

CONTENTS

Preface	xiii
About the Companion Website	xvii
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	5
1.3.4 History	5
1.3.5 Environmental Sciences	7
1.3.6 Industrial Production	7
1.3.7 The Space Shuttle Challenger	8
1.3.8 Cost of Health Care	8
1.4 Steps in Regression Analysis	9
1.4.1 Statement of the Problem	9
1.4.2 Selection of Potentially Relevant Variables	10
1.4.3 Data Collection	10
1.4.4 Model Specification	11
1.4.5 Method of Fitting	13
1.4.6 Model Fitting	14
1.4.7 Model Criticism and Selection	15
1.4.8 Objectives of Regression Analysis	16

1.5	Scope and Organization of the Book	16
	Exercises	19
2	A Brief Introduction to R	21
2.1	What Is R and RStudio?	21
2.2	Installing R and RStudio	22
2.3	Getting Started With R	23
2.3.1	Command Level Prompt	23
2.3.2	Calculations Using R	24
2.3.3	Editing Your R Code	27
2.3.4	Best Practice: Object Names in R	27
2.4	Data Values and Objects in R	28
2.4.1	Types of Data Values in R	28
2.4.2	Types (Structures) of Objects in R	31
2.4.3	Object Attributes	37
2.4.4	Testing (Checking) Object Type	37
2.4.5	Changing Object Type	38
2.5	R Packages (Libraries)	39
2.5.1	Installing R Packages	39
2.5.2	Name Spaces	40
2.5.3	Updating R	40
2.5.4	Datasets in R Packages	40
2.6	Importing (Reading) Data into R Workspace	41
2.6.1	Best Practice: Working Directory	41
2.6.2	Reading ASCII (Text) Files	42
2.6.3	Reading CSV Files	44
2.6.4	Reading Excel Files	44
2.6.5	Reading Files from the Internet	45
2.7	Writing (Exporting) Data to Files	46
2.7.1	Diverting Normal R Output to a File	46
2.7.2	Saving Graphs in Files	46
2.7.3	Exporting Data to Files	47
2.8	Some Arithmetic and Other Operators	48
2.8.1	Vectors	48
2.8.2	Matrix Computations	50
2.9	Programming in R	55
2.9.1	Best Practice: Script Files	55
2.9.2	Some Useful Commands or Functions	56
2.9.3	Conditional Execution	56
2.9.4	Loops	58
2.9.5	Functions and Functionals	59
2.9.6	User-Defined Functions	61
2.10	Bibliographic Notes	66
	Exercises	67

3	Simple Linear Regression	71
3.1	Introduction	71
3.2	Covariance and Correlation Coefficient	71
3.3	Example: Computer Repair Data	77
3.4	The Simple Linear Regression Model	79
3.5	Parameter Estimation	80
3.6	Tests of Hypotheses	85
3.7	Confidence Intervals	90
3.8	Predictions	91
3.9	Measuring the Quality of Fit	93
3.10	Regression Line Through the Origin	97
3.11	Trivial Regression Models	100
3.12	Bibliographic Notes	101
	Exercises	101
4	Multiple Linear Regression	109
4.1	Introduction	109
4.2	Description of the Data and Model	109
4.3	Example: Supervisor Performance Data	110
4.4	Parameter Estimation	112
4.5	Interpretations of Regression Coefficients	114
4.6	Centering and Scaling	117
	4.6.1 Centering and Scaling in Intercept Models	118
	4.6.2 Scaling in No-Intercept Models	119
4.7	Properties of the Least Squares Estimators	121
4.8	Multiple Correlation Coefficient	122
4.9	Inference for Individual Regression Coefficients	123
4.10	Tests of Hypotheses in a Linear Model	126
	4.10.1 Testing All Regression Coefficients Equal to Zero	128
	4.10.2 Testing a Subset of Regression Coefficients Equal to Zero	131
	4.10.3 Testing the Equality of Regression Coefficients	134
	4.10.4 Estimating and Testing of Regression Parameters Under Constraints	136
4.11	Predictions	138
4.12	Summary	139
	Exercises	139
	Appendix: 4.A Multiple Regression in Matrix Notation	146
5	Regression Diagnostics: Detection of Model Violations	151
5.1	Introduction	151
5.2	The Standard Regression Assumptions	152
5.3	Various Types of Residuals	155

5.4	Graphical Methods	157
5.5	Graphs Before Fitting a Model	160
5.5.1	One-Dimensional Graphs	160
5.5.2	Two-Dimensional Graphs	161
5.5.3	Rotating Plots	163
5.5.4	Dynamic Graphs	164
5.6	Graphs After Fitting a Model	164
5.7	Checking Linearity and Normality Assumptions	165
5.8	Leverage, Influence, and Outliers	166
5.8.1	Outliers in the Response Variable	168
5.8.2	Outliers in the Predictors	169
5.8.3	Masking and Swamping Problems	169
5.9	Measures of Influence	172
5.9.1	Cook's Distance	173
5.9.2	Welsch and Kuh Measure	174
5.9.3	Hadi's Influence Measure	174
5.10	The Potential-Residual Plot	177
5.11	Regression Diagnostics in R	178
5.12	What to Do with the Outliers?	178
5.13	Role of Variables in a Regression Equation	180
5.13.1	Added-Variable Plot	180
5.13.2	Residual Plus Component Plot	181
5.14	Effects of an Additional Predictor	184
5.15	Robust Regression	186
	Exercises	186
6	Qualitative Variables as Predictors	193
6.1	Introduction	193
6.2	Salary Survey Data	194
6.3	Interaction Variables	197
6.4	Systems of Regression Equations: Comparing Two Groups	202
6.4.1	Models with Different Slopes and Different Intercepts	203
6.4.2	Models with Same Slope and Different Intercepts	210
6.4.3	Models with Same Intercept and Different Slopes	212
6.5	Other Applications of Indicator Variables	213
6.6	Seasonality	214
6.7	Stability of Regression Parameters Over Time	215
	Exercises	218
7	Transformation of Variables	225
7.1	Introduction	225
7.2	Transformations to Achieve Linearity	227
7.3	Bacteria Deaths Due to X-Ray Radiation	230

