
Contents

Preface	VII
1 Introduction	1
1.1 Example: Treatment of Back Pain	1
1.2 The Family of Multipredictor Regression Methods	2
1.3 Motivation for Multipredictor Regression	3
1.3.1 Prediction	3
1.3.2 Isolating the Effect of a Single Predictor	3
1.3.3 Understanding Multiple Predictors	3
1.4 Guide to the Book	4
2 Exploratory and Descriptive Methods	7
2.1 Data Checking	7
2.2 Types Of Data	8
2.3 One-Variable Descriptions	8
2.3.1 Numerical Variables	9
2.3.2 Categorical Variables	16
2.4 Two-Variable Descriptions	17
2.4.1 Outcome Versus Predictor Variables	18
2.4.2 Continuous Outcome Variable	18
2.4.3 Categorical Outcome Variable	21
2.5 Multivariable Descriptions	23
2.6 Problems	26
3 Basic Statistical Methods	29
3.1 <i>t</i> -Test and Analysis of Variance	29
3.1.1 <i>t</i> -Test	30
3.1.2 One- and Two-Sided Hypothesis Tests	30
3.1.3 Paired <i>t</i> -Test	31
3.1.4 One-Way Analysis of Variance (ANOVA)	32
3.1.5 Pairwise Comparisons in ANOVA	33

3.1.6	Multi-Way ANOVA and ANCOVA	33
3.1.7	Robustness to Violations of Assumptions	33
3.2	Correlation Coefficient	35
3.3	Simple Linear Regression Model	36
3.3.1	Systematic Part of the Model	36
3.3.2	Random Part of the Model	38
3.3.3	Assumptions About the Predictor	38
3.3.4	Ordinary Least Squares Estimation	39
3.3.5	Fitted Values and Residuals	40
3.3.6	Sums of Squares	40
3.3.7	Standard Errors of the Regression Coefficients	41
3.3.8	Hypothesis Tests and Confidence Intervals	42
3.3.9	Slope, Correlation Coefficient, and R^2	43
3.4	Contingency Table Methods for Binary Outcomes	44
3.4.1	Measures of Risk and Association for Binary Outcomes	44
3.4.2	Tests of Association in Contingency Tables	47
3.4.3	Predictors With Multiple Categories	49
3.4.4	Analyses Involving Multiple Categorical Predictors	51
3.5	Basic Methods for Survival Analysis	54
3.5.1	Right Censoring	54
3.5.2	Kaplan–Meier Estimator of the Survival Function	55
3.5.3	Interpretation of Kaplan–Meier Curves	57
3.5.4	Median Survival	58
3.5.5	Cumulative Incidence Function	59
3.5.6	Comparing Groups Using the Logrank Test	60
3.6	Bootstrap Confidence Intervals	62
3.7	Interpretation of Negative Findings	63
3.8	Further Notes and References	65
3.9	Problems	65
3.10	Learning Objectives	67
4	Linear Regression	69
4.1	Example: Exercise and Glucose	70
4.2	Multiple Linear Regression Model	72
4.2.1	Systematic Part of the Model	72
4.2.2	Random Part of the Model	73
4.2.3	Generalization of R^2 and r	75
4.2.4	Standardized Regression Coefficients	75
4.3	Categorical Predictors	76
4.3.1	Binary Predictors	76
4.3.2	Multilevel Categorical Predictors	77
4.3.3	The F -Test	79
4.3.4	Multiple Pairwise Comparisons Between Categories	80
4.3.5	Testing for Trend Across Categories	82
4.4	Confounding	83

