

Eric Vittinghoff • David V. Glidden
Stephen C. Shiboski • Charles E. McCulloch

Regression Methods in Biostatistics

Linear, Logistic, Survival, and Repeated
Measures Models

Second edition

Contents

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	4
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	9
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	17
2.4.2	Continuous Outcome Variable	18
2.4.3	Categorical Outcome Variable	21
2.5	Multivariable Descriptions	22
2.6	Summary	25
2.7	Problems	25
3	Basic Statistical Methods	27
3.1	t -Test and Analysis of Variance	27
3.1.1	t -Test	28
3.1.2	One- and Two-Sided Hypothesis Tests	28
3.1.3	Paired t -Test	29
3.1.4	One-Way Analysis of Variance	30
3.1.5	Pairwise Comparisons in ANOVA	30
3.1.6	Multi-way ANOVA and ANCOVA	31
3.1.7	Robustness to Violations of Normality Assumption	31

3.1.8	Nonparametric Alternatives	32
3.1.9	Equal Variance Assumption	32
3.2	Correlation Coefficient	33
3.2.1	Spearman Rank Correlation Coefficient	34
3.2.2	Kendall's τ	34
3.3	Simple Linear Regression Model	35
3.3.1	Systematic Part of the Model.....	35
3.3.2	Random Part of the Model	36
3.3.3	Assumptions About the Predictor	37
3.3.4	Ordinary Least Squares Estimation	38
3.3.5	Fitted Values and Residuals	39
3.3.6	Sums of Squares	39
3.3.7	Standard Errors of the Regression Coefficients	40
3.3.8	Hypothesis Tests and Confidence Intervals	40
3.3.9	Slope, Correlation Coefficient, and R^2	42
3.4	Contingency Table Methods for Binary Outcomes.....	42
3.4.1	Measures of Risk and Association for Binary Outcomes	43
3.4.2	Tests of Association in Contingency Tables	46
3.4.3	Predictors with Multiple Categories	48
3.4.4	Analyses Involving Multiple Categorical Predictors	50
3.4.5	Collapsibility of Standard Measures of Association ...	52
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 Event Function	59
3.5.6	Comparing Groups Using the Logrank Test	60
3.6	Bootstrap Confidence Intervals.....	62
3.7	Interpretation of Negative Findings	64
3.8	Further Notes and References	65
3.9	Problems	65
3.10	Learning Objectives.....	66
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	81
4.3.4	Multiple Pairwise Comparisons Between Categories ..	81
4.3.5	Testing for Trend Across Categories	84
4.4	Confounding	89
4.4.1	Range of Confounding Patterns	90
4.4.2	Confounding Is Difficult to Rule Out	91
4.4.3	Adjusted Versus Unadjusted $\hat{\beta}$ s	92
4.4.4	Example: BMI and LDL.....	93
4.5	Mediation.....	94
4.5.1	Indirect Effects via the Mediator	95
4.5.2	Overall and Direct Effects	95
4.5.3	Percent Explained.....	96
4.5.4	Example: BMI, Exercise, and Glucose	96
4.5.5	Pitfalls in Evaluating Mediation.....	97
4.6	Interaction	99
4.6.1	Example: Hormone Therapy and Statin Use	100
4.6.2	Example: BMI and Statin Use.....	102
4.6.3	Interaction and Scale.....	105
4.6.4	Example: Hormone Therapy and Baseline LDL	106
4.6.5	Details	107
4.7	Checking Model Assumptions and Fit	108
4.7.1	Linearity	109
4.7.2	Normality.....	116
4.7.3	Constant Variance.....	119
4.7.4	Outlying, High Leverage, and Influential Points	124
4.7.5	Interpretation of Results for Log Transformed Variables.....	128
4.7.6	When to Use Transformations.....	129
4.8	Sample Size, Power, and Detectable Effects.....	130
4.8.1	Calculations Using Standard Errors Based on Published Data.....	133
4.9	Summary	135
4.10	Further Notes and References	135
4.10.1	Generalized Additive Models	136
4.11	Problems	136
4.12	Learning Objectives.....	138
5	Logistic Regression	139
5.1	Single Predictor Models	140
5.1.1	Interpretation of Regression Coefficients	144
5.1.2	Categorical Predictors	146
5.2	Multipredictor Models.....	150
5.2.1	Likelihood Ratio Tests.....	154
5.2.2	Confounding	156

