

Frank E. Harrell, Jr.

Regression Modeling Strategies

With Applications to Linear Models,
Logistic and Ordinal Regression,
and Survival Analysis

Second Edition

 Springer

Contents

Typographical Conventions	xxv
1 Introduction	1
1.1 Hypothesis Testing, Estimation, and Prediction	1
1.2 Examples of Uses of Predictive Multivariable Modeling	3
1.3 Prediction vs. Classification	4
1.4 Planning for Modeling	6
1.4.1 Emphasizing Continuous Variables	8
1.5 Choice of the Model	8
1.6 Further Reading	11
2 General Aspects of Fitting Regression Models	13
2.1 Notation for Multivariable Regression Models	13
2.2 Model Formulations	14
2.3 Interpreting Model Parameters	15
2.3.1 Nominal Predictors	16
2.3.2 Interactions	16
2.3.3 Example: Inference for a Simple Model	17
2.4 Relaxing Linearity Assumption for Continuous Predictors ..	18
2.4.1 Avoiding Categorization	18
2.4.2 Simple Nonlinear Terms	21
2.4.3 Splines for Estimating Shape of Regression Function and Determining Predictor Transformations	22
2.4.4 Cubic Spline Functions	23
2.4.5 Restricted Cubic Splines	24
2.4.6 Choosing Number and Position of Knots	26
2.4.7 Nonparametric Regression	28
2.4.8 Advantages of Regression Splines over Other Methods	30

2.5	Recursive Partitioning: Tree-Based Models	30
2.6	Multiple Degree of Freedom Tests of Association	31
2.7	Assessment of Model Fit	33
	2.7.1 Regression Assumptions	33
	2.7.2 Modeling and Testing Complex Interactions	36
	2.7.3 Fitting Ordinal Predictors	38
	2.7.4 Distributional Assumptions	39
2.8	Further Reading	40
2.9	Problems	42
3	Missing Data	45
	3.1 Types of Missing Data	45
	3.2 Prelude to Modeling	46
	3.3 Missing Values for Different Types of Response Variables	47
	3.4 Problems with Simple Alternatives to Imputation	47
	3.5 Strategies for Developing an Imputation Model	49
	3.6 Single Conditional Mean Imputation	52
	3.7 Predictive Mean Matching	52
	3.8 Multiple Imputation	53
	3.8.1 The <code>aregImpute</code> and Other Chained Equations Approaches	55
	3.9 Diagnostics	56
	3.10 Summary and Rough Guidelines	56
	3.11 Further Reading	58
	3.12 Problems	59
4	Multivariable Modeling Strategies	63
	4.1 Prespecification of Predictor Complexity Without Later Simplification	64
	4.2 Checking Assumptions of Multiple Predictors Simultaneously	67
	4.3 Variable Selection	67
	4.4 Sample Size, Overfitting, and Limits on Number of Predictors	72
	4.5 Shrinkage	75
	4.6 Collinearity	78
	4.7 Data Reduction	79
	4.7.1 Redundancy Analysis	80
	4.7.2 Variable Clustering	81
	4.7.3 Transformation and Scaling Variables Without Using Y	81
	4.7.4 Simultaneous Transformation and Imputation	83
	4.7.5 Simple Scoring of Variable Clusters	85
	4.7.6 Simplifying Cluster Scores	87
	4.7.7 How Much Data Reduction Is Necessary?	87

