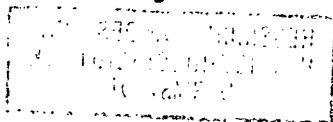


# Introduction to Linear Regression Analysis

Third Edition



DOUGLAS C. MONTGOMERY

Arizona State University

ELIZABETH A. PECK

The Coca-Cola Company

G. GEOFFREY VINING

Virginia Tech



A Wiley Interscience Publication

**JOHN WILEY & SONS, INC.**

New York • Chichester • Weinheim • Brisbane • Singapore • Toronto

# Contents

|                                                                                      |             |
|--------------------------------------------------------------------------------------|-------------|
| <b>Preface</b>                                                                       | <b>xiii</b> |
| <b>1. Introduction</b>                                                               | <b>1</b>    |
| 1.1 Regression and Model Building, 1                                                 |             |
| 1.2 Data Collection, 7                                                               |             |
| 1.3 Uses of Regression, 11                                                           |             |
| 1.4 Role of the Computer, 12                                                         |             |
| <b>2. Simple Linear Regression</b>                                                   | <b>13</b>   |
| 2.1 Simple Linear Regression Model, 13                                               |             |
| 2.2 Least-Squares Estimation of the Parameters, 14                                   |             |
| 2.2.1 Estimation of $\beta_0$ and $\beta_1$ , 14                                     |             |
| 2.2.2 Properties of the Least-Squares Estimators and the Fitted Regression Model, 20 |             |
| 2.2.3 Estimation of $\sigma^2$ , 22                                                  |             |
| 2.2.4 An Alternate Form of the Model, 24                                             |             |
| 2.3 Hypothesis Testing on the Slope and Intercept, 24                                |             |
| 2.3.1 Use of $t$ -Tests, 25                                                          |             |
| 2.3.2 Testing Significance of Regression, 26                                         |             |
| 2.3.3 The Analysis of Variance, 28                                                   |             |
| 2.4 Interval Estimation in Simple Linear Regression, 32                              |             |
| 2.4.1 Confidence Intervals on $\beta_0$ , $\beta_1$ , and $\sigma^2$ , 32            |             |
| 2.4.2 Interval Estimation of the Mean Response, 34                                   |             |
| 2.5 Prediction of New Observations, 37                                               |             |
| 2.6 Coefficient of Determination, 39                                                 |             |
| 2.7 Some Considerations in the Use of Regression, 41                                 |             |
| 2.8 Regression Through the Origin, 44                                                |             |

- 2.9 Estimation by Maximum Likelihood, 50
- 2.10 Case Where the Regressor  $x$  is Random, 52
  - 2.10.1  $x$  and  $y$  Jointly Distributed, 52
  - 2.10.2  $x$  and  $y$  Jointly Normally Distributed: The Correlation Model, 53
- Problems, 58

### 3. Multiple Linear Regression

67

- 3.1 Multiple Regression Models, 67
- 3.2 Estimation of the Model Parameters, 71
  - 3.2.1 Least-Squares Estimation of the Regression Coefficients, 71
  - 3.2.2 A Geometrical Interpretation of Least Squares, 81
  - 3.2.3 Properties of the Least-Squares Estimators, 82
  - 3.2.4 Estimation of  $\sigma^2$ , 82
  - 3.2.5 Inadequacy of Scatter Diagrams in Multiple Regression, 84
  - 3.2.6 Maximum-Likelihood Estimation, 85
- 3.3 Hypothesis Testing in Multiple Linear Regression, 87
  - 3.3.1 Test for Significance of Regression, 87
  - 3.3.2 Tests on Individual Regression Coefficients, 91
  - 3.3.3 Special Case of Orthogonal Columns in  $\mathbf{X}$ , 96
  - 3.3.4 Testing the General Linear Hypothesis, 98
- 3.4 Confidence Intervals in Multiple Regression, 101
  - 3.4.1 Confidence Intervals on the Regression Coefficients, 102
  - 3.4.2 Confidence Interval Estimation of the Mean Response, 103
  - 3.4.3 Simultaneous Confidence Intervals on Regression Coefficients, 104
- 3.5 Prediction of New Observations, 108
- 3.6 Hidden Extrapolation in Multiple Regression, 109
- 3.7 Standardized Regression Coefficients, 112
- 3.8 Multicollinearity, 117
- 3.9 Why Do Regression Coefficients have the Wrong Sign?, 120
- Problems, 122

