Methods and Applications of Linear Models
Regression and the Analysis of Variance
Second Edition
This Page Intentionally Left Blank
Methods and Applications of Linear Models
Regression and the Analysis of Variance

Second Edition

Ronald R. Hocking
Texas A&M University
This edition is dedicated to the choir angel.
## Contents

Preface to the Second Edition ........................................... xvii
Preface to the First Edition ........................................... xix

### Part I Regression Models ........................................... 1

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Introduction to Linear Models</td>
<td>3</td>
</tr>
<tr>
<td>1.1</td>
<td>Background Information</td>
<td>3</td>
</tr>
<tr>
<td>1.2</td>
<td>Mathematical and Statistical Models</td>
<td>5</td>
</tr>
<tr>
<td>1.3</td>
<td>Definition of the Linear Model</td>
<td>7</td>
</tr>
<tr>
<td>1.4</td>
<td>Examples of Regression Models</td>
<td>12</td>
</tr>
<tr>
<td>1.4.1</td>
<td>Single-Variable, Regression Model</td>
<td>12</td>
</tr>
<tr>
<td>1.4.2</td>
<td>Regression Models with Several Inputs</td>
<td>16</td>
</tr>
<tr>
<td>1.4.3</td>
<td>Discrete Response Variables</td>
<td>19</td>
</tr>
<tr>
<td>1.4.4</td>
<td>Multivariate Linear Models</td>
<td>19</td>
</tr>
<tr>
<td>1.5</td>
<td>Concluding Comments</td>
<td>20</td>
</tr>
</tbody>
</table>

### Exercises

<table>
<thead>
<tr>
<th>Exercises</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>20</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Regression on Functions of One Variable</td>
<td>23</td>
</tr>
<tr>
<td>2.1</td>
<td>Simple Linear Regression Model</td>
<td>23</td>
</tr>
<tr>
<td>2.2</td>
<td>Parameter Estimation</td>
<td>27</td>
</tr>
<tr>
<td>2.2.1</td>
<td>Least Squares Estimation</td>
<td>27</td>
</tr>
<tr>
<td>2.2.2</td>
<td>Maximum Likelihood Estimation</td>
<td>31</td>
</tr>
<tr>
<td>2.2.3</td>
<td>Coded Data: Centering and Scaling</td>
<td>32</td>
</tr>
<tr>
<td>2.3</td>
<td>Properties of the Estimators</td>
<td>33</td>
</tr>
<tr>
<td>2.4</td>
<td>Analysis of the Simple Linear Regression Model</td>
<td>36</td>
</tr>
<tr>
<td>2.4.1</td>
<td>Parameter Estimation</td>
<td>36</td>
</tr>
<tr>
<td>2.4.2</td>
<td>Inference on the Parameters of the Model</td>
<td>38</td>
</tr>
<tr>
<td>2.4.3</td>
<td>Prediction and Prediction Intervals</td>
<td>46</td>
</tr>
</tbody>
</table>
2.5 Examining the Data and the Model
   2.5.1 Residuals 50
   2.5.2 Outliers, Extreme Points, and Influence 53
   2.5.3 Normality, Independence and Variance Homogeneity 59
2.6 Test for Lack of Fit 63
2.7 Polynomial Regression Models 66
   2.7.1 Quadratic Model 67
   2.7.2 Higher-Ordered Polynomial Models 72
   2.7.3 Orthogonal Polynomials 73
   2.7.4 Regression Through the Origin 74
   Exercises 75

3 Transforming the Data 83
   3.1 Need for Transformations 83
   3.2 Weighted Least Squares 84
   3.3 Variance Stabilizing Transformations 87
3.4 Transformations to Achieve a Linear Model 88
   3.4.1 Transforming the Dependent Variable 88
   3.4.2 Transforming the Predictors 94
3.5 Analysis of the Transformed Model 95
3.6 Transformations with Forbes Data 95
   Exercises 98

