Contents

Preface . . . xiii

1 Introduction 1

1.1 Algebraic notation . . . 1

1.2 Descriptive statistics . . . 3

1.2.1 Dot plot . . . 3

1.2.2 Sample mean . . . 4

1.2.3 Residual . . . 4

1.2.4 Sample variance . . . 5

1.2.5 Sample standard deviation . . . 5

1.2.6 Percentile and median . . . 6

1.2.7 Box plot . . . 6

1.2.8 Histogram . . . 7

1.2.9 Scatter plot . . . 8

1.3 The Stata statistical software package . . . 8

1.3.1 Downloading data from my web site . . . 9

1.3.2 Creating histograms with Stata . . . 10

1.3.3 Stata command syntax . . . 13

1.3.4 Obtaining interactive help from Stata . . . 14

1.3.5 Stata log files . . . 15

1.3.6 Stata graphics and schemes . . . 16

1.3.7 Stata do files . . . 17

1.3.8 Stata pulldown menus . . . 17

1.3.9 Displaying other descriptive statistics with Stata . . . 22

1.4 Inferential statistics . . . 24

1.4.1 Probability density function . . . 25

1.4.2 Mean, variance and standard deviation . . . 25

1.4.3 Normal distribution . . . 26

1.4.4 Expected value . . . 27

1.4.5 Standard error . . . 27

1.4.6 Null hypothesis, alternative hypothesis and P-value 28
1.4.7 95% confidence interval 29
1.4.8 Statistical power 30
1.4.9 The z and Student's t distributions 31
1.4.10 Paired t test 33
1.4.11 Performing paired t tests with Stata 34
1.4.12 Independent t test using a pooled standard error estimate 36
1.4.13 Independent t test using separate standard error estimates 38
1.4.14 Independent t tests using Stata 39
1.4.15 The chi-squared distribution 41
1.5 Overview of methods discussed in this text 42
1.5.1 Models with one response per patient . . . 42
1.5.2 Models with multiple responses per patient 44
1.6 Additional reading 44
1.7 Exercises 45

2 Simple linear regression 47
2.1 Sample covariance 48
2.2 Sample correlation coefficient 49
2.3 Population covariance and correlation coefficient . 50
2.4 Conditional expectation 51
2.5 Simple linear regression model 51
2.6 Fitting the linear regression model 53
2.7 Historical trivia: origin of the term *regression* . . . 54
2.8 Determining the accuracy of linear regression estimates 55
2.9 Ethylene glycol poisoning example 57
2.10 95% confidence interval for $y[x] = \alpha + \beta x$ evaluated at x 58
2.11 95% prediction interval for the response of a new patient 59
2.12 Simple linear regression with Stata 60
2.13 Lowess regression 67
2.14 Plotting a lowess regression curve in Stata 69
2.15 Residual analyses 69
2.16 Studentized residual analysis using Stata 73
2.17 Transforming the x and y variables 75

2.17.1 Stabilizing the variance 75
2.17.2 Correcting for non-linearity 75
2.17.3 Example: research funding and morbidity for 29 diseases 77
2.18 Analyzing transformed data with Stata 79
2.19 Testing the equality of regression slopes 84
2.19.1 Example: the Framingham Heart Study . . 85
2.20 Comparing slope estimates with Stata 87
2.21 Density-distribution sunflower plots 91
2.22 Creating density-distribution sunflower plots with Stata 93
2.23 Additional reading 96
2.24 Exercises 97

3 Multiple linear regression 100
3.1 The model 100
3.2 Confounding variables 101
3.3 Estimating the parameters for a multiple linear regression model 102
3.4 R^2 statistic for multiple regression models 102
3.5 Expected response in the multiple regression model 103
3.6 The accuracy of multiple regression parameter estimates 103
3.7 Hypothesis tests 104
3.8 Leverage 105
3.9 95% confidence interval for $\hat{y}_i$ 105
3.10 95% prediction intervals 106
3.11 Example: the Framingham Heart Study 106
3.11.1 Preliminary univariate analyses 107
3.12 Scatterplot matrix graphs 109
3.12.1 Producing scatterplot matrix graphs with Stata 110
3.13 Modeling interaction in multiple linear regression . 111
3.13.1 The Framingham example 111
3.14 Multiple regression modeling of the Framingham data 113
3.15 Intuitive understanding of a multiple regression model 114
3.15.1 The Framingham example 114
3.16 Calculating 95% confidence and prediction intervals 118