|                                                                       |            |
|-----------------------------------------------------------------------|------------|
| <b>4. Model Adequacy Checking</b>                                     | <b>131</b> |
| 4.1 Introduction, 131                                                 |            |
| 4.2 Residual Analysis, 132                                            |            |
| 4.2.1 Definition of Residuals, 132                                    |            |
| 4.2.2 Methods for Scaling Residuals, 132                              |            |
| 4.2.3 Residual Plots, 138                                             |            |
| 4.2.4 Partial Regression and Partial Residual Plots, 146              |            |
| 4.2.5 Other Residual Plotting and Analysis Methods, 150               |            |
| 4.3 The PRESS Statistic, 152                                          |            |
| 4.4 Detection and Treatment of Outliers, 154                          |            |
| 4.5 Lack of Fit of the Regression Model, 158                          |            |
| 4.5.1 A Formal Test for Lack of Fit, 158                              |            |
| 4.5.2 Estimation of Pure Error from<br>Near-Neighbors, 162            |            |
| Problems, 166                                                         |            |
| <b>5. Transformations and Weighting to Correct Model Inadequacies</b> | <b>173</b> |
| 5.1 Introduction, 173                                                 |            |
| 5.2 Variance-Stabilizing Transformations, 174                         |            |
| 5.3 Transformations to Linearize the Model, 178                       |            |
| 5.4 Analytical Methods for Selecting a Transformation, 186            |            |
| 5.4.1 Transformations on $y$ : The Box–Cox Method, 186                |            |
| 5.4.2 Transformations on the Regressor Variables, 189                 |            |
| 5.5 Generalized and Weighted Least Squares, 193                       |            |
| 5.5.1 Generalized Least Squares, 193                                  |            |
| 5.5.2 Weighted Least Squares, 195                                     |            |
| 5.5.3 Some Practical Issues, 196                                      |            |
| Problems, 200                                                         |            |
| <b>6. Diagnostics for Leverage and Influence</b>                      | <b>207</b> |
| 6.1 Importance of Detecting Influential Observations, 207             |            |
| 6.2 Leverage, 209                                                     |            |
| 6.3 Measures of Influence: Cook's $D$ , 210                           |            |
| 6.4 Measures of Influence: $DFFITS$ and $DFBETAS$ , 213               |            |
| 6.5 A Measure of Model Performance, 216                               |            |
| 6.6 Detecting Groups of Influential Observations, 217                 |            |

|           |                                                                          |            |
|-----------|--------------------------------------------------------------------------|------------|
| 6.7       | Treatment of Influential Observations, 218                               |            |
|           | Problems, 219                                                            |            |
| <b>7.</b> | <b>Polynomial Regression Models</b>                                      | <b>221</b> |
| 7.1       | Introduction, 221                                                        |            |
| 7.2       | Polynomial Models in One Variable, 221                                   |            |
| 7.2.1     | Basic Principles, 221                                                    |            |
| 7.2.2     | Piecewise Polynomial Fitting (Splines), 228                              |            |
| 7.2.3     | Polynomial and Trigonometric Terms, 236                                  |            |
| 7.3       | Nonparametric Regression, 237                                            |            |
| 7.3.1     | Kernel Regression, 238                                                   |            |
| 7.3.2     | Locally Weighted Regression (Loess), 239                                 |            |
| 7.3.3     | Final Cautions, 243                                                      |            |
| 7.4       | Polynomial Models in Two or More Variables, 244                          |            |
| 7.5       | Orthogonal Polynomials, 253                                              |            |
|           | Problems, 258                                                            |            |
| <b>8.</b> | <b>Indicator Variables</b>                                               | <b>265</b> |
| 8.1       | The General Concept of Indicator Variables, 265                          |            |
| 8.2       | Comments on the Use of Indicator Variables, 279                          |            |
| 8.2.1     | Indicator Variables versus Regression on<br>Allocated Codes, 279         |            |
| 8.2.2     | Indicator Variables as a Substitute for a<br>Quantitative Regressor, 280 |            |
| 8.3       | Regression Approach to Analysis of Variance, 281                         |            |
|           | Problems, 287                                                            |            |
| <b>9.</b> | <b>Variable Selection and Model Building</b>                             | <b>291</b> |
| 9.1       | Introduction, 291                                                        |            |
| 9.1.1     | The Model-Building Problem, 291                                          |            |
| 9.1.2     | Consequences of Model Misspecification, 292                              |            |
| 9.1.3     | Criteria for Evaluating Subset Regression<br>Models, 296                 |            |
| 9.2       | Computational Techniques for Variable Selection, 302                     |            |
| 9.2.1     | All Possible Regressions, 302                                            |            |
| 9.2.2     | Stepwise Regression Methods, 310                                         |            |
| 9.3       | Some Final Recommendations for Practice, 317                             |            |
|           | Problems, 318                                                            |            |

