Methods and Applications of Linear Models
Regression and the Analysis of Variance
Third Edition

RONALD R. HOCKING
PenHock Statistical Consultants
Ishpeming, Michigan

WILEY
Contents

Preface to the Third Edition xvii
Preface to the Second Edition xix
Preface to the First Edition xxii

PART I REGRESSION 1

1 Introduction to Linear Models 3

1.1 Background Information, 3
1.2 Mathematical and Statistical Models, 5
1.3 Definition of the Linear Model, 8
1.4 Examples of Regression Models, 13

1.4.1 Single-Variable, Regression Model, 13
1.4.2 Regression Models with Several Inputs, 17
1.4.3 Discrete Response Variables, 20
1.4.4 Multivariate Linear Models, 20

1.5 Concluding Comments, 21
Exercises, 21

2 Regression on Functions of One Variable 23

2.1 The Simple Linear Regression Model, 23
2.2 Parameter Estimation, 25

2.2.1 Least Squares Estimation, 25
2.2.2 Maximum Likelihood Estimation, 31
2.2.3 Coded Data: Centering and Scaling, 32
2.2.4 The Analysis of Variance Table, 33
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.3 Properties of the Estimators and Test Statistics</td>
<td>34</td>
</tr>
<tr>
<td>2.3.1 Moments of Linear Functions of Random Variables</td>
<td>35</td>
</tr>
<tr>
<td>2.3.2 Moments of Least Squares Estimators</td>
<td>35</td>
</tr>
<tr>
<td>2.3.3 Distribution of the Least Squares Estimators</td>
<td>36</td>
</tr>
<tr>
<td>2.3.4 The Distribution of Test Statistics</td>
<td>38</td>
</tr>
<tr>
<td>2.4 The Analysis of Simple Linear Regression Models</td>
<td>39</td>
</tr>
<tr>
<td>2.4.1 Two Numerical Examples</td>
<td>39</td>
</tr>
<tr>
<td>2.4.2 A Test for Lack-of-Fit</td>
<td>42</td>
</tr>
<tr>
<td>2.4.3 Inference on the Parameters of the Model</td>
<td>45</td>
</tr>
<tr>
<td>2.4.4 Prediction and Prediction Intervals</td>
<td>48</td>
</tr>
<tr>
<td>2.5 Examining the Data and the Model</td>
<td>50</td>
</tr>
<tr>
<td>2.5.1 Residuals</td>
<td>50</td>
</tr>
<tr>
<td>2.5.2 Outliers, Extreme Points, and Influence</td>
<td>54</td>
</tr>
<tr>
<td>2.5.3 Normality, Independence, and Variance Homogeneity</td>
<td>59</td>
</tr>
<tr>
<td>2.6 Polynomial Regression Models</td>
<td>63</td>
</tr>
<tr>
<td>2.6.1 The Quadratic Model</td>
<td>64</td>
</tr>
<tr>
<td>2.6.2 Higher Ordered Polynomial Models</td>
<td>70</td>
</tr>
<tr>
<td>2.6.3 Orthogonal Polynomials</td>
<td>71</td>
</tr>
<tr>
<td>2.6.4 Regression through the Origin</td>
<td>72</td>
</tr>
<tr>
<td>Exercises</td>
<td>72</td>
</tr>
<tr>
<td>3 Transforming the Data</td>
<td>81</td>
</tr>
<tr>
<td>3.1 The Need for Transformations</td>
<td>81</td>
</tr>
<tr>
<td>3.2 Weighted Least Squares</td>
<td>82</td>
</tr>
<tr>
<td>3.3 Variance Stabilizing Transformations</td>
<td>85</td>
</tr>
<tr>
<td>3.4 Transformations to Achieve a Linear Model</td>
<td>86</td>
</tr>
<tr>
<td>3.4.1 Transforming the Dependent Variable</td>
<td>86</td>
</tr>
<tr>
<td>3.4.2 Transforming the Predictors</td>
<td>91</td>
</tr>
<tr>
<td>3.5 Analysis of the Transformed Model</td>
<td>92</td>
</tr>
<tr>
<td>3.5.1 Transformations with Forbes Data</td>
<td>93</td>
</tr>
<tr>
<td>Exercises</td>
<td>95</td>
</tr>
<tr>
<td>4 Regression on Functions of Several Variables</td>
<td>99</td>
</tr>
<tr>
<td>4.1 The Multiple Linear Regression Model</td>
<td>99</td>
</tr>
<tr>
<td>4.2 Preliminary Data Analysis</td>
<td>100</td>
</tr>
<tr>
<td>4.3 Analysis of the Multiple Linear Regression Model</td>
<td>103</td>
</tr>
<tr>
<td>4.3.1 Fitting the Model in Centered Form</td>
<td>104</td>
</tr>
<tr>
<td>4.3.2 Estimation and Analysis of the Original Data</td>
<td>105</td>
</tr>
<tr>
<td>4.3.3 Model Assessment and Residual Analysis</td>
<td>107</td>
</tr>
<tr>
<td>4.3.4 Prediction</td>
<td>108</td>
</tr>
<tr>
<td>4.3.5 Transforming the Response</td>
<td>110</td>
</tr>
</tbody>
</table>
4.4 Partial Correlation and Added-Variable Plots, 113
   4.4.1 Partial Correlation, 113
   4.4.2 Added-Variable Plots, 115
   4.4.3 Simple Versus Partial Correlation, 118
4.5 Variable Selection, 119
   4.5.1 The Case of Orthogonal Predictors, 120
   4.5.2 Criteria for Deletion of Variables, 121
   4.5.3 Nonorthogonal Predictors, 123
   4.5.4 Computational Considerations, 125
   4.5.5 Selection Strategies, 127
4.6 Model Specification, 130
   4.6.1 Application to Subset Selection, 134
   4.6.2 Improved Mean Squared Error, 135
   4.6.3 Development of the $C_p$ Statistic, 136
Exercises, 137