4.8	Other Approaches to Predictive Modeling	89
4.9	Overly Influential Observations	90
4.10	Comparing Two Models	92
4.11	Improving the Practice of Multivariable Prediction	94
4.12	Summary: Possible Modeling Strategies	94
4.12.1	Developing Predictive Models	95
4.12.2	Developing Models for Effect Estimation	98
4.12.3	Developing Models for Hypothesis Testing	99
4.13	Further Reading	100
4.14	Problems	102
5	Describing, Resampling, Validating, and Simplifying the Model	103
5.1	Describing the Fitted Model	103
5.1.1	Interpreting Effects	103
5.1.2	Indexes of Model Performance	104
5.2	The Bootstrap	106
5.3	Model Validation	109
5.3.1	Introduction	109
5.3.2	Which Quantities Should Be Used in Validation?	110
5.3.3	Data-Splitting	111
5.3.4	Improvements on Data-Splitting: Resampling	112
5.3.5	Validation Using the Bootstrap	114
5.4	Bootstrapping Ranks of Predictors	117
5.5	Simplifying the Final Model by Approximating It	118
5.5.1	Difficulties Using Full Models	118
5.5.2	Approximating the Full Model	119
5.6	Further Reading	121
5.7	Problem	124
6	R Software	127
6.1	The R Modeling Language	128
6.2	User-Contributed Functions	129
6.3	The rms Package	130
6.4	Other Functions	141
6.5	Further Reading	142
7	Modeling Longitudinal Responses using Generalized Least Squares	143
7.1	Notation and Data Setup	143
7.2	Model Specification for Effects on $E(Y)$	144
7.3	Modeling Within-Subject Dependence	144
7.4	Parameter Estimation Procedure	147
7.5	Common Correlation Structures	147
7.6	Checking Model Fit	148

7.7	Sample Size Considerations	148
7.8	R Software	149
7.9	Case Study	149
7.9.1	Graphical Exploration of Data	150
7.9.2	Using Generalized Least Squares	151
7.10	Further Reading	158
8	Case Study in Data Reduction	161
8.1	Data	161
8.2	How Many Parameters Can Be Estimated?	164
8.3	Redundancy Analysis	164
8.4	Variable Clustering	166
8.5	Transformation and Single Imputation Using <code>transcan</code>	167
8.6	Data Reduction Using Principal Components	170
8.6.1	Sparse Principal Components	175
8.7	Transformation Using Nonparametric Smoothers	176
8.8	Further Reading	177
8.9	Problems	178
9	Overview of Maximum Likelihood Estimation	181
9.1	General Notions—Simple Cases	181
9.2	Hypothesis Tests	185
9.2.1	Likelihood Ratio Test	185
9.2.2	Wald Test	186
9.2.3	Score Test	186
9.2.4	Normal Distribution—One Sample	187
9.3	General Case	188
9.3.1	Global Test Statistics	189
9.3.2	Testing a Subset of the Parameters	190
9.3.3	Tests Based on Contrasts	192
9.3.4	Which Test Statistics to Use When	193
9.3.5	Example: Binomial—Comparing Two Proportions	194
9.4	Iterative ML Estimation	195
9.5	Robust Estimation of the Covariance Matrix	196
9.6	Wald, Score, and Likelihood-Based Confidence Intervals	198
9.6.1	Simultaneous Wald Confidence Regions	199
9.7	Bootstrap Confidence Regions	199
9.8	Further Use of the Log Likelihood	203
9.8.1	Rating Two Models, Penalizing for Complexity	203
9.8.2	Testing Whether One Model Is Better than Another	204
9.8.3	Unitless Index of Predictive Ability	205
9.8.4	Unitless Index of Adequacy of a Subset of Predictors	207
9.9	Weighted Maximum Likelihood Estimation	208
9.10	Penalized Maximum Likelihood Estimation	209