4.4.1	Causal Effects and Counterfactuals	84
4.4.2	A Linear Model for the Counterfactual Experiment	85
4.4.3	Confounding of Causal Effects	87
4.4.4	Randomization Assumption	88
4.4.5	Conditions for Confounding of Causal Effects	89
4.4.6	Control of Confounding	89
4.4.7	Range of Confounding Patterns	90
4.4.8	Diagnostics for Confounding in a Sample	91
4.4.9	Confounding Is Difficult To Rule Out	92
4.4.10	Adjusted vs. Unadjusted $\hat{\beta}$ s	93
4.4.11	Example: BMI and LDL	93
4.5	Mediation	95
4.5.1	Modeling Mediation	96
4.5.2	Confidence Intervals for Measures of Mediation	97
4.5.3	Example: BMI, Exercise, and Glucose	97
4.6	Interaction	98
4.6.1	Causal Effects and Interaction	99
4.6.2	Modeling Interaction	100
4.6.3	Overall Causal Effect in the Presence of Interaction	100
4.6.4	Example: Hormone Therapy and Statin Use	101
4.6.5	Example: BMI and Statin Use	103
4.6.6	Interaction and Scale	105
4.6.7	Example: Hormone Therapy and Baseline LDL	106
4.6.8	Details	108
4.7	Checking Model Assumptions and Fit	109
4.7.1	Linearity	109
4.7.2	Normality	114
4.7.3	Constant Variance	117
4.7.4	Outlying, High Leverage, and Influential Points	121
4.7.5	Interpretation of Results for Log-Transformed Variables	125
4.7.6	When to Use Transformations	127
4.8	Summary	127
4.9	Further Notes and References	127
4.10	Problems	128
4.11	Learning Objectives	131
5	Predictor Selection	133
5.1	Diagramming the Hypothesized Causal Model	135
5.2	Prediction	137
5.2.1	Bias–Variance Trade-off	137
5.2.2	Estimating Prediction Error	138
5.2.3	Screening Candidate Models	139
5.2.4	Classification and Regression Trees (CART)	139
5.3	Evaluating a Predictor of Primary Interest	140
5.3.1	Including Predictors for Face Validity	141

5.3.2	Selecting Predictors on Statistical Grounds	141
5.3.3	Interactions With the Predictor of Primary Interest	141
5.3.4	Example: Incontinence as a Risk Factor for Falling	142
5.3.5	Randomized Experiments	142
5.4	Identifying Multiple Important Predictors	144
5.4.1	Ruling Out Confounding Is Still Central	145
5.4.2	Cautious Interpretation Is Also Key	146
5.4.3	Example: Risk Factors for Coronary Heart Disease	146
5.4.4	Allen–Cady Modified Backward Selection	147
5.5	Some Details	147
5.5.1	Collinearity	147
5.5.2	Number of Predictors	149
5.5.3	Alternatives to Backward Selection	150
5.5.4	Model Selection and Checking	151
5.5.5	Model Selection Complicates Inference	152
5.6	Summary	153
5.7	Further Notes and References	154
5.8	Problems	155
5.9	Learning Objectives	156
6	Logistic Regression	157
6.1	Single Predictor Models	158
6.1.1	Interpretation of Regression Coefficients	162
6.1.2	Categorical Predictors	164
6.2	Multipredictor Models	167
6.2.1	Likelihood Ratio Tests	170
6.2.2	Confounding	173
6.2.3	Interaction	175
6.2.4	Prediction	180
6.2.5	Prediction Accuracy	181
6.3	Case-Control Studies	183
6.3.1	Matched Case-Control Studies	187
6.4	Checking Model Assumptions and Fit	188
6.4.1	Outlying and Influential Points	188
6.4.2	Linearity	190
6.4.3	Model Adequacy	192
6.4.4	Technical Issues in Logistic Model Fitting	195
6.5	Alternative Strategies for Binary Outcomes	196
6.5.1	Infectious Disease Transmission Models	196
6.5.2	Regression Models Based on Excess and Relative Risks	198
6.5.3	Nonparametric Binary Regression	200
6.5.4	More Than Two Outcome Levels	201
6.6	Likelihood	203
6.7	Summary	206
6.8	Further Notes and References	207

