Contents

Preface xv
Acknowledgments xix
About the Author xxi

I The Basics 1

1 Why Use Regression Models? 3
  1.1 Why Use Simple Regression Models? 3
  1.2 Why Use Multiple Regression Models? 4
  1.3 Some Basic Notation 6

2 An Introductory Example 9
  2.1 A Single Line Model 9
  2.2 Fitting a Single Line Model 11
  2.3 Taking Uncertainty into Account 13
  2.4 A Two-Line Model 14
  2.5 How to Perform These Steps with Stata 17
  2.6 Exercise 5-HIAA and Serotonin 19
  2.7 Exercise Haemoglobin 19
  2.8 Exercise Scaling of Variables 19

3 The Classical Multiple Regression Model 21

4 Adjusted Effects 23
  4.1 Adjusting for Confounding 23
  4.2 Adjusting for Imbalances 26
  4.3 Exercise Physical Activity in Schoolchildren 27

5 Inference for the Classical Multiple Regression Model 29
  5.1 The Traditional and the Modern Way of Inference 29
  5.2 How to Perform the Modern Way of Inference with Stata 34
  5.3 How Valid and Good are Least Squares Estimates? 35
  5.4 A Note on the Use and Interpretation of p-Values in Regression Analyses 35
6 Logistic Regression
6.1 The Definition of the Logistic Regression Model 39
6.2 Analysing a Dose Response Experiment by Logistic Regression 40
6.3 How to Fit a Dose Response Model with Stata 44
6.4 Estimating Odds Ratios and Adjusted Odds Ratios Using Logistic Regression 45
6.5 How to Compute (Adjusted) Odds Ratios Using Logistic Regression in Stata 49
6.6 Exercise Allergy in Children 50
6.7 More on Logit Scale and Odds Scale 51

7 Inference for the Logistic Regression Model 55
7.1 The Maximum Likelihood Principle 55
7.2 Properties of the ML Estimates for Logistic Regression 56
7.3 Inference for a Single Regression Parameter 57
7.4 How to Perform Wald Tests and Likelihood Ratio Tests in Stata 58

8 Categorical Covariates 63
8.1 Incorporating Categorical Covariates in a Regression Model 63
8.2 Some Technicalities in Using Categorical Covariates 65
8.3 Testing the Effect of a Categorical Covariate 67
8.4 The Handling of Categorical Covariates in Stata 68
8.5 Presenting Results of a Regression Analysis Involving Categorical Covariates in a Table 73
8.6 Exercise Physical Occupation and Back Pain 76
8.7 Exercise Odds Ratios and Categorical covariates 77

9 Handling Ordered Categories: A First Lesson in Regression Modelling Strategies 79

10 The Cox Proportional Hazards Model 85
10.1 Modelling the Risk of Dying 85
10.2 Modelling the Risk of Dying in Continuous Time 87
10.3 Using the Cox Proportional Hazards Model to Quantify the Difference in Survival Between Groups 90
10.4 How to Fit a Cox Proportional Hazards Model with Stata 91
10.5 Exercise Prognostic Factors in Breast Cancer Patients—Part 1 94

11 Common Pitfalls in Using Regression Models 97
11.1 Association versus Causation 97
11.2 Difference between Subjects versus Difference within Subjects 99
11.3 Real-World Models versus Statistical Models 100
11.4 Relevance versus Significance 102
11.5 Exercise Prognostic Factors in Breast Cancer Patients—Part 2 104
# II Advanced Topics and Techniques

## 12 Some Useful Technicalities

12.1 Illustrating Models by Using Model-Based Predictions 109  
12.2 How to Work with Predictions in Stata 110  
12.3 Residuals and the Standard Deviation of the Error Term 116  
12.4 Working with Residuals and the RMSE in Stata 118  
12.5 Linear and Nonlinear Functions of Regression Parameters 119  
12.6 Transformations of Regression Parameters 120  
12.7 Centering of Covariate Values 121  
12.8 Exercise *Paternal Smoking versus Maternal Smoking* 122

## 13 Comparing Regression Coefficients

13.1 Comparing Regression Coefficients among Continuous Covariates 123  
13.2 Comparing Regression Coefficients among Binary Covariates 127  
13.3 Measuring the Impact of Changing Covariate Values 128  
13.4 Translating Regression Coefficients 130  
13.5 How to Compare Regression Coefficients in Stata 131  
13.6 Exercise *Health in Young People* 137