5 Collinearity in Multiple Linear Regression 142
5.1 The Collinearity Problem, 142
   5.1.1 Introduction, 142
   5.1.2 A Simple Example, 143
   5.1.3 The Picket Fence, 147
   5.1.4 Rotation of Coordinates, 149
5.2 An Example with Collinearity, 150
   5.2.1 Preliminary Data Analysis, 151
   5.2.2 Initial Regression Analysis, 152
5.3 Collinearity Diagnostics, 156
   5.3.1 Variance Inflation Factors, 156
   5.3.2 Eigenvalues, Eigenvectors, and Principal Component Plots, 159
5.4 Remedial Solutions: Biased Estimators, 166
   5.4.1 Variable Deletion, 168
   5.4.2 Regression on Principal Components, 170
   5.4.3 Ridge Regression, 174
Exercises, 178

6 Influential Observations in Multiple Linear Regression 182
6.1 The Influential Data Problem, 182
6.2 The Hat Matrix, 183
   6.2.1 The Centered and Uncentered Hat Matrices, 183
   6.2.2 Properties of the Hat Matrices, 185
6.3 The Effects of Deleting Observations, 188
   6.3.1 Estimation of $\beta$, 189
6.3.2 Computation of Residuals, 190
6.3.3 Computation of Predicted Values, 190
6.3.4 Estimation of the Error Variance, $\sigma^2$, 191
6.3.5 Elements of the Hat Matrix, 191
6.3.6 The Determinant of $X'X$, 192
6.3.7 Deletion of More Than One Case, 192
6.4 Numerical Measures of Influence, 192
6.4.1 The Diagonal Elements of the Hat Matrix, 193
6.4.2 Residuals, 193
6.4.3 The Mean Square Ratio, 195
6.4.4 Cook's Distance, 195
6.4.5 Other Indicators of Influential Data, 196
6.5 The Dilemma Data, 197
6.6 Plots for Identifying Unusual Cases, 201
6.6.1 The Projection Ellipse, 201
6.6.2 The Augmented Hat Matrix, 204
6.6.3 Multiple Extremes: The Masking Problem, 206
6.7 Robust/Resistant Methods in Regression Analysis, 209
6.7.1 $M$-Estimation, 209
6.7.2 Iterative, Reweighted Least Squares, 211
6.7.3 Regression with Bounded Influence, 212
Exercises, 213