4 Regression on Functions of Several Variables 103
   4.1 Multiple Linear Regression Model 103
   4.2 Preliminary Data Analysis 104
4.3 Analysis of the Multiple Linear Regression Model 108
   4.3.1 Fitting the Model in Centered Form 109
   4.3.2 Estimation and Analysis of the Original Data 110
   4.3.3 Model Assessment and Residual Analysis 112
   4.3.4 Prediction 113
   4.3.5 Transforming the Response 115
4.4 Partial Correlation and Added-Variable Plots 119
   4.4.1 Partial Correlation 119
   4.4.2 Added-Variable Plots 121
   4.4.3 Simple Versus Partial Correlation 123
4.5 Variable Selection 126
  4.5.1 Orthogonal Predictors 126
  4.5.2 Criteria for Deletion of Variables 128
  4.5.3 Non-Orthogonal Predictors 130
  4.5.4 Computational Considerations 132
  4.5.5 Selection Strategies 134

4.6 Model Specification 138
Exercises 145

5 Collinearity in Multiple Linear Regression 151

5.1 Collinearity Problem 151
  5.1.1 Introduction 151
  5.1.2 Simple Example 152
  5.1.3 The Picket Fence 155
  5.1.4 Rotation of Coordinates 157

5.2 Example With Collinearity 159
  5.2.1 Preliminary Data Analysis 160
  5.2.2 Initial Regression Analysis 162

5.3 Collinearity Diagnostics 166
  5.3.1 Variance Inflation Factors 166
  5.3.2 Eigenvalues, Eigenvectors and Principal Component Plots 170

5.4 Remedial Solutions: Biased Estimators 176
  5.4.1 Variable Deletion 178
  5.4.2 Regression on Principal Components 180
  5.4.3 Ridge Regression 184
Exercises 188

6 Influential Observations in Multiple Linear Regression 193

6.1 Influential Data Problem 193

6.2 Hat Matrix 194
  6.2.1 Centered and Uncentered Hat Matrices 194
  6.2.2 Properties of the Hat Matrices 196

6.3 Effects of Deleting Observations 199

6.4 Numerical Measures of Influence 203
  6.4.1 Diagonal Elements of the Hat Matrix 204
6.4.2 Residuals 204
6.4.3 Mean Square Ratio 206
6.4.4 Cook's Distance 206
6.4.5 Other Indicators of Influential Data 207
6.5 Dilemma Data 208
6.6 Plots for Identifying Unusual Cases 213
6.6.1 Projection Ellipse 213
6.6.2 Augmented Hat Matrix 216
6.6.3 Multiple Extremes: The Masking Problem 218
6.7 Robust/Resistant Methods in Regression Analysis 221
6.7.1 M-Estimation 222
6.7.2 Iterative, Reweighted Least Squares 223
6.7.3 Regression with Bounded Influence 224
Exercises 225

7 Polynomial Models and Qualitative Predictors 229
7.1 Polynomial Models 229
7.1.1 Quadratic Model with Two Predictors 229
7.1.2 Quadratic Surfaces 231
7.2 Analysis of Response Surfaces 234
7.2.1 Analysis with First-Order Models 234
7.2.2 Analysis with Second-Order Models 237
7.3 Models with Qualitative Predictors 239
7.3.1 Indicator Variables to Identify Groups of Data 239
7.3.2 Indicator Variables to Fit Segmented Polynomials 255
Exercises 263

8 Additional Topics 271
8.1 Non-Linear Regression Models 271
8.1.1 Some Linearizeable Functions 271
8.1.2 Modified Gauss-Newton Method 274
8.2 Non-Parametric Model-Fitting Methods 277
8.2.1 Locally Weighted-Average Predictors 277
8.2.2 Projection Pursuit Regression 280
8.3 Logistic Regression 282
8.4 Random Input Variables 285
8.5 Errors in the Inputs 288
8.6 Calibration  
Exercises  