## 14 Power and Sample Size

14.1 The Power of a Regression Analysis 139  
14.2 Determinants of Power in Regression Models with a Single Covariate 140  
14.3 Determinants of Power in Regression Models with Several Covariates 148  
14.4 Power and Sample Size Calculations When a Sample from the Covariate Distribution Is Given 152  
14.5 Power and Sample Size Calculations Given a Sample from the Covariate Distribution with Stata 154  
14.6 The Choice of the Values of the Regression Parameters in a Simulation Study 165  
14.7 Simulating a Covariate Distribution 166  
14.8 Simulating a Covariate Distribution with Stata 169  
14.9 Choosing the Parameters to Simulate a Covariate Distribution 177  
14.10 Necessary Sample Sizes to Justify Asymptotic Methods 178  
14.11 Exercise *Power Considerations for a Study on Neck Pain* 178  
14.12 Exercise *Choosing between Two Outcomes* 179

## 15 Selection of the Sample

15.1 Selection in Dependence on the Covariates 181  
15.2 Selection in Dependence on the Outcome 183  
15.3 Sampling in Dependence on Covariate Values 185

## 16 Selection of Covariates

16.1 Fitting Regression Models with Correlated Covariates 187  
16.2 The “Adjustment versus Power” Dilemma 189
### 16 Adjusting for Covariates

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>16.3 The “Adjustment Makes Effects Small” Dilemma</td>
<td>191</td>
</tr>
<tr>
<td>16.4 Adjusting for Mediators</td>
<td>193</td>
</tr>
<tr>
<td>16.5 Adjusting for Confounding—A Useful Academic Game</td>
<td>196</td>
</tr>
<tr>
<td>16.6 Adjusting for Correlated Confounders</td>
<td>198</td>
</tr>
<tr>
<td>16.7 Including Predictive Covariates</td>
<td>199</td>
</tr>
<tr>
<td>16.8 Automatic Variable Selection</td>
<td>201</td>
</tr>
<tr>
<td>16.9 How to Choose Relevant Sets of Covariates</td>
<td>202</td>
</tr>
<tr>
<td>16.10 Preparing the Selection of Covariates: Analysing the Association Among Covariates</td>
<td>206</td>
</tr>
<tr>
<td>16.11 Preparing the Selection of Covariates: Univariate Analyses?</td>
<td>206</td>
</tr>
<tr>
<td>16.12 Exercise <em>Vocabulary Size in Young Children—Part 1</em></td>
<td>207</td>
</tr>
<tr>
<td>16.13 Preprocessing of the Covariate Space</td>
<td>208</td>
</tr>
<tr>
<td>16.14 How to Preprocess the Covariate Space with Stata</td>
<td>210</td>
</tr>
<tr>
<td>16.15 Exercise <em>Vocabulary Size in Young Children—Part 2</em></td>
<td>219</td>
</tr>
<tr>
<td>16.16 What Is a Confounder?</td>
<td>219</td>
</tr>
</tbody>
</table>

### 17 Modelling Nonlinear Effects

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>17.1 Quadratic Regression</td>
<td>221</td>
</tr>
<tr>
<td>17.2 Polynomial Regression</td>
<td>225</td>
</tr>
<tr>
<td>17.3 Splines</td>
<td>225</td>
</tr>
<tr>
<td>17.4 Fractional Polynomials</td>
<td>229</td>
</tr>
<tr>
<td>17.5 Gain in Power by Modelling Nonlinear Effects?</td>
<td>230</td>
</tr>
<tr>
<td>17.6 Demonstrating the Effect of a Covariate</td>
<td>232</td>
</tr>
<tr>
<td>17.7 Demonstrating a Nonlinear Effect</td>
<td>233</td>
</tr>
<tr>
<td>17.8 Describing the Shape of a Nonlinear Effect</td>
<td>234</td>
</tr>
<tr>
<td>17.9 Detecting Nonlinearity by Analysis of Residuals</td>
<td>237</td>
</tr>
<tr>
<td>17.10 Judging of Nonlinearity May Require Adjustment</td>
<td>237</td>
</tr>
<tr>
<td>17.11 How to Model Nonlinear Effects in Stata</td>
<td>238</td>
</tr>
<tr>
<td>17.12 The Impact of Ignoring Nonlinearity</td>
<td>254</td>
</tr>
<tr>
<td>17.13 Modelling the Nonlinear Effect of Confounders</td>
<td>255</td>
</tr>
<tr>
<td>17.14 Nonlinear Models</td>
<td>257</td>
</tr>
<tr>
<td>17.15 Exercise <em>Serum Markers for AMI</em></td>
<td>258</td>
</tr>
</tbody>
</table>

### 18 Transformation of Covariates

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>18.1 Transformations to Obtain a Linear Relationship</td>
<td>259</td>
</tr>
<tr>
<td>18.2 Transformation of Skewed Covariates</td>
<td>262</td>
</tr>
<tr>
<td>18.3 To Categorise or Not to Categorise</td>
<td>264</td>
</tr>
</tbody>
</table>