5.2.3	Mediation.....	158
5.2.4	Interaction	160
5.2.5	Prediction.....	165
5.2.6	Prediction Accuracy	166
5.3	Case-Control Studies	168
5.3.1	Matched Case-Control Studies	171
5.4	Checking Model Assumptions and Fit	173
5.4.1	Linearity	173
5.4.2	Outlying and Influential Points.....	175
5.4.3	Model Adequacy	177
5.4.4	Technical Issues in Logistic Model Fitting	179
5.5	Alternative Strategies for Binary Outcomes	180
5.5.1	Infectious Disease Transmission Models	181
5.5.2	Pooled Logistic Regression	183
5.5.3	Regression Models Based on Risk Differences and Relative Risks.....	186
5.5.4	Exact Logistic Regression	188
5.5.5	Nonparametric Binary Regression	189
5.5.6	More Than Two Outcome Levels	190
5.6	Likelihood	192
5.7	Sample Size, Power, and Detectable Effects.....	194
5.8	Summary	199
5.9	Further Notes and References	200
5.10	Problems	200
5.11	Learning Objectives.....	202
6	Survival Analysis	203
6.1	Survival Data	203
6.1.1	Why Linear and Logistic Regression Would not Work.....	203
6.1.2	Hazard Function	204
6.1.3	Hazard Ratio	205
6.1.4	Proportional Hazards Assumption	207
6.2	Cox Proportional Hazards Model	207
6.2.1	Proportional Hazards Models	207
6.2.2	Parametric Versus Semi-parametric Models	208
6.2.3	Hazard Ratios, Risk, and Survival Times	211
6.2.4	Hypothesis Tests and Confidence Intervals.....	212
6.2.5	Binary Predictors	213
6.2.6	Multilevel Categorical Predictors	213
6.2.7	Continuous Predictors	217
6.2.8	Confounding	218
6.2.9	Mediation.....	219
6.2.10	Interaction	220
6.2.11	Model Building	222

6.2.12	Adjusted Survival Curves for Comparing Groups.....	222
6.2.13	Predicted Survival for Specific Covariate Patterns	224
6.3	Extensions to the Cox Model	225
6.3.1	Time-Dependent Covariates	225
6.3.2	Stratified Cox Model	228
6.4	Checking Model Assumptions and Fit	231
6.4.1	Log-Linearity of the Hazard Function	231
6.4.2	Proportional Hazards	232
6.5	Competing Risks Data	239
6.5.1	What Are Competing Risks Data?	239
6.5.2	Notation for Competing Risks Data.....	240
6.5.3	Summaries for Competing Risk Data	241
6.6	Some Details	247
6.6.1	Bootstrap Confidence Intervals	247
6.6.2	Prediction.....	248
6.6.3	Adjusting for Nonconfounding Covariates	248
6.6.4	Independent Censoring	249
6.6.5	Interval Censoring	249
6.6.6	Left-Truncation	250
6.7	Sample Size, Power, and Detectable Effects.....	252
6.8	Summary	256
6.9	Further Notes and References	256
6.10	Problems	257
6.11	Learning Objectives.....	259
7	Repeated Measures and Longitudinal Data Analysis.....	261
7.1	A Simple Repeated Measures Example: Fecal Fat	262
7.1.1	Model Equations for the Fecal Fat Example	264
7.1.2	Correlations Within Subjects	264
7.1.3	Estimates of the Effects of Pill Type	266
7.2	Hierarchical Data	267
7.2.1	Example: Treatment of Back Pain	267
7.2.2	Example: Physician Profiling	267
7.2.3	Analysis Strategies for Hierarchical Data	268
7.3	Longitudinal Data	270
7.3.1	Analysis Strategies for Longitudinal Data.....	271
7.3.2	Analyzing Change Scores	273
7.4	Generalized Estimating Equations	276
7.4.1	Example: Birthweight and Birth Order Revisited	277
7.4.2	Correlation Structures	279
7.4.3	Working Correlation and Robust Standard Errors.....	281
7.4.4	Tests and Confidence Intervals	282
7.4.5	Use of <code>xtgee</code> for Clustered Logistic Regression	284
7.5	Random Effects Models	284
7.6	Re-Analysis of the Georgia Babies Data Set	286