<table>
<thead>
<tr>
<th>Part II</th>
<th>Analysis of Variance Models</th>
<th>293</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>9</th>
<th>Introduction to Analysis of Variance Models</th>
<th>295</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.1</td>
<td>Background Information</td>
<td>295</td>
</tr>
<tr>
<td>9.2</td>
<td>Cell Means Model</td>
<td>296</td>
</tr>
<tr>
<td>9.3</td>
<td>Fixed Effects Models</td>
<td>299</td>
</tr>
<tr>
<td>9.3.1</td>
<td>One-Way Classification Model</td>
<td>299</td>
</tr>
<tr>
<td>9.3.2</td>
<td>Two-Way Classification Model</td>
<td>300</td>
</tr>
<tr>
<td>9.3.3</td>
<td>Constrained Cell Means Model</td>
<td>303</td>
</tr>
<tr>
<td>9.4</td>
<td>Mixed Effects Models</td>
<td>304</td>
</tr>
<tr>
<td>9.5</td>
<td>Concluding Comments</td>
<td>306</td>
</tr>
<tr>
<td></td>
<td>Exercises</td>
<td>306</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>10</th>
<th>Fixed Effects Models I: One-Way Classification of Means</th>
<th>309</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.1</td>
<td>Introduction</td>
<td>309</td>
</tr>
<tr>
<td>10.2</td>
<td>One-Way Classification: Balanced Data</td>
<td>310</td>
</tr>
<tr>
<td>10.2.1</td>
<td>Parameter Estimation</td>
<td>311</td>
</tr>
<tr>
<td>10.2.2</td>
<td>Hypothesis of Equal Means</td>
<td>312</td>
</tr>
<tr>
<td>10.2.3</td>
<td>Simultaneous Inferences About the Population Means</td>
<td>316</td>
</tr>
<tr>
<td>10.2.4</td>
<td>Simultaneous Acceptance and Confidence Ellipses</td>
<td>321</td>
</tr>
<tr>
<td>10.2.5</td>
<td>Orthogonal Contrasts</td>
<td>325</td>
</tr>
<tr>
<td>10.2.6</td>
<td>Reparameterizations of the One-Way Model</td>
<td>328</td>
</tr>
<tr>
<td>10.3</td>
<td>One-Way Classification: Unbalanced Data</td>
<td>331</td>
</tr>
<tr>
<td>10.4</td>
<td>Analysis of Covariance</td>
<td>334</td>
</tr>
<tr>
<td></td>
<td>Exercises</td>
<td>336</td>
</tr>
</tbody>
</table>
11 Fixed Effects Models II: Two-Way Classification of Means

11.1 Unconstrained Model: Balanced Data 343
   11.1.1 Parameter Estimation 345
   11.1.2 Tests of Hypotheses 346
   11.1.3 Simultaneous Inference 353
   11.1.4 Reparameterizations of the Two-Factor Model 355
   11.1.5 Test for Interaction with One Observation per Cell 360

11.2 Unconstrained Model: Unbalanced Data 363
   11.2.1 Discussion in Terms of the Cell Means Model 363
   11.2.2 Reparameterized Model 366

11.3 No-Interaction Model: Balanced Data 375
   11.3.1 Parameter Estimation 375
   11.3.2 Tests of Hypotheses 376
   11.3.3 Simultaneous Inference 378
   11.3.4 Reparameterization of the No-Interaction Model 379

11.4 No-Interaction Model: Unbalanced Data 380
   11.4.1 Missing Cells: Estimation 381
   11.4.2 Missing Cells: Testing Hypotheses 385
   11.4.3 Connected Designs 388

11.5 Non-Homogeneous Experimental Units: The Concept of Blocking 389
   11.5.1 Model for the Randomized, Complete Block Design 390
   11.5.2 Inferences on Parameters 391

Exercises 394

12 Fixed Effects Models III: Multiple Crossed and Nested Factors 401

12.1 Three-Factor Cross-Classified Model 401
   12.1.1 Tests of Hypotheses 402
   12.1.2 Reparameterized Model 404
   12.1.3 Estimability and Testability with Missing Cells 406