9.11	Further Reading	213
9.12	Problems	216
10	Binary Logistic Regression	219
10.1	Model	219
10.1.1	Model Assumptions and Interpretation of Parameters	221
10.1.2	Odds Ratio, Risk Ratio, and Risk Difference	224
10.1.3	Detailed Example	225
10.1.4	Design Formulations	230
10.2	Estimation	231
10.2.1	Maximum Likelihood Estimates	231
10.2.2	Estimation of Odds Ratios and Probabilities	232
10.2.3	Minimum Sample Size Requirement	233
10.3	Test Statistics	234
10.4	Residuals	235
10.5	Assessment of Model Fit	236
10.6	Collinearity	255
10.7	Overly Influential Observations	255
10.8	Quantifying Predictive Ability	256
10.9	Validating the Fitted Model	259
10.10	Describing the Fitted Model	264
10.11	R Functions	269
10.12	Further Reading	271
10.13	Problems	273
11	Binary Logistic Regression Case Study 1	275
11.1	Overview	275
11.2	Background	275
11.3	Data Transformations and Single Imputation	276
11.4	Regression on Original Variables, Principal Components and Pretransformations	277
11.5	Description of Fitted Model	278
11.6	Backwards Step-Down	280
11.7	Model Approximation	287
12	Logistic Model Case Study 2: Survival of Titanic Passengers	291
12.1	Descriptive Statistics	291
12.2	Exploring Trends with Nonparametric Regression	294
12.3	Binary Logistic Model With Casewise Deletion of Missing Values	296
12.4	Examining Missing Data Patterns	302
12.5	Multiple Imputation	304
12.6	Summarizing the Fitted Model	307

13 Ordinal Logistic Regression	311
13.1 Background	311
13.2 Ordinality Assumption	312
13.3 Proportional Odds Model	313
13.3.1 Model	313
13.3.2 Assumptions and Interpretation of Parameters	313
13.3.3 Estimation	314
13.3.4 Residuals	314
13.3.5 Assessment of Model Fit	315
13.3.6 Quantifying Predictive Ability	318
13.3.7 Describing the Fitted Model	318
13.3.8 Validating the Fitted Model	318
13.3.9 R Functions	319
13.4 Continuation Ratio Model	319
13.4.1 Model	319
13.4.2 Assumptions and Interpretation of Parameters	320
13.4.3 Estimation	320
13.4.4 Residuals	321
13.4.5 Assessment of Model Fit	321
13.4.6 Extended CR Model	321
13.4.7 Role of Penalization in Extended CR Model	322
13.4.8 Validating the Fitted Model	322
13.4.9 R Functions	323
13.5 Further Reading	324
13.6 Problems	324
14 Case Study in Ordinal Regression, Data Reduction, and Penalization	327
14.1 Response Variable	328
14.2 Variable Clustering	329
14.3 Developing Cluster Summary Scores	330
14.4 Assessing Ordinality of Y for each X , and Unadjusted Checking of PO and CR Assumptions	333
14.5 A Tentative Full Proportional Odds Model	333
14.6 Residual Plots	336
14.7 Graphical Assessment of Fit of CR Model	338
14.8 Extended Continuation Ratio Model	340
14.9 Penalized Estimation	342
14.10 Using Approximations to Simplify the Model	348
14.11 Validating the Model	353
14.12 Summary	355
14.13 Further Reading	356
14.14 Problems	357

15	Regression Models for Continuous Y and Case Study in Ordinal Regression	359
15.1	The Linear Model	359
15.2	Quantile Regression.....	360
15.3	Ordinal Regression Models for Continuous Y	361
15.3.1	Minimum Sample Size Requirement	363
15.4	Comparison of Assumptions of Various Models	364
15.5	Dataset and Descriptive Statistics	365
15.5.1	Checking Assumptions of OLS and Other Models... ..	368
15.6	Ordinal Regression Applied to HbA_{1c}	370
15.6.1	Checking Fit for Various Models Using Age.....	370
15.6.2	Examination of BMI.....	374
15.6.3	Consideration of All Body Size Measurements.....	375
16	Transform-Both-Sides Regression	389
16.1	Background.....	389
16.2	Generalized Additive Models.....	390
16.3	Nonparametric Estimation of Y -Transformation	390
16.4	Obtaining Estimates on the Original Scale.....	391
16.5	R Functions.....	392
16.6	Case Study	393
17	Introduction to Survival Analysis	399
17.1	Background.....	399
17.2	Censoring, Delayed Entry, and Truncation	401
17.3	Notation, Survival, and Hazard Functions	402
17.4	Homogeneous Failure Time Distributions	407
17.5	Nonparametric Estimation of S and Λ	409
17.5.1	Kaplan–Meier Estimator	409
17.5.2	Altschuler–Nelson Estimator	413
17.6	Analysis of Multiple Endpoints.....	413
17.6.1	Competing Risks	414
17.6.2	Competing Dependent Risks	414
17.6.3	State Transitions and Multiple Types of Nonfatal Events	416
17.6.4	Joint Analysis of Time and Severity of an Event.....	417
17.6.5	Analysis of Multiple Events	417
17.7	R Functions.....	418
17.8	Further Reading.....	420
17.9	Problems	421
18	Parametric Survival Models	423
18.1	Homogeneous Models (No Predictors).....	423
18.1.1	Specific Models	423
18.1.2	Estimation	424
18.1.3	Assessment of Model Fit	426