3.17 Multiple linear regression with Stata 119
3.18 Automatic methods of model selection 124
3.18.1 Forward selection using Stata 124
3.18.2 Backward selection 127
3.18.3 Forward stepwise selection 128
3.18.4 Backward stepwise selection 128
3.18.5 Pros and cons of automated model selection 128
3.19 Collinearity 129
3.20 Residual analyses 131
3.21 Influence 132
3.21.1 $\Delta\hat{\beta}$ influence statistic 132
3.21.2 Cook's distance 133
3.21.3 The Framingham example 133
3.22 Residual and influence analyses using Stata 135
3.23 Using multiple linear regression for non-linear models 138
3.24 Building non-linear models with restricted cubic splines 139
3.24.1 Choosing the knots for a restricted cubic spline model 142
3.25 The SUPPORT Study of hospitalized patients . . . 142
3.25.1 Modeling length-of-stay and MAP using restricted cubic splines 144
3.25.2 Using Stata for non-linear models with restricted cubic splines 149
3.26 Additional reading 159
3.27 Exercises 160

4 Simple logistic regression 164
4.1 Example: APACHE score and mortality in patients with sepsis 164
4.2 Sigmoidal family of logistic regression curves . . . 164
4.3 The log odds of death given a logistic probability function 166
4.4 The binomial distribution 167
4.5 Simple logistic regression model 168
4.6 Generalized linear model 168
4.7 Contrast between logistic and linear regression . . 169
4.8 Maximum likelihood estimation 169
4.8.1 Variance of maximum likelihood parameter estimates 171

4.9 Statistical tests and confidence intervals 171
4.9.1 Likelihood ratio tests 172
4.9.2 Quadratic approximations to the log likelihood ratio function 172
4.9.3 Score tests 173
4.9.4 Wald tests and confidence intervals 174
4.9.5 Which test should you use? 174
4.10 Sepsis example 176
4.11 Logistic regression with Stata 176
4.12 Odds ratios and the logistic regression model . . . 180
4.13 95% CI for the OR associated with a unit increase in x 180
4.13.1 Calculating this odds ratio with Stata . . . 181
4.14 Logistic regression with grouped response data . . 182
4.15 95% confidence interval for $\pi[x]$ 182
4.16 Exact $100(1-\alpha)$% confidence intervals for proportions 183
4.17 Example: the Ibuprofen in Sepsis Study 184
4.18 Logistic regression with grouped data using Stata . 187
4.19 Simple 2 × 2 case-control studies 193
4.19.1 Example: The Ille-et-Vilaine Study of esophageal cancer and alcohol 193
4.19.2 Review of classical case-control theory . . . 194
4.19.3 95% confidence interval for the odds ratio: Woolf's method 195
4.19.4 Test of the null hypothesis that the odds ratio equals one 196
4.19.5 Test of the null hypothesis that two proportions are equal 197
4.20 Logistic regression models for 2 × 2 contingency tables 197
4.20.1 Nuisance parameters 198
4.20.2 95% confidence interval for the odds ratio: logistic regression 198
4.21 Creating a Stata data file 198
4.22 Analyzing case-control data with Stata 201
4.23 Regressing disease against exposure 203
4.24 Additional reading 204
4.25 Exercises 205

