4	49.48	109.0
16	0	1	0	0	0	12.4	48.74	89.5
17	1	0	0	0	0	6.0	23.21	42.0
18	1	0	0	0	0	9.0	28.64	65.0
19	1	0	0	0	0	9.0	44.95	115.0
20	0	0	0	0	1	12.4	23.77	49.5
21	1	0	0	0	0	7.5	20.21	36.5
22	1	0	0	0	0	14.0	32.62	109.0
23	1	0	0	0	0	7.0	17.84	45.0
24	1	0	0	0	0	9.0	22.82	58.0
25	1	0	0	0	0	12.0	29.48	89.0
26	1	0	0	0	0	5.5	15.61	30.0
27	1	0	0	0	0	6.0	13.25	31.0
28	1	0	0	0	0	12	45.78	119.0
29	0	0	0	1	0	5.5	26.53	22.0
30	1	0	0	0	0	14.2	37.11	109.0
31	0	0	1	0	0	11.0	45.12	99.0
32	0	0	0	0	1	16.0	26.09	99.0
33	0	1	0	0	0	13.5	68.63	179.0
34	0	0	1	0	0	11.1	33.71	99.0
35	0	0	1	0	0	9.8	44.45	89.0
36	1	0	0	0	0	10.0	23.74	75.0
37	0	1	0	0	0	13.0	86.42	199.0
38	1	0	0	0	0	13.0	39.71	93.0
39	0	0	0	0	1	11.7	26.52	65.0
40	0	0	0	1	0	12.3	33.89	74.0
41	0	0	0	1	0	19.5	64.30	165.0
42	1	0	0	0	0	15.2	22.55	99.0
43	0	0	0	0	1	10.0	31.86	43.5
44	1	0	0	0	0	11.0	53.18	94.0
45	0	0	0	1	0	17.8	74.48	189.0
46	0	0	0	1	0	11.5	34.16	75.0
47	0	0	0	1	0	12.7	31.46	59.5
48	1	0	0	0	0	8.0	21.34	42.0
49	0	0	0	1	0	7.5	20.83	23.0
50	1	0	0	0	0	9.0	20.59	52.5
51	0	1	0	0	0	14.0	33.70	99.0
52	0	1	0	0	0	12.4	32.90	89.0
53	0	0	1	0	0	8.8	27.76	65.0
54	1	0	0	0	0	8.5	30.20	54.5
55	0	0	0	0	1	6.0	20.85	24.5
56	0	0	0	0	1	11.0	26.25	52.0
57	0	0	0	0	1	11.1	21.87	62.5
58	0	0	0	0	1	14.5	23.88	89.0
59	0	0	0	0	1	5.0	16.66	21.5

Table 1.11 Number of O-rings Damaged and Temperature (Degrees Fahrenheit) at Time of Launch for 23 Flights of Space Shuttle *Challenger*

Flight	Damaged	Temperature	Flight	Damaged	Temperature
1	2	53	13	1	70
2	1	57	14	1	70
3	1	58	15	0	72
4	1	63	16	0	73
5	0	66	17	0	75
6	0	67	18	2	75
7	0	67	19	0	76
8	0	67	20	0	78
9	0	68	21	0	79
10	0	69	22	0	81
11	0	70	23	0	76
12	0	70			

1.3.7 The Space Shuttle Challenger

The explosion of the space shuttle Challenger in 1986 killing the crew was a shattering tragedy. To look into the case a Presidential Commission was appointed. The O-rings in the booster rockets, which are used in space launching, play a very important part in its safety. The rigidity of the O-rings is thought to be affected by the temperature at launching. There are six O-rings in a booster rocket. Table 1.11 gives the number of rings damaged and the temperature at launchings of the 23 flights. The data set can also be found at the book's Website. The analysis performed before the launch did not include the launches in which no O-ring was damaged and came to the wrong conclusion. A detailed discussion of the problem is found in The *Flight of the Space Shuttle Challenger* in Chatterjee, Handcock, and Simonoff (1995, pp. 33–35). Note here that the response variable is a proportion bounded between 0 and 1.

1.3.8 Cost of Health Care

The cost of delivery of health care has become an important concern. Getting data on this topic is extremely difficult because it is highly proprietary. These data were collected by the Department of Health and Social Services of the State of New Mexico and cover 52 of the 60 licensed facilities in New Mexico in 1988. The variables in these data are the characteristics which describe the facilities size, volume of usage, expenditures, and revenue. The location of the facility is also indicated, whether it is in the rural or nonrural area. Specific definitions of the variables are given in Table 1.12 and the data are given in Table 1.13 and also at the book's Website. There are several ways of looking at a body of data and extracting various kinds of information. For example, (a) Are rural facilities different from