18.2	Parametric Proportional Hazards Models	427
18.2.1	Model	427
18.2.2	Model Assumptions and Interpretation of Parameters	428
18.2.3	Hazard Ratio, Risk Ratio, and Risk Difference	430
18.2.4	Specific Models	431
18.2.5	Estimation	432
18.2.6	Assessment of Model Fit	434
18.3	Accelerated Failure Time Models	436
18.3.1	Model	436
18.3.2	Model Assumptions and Interpretation of Parameters	436
18.3.3	Specific Models	437
18.3.4	Estimation	438
18.3.5	Residuals	440
18.3.6	Assessment of Model Fit	440
18.3.7	Validating the Fitted Model	446
18.4	Buckley–James Regression Model	447
18.5	Design Formulations	447
18.6	Test Statistics	447
18.7	Quantifying Predictive Ability	447
18.8	Time-Dependent Covariates	447
18.9	R Functions	448
18.10	Further Reading	450
18.11	Problems	451
19	Case Study in Parametric Survival Modeling and Model Approximation	453
19.1	Descriptive Statistics	453
19.2	Checking Adequacy of Log-Normal Accelerated Failure Time Model	458
19.3	Summarizing the Fitted Model	466
19.4	Internal Validation of the Fitted Model Using the Bootstrap	466
19.5	Approximating the Full Model	469
19.6	Problems	473
20	Cox Proportional Hazards Regression Model	475
20.1	Model	475
20.1.1	Preliminaries	475
20.1.2	Model Definition	476
20.1.3	Estimation of β	476
20.1.4	Model Assumptions and Interpretation of Parameters	478
20.1.5	Example	478

20.1.6	Design Formulations	480
20.1.7	Extending the Model by Stratification	481
20.2	Estimation of Survival Probability and Secondary Parameters	483
20.3	Sample Size Considerations	486
20.4	Test Statistics	486
20.5	Residuals	487
20.6	Assessment of Model Fit	487
20.6.1	Regression Assumptions	487
20.6.2	Proportional Hazards Assumption	494
20.7	What to Do When PH Fails	501
20.8	Collinearity	503
20.9	Overly Influential Observations	504
20.10	Quantifying Predictive Ability	504
20.11	Validating the Fitted Model	506
20.11.1	Validation of Model Calibration	506
20.11.2	Validation of Discrimination and Other Statistical Indexes	507
20.12	Describing the Fitted Model	509
20.13	R Functions	513
20.14	Further Reading	517
21	Case Study in Cox Regression	521
21.1	Choosing the Number of Parameters and Fitting the Model	521
21.2	Checking Proportional Hazards	525
21.3	Testing Interactions	527
21.4	Describing Predictor Effects	527
21.5	Validating the Model	529
21.6	Presenting the Model	530
21.7	Problems	531
A	Datasets, R Packages, and Internet Resources	535
	References	539
	Index	571