

Contents

Preface	xi
1 Introduction	1
1.1 Before You Start	1
1.2 Initial Data Analysis	2
1.3 When to Use Linear Modeling	7
1.4 History	8
2 Estimation	13
2.1 Linear Model	13
2.2 Matrix Representation	14
2.3 Estimating β	15
2.4 Least Squares Estimation	16
2.5 Examples of Calculating β	17
2.6 Example	17
2.7 QR Decomposition	20
2.8 Gauss–Markov Theorem	22
2.9 Goodness of Fit	23
2.10 Identifiability	26
2.11 Orthogonality	28
3 Inference	33
3.1 Hypothesis Tests to Compare Models	33
3.2 Testing Examples	35
3.3 Permutation Tests	40
3.4 Sampling	42
3.5 Confidence Intervals for β	43
3.6 Bootstrap Confidence Intervals	46
4 Prediction	51
4.1 Confidence Intervals for Predictions	51
4.2 Predicting Body Fat	52
4.3 Autoregression	54
4.4 What Can Go Wrong with Predictions?	56

5 Explanation	59
5.1 Simple Meaning	59
5.2 Causality	61
5.3 Designed Experiments	62
5.4 Observational Data	63
5.5 Matching	65
5.6 Covariate Adjustment	68
5.7 Qualitative Support for Causation	69
6 Diagnostics	73
6.1 Checking Error Assumptions	73
6.1.1 Constant Variance	73
6.1.2 Normality	78
6.1.3 Correlated Errors	81
6.2 Finding Unusual Observations	83
6.2.1 Leverage	83
6.2.2 Outliers	85
6.2.3 Influential Observations	89
6.3 Checking the Structure of the Model	92
6.4 Discussion	96
7 Problems with the Predictors	99
7.1 Errors in the Predictors	99
7.2 Changes of Scale	103
7.3 Collinearity	106
8 Problems with the Error	113
8.1 Generalized Least Squares	113
8.2 Weighted Least Squares	116
8.3 Testing for Lack of Fit	119
8.4 Robust Regression	123
8.4.1 M-Estimation	123
8.4.2 Least Trimmed Squares	126
9 Transformation	133
9.1 Transforming the Response	133
9.2 Transforming the Predictors	137
9.3 Broken Stick Regression	137
9.4 Polynomials	139
9.5 Splines	141
9.6 Additive Models	144
9.7 More Complex Models	145

10 Model Selection	150
10.1 Hierarchical Models	150
10.2 Testing-Based Procedures	151
10.3 Criterion-Based Procedures	153
10.4 Summary	159
11 Shrinkage Methods	161
11.1 Principal Components	161
11.2 Partial Least Squares	172
11.3 Ridge Regression	174
11.4 Lasso	177
12 Insurance Redlining — A Complete Example	183
12.1 Ecological Correlation	183
12.2 Initial Data Analysis	185
12.3 Full Model and Diagnostics	188
12.4 Sensitivity Analysis	190
12.5 Discussion	193
13 Missing Data	197
13.1 Types of Missing Data	197
13.2 Deletion	198
13.3 Single Imputation	200
13.4 Multiple Imputation	202
14 Categorical Predictors	205
14.1 A Two-Level Factor	205
14.2 Factors and Quantitative Predictors	209
14.3 Interpretation with Interaction Terms	212
14.4 Factors With More Than Two Levels	213
14.5 Alternative Codings of Qualitative Predictors	219
15 One Factor Models	223
15.1 The Model	223
15.2 An Example	224
15.3 Diagnostics	227
15.4 Pairwise Comparisons	228
15.5 False Discovery Rate	230
16 Models with Several Factors	235
16.1 Two Factors with No Replication	235
16.2 Two Factors with Replication	239
16.3 Two Factors with an Interaction	243
16.4 Larger Factorial Experiments	246

17 Experiments with Blocks	24
17.1 Randomized Block Design	252
17.2 Latin Squares	256
17.3 Balanced Incomplete Block Design	259
A About R	265
Bibliography	267
Bibliography	267
Index	271