7 Polynomial Models and Qualitative Predictors
7.1 Polynomial Models, 216
7.1.1 The Quadratic Model with Two Predictors, 216
7.1.2 Quadratic Surfaces, 217
7.2 The Analysis of Response Surfaces, 220
7.2.1 Analysis with First-Order Models, 221
7.2.2 Analysis with Second-Order Models, 223
7.3 Models with Qualitative Predictors, 225
7.3.1 Indicator Variables to Identify Groups of Data, 226
7.3.2 Indicator Variables to Fit Segmented Polynomials, 240
Exercises, 247

8 Additional Topics
8.1 Nonlinear Regression Models, 254
8.1.1 Some Linearizeable Functions, 255
8.1.2 The Modified Gauss–Newton Method, 258
8.2 Nonparametric Model-Fitting Methods, 260
  8.2.1 Locally Weighted Average Predictors, 261
  8.2.2 Projection Pursuit Regression, 263
8.3 Generalized Linear Models, 265
  8.3.1 Logistic Regression, 266
  8.3.2 Poisson Regression, 272
8.4 Random Input Variables, 274
8.5 Errors in the Inputs, 276
8.6 Calibration, 277
Exercises, 278

PART II THE ANALYSIS OF VARIANCE 283

9 Classification Models I: Introduction 285
  9.1 Background Information, 285
  9.2 The One-Way Classification Model, 286
    9.2.1 The Cell Means Model, 288
    9.2.2 Specification of Hypotheses, 290
    9.2.3 The Numerator Sum of Squares, 291
    9.2.4 Pairwise Comparisons of Means, 292
    9.2.5 Orthogonal Contrasts Among the Means, 294
    9.2.6 The Acceptance Ellipsoid, 297
    9.2.7 A Reparameterized Model, 300
    9.2.8 The Analysis of Covariance, 302
  9.3 The Two-Way Classification Model: Balanced Data, 304
    9.3.1 The Cell Means Model for the Two-Way Model, 304
    9.3.2 Hypotheses for the Two-Way Model, 307
    9.3.3 Simultaneous Inference on Marginal the Means, 313
    9.3.4 Reparameterizations of the Two-Factor Model, 315
    9.3.5 Parameter Estimation and Hypothesis Testing, 318
    9.3.6 A Non-Full Rank Model, 321
  9.4 The Two-Way Classification Model: Unbalanced Data, 322
    9.4.1 Discussion in Terms of the Cell Means Model, 322
    9.4.2 Discussion in Terms of the Reparameterized Model, 323
    9.4.3 The Case of Zero Cell Frequencies, 327
  9.5 The Two-Way Classification Model: No Interaction, 334
    9.5.1 Parameter Estimation, 334
    9.5.2 Tests of Hypotheses, 335
    9.5.3 Simultaneous Inference, 338
9.5.4 The No-Interaction Model: Unbalanced Data, 338
9.5.5 Missing Cells: Estimation, 339
9.5.6 Missing Cells: Testing Hypotheses, 343
9.5.7 Connected Designs, 346
9.6 Concluding Comments, 347
Exercises, 347

10 The Mathematical Theory of Linear Models 359
10.1 The Distribution of Linear and Quadratic Forms, 359
  10.1.1 The Distribution of Linear Functions of Normal Variables, 362
  10.1.2 The Distribution of Quadratic Functions of Normal Variables, 362
10.2 Estimation and Inference for Linear Models, 368
  10.2.1 Estimation of Parameters, 368
  10.2.2 Optimality Properties of the Estimators, 370
  10.2.3 Estimation for the Constrained Model, 372
  10.2.4 A Partitioned Form of the Model, 377
  10.2.5 Reparameterized Models, 378
10.3 Tests of Linear Hypotheses on $\beta$, 380
  10.3.1 Unconstrained Model, 381
  10.3.2 The Constrained Model, 390
10.4 Confidence Regions and Intervals, 392
  10.4.1 Confidence Regions, 392
  10.4.2 Confidence Intervals, 393
Exercises, 395