|                                                          |            |
|----------------------------------------------------------|------------|
| <b>10. Multicollinearity</b>                             | <b>325</b> |
| 10.1 Introduction,                                       | 325        |
| 10.2 Sources of Multicollinearity,                       | 325        |
| 10.3 Effects of Multicollinearity,                       | 328        |
| 10.4 Multicollinearity Diagnostics,                      | 334        |
| 10.4.1 Examination of the Correlation Matrix,            | 334        |
| 10.4.2 Variance Inflation Factors,                       | 337        |
| 10.4.3 Eigensystem Analysis of $\mathbf{X}'\mathbf{X}$ , | 339        |
| 10.4.4 Other Diagnostics,                                | 343        |
| 10.5 Methods for Dealing with Multicollinearity,         | 345        |
| 10.5.1 Collecting Additional Data,                       | 345        |
| 10.5.2 Model Respecification,                            | 346        |
| 10.5.3 Ridge Regression,                                 | 348        |
| 10.5.4 Other Methods,                                    | 363        |
| 10.5.5 Comparison and Evaluation of Biased Estimators,   | 375        |
| Problems,                                                | 378        |
| <b>11. Robust Regression</b>                             | <b>382</b> |
| 11.1 The Need for Robust Regression,                     | 382        |
| 11.2 <i>M</i> -Estimators,                               | 386        |
| 11.3 Properties of Robust Estimators,                    | 400        |
| 11.3.1 Breakdown Point,                                  | 400        |
| 11.3.2 Efficiency,                                       | 401        |
| 11.4 Survey of Other Robust Regression Estimators,       | 401        |
| 11.4.1 High-Breakdown-Point Estimators,                  | 401        |
| 11.4.2 Bounded Influence Estimators,                     | 406        |
| 11.4.3 Other Procedures,                                 | 407        |
| 11.4.4 Computing Robust Regression Estimators,           | 409        |
| Problems,                                                | 410        |
| <b>12. Introduction to Nonlinear Regression</b>          | <b>414</b> |
| 12.1 Linear and Nonlinear Regression Models,             | 414        |
| 12.1.1 Linear Regression Models,                         | 414        |
| 12.1.2 Nonlinear Regression Models,                      | 415        |
| 12.2 Nonlinear Least Squares,                            | 416        |
| 12.3 Transformation to a Linear Model,                   | 420        |