5 Multiple logistic regression 207
5.1 Mantel–Haenszel estimate of an age-adjusted odds ratio 207
5.2 Mantel–Haenszel χ^2 statistic for multiple 2×2 tables 209
5.3 95% confidence interval for the age-adjusted odds ratio 210
5.4 Breslow–Day–Tarone test for homogeneity 211
5.5 Calculating the Mantel–Haenszel odds ratio using Stata 213
5.6 Multiple logistic regression model 216
5.6.1 Likelihood ratio test of the influence of the covariates on the response variable 217
5.7 95% confidence interval for an adjusted odds ratio 218
5.8 Logistic regression for multiple 2×2 contingency tables 218
5.9 Analyzing multiple 2×2 tables with Stata 220
5.10 Handling categorical variables in Stata 222
5.11 Effect of dose of alcohol on esophageal cancer risk 224
5.11.1 Analyzing Model (5.25) with Stata 226
5.12 Effect of dose of tobacco on esophageal cancer risk 227
5.13 Deriving odds ratios from multiple parameters . . 228
5.14 The standard error of a weighted sum of regression coefficients 229
5.15 Confidence intervals for weighted sums of coefficients 229
5.16 Hypothesis tests for weighted sums of coefficients . 230
5.17 The estimated variance–covariance matrix 230
5.18 Multiplicative models of two risk factors 231
5.19 Multiplicative model of smoking, alcohol, and esophageal cancer 232
5.20 Fitting a multiplicative model with Stata 234
5.21 Model of two risk factors with interaction 238
5.22 Model of alcohol, tobacco, and esophageal cancer . 239
5.23 Fitting a model with interaction using Stata 241
5.24 Model fitting: nested models and model deviance . 245
5.25 Effect modifiers and confounding variables 246
5.26 Goodness-of-fit tests 247
5.26.1 The Pearson χ^2 goodness-of-fit statistic . . 247
5.27 Hosmer–Lemeshow goodness-of-fit test 249

5.27.1 An example: the Ille-et-Vilaine cancer data set . . . 249
5.28 Residual and influence analysis . . . 251
5.28.1 Standardized Pearson residual . . . 252
5.28.2 $\Delta\hat{\beta}_j$ influence statistic . . . 252
5.28.3 Residual plots of the Ille-et-Vilaine data on esophageal cancer . . . 253
5.29 Using Stata for goodness-of-fit tests and residual analyses . . . 256
5.30 Frequency matched case-control studies . . . 264
5.31 Conditional logistic regression . . . 264
5.32 Analyzing data with missing values . . . 265
5.32.1 Imputing data that is missing at random . 265
5.32.2 Cardiac output in the Ibuprofen in Sepsis Study . . . 267
5.32.3 Modeling missing values with Stata . . . 270
5.33 Logistic regression using restricted cubic splines . . 271
5.33.1 Odds ratios from restricted cubic spline models . . . 273
5.33.2 95% confidence intervals for $\hat{\psi}[x]$ . . . 273
5.34 Modeling hospital mortality in the SUPPORT Study . . . 274
5.35 Using Stata for logistic regression with restricted cubic splines . . . 278
5.36 Regression methods with a categorical response variable . . . 285
5.36.1 Proportional odds logistic regression . . . 285
5.36.2 Polytomous logistic regression . . . 287
5.37 Additional reading . . . 289
5.38 Exercises . . . 290

6 Introduction to survival analysis 295
6.1 Survival and cumulative mortality functions . . . 295
6.2 Right censored data . . . 297
6.3 Kaplan–Meier survival curves . . . 298
6.4 An example: Genetic risk of recurrent intracerebral hemorrhage . . . 299
6.5 95% confidence intervals for survival functions . . . 302
6.6 Cumulative mortality function . . . 303
6.7 Censoring and bias . . . 304

6.8 Log-rank test 305
6.9 Using Stata to derive survival functions and the log-rank test 308
6.10 Log-rank test for multiple patient groups 314
6.11 Hazard functions 314
6.12 Proportional hazards 315
6.13 Relative risks and hazard ratios 317
6.14 Proportional hazards regression analysis 318
6.15 Hazard regression analysis of the intracerebral hemorrhage data 319
6.16 Proportional hazards regression analysis with Stata 320
6.17 Tied failure times 321
6.18 Additional reading 321
6.19 Exercises 322

7 Hazard regression analysis 324
7.1 Proportional hazards model 324
7.2 Relative risks and hazard ratios 324
7.3 95% confidence intervals and hypothesis tests . . . 326
7.4 Nested models and model deviance 326
7.5 An example: The Framingham Heart Study 326
7.5.1 Kaplan–Meier survival curves for DBP . . . 327
7.5.2 Simple hazard regression model for CHD risk and DBP 328
7.5.3 Restricted cubic spline model of CHD risk and DBP 329
7.5.4 Categorical hazard regression model of CHD risk and DBP 332
7.5.5 Simple hazard regression model of CHD risk and gender 334
7.5.6 Multiplicative model of DBP and gender on risk of CHD 335
7.5.7 Using interaction terms to model the effects of gender and DBP on CHD 336
7.5.8 Adjusting for confounding variables 337
7.5.9 Interpretation 339
7.5.10 Alternative models 340
7.6 Proportional hazards regression analysis using Stata 341
7.7 Stratified proportional hazards models 357
7.8 Survival analysis with ragged study entry 358