6.9	Problems	207
6.10	Learning Objectives	209
7	Survival Analysis	211
7.1	Survival Data	211
7.1.1	Why Linear and Logistic Regression Won't Work	211
7.1.2	Hazard Function	212
7.1.3	Hazard Ratio	213
7.1.4	Proportional Hazards Assumption	215
7.2	Cox Proportional Hazards Model	215
7.2.1	Proportional Hazards Models	215
7.2.2	Parametric vs. Semi-Parametric Models	216
7.2.3	Hazard Ratios, Risk, and Survival Times	219
7.2.4	Hypothesis Tests and Confidence Intervals	219
7.2.5	Binary Predictors	221
7.2.6	Multilevel Categorical Predictors	221
7.2.7	Continuous Predictors	224
7.2.8	Confounding	226
7.2.9	Mediation	227
7.2.10	Interaction	227
7.2.11	Adjusted Survival Curves for Comparing Groups	229
7.2.12	Predicted Survival for Specific Covariate Patterns	231
7.3	Extensions to the Cox Model	231
7.3.1	Time-Dependent Covariates	231
7.3.2	Stratified Cox Model	234
7.4	Checking Model Assumptions and Fit	238
7.4.1	Log-Linearity	238
7.4.2	Proportional Hazards	238
7.5	Some Details	245
7.5.1	Bootstrap Confidence Intervals	245
7.5.2	Prediction	246
7.5.3	Adjusting for Non-Confounding Covariates	246
7.5.4	Independent Censoring	247
7.5.5	Interval Censoring	247
7.5.6	Left Truncation	248
7.6	Summary	249
7.7	Further Notes and References	249
7.8	Problems	250
7.9	Learning Objectives	251
8	Repeated Measures Analysis	253
8.1	A Simple Repeated Measures Example: Fecal Fat	254
8.1.1	Model Equations for the Fecal Fat Example	256
8.1.2	Correlations Within Subjects	257
8.1.3	Estimates of the Effects of Pill Type	259

8.2	Hierarchical Data	259
8.2.1	Analysis Strategies for Hierarchical Data	259
8.3	Longitudinal Data	262
8.3.1	Analysis Strategies for Longitudinal Data	262
8.3.2	Example: Birthweight and Birth Order	262
8.3.3	When To Use Repeated Measures Analyses	265
8.4	Generalized Estimating Equations	266
8.4.1	Birthweight and Birth Order Revisited	266
8.4.2	Correlation Structures	268
8.4.3	Working Correlation and Robust Standard Errors	270
8.4.4	Hypothesis Tests and Confidence Intervals	271
8.4.5	Use of <code>xtgee</code> for Clustered Logistic Regression	273
8.5	Random Effects Models	274
8.5.1	Re-Analysis of Birthweight and Birth Order	276
8.5.2	Prediction	278
8.5.3	Logistic Model for Low Birthweight	279
8.5.4	Marginal Versus Conditional Models	281
8.6	Example: Cardiac Injury Following Brain Hemorrhage	281
8.6.1	Bootstrap Confidence Intervals	283
8.7	Summary	286
8.8	Further Notes and References	286
8.9	Problems	287
8.10	Learning Objectives	288
9	Generalized Linear Models	291
9.1	Example: Treatment for Depression	291
9.1.1	Statistical Issues	292
9.1.2	Model for the Mean Response	292
9.1.3	Choice of Distribution	293
9.1.4	Interpreting the Parameters	294
9.1.5	Further Notes	295
9.2	Example: Costs of Phototherapy	295
9.2.1	Model for the Mean Response	296
9.2.2	Choice of Distribution	297
9.2.3	Interpreting the Parameters	297
9.3	Generalized Linear Models	297
9.3.1	Example: Risky Drug Use Behavior	298
9.3.2	Relationship of Mean to Variance	300
9.3.3	Nonlinear Models	300
9.4	Summary	301
9.5	Further Notes and References	301
9.6	Problems	302
9.7	Learning Objectives	303

10	Complex Surveys	305
10.1	Example: NHANES	307
10.2	Probability Weights	307
10.3	Variance Estimation	310
10.3.1	Design Effects	312
10.3.2	Simplification of Correlation Structure	313
10.3.3	Other Methods of Variance Estimation	313
10.4	Summary	314
10.5	Further Notes and References	314
10.6	Problems	315
10.7	Learning Objectives	316
11	Summary	317
11.1	Introduction	317
11.2	Selecting Appropriate Statistical Methods	318
11.3	Planning and Executing a Data Analysis	319
11.3.1	Analysis Plans	319
11.3.2	Choice of Software	320
11.3.3	Record Keeping and Organization	320
11.3.4	Data Security	320
11.3.5	Consulting a Statistician	321
11.3.6	Use of Internet Resources	321
11.4	Further Notes and References	321
	References	323
	Index	333