x CONTENTS

7.3.1	Inadequacy of a Linear Model	231
7.3.2	Logarithmic Transformation for Achieving Linearity	233
7.4	Transformations to Stabilize Variance	234
7.5	Detection of Heteroscedastic Errors	239
7.6	Removal of Heteroscedasticity	241
7.7	Weighted Least Squares	243
7.8	Logarithmic Transformation of Data	244
7.9	Power Transformation	247
7.10	Summary	249
	Exercises	250
8	Weighted Least Squares	257
8.1	Introduction	257
8.2	Heteroscedastic Models	258
8.2.1	Supervisors Data	259
8.2.2	College Expense Data	260
8.3	Two-Stage Estimation	262
8.4	Education Expenditure Data	264
8.5	Fitting a Dose–Response Relationship Curve	273
	Exercises	275
9	The Problem of Correlated Errors	277
9.1	Introduction: Autocorrelation	277
9.2	Consumer Expenditure and Money Stock	278
9.3	Durbin–Watson Statistic	281
9.4	Removal of Autocorrelation by Transformation	283
9.5	Iterative Estimation with Autocorrelated Errors	287
9.6	Autocorrelation and Missing Variables	288
9.7	Analysis of Housing Starts	289
9.8	Limitations of the Durbin–Watson Statistic	292
9.9	Indicator Variables to Remove Seasonality	294
9.10	Regressing Two Time Series	296
	Exercises	298
10	Analysis of Collinear Data	301
10.1	Introduction	301
10.2	Effects of Collinearity on Inference	302
10.3	Effects of Collinearity on Forecasting	308
10.4	Detection of Collinearity	311
10.4.1	Simple Signs of Collinearity	312
10.4.2	Variance Inflation Factors	315
10.4.3	The Condition Indices	318
	Exercises	321

11	Working With Collinear Data	327
11.1	Introduction	327
11.2	Principal Components	328
11.3	Computations Using Principal Components	332
11.4	Imposing Constraints	335
11.5	Searching for Linear Functions of the β 's	338
11.6	Biased Estimation of Regression Coefficients	342
11.7	Principal Components Regression	343
11.8	Reduction of Collinearity in the Estimation Data	345
11.9	Constraints on the Regression Coefficients	347
11.10	Principal Components Regression: A Caution	348
11.11	Ridge Regression	351
11.12	Estimation by the Ridge Method	353
11.13	Ridge Regression: Some Remarks	358
11.14	Summary	359
11.15	Bibliographic Notes	360
	Exercises	360
	Appendix: 11.A Principal Components	363
	Appendix: 11.B Ridge Regression	365
	Appendix: 11.C Surrogate Ridge Regression	369
12	Variable Selection Procedures	371
12.1	Introduction	371
12.2	Formulation of the Problem	372
12.3	Consequences of Variables Deletion	373
12.4	Uses of Regression Equations	374
	12.4.1 Description and Model Building	374
	12.4.2 Estimation and Prediction	374
	12.4.3 Control	375
12.5	Criteria for Evaluating Equations	376
	12.5.1 Residual Mean Square	376
	12.5.2 Mallows C_p	376
	12.5.3 Information Criteria	377
12.6	Collinearity and Variable Selection	379
12.7	Evaluating All Possible Equations	379
12.8	Variable Selection Procedures	380
	12.8.1 Forward Selection Procedure	381
	12.8.2 Backward Elimination Procedure	381
	12.8.3 Stepwise Method	382
12.9	General Remarks on Variable Selection Methods	382
12.10	A Study of Supervisor Performance	383
12.11	Variable Selection with Collinear Data	388
12.12	The Homicide Data	388
12.13	Variable Selection Using Ridge Regression	391

xii CONTENTS

12.14	Selection of Variables in an Air Pollution Study	392
12.15	A Possible Strategy for Fitting Regression Models	398
12.16	Bibliographic Notes	400
	Exercises	400
	Appendix: 12.A Effects of Incorrect Model Specifications	404
13	Logistic Regression	409
13.1	Introduction	409
13.2	Modeling Qualitative Data	410
13.3	The Logit Model	410
13.4	Example: Estimating Probability of Bankruptcies	413
13.5	Logistic Regression Diagnostics	415
13.6	Determination of Variables to Retain	417
13.7	Judging the Fit of a Logistic Regression	420
13.8	The Multinomial Logit Model	422
	13.8.1 Multinomial Logistic Regression	422
	13.8.2 Example: Determining Chemical Diabetes	423
	13.8.3 Ordinal Logistic Regression	426
	13.8.4 Example: Determining Chemical Diabetes Revisited	426
13.9	Classification Problem: Another Approach	428
	Exercises	430
14	Further Topics	433
14.1	Introduction	433
14.2	Generalized Linear Model	434
14.3	Poisson Regression Model	435
14.4	Introduction of New Drugs	436
14.5	Robust Regression	437
14.6	Fitting a Quadratic Model	438
14.7	Distribution of PCB in U.S. Bays	440
	Exercises	443
	References	445
	Index	455