

Linear Models with Python

Julian J. Faraway



CRC Press

Taylor & Francis Group

Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an *informa* business

A CHAPMAN & HALL BOOK

Contents

Preface	ix
1 Introduction	1
1.1 Before You Start	1
1.2 Initial Data Analysis	2
1.3 When to Use <i>Linear Modeling</i>	6
1.4 History	7
2 Estimation	15
2.1 Linear Model	15
2.2 Matrix Representation	16
2.3 Estimating β	17
2.4 Least Squares Estimation	18
2.5 Examples of Calculating $\hat{\beta}$	19
2.6 Example	19
2.7 Computing Least Squares Estimates	22
2.8 Gauss–Markov Theorem	24
2.9 Goodness of Fit	26
2.10 Identifiability	28
2.11 Orthogonality	31
3 Inference	37
3.1 Hypothesis Tests to Compare Models	37
3.2 Testing Examples	39
3.3 Permutation Tests	44
3.4 Sampling	45
3.5 Confidence Intervals for β	47
3.6 Bootstrap Confidence Intervals	48
4 Prediction	53
4.1 Confidence Intervals for Predictions	53
4.2 Predicting Body Fat	54
4.3 Autoregression	56
4.4 What Can Go Wrong with Predictions?	58

5	Explanation	61
5.1	Simple Meaning	61
5.2	Causality	63
5.3	Designed Experiments	64
5.4	Observational Data	65
5.5	Matching	67
5.6	Covariate Adjustment	70
5.7	Qualitative Support for Causation	71
6	Diagnostics	75
6.1	Checking Error Assumptions	75
6.1.1	Constant Variance	75
6.1.2	Normality	80
6.1.3	Correlated Errors	83
6.2	Finding Unusual Observations	85
6.2.1	Leverage	85
6.2.2	Outliers	87
6.2.3	Influential Observations	91
6.3	Checking the Structure of the Model	93
6.4	Discussion	96
7	Problems with the Predictors	101
7.1	Errors in the Predictors	101
7.2	Changes of Scale	105
7.3	Collinearity	108
8	Problems with the Error	115
8.1	Generalized Least Squares	115
8.2	Weighted Least Squares	117
8.3	Testing for Lack of Fit	121
8.4	Robust Regression	125
8.4.1	M-Estimation	125
8.4.2	High Breakdown Estimators	128
9	Transformation	135
9.1	Transforming the Response	135
9.2	Transforming the Predictors	140
9.3	Broken Stick Regression	140
9.4	Polynomials	142
9.5	Splines	148
9.6	Additive Models	150
9.7	More Complex Models	152

10 Model Selection	155
10.1 Hierarchical Models	156
10.2 Hypothesis Testing-Based Procedures	156
10.3 Criterion-Based Procedures	160
10.4 Sample Splitting	163
10.5 Crossvalidation	167
10.6 Summary	169
11 Shrinkage Methods	173
11.1 Principal Components	173
11.2 Partial Least Squares	184
11.3 Ridge Regression	187
11.4 Lasso	191
11.5 Other Methods	194
12 Insurance Redlining — A Complete Example	197
12.1 Ecological Correlation	197
12.2 Initial Data Analysis	199
12.3 Full Model and Diagnostics	202
12.4 Sensitivity Analysis	204
12.5 Discussion	207
13 Missing Data	211
13.1 Types of Missing Data	211
13.2 Representation and Detection of Missing Values	212
13.3 Deletion	213
13.4 Single Imputation	215
13.5 Multiple Imputation	217
13.6 Discussion	219
14 Categorical Predictors	221
14.1 A Two-Level Factor	221
14.2 Factors and Quantitative Predictors	225
14.3 Interpretation with Interaction Terms	228
14.4 Factors with More Than Two Levels	230
14.5 Alternative Codings of Qualitative Predictors	235
15 One-Factor Models	241
15.1 The Model	241
15.2 An Example	242
15.3 Diagnostics	245
15.4 Pairwise Comparisons	246
15.5 False Discovery Rate	248

16 Models with Several Factors	253
16.1 Two Factors with No Replication	253
16.2 Two Factors with Replication	257
16.3 Two Factors with an Interaction	262
16.4 Larger Factorial Experiments	266
17 Experiments with Blocks	273
17.1 Randomized Block Design	274
17.2 Latin Squares	278
17.3 Balanced Incomplete Block Design	282
A About Python	289
Bibliography	291
Index	295