7.7	Analysis of the SOF BMD Data.....	288
7.7.1	Time Varying Predictors	289
7.7.2	Separating Between- and Within-Cluster Information .	291
7.7.3	Prediction.....	293
7.7.4	A Logistic Analysis	294
7.8	Marginal Versus Conditional Models	295
7.9	Example: Cardiac Injury Following Brain Hemorrhage	296
7.9.1	Bootstrap Analysis.....	298
7.10	Power and Sample Size for Repeated Measures Designs	301
7.10.1	Between-Cluster Predictor	301
7.10.2	Within-Cluster Predictor.....	303
7.11	Summary	304
7.12	Further Notes and References	305
7.12.1	Missing Data	305
7.12.2	Computing	306
7.13	Problems	306
7.14	Learning Objectives.....	308
8	Generalized Linear Models.....	309
8.1	Example: Treatment for Depression	309
8.1.1	Statistical Issues	310
8.1.2	Model for the Mean Response	311
8.1.3	Choice of Distribution	312
8.1.4	Interpreting the Parameters	312
8.1.5	Further Notes.....	313
8.2	Example: Costs of Phototherapy	314
8.2.1	Model for the Mean Response	315
8.2.2	Choice of Distribution	315
8.2.3	Interpreting the Parameters.....	316
8.3	Generalized Linear Models.....	316
8.3.1	Example: Risky Drug Use Behavior	317
8.3.2	Modeling Data with Many Zeros	318
8.3.3	Example: A Randomized Trial to Reduce Risk of Fracture	321
8.3.4	Relationship of Mean to Variance	323
8.3.5	Non-Linear Models	324
8.4	Sample Size for the Poisson Model	325
8.5	Summary	328
8.6	Further Notes and References	328
8.7	Problems	329
8.8	Learning Objectives.....	330
9	Strengthening Causal Inference.....	331
9.1	Potential Outcomes and Causal Effects	332
9.1.1	Average Causal Effects	332
9.1.2	Marginal Structural Model	333

9.1.3	Fundamental Problem of Causal Inference	333
9.1.4	Randomization Assumption	334
9.1.5	Conditional Independence	334
9.1.6	Marginal and Conditional Means	335
9.1.7	Potential Outcomes Estimation	336
9.1.8	Inverse Probability Weighting	337
9.2	Regression as a Basis for Causal Inference	337
9.2.1	No Unmeasured Confounders	338
9.2.2	Correct Model Specification	338
9.2.3	Overlap and the Positivity Assumption	338
9.2.4	Lack of Overlap and Model Misspecification	339
9.2.5	Adequate Sample Size and Number of Events	341
9.2.6	Example: Phototherapy for Neonatal Jaundice	341
9.3	Marginal Effects and Potential Outcomes Estimation	344
9.3.1	Marginal and Conditional Effects	344
9.3.2	Contrasting Conditional and Marginal Effects	346
9.3.3	When Marginal and Conditional Odds-Ratios Differ	346
9.3.4	Potential Outcomes Estimation	347
9.3.5	Marginal Effects in Longitudinal Data	350
9.4	Propensity Scores	352
9.4.1	Estimation of Propensity Scores	352
9.4.2	Effect Estimation Using Propensity Scores	355
9.4.3	Inverse Probability Weights	356
9.4.4	Checking for Propensity Score/Exposure Interaction ..	358
9.4.5	Addressing Positivity Violations Using Restriction	359
9.4.6	Average Treatment Effect in the Treated (ATT)	360
9.4.7	Recommendations for Using Propensity Scores	362
9.5	Time-Dependent Treatments	364
9.5.1	Models Using Time-dependent IP Weights	365
9.5.2	Implementation	367
9.5.3	Drawbacks and Difficulties	368
9.5.4	Focusing on New Users	369
9.5.5	Nested New-User Cohorts	370
9.6	Mediation	370
9.7	Instrumental Variables	373
9.7.1	Vulnerabilities	375
9.7.2	Structural Equations and Instrumental Variables	377
9.7.3	Checking IV Assumptions	377
9.7.4	Example: Effect of Hormone Therapy on Change in LDL	378
9.7.5	Extension to Binary Exposures and Outcomes	379
9.7.6	Example: Phototherapy for Neonatal Jaundice	380
9.7.7	Interpretation of IV Estimates	382
9.8	Trials with Incomplete Adherence to Treatment	382
9.8.1	Intention-to-Treat	382