### 19 Effect Modification and Interactions

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>19.1 Modelling Effect Modification</td>
<td>269</td>
</tr>
<tr>
<td>19.2 Adjusted Effect Modifications</td>
<td>274</td>
</tr>
<tr>
<td>19.3 Interactions</td>
<td>276</td>
</tr>
<tr>
<td>19.4 Modelling Effect Modifications in Several Covariates</td>
<td>280</td>
</tr>
<tr>
<td>19.5 The Effect of a Covariate in the Presence of Interactions</td>
<td>281</td>
</tr>
<tr>
<td>19.6 Interactions as Deviations from Additivity</td>
<td>282</td>
</tr>
</tbody>
</table>
24.2 Judging the Usefulness of a Risk Score 352
24.3 The Precision of Risk Score Values 353
24.4 The Overall Precision of a Risk Score 356
24.5 Using Stata’s predict Command to Compute Risk Scores 357
24.6 Categorisation of Risk Scores 368
24.7 Exercise Computing Risk Scores for Breast Cancer Patients 369

25 Construction of Predictors 371
25.1 From Risk Scores to Predictors 371
25.2 Predictions and Prediction Intervals for a Continuous Outcome 371
25.3 Predictions for a Binary Outcome 373
25.4 Construction of Predictions for Time-to-Event Data 376
25.5 How to Construct Predictions with Stata 378
25.6 The Overall Precision of a Predictor 382

26 Evaluating the Predictive Performance 383
26.1 The Predictive Performance of an Existing Predictor 383
26.2 How to Assess the Predictive Performance of an Existing Predictor in Stata 385
26.3 Estimating the Predictive Performance of a New Predictor 387
26.4 How to Assess the Predictive Performance via Cross-Validation in Stata 389
26.5 Exercise Assessing the Predictive Performance of a Prognostic Score in Breast Cancer Patients 392

27 Outlook: Construction of Parsimonious Predictors 393

IV Miscellaneous 395

28 Alternatives to Regression Modelling 397
28.1 Stratification 397
28.2 Measures of Association: Correlation Coefficients 399
28.3 Measures of Association: The Odds Ratio 400
28.4 Propensity Scores 402
28.5 Classification and Regression Trees 404

29 Specific Regression Models 407
29.1 Probit Regression for Binary Outcomes 407
29.2 Generalised Linear Models 408
29.3 Regression Models for Count Data 409
29.4 Regression Models for Ordinal Outcome Data 411
29.5 Quantile Regression and Robust Regression 412
29.6 ANOVA and Regression 414
30 Specific Usages of Regression Models 415
30.1 Logistic Regression for the Analysis of Case-Control Studies 415
30.2 Logistic Regression for the Analysis of Matched Case-Control Studies 417
30.3 Adjusting for Baseline Values in Randomised Clinical Trials 418
30.4 Assessing Predictive Factors 421
30.5 Incorporating Time-Varying Covariates in a Cox Model 422
30.6 Time-Dependent Effects in a Cox Model 424
30.7 Using the Cox Model in the Presence of Competing Risks 426
30.8 Using the Cox Model to Analyse Multi-State Models 427

31 What Is a Good Model? 429
31.1 Does the Model Fit the Data? 429
31.2 How Good Are Predictions? 430
31.3 Explained Variation 431
31.4 Goodness of Fit 432
31.5 Model Stability 434
31.6 The Usefulness of a Model 435

32 Final Remarks on the Role of Prespecified Models and Model Development 439

V Mathematical Details 443

A Mathematics Behind the Classical Linear Regression Model 445
A.1 Computing Regression Parameters in Simple Linear Regression 445
A.2 Computing Regression Parameters in the Classical Multiple Regression Model 446
A.3 Estimation of the Standard Error 448
A.4 Construction of Confidence Intervals and p-Values 450

B Mathematics Behind the Logistic Regression Model 453
B.1 The Least Squares Principle as a Maximum Likelihood Principle 453
B.2 Maximising the Likelihood of a Logistic Regression Model 454
B.3 Estimating the Standard Error of the ML Estimates 457
B.4 Testing Composite Hypotheses 458

C The Modern Way of Inference 461
C.1 Robust Estimation of Standard Errors 461
C.2 Robust Estimation of Standard Errors in the Presence of Clustering 461

D Mathematics for Risk Scores and Predictors 463
D.1 Computing Individual Survival Probabilities after Fitting a Cox Model 463
D.2 Standard Errors for Risk Scores 463