7.8.1 Kaplan–Meier survival curve and the log-rank test with ragged entry 359
7.8.2 Age, sex, and CHD in the Framingham Heart Study 359
7.8.3 Proportional hazards regression analysis with ragged entry 360
7.8.4 Survival analysis with ragged entry using Stata 361
7.9 Predicted survival, log-log plots and the proportional hazards assumption 364
7.9.1 Evaluating the proportional hazards assumption with Stata 366
7.10 Hazard regression models with time-dependent covariates 368
7.10.1 Testing the proportional hazards assumption 370
7.10.2 Modeling time-dependent covariates with Stata 371
7.11 Additional reading 379
7.12 Exercises 380

8 Introduction to Poisson regression 383
8.1 Elementary statistics involving rates 383
8.2 Calculating relative risks from incidence data using Stata 384
8.3 The binomial and Poisson distributions 386
8.4 Simple Poisson regression for 2×2 tables 387
8.5 Poisson regression and the generalized linear model 388
8.6 Contrast between Poisson, logistic, and linear regression 389
8.7 Simple Poisson regression with Stata 389
8.8 Poisson regression and survival analysis 391
8.8.1 Recoding survival data on patients as patient-year data 391
8.8.2 Converting survival records to person-years of follow-up using Stata 393
8.9 Converting the Framingham survival data set to person-time data 396
8.10 Simple Poisson regression with multiple data records 402
8.11 Poisson regression with a classification variable . . 403

8.12 Applying simple Poisson regression to the Framingham data 404
8.13 Additional reading 407
8.14 Exercises 407

9 Multiple Poisson regression 411
9.1 Multiple Poisson regression model 411
9.2 An example: The Framingham Heart Study 414
9.2.1 A multiplicative model of gender, age and coronary heart disease 415
9.2.2 A model of age, gender and CHD with interaction terms 417
9.2.3 Adding confounding variables to the model 420
9.3 Using Stata to perform Poisson regression 422
9.4 Residual analyses for Poisson regression models . . 432
9.4.1 Deviance residuals 433
9.5 Residual analysis of Poisson regression models using Stata 434
9.6 Additional reading 436
9.7 Exercises 437

10 Fixed effects analysis of variance 439
10.1 One-way analysis of variance 439
10.2 Multiple comparisons 441
10.3 Reformulating analysis of variance as a linear regression model 444
10.4 Non-parametric methods 445
10.5 Kruskal–Wallis Test 445
10.6 Example: A polymorphism in the estrogen receptor gene 446
10.7 User contributed software in Stata 450
10.8 One-way analyses of variance using Stata 450
10.9 Two-way analysis of variance 457
10.10Additional reading 458
10.11Exercises 459

11 Repeated-measures analysis of variance 461
11.1 Example: Effect of race and dose of isoproterenol on blood flow 461

11.2 Exploratory analysis of repeated measures data using Stata 464
11.3 Response feature analysis 469
11.4 Example: The isoproterenol data set 471
11.5 Response feature analysis using Stata 473
11.6 The area-under-the-curve response feature 478
11.7 Generalized estimating equations 479
11.8 Common correlation structures 480
11.9 GEE analysis and the Huber–White sandwich estimator 482
11.10 Example: Analyzing the isoproterenol data with GEE 483
11.11 Using Stata to analyze the isoproterenol data set using GEE 487
11.12 GEE analyses with logistic or Poisson models . . . 491
11.13 Additional reading 492
11.14 Exercises 492

Appendices 494

A Summary of Statistical Models Discussed in this Text 495
A.1. Models for continuous response variables with one response per patient. 496
A.2. Models for dichotomous or categorical response variables with one response per patient. 497
A.3. Models for survival data. 498
A.4. Models for response variables that are event rates or the number of events during a specified number of patient-years of follow-up. 499
A.5. Models with multiple observations per patient or matched or clustered patients. 499

B Summary of Stata Commands Used in this Text 500
B.1. Data manipulation and description 501
B.2. Analysis commands 503
B.3. Graph commands 507
B.4. Common options for graph commands 511
B.5. Post-estimation commands 513
B.6. Command prefixes 515

B.7. Command qualifiers 516
B.8. Logical and relational operators and system variables517
B.9. Functions 518

References **519**