9.8.2	As-Treated Comparisons by Treatment Received	384
9.8.3	Instrumental Variables	385
9.8.4	Principal Stratification	385
9.9	Summary	387
9.10	Further Notes and References	387
9.11	Problems	391
9.12	Learning Objectives.....	394
10	Predictor Selection	395
10.1	Prediction.....	396
10.1.1	Bias–Variance Trade-off and Overfitting	397
10.1.2	Measures of Prediction Error.....	397
10.1.3	Optimism-Corrected Estimates of Prediction Error	398
10.1.4	Minimizing Prediction Error Without Overfitting	401
10.1.5	Point Scores	404
10.1.6	Example: Risk Stratification of Patients with Heart Disease	405
10.2	Evaluating a Predictor of Primary Interest.....	407
10.2.1	Including Predictors for Face Validity	408
10.2.2	Selecting Predictors on Statistical Grounds	408
10.2.3	Interactions With the Predictor of Primary Interest	409
10.2.4	Example: Incontinence as a Risk Factor for Falling ...	409
10.2.5	Directed Acyclic Graphs	410
10.2.6	Randomized Experiments	416
10.3	Identifying Multiple Important Predictors	418
10.3.1	Ruling Out Confounding Is Still Central	418
10.3.2	Cautious Interpretation Is Also Key	419
10.3.3	Example: Risk Factors for Coronary Heart Disease ...	420
10.3.4	Allen–Cady Modified Backward Selection	420
10.4	Some Details	421
10.4.1	Collinearity	421
10.4.2	Number of Predictors	422
10.4.3	Alternatives to Backward Selection	424
10.4.4	Model Selection and Checking	425
10.4.5	Model Selection Complicates Inference	425
10.5	Summary	427
10.6	Further Notes and References	427
10.7	Problems	428
10.8	Learning Objectives.....	429
11	Missing Data	431
11.1	Why Missing Data Can Be a Problem	432
11.1.1	Missing Predictor in Linear Regression	432
11.1.2	Missing Outcome in Longitudinal Data	434

11.2	Classifications of Missing Data	437
11.2.1	Mechanisms for Missing Data	438
11.3	Simple Approaches to Handling Missing Data	442
11.3.1	Include a Missing Data Category	442
11.3.2	Last Observation or Baseline Carried Forward	442
11.4	Methods for Handling Missing Data	444
11.5	Missing Data in the Predictors and Multiple Imputation	444
11.5.1	Remarks About Using Multiple Imputation	446
11.5.2	Approaches to Multiple Imputation	447
11.5.3	Multiple Imputation for HERS	449
11.6	Deciding Which Missing Data Mechanism May Be Applicable	451
11.7	Missing Outcomes, Missing Completely at Random	452
11.8	Missing Outcomes, Covariate-Dependent Missing Completely at Random	452
11.9	Missing Outcomes for Longitudinal Studies, Missing at Random	453
11.9.1	ML and MAR	455
11.9.2	Multiple Imputation	456
11.9.3	Inverse Probability Weighting	456
11.10	Technical Details About Maximum Likelihood and Data Which are Missing at Random	458
11.10.1	An Example of the EM Algorithm	458
11.10.2	The EM Algorithm Imputes the Missing Data	460
11.10.3	ML Versus MI with Missing Outcomes	461
11.11	Methods for Data that are Missing Not at Random	461
11.11.1	Pattern Mixture Models	461
11.11.2	Multiple Imputation Under MNAR	463
11.11.3	Joint Modeling of Outcomes and the Dropout Process	463
11.12	Summary	463
11.13	Further Notes and References	464
11.14	Problems	465
11.15	Learning Objectives	467
12	Complex Surveys	469
12.1	Overview of Complex Survey Designs	470
12.2	Inverse Probability Weighting	471
12.2.1	Accounting for Inverse Probability Weights in the Analysis	473
12.2.2	Inverse Probability Weights and Missing Data	473
12.3	Clustering and Stratification	474
12.3.1	Design Effects	474
12.4	Example: Diabetes in NHANES	475

12.5	Some Details	477
12.5.1	Ignoring Secondary Levels of Clustering	477
12.5.2	Other Methods of Variance Estimation	477
12.5.3	Model Checking	478
12.5.4	Postestimation Capabilities in Stata.....	478
12.5.5	Other Statistical Packages for Complex Surveys	479
12.6	Summary	479
12.7	Further Notes and References	479
12.8	Problems	480
12.9	Learning Objectives.....	480
13	Summary	481
13.1	Introduction	481
13.2	Selecting Appropriate Statistical Methods.....	482
13.3	Planning and Executing a Data Analysis	483
13.3.1	Analysis Plans	483
13.3.2	Choice of Software	484
13.3.3	Data Preparation	484
13.3.4	Record Keeping and Reproducibility of Results	484
13.3.5	Data Security	485
13.3.6	Consulting a Statistician	485
13.3.7	Use of Internet Resources	486
13.4	Further Notes and References	486
13.4.1	Multiple Hypothesis Tests	486
13.4.2	Statistical Learning	487
	References.....	489
	Index	501