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
Exercises 23

2 Simple Linear Regression

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

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
<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>Transformation of Variables</td>
<td>163</td>
</tr>
<tr>
<td>6.1</td>
<td>Introduction</td>
<td>163</td>
</tr>
<tr>
<td>6.2</td>
<td>Transformations to Achieve Linearity</td>
<td>165</td>
</tr>
<tr>
<td>6.3</td>
<td>Bacteria Deaths Due to X-Ray Radiation</td>
<td>167</td>
</tr>
<tr>
<td>6.3.1</td>
<td>Inadequacy of a Linear Model</td>
<td>168</td>
</tr>
<tr>
<td>6.3.2</td>
<td>Logarithmic Transformation for Achieving Linearity</td>
<td>170</td>
</tr>
<tr>
<td>6.4</td>
<td>Transformations to Stabilize Variance</td>
<td>171</td>
</tr>
<tr>
<td>6.5</td>
<td>Detection of Heteroscedastic Errors</td>
<td>176</td>
</tr>
<tr>
<td>6.6</td>
<td>Removal of Heteroscedasticity</td>
<td>178</td>
</tr>
<tr>
<td>6.7</td>
<td>Weighted Least Squares</td>
<td>179</td>
</tr>
<tr>
<td>6.8</td>
<td>Logarithmic Transformation of Data</td>
<td>180</td>
</tr>
<tr>
<td>6.9</td>
<td>Power Transformation</td>
<td>181</td>
</tr>
<tr>
<td>6.10</td>
<td>Summary</td>
<td>185</td>
</tr>
<tr>
<td></td>
<td>Exercises</td>
<td>186</td>
</tr>
<tr>
<td>7</td>
<td>Weighted Least Squares</td>
<td>191</td>
</tr>
<tr>
<td>7.1</td>
<td>Introduction</td>
<td>191</td>
</tr>
<tr>
<td>7.2</td>
<td>Heteroscedastic Models</td>
<td>192</td>
</tr>
<tr>
<td>7.2.1</td>
<td>Supervisors Data</td>
<td>192</td>
</tr>
<tr>
<td>7.2.2</td>
<td>College Expense Data</td>
<td>194</td>
</tr>
<tr>
<td>7.3</td>
<td>Two-Stage Estimation</td>
<td>195</td>
</tr>
<tr>
<td>7.4</td>
<td>Education Expenditure Data</td>
<td>197</td>
</tr>
<tr>
<td>7.5</td>
<td>Fitting a Dose-Response Relationship Curve</td>
<td>206</td>
</tr>
<tr>
<td></td>
<td>Exercises</td>
<td>208</td>
</tr>
<tr>
<td>8</td>
<td>The Problem of Correlated Errors</td>
<td>209</td>
</tr>
<tr>
<td>8.1</td>
<td>Introduction: Autocorrelation</td>
<td>209</td>
</tr>
<tr>
<td>8.2</td>
<td>Consumer Expenditure and Money Stock</td>
<td>210</td>
</tr>
<tr>
<td>8.3</td>
<td>Durbin-Watson Statistic</td>
<td>212</td>
</tr>
<tr>
<td>8.4</td>
<td>Removal of Autocorrelation by Transformation</td>
<td>214</td>
</tr>
<tr>
<td>8.5</td>
<td>Iterative Estimation with Autocorrelated Errors</td>
<td>216</td>
</tr>
<tr>
<td>8.6</td>
<td>Autocorrelation and Missing Variables</td>
<td>217</td>
</tr>
<tr>
<td>8.7</td>
<td>Analysis of Housing Starts</td>
<td>218</td>
</tr>
<tr>
<td>8.8</td>
<td>Limitations of the Durbin-Watson Statistic</td>
<td>222</td>
</tr>
<tr>
<td>8.9</td>
<td>Indicator Variables to Remove Seasonality</td>
<td>223</td>
</tr>
<tr>
<td>8.10</td>
<td>Regressing Two Time Series</td>
<td>226</td>
</tr>
<tr>
<td></td>
<td>Exercises</td>
<td>228</td>
</tr>
<tr>
<td>9</td>
<td>Analysis of Collinear Data</td>
<td>233</td>
</tr>
<tr>
<td>9.1</td>
<td>Introduction</td>
<td>233</td>
</tr>
<tr>
<td>9.2</td>
<td>Effects of Collinearity on Inference</td>
<td>234</td>
</tr>
<tr>
<td>9.3</td>
<td>Effects of Collinearity on Forecasting</td>
<td>240</td>
</tr>
</tbody>
</table>
9.4 Detection of Collinearity
9.4.1 Simple Signs of Collinearity
9.4.2 Variance Inflation Factors
9.4.3 The Condition Indices
Exercises

10 Working With Collinear Data
10.1 Introduction
10.2 Principal Components
10.3 Computations Using Principal Components
10.4 Imposing Constraints
10.5 Searching for Linear Functions of the $\beta$'s
10.6 Biased Estimation of Regression Coefficients
10.7 Principal Components Regression
10.8 Reduction of Collinearity in the Estimation Data
10.9 Constraints on the Regression Coefficients
10.10 Principal Components Regression: A Caution
10.11 Ridge Regression
10.12 Estimation by the Ridge Method
10.13 Ridge Regression: Some Remarks
10.14 Summary
10.15 Bibliographic Notes
Exercises
Appendix 10.A: Principal Components
Appendix 10.B: Ridge Regression
Appendix 10.C: Surrogate Ridge Regression

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