|            |                                                                      |            |
|------------|----------------------------------------------------------------------|------------|
| 12.4       | Parameter Estimation in a Nonlinear System, 423                      |            |
| 12.4.1     | Linearization, 423                                                   |            |
| 12.4.2     | Other Parameter Estimation Methods, 431                              |            |
| 12.4.3     | Starting Values, 432                                                 |            |
| 12.4.4     | Computer Programs, 433                                               |            |
| 12.5       | Statistical Inference in Nonlinear Regression, 434                   |            |
| 12.6       | Examples of Nonlinear Regression Models, 437                         |            |
|            | Problems, 438                                                        |            |
| <b>13.</b> | <b>Generalized Linear Models</b>                                     | <b>443</b> |
| 13.1       | Introduction, 443                                                    |            |
| 13.2       | Logistic Regression Models, 444                                      |            |
| 13.2.1     | Models with a Binary Response Variable, 444                          |            |
| 13.2.2     | Estimating the Parameters in a Logistic Regression Model, 447        |            |
| 13.2.3     | Interpretation of the Parameters in a Logistic Regression Model, 450 |            |
| 13.2.4     | Hypothesis Tests on Model Parameters, 453                            |            |
| 13.3       | Poisson Regression, 459                                              |            |
| 13.4       | The Generalized Linear Model, 466                                    |            |
| 13.4.1     | Link Functions and Linear Predictors, 467                            |            |
| 13.4.2     | Parameter Estimation and Inference in the GLM, 468                   |            |
| 13.4.3     | Prediction and Estimation with the GLM, 472                          |            |
| 13.4.4     | Residual Analysis in the GLM, 474                                    |            |
| 13.4.5     | Overdispersion, 475                                                  |            |
|            | Problems, 477                                                        |            |
| <b>14.</b> | <b>Other Topics in the Use of Regression Analysis</b>                | <b>488</b> |
| 14.1       | Regression Models with Autocorrelation Errors, 488                   |            |
| 14.1.1     | Source and Effects of Autocorrelation, 488                           |            |
| 14.1.2     | Detecting the Presence of Autocorrelation, 489                       |            |
| 14.1.3     | Parameter Estimation Methods, 494                                    |            |
| 14.2       | Effect of Measurement Errors in the Regressors, 500                  |            |
| 14.2.1     | Simple Linear Regression, 501                                        |            |
| 14.2.2     | The Berkson Model, 502                                               |            |
| 14.3       | Inverse Estimation—The Calibration Problem, 503                      |            |
| 14.4       | Bootstrapping in Regression, 508                                     |            |
| 14.4.1     | Bootstrap Sampling in Regression, 509                                |            |

|                                                                                 |            |
|---------------------------------------------------------------------------------|------------|
| 14.4.2 Bootstrap Confidence Intervals,                                          | 510        |
| 14.5 Classification and Regression Trees (CART),                                | 516        |
| 14.6 Neural Networks,                                                           | 518        |
| 14.7 Designed Experiments for Regression,                                       | 521        |
| Problems,                                                                       | 524        |
| <b>15. Validation of Regression Models</b>                                      | <b>529</b> |
| 15.1 Introduction,                                                              | 529        |
| 15.2 Validation Techniques,                                                     | 530        |
| 15.2.1 Analysis of Model Coefficients and Predicted Values,                     | 530        |
| 15.2.2 Collecting Fresh Data—Confirmation Runs,                                 | 532        |
| 15.2.3 Data Splitting,                                                          | 534        |
| 15.3 Data from Planned Experiments,                                             | 545        |
| Problems,                                                                       | 545        |
| <b>APPENDIX A. Statistical Tables</b>                                           | <b>549</b> |
| <b>APPENDIX B. Data Sets For Exercises</b>                                      | <b>567</b> |
| <b>APPENDIX C. Supplemental Technical Material</b>                              | <b>582</b> |
| C.1 Background on Basic Test Statistics,                                        | 582        |
| C.2 Background from the Theory of Linear Models,                                | 585        |
| C.3 Important Results on $SS_R$ and $SS_{Res}$ ,                                | 588        |
| C.4 The Gauss–Markov Theorem, $\text{Var}(\varepsilon) = \sigma^2 \mathbf{I}$ , | 594        |
| C.5 Computational Aspects of Multiple Regression,                               | 595        |
| C.6 A Result on the Inverse of a Matrix,                                        | 597        |
| C.7 Development of the PRESS Statistic,                                         | 598        |
| C.8 Development of $S^2_{(i)}$ ,                                                | 600        |
| C.9 An Outlier Test Based on $R$ -Student,                                      | 601        |
| C.10 The Gauss–Markov Theorem, $\text{Var}(\varepsilon) = \mathbf{V}$ ,         | 604        |
| C.11 The Bias in $MS_{Res}$ When the Model is Underspecified,                   | 606        |
| C.12 Computation of Influence Diagnostics,                                      | 608        |
| C.13 Generalized Linear Models,                                                 | 610        |
| <b>References</b>                                                               | <b>621</b> |
| <b>Index</b>                                                                    | <b>637</b> |