11 Classification Models II: Multiple Crossed and Nested Factors 405
11.1 The Three-Factor Cross-Classified Model, 406
  11.1.1 The Analysis with Balanced Data, 406
  11.1.2 Tests of Hypotheses, 406
  11.1.3 Unbalanced Data, $n_{ijk} \neq 0$, 410
  11.1.4 Estimability and Testability with Missing Cells, 411
11.2 A General Structure for Balanced, Factorial Models, 412
  11.2.1 The Hypotheses, 413
  11.2.2 Numerator Sums of Squares, 414
  11.2.3 The Reparameterized Model, 415
  11.2.4 The Sum of Squares Identity, 416
11.3 The Twofold Nested Model, 417
11.4 A General Structure for Balanced, Nested Models, 426
  11.4.1 The Hypotheses, 426
11.4.2 The Hypothesis Sums of Squares, 427
11.4.3 The Reparameterized Model, 428
11.5 A Three-Factor, Nested-Factorial Model, 429
11.5.1 The Analysis with Balanced Data, 429
11.5.2 The Analysis with Unbalanced Data, 432
11.6 A General Structure for Balanced, Nested-Factorial Models, 434
11.6.1 The Hypotheses, 435
11.6.2 The Reparameterized Model, 436
11.6.3 The Hypotheses Sums of Squares, 437
11.6.4 The Sum of Squares Identity, 438
11.6.5 Summary, 438
Exercises, 438

12 Mixed Models I: The AOV Method with Balanced Data 443

12.1 Introduction, 443
12.2 Examples of the Analysis of Mixed Models, 444
  12.2.1 The One-Way Classification, Random Model, 444
  12.2.2 The Two-Way Classification, Mixed Model, 451
  12.2.3 The Three-Factor, Nested-Factorial Model, 458
  12.2.4 The Randomized Block Design, 461
12.3 The General Analysis for Balanced, Mixed Models, 464
  12.3.1 A Description of the Model, 465
  12.3.2 Parameter Estimation and Inference, 466
12.4 Additional Examples, 479
  12.4.1 The Twofold Nested, Random Model, 479
  12.4.2 Mixed Models for Split-Plot Designs, 481
  12.4.3 Repeated Measures Designs, 486
  12.4.4 Longitudinal Studies, 487
12.5 Alternative Developments of Mixed Models, 487
  12.5.1 The Graybill Mixed Model, 488
  12.5.2 The Scheffé Mixed Model, 491
  12.5.3 A Randomization Theory, 492
Exercises, 493

13 Mixed Models II: The AVE Method with Balanced Data 499

13.1 Introduction, 499
13.2 The Two-Way Cross-Classification Model, 500
  13.2.1 The Mixed Model, 500
  13.2.2 The Random Model, 508
13.3 The Three-Factor, Cross-Classification Model, 511
13.4 Nested Models, 515
13.5 Nested-Factorial Models, 518
13.6 A General Description of the AVE Table, 524
   13.6.1 The AVE Table for Factorial Models, 525
   13.6.2 The AVE Table for Nested Models, 529
   13.6.3 The AVE Table for Nested-Factorial Models, 529
   13.6.4 The AVE Method for General Mixed Effects Models, 530
13.7 Additional Examples, 531
13.8 The Computational Procedure for the AVE Method, 537
   Exercises, 537

