

Part of the core of statistics, linear models are used to make predictions and explain the relationship between the response and the predictors. Understanding linear models is crucial to a broader competence in the practice of statistics. **Linear Models with R, Second Edition** explains how to use linear models in physical science, engineering, social science, and business applications. The book incorporates several improvements that reflect how the world of R has greatly expanded since the publication of the first edition.

### New to the Second Edition

- Reorganized material on interpreting linear models, which distinguishes the main applications of prediction and explanation and introduces elementary notions of causality
- Additional topics, including QR decomposition, splines, additive models, Lasso, multiple imputation, and false discovery rates
- Extensive use of the ggplot2 graphics package in addition to base graphics

Like its widely praised, best-selling predecessor, this edition combines statistics and R to seamlessly give a coherent exposition of the practice of linear modeling. The text offers up-to-date insight on essential data analysis topics, from estimation, inference, and prediction to missing data, factorial models, and block designs. Numerous examples illustrate how to apply the different methods using R.

### Features

- Demonstrates the flexibility of linear models in many examples
- Assumes basic knowledge of R and statistics
- Emphasizes intuition over rigorous proofs
- Presents exercises at the end of each chapter
- Includes datasets and R commands



**CRC Press**

Taylor & Francis Group  
an informa business  
[www.crcpress.com](http://www.crcpress.com)

6000 Broken Sound Parkway, NW  
Suite 300, Boca Raton, FL 33487  
711 Third Avenue  
New York, NY 10017  
2 Park Square, Milton Park  
Abingdon, Oxon OX14 4RN, UK

K14039

ISBN: 978-1-4398-8733-2



9 781439 887332

# Contents

---

## Preface

|                                           |           |
|-------------------------------------------|-----------|
|                                           | <b>xi</b> |
| <b>1 Introduction</b>                     | <b>1</b>  |
| 1.1 Before You Start                      | 1         |
| 1.2 Initial Data Analysis                 | 2         |
| 1.3 When to Use Linear Modeling           | 7         |
| 1.4 History                               | 8         |
| <b>2 Estimation</b>                       | <b>13</b> |
| 2.1 Linear Model                          | 13        |
| 2.2 Matrix Representation                 | 14        |
| 2.3 Estimating $\beta$                    | 15        |
| 2.4 Least Squares Estimation              | 16        |
| 2.5 Examples of Calculating $\hat{\beta}$ | 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        |
| <b>3 Inference</b>                        | <b>33</b> |
| 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 $\beta$      | 43        |
| 3.6 Bootstrap Confidence Intervals        | 46        |
| <b>4 Prediction</b>                       | <b>51</b> |
| 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        |

|          |                                     |            |
|----------|-------------------------------------|------------|
| <b>5</b> | <b>Explanation</b>                  | <b>59</b>  |
| 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         |
| <b>6</b> | <b>Diagnostics</b>                  | <b>73</b>  |
| 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         |
| <b>7</b> | <b>Problems with the Predictors</b> | <b>99</b>  |
| 7.1      | Errors in the Predictors            | 99         |
| 7.2      | Changes of Scale                    | 103        |
| 7.3      | Collinearity                        | 106        |
| <b>8</b> | <b>Problems with the Error</b>      | <b>113</b> |
| 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        |
| <b>9</b> | <b>Transformation</b>               | <b>133</b> |
| 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        |

|                                                    |            |
|----------------------------------------------------|------------|
| <b>10 Model Selection</b>                          | <b>149</b> |
| 10.1 Hierarchical Models                           | 150        |
| 10.2 Testing-Based Procedures                      | 151        |
| 10.3 Criterion-Based Procedures                    | 153        |
| 10.4 Summary                                       | 159        |
| <b>11 Shrinkage Methods</b>                        | <b>161</b> |
| 11.1 Principal Components                          | 161        |
| 11.2 Partial Least Squares                         | 172        |
| 11.3 Ridge Regression                              | 174        |
| 11.4 Lasso                                         | 177        |
| <b>12 Insurance Redlining — A Complete Example</b> | <b>183</b> |
| 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        |
| <b>13 Missing Data</b>                             | <b>197</b> |
| 13.1 Types of Missing Data                         | 197        |
| 13.2 Deletion                                      | 198        |
| 13.3 Single Imputation                             | 200        |
| 13.4 Multiple Imputation                           | 202        |
| <b>14 Categorical Predictors</b>                   | <b>205</b> |
| 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        |
| <b>15 One Factor Models</b>                        | <b>223</b> |
| 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        |
| <b>16 Models with Several Factors</b>              | <b>235</b> |
| 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        |

|                                       |            |
|---------------------------------------|------------|
| <b>17 Experiments with Blocks</b>     | <b>251</b> |
| 17.1 Randomized Block Design          | 252        |
| 17.2 Latin Squares                    | 256        |
| 17.3 Balanced Incomplete Block Design | 259        |
| <b>A About R</b>                      | <b>265</b> |
| <b>Bibliography</b>                   | <b>267</b> |
| <b>Bibliography</b>                   | <b>267</b> |
| <b>Index</b>                          | <b>271</b> |