14 Mixed Models III: Unbalanced Data 543
14.1 Introduction, 543
14.2 Parameter Estimation: Likelihood Methods, 545
   14.2.1 Maximum Likelihood Estimation, 545
   14.2.2 Restricted Maximum Likelihood Estimation, 549
   14.2.3 Minimum Norm Quadratic Unbiased Estimators, 552
   14.2.4 A Numerical Illustration of the Methods, 552
14.3 ML and REML Estimates with Balanced Data, 554
   14.3.1 ML Estimation with Balanced Data, 554
   14.3.2 REML Estimation with Balanced Data, 558
14.4 The EM Algorithm for REML Estimation, 558
   14.4.1 A Review of the EM Algorithm, 559
   14.4.2 The EM Algorithm Applied to REML Estimation, 560
   14.4.3 The Estimation of Fixed Effects, 565
   14.4.4 Inferences on Variance Components and Fixed Effects, 566
   14.4.5 Numerical Examples to Illustrate the EM-AOV Algorithm, 568
14.5 Diagnostic Analysis with the EM Algorithm, 572
   14.5.1 Numerical Examples to Illustrate the Diagnostics, 572
   14.5.2 The Computation of Individual, Pseudo Degrees of Freedom, 573
   14.5.3 Additional Numerical Examples, 575
14.6 Models with Covariates, 581
   14.6.1 The Development of the Analysis, 581
   14.6.2 A Numerical Example, 583
14.7 Summary, 585
   Exercises, 585
CONTENTS

15 Simultaneous Inference: Tests and Confidence Intervals 591
15.1 Simultaneous Tests, 591
  15.1.1 Simultaneous Tests: General Methods, 592
  15.1.2 Simultaneous Tests: Cell Means Models, 603
15.2 Simultaneous Confidence Intervals, 610
  15.2.1 The Bonferroni Confidence Intervals, 611
  15.2.2 The Scheffé Confidence Intervals, 611
  15.2.3 Tukey Studentized-Range Intervals, 611
Exercises, 612

Appendix A Mathematics 615
A.I Matrix Algebra, 615
  A.I.1 Notation, 615
  A.I.2 The Rank of a Matrix, 616
  A.I.3 The Trace of a Matrix, 617
  A.I.4 Eigenvalues and Eigenvectors, 617
  A.I.5 Quadratic Forms and Definite Matrices, 618
  A.I.6 Special Matrices, 619
  A.I.7 The Diagonalization of Matrices, 620
  A.I.8 Kronecker Products of Matrices, 620
  A.I.9 Factorization of Matrices, 621
  A.I.10 Matrix Inversion, 622
  A.I.11 The Solution of Linear Equations, 624
  A.I.12 Generalized Inverses, 627
  A.I.13 Cauchy–Schwartz Inequalities, 630
A.II Optimization, 630
  A.II.1 The Differentiation of Matrices and Determinants, 630
  A.II.2 The Differentiation of a Function with Respect to
        a Vector, 631
  A.II.3 The Optimization of a Function, 632

Appendix B Statistics 634
B.I Distributions, 634
  B.I.1 The Normal Distribution, 634
  B.I.2 The $\chi^2$-Distribution, 637
  B.I.3 The $t$-Distribution, 638
  B.I.4 The $F$-distribution, 639
B.II The Distribution of Quadratic Forms, 639
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>B.III Estimation</td>
<td>642</td>
</tr>
<tr>
<td>B.III.1 Maximum Likelihood Estimation</td>
<td>642</td>
</tr>
<tr>
<td>B.III.2 Constrained Maximum Likelihood Estimation</td>
<td>642</td>
</tr>
<tr>
<td>B.III.3 Complete, Sufficient Statistics</td>
<td>643</td>
</tr>
<tr>
<td>B.IV Tests of Hypotheses and Confidence Regions</td>
<td>643</td>
</tr>
<tr>
<td>B.IV.1 Tests of Hypotheses</td>
<td>643</td>
</tr>
<tr>
<td>B.IV.2 Confidence Intervals and Regions</td>
<td>644</td>
</tr>
</tbody>
</table>

**Appendix C** Data Tables  

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>C.I Downloading Data Files from FTP Server</td>
<td>645</td>
</tr>
<tr>
<td>C.II Listing of Data Set Files</td>
<td>645</td>
</tr>
</tbody>
</table>

**Appendix D** Statistical Tables  

References  

Index