12.2 General Structure for Balanced, Factorial Models 407

12.3 Two-Fold Nested Model 412
   12.3.1 Analysis with Balanced Data 413
   12.3.2 Analysis with Unbalanced Data 417

12.4 General Structure for Balanced, Nested Models 420
## 12.5 Three-Factor, Nested-Factorial Model
12.5.1 Analysis with Balanced Data 422
12.5.2 Analysis with Unbalanced Data 425

## 12.6 General Structure for Balanced, Nested-Factorial Models
Exercises 431

## 13 Mixed Models I: The AOV Method with Balanced Data

### 13.1 Introduction 437

### 13.2 One-Way Classification, Random Model 438

### 13.3 Two-Way Classification, Mixed Model 443

### 13.4 Three-Factor, Nested-Factorial Model 448

### 13.5 General Analysis for Balanced, Mixed Models 452
13.5.1 Description of the Model 452
13.5.2 Parameter Estimation 453
13.5.3 Properties of the Estimators and Inferential Results 460

### 13.6 Additional Examples 466
13.6.1 Two-Fold Nested, Random Model 466
13.6.2 Randomized Block Design 468
13.6.3 Mixed Model for Split-Plot Designs 471
13.6.4 Repeated Measures Designs 475
13.6.5 Longitudinal Studies 477

### 13.7 Alternative Development of Mixed Models 477
13.7.1 Graybill Mixed Model 477
13.7.2 Scheffe Mixed Model 480
13.7.3 Randomization Theory 481
Exercises 483

## 14 Mixed Models II: The AVE Method with Balanced Data

### 14.1 Introduction 489

### 14.2 Two-Way Cross-Classification Model 490

### 14.3 Two-Fold Nested Model 504

### 14.4 Three Factor, Nested-Factorial Model 505

### 14.5 General Description of the AVE Table 511
14.5.1 AVE Table for Factorial Models 512
14.5.2 AVE Table for Nested Models 517
14.5.3 AVE Table for Nested-Factorial Models 518
14.5.4 AVE Method for General Mixed Effects Models 519

14.6 Additional Examples 520
14.7 Computational Procedure for the AVE Method 530
14.8 Properties of the AVE Estimates 531
  14.8.1 Diagnostic Analysis for the Two-Way Classification Model 532
  14.8.2 Confidence Intervals 533
Exercises 533

15 Mixed Models III: Unbalanced Data 539
15.1 Introduction 539
15.2 Parameter Estimation: Likelihood Methods 541
  15.2.1 Maximum Likelihood Estimation 541
  15.2.2 Restricted Maximum Likelihood Estimation 544
  15.2.3 Minimum Norm Quadratic Unbiased Estimators 547
  15.2.4 Numerical Illustration of the Methods 548
15.3 ML and REML Estimates with Balanced Data 550
  15.3.1 ML Estimation with Balanced Data 550
  15.3.2 REML Estimates with Balanced Data 554
15.4 EM Algorithm for REML Estimation 554
  15.4.1 Review of the EM Algorithm 555
  15.4.2 EM Algorithm Applied to REML Estimation 555
  15.4.3 Estimation of Fixed Effects 561
  15.4.4 Inferences on Variance Components and Fixed Effects 562
  15.4.5 Numerical Examples to Illustrate the EM-AOV Algorithm 571
15.5 EM Algorithm Applied to the AVE Method 568
  15.5.1 Description of the Method 568
  15.5.2 Numerical Examples of the EM-AVE Method 571
15.6 Models with Covariates 578
  15.6.1 Development of the Analysis 579
  15.6.2 Numerical Example 580
15.7 Summary 582
Exercises 582
Part III  Mathematical Theory of Linear Models  589

16  Distribution of Linear and Quadratic Forms  591
16.1  Linear and Quadratic Forms  591
16.2  Multivariate Normal Distribution  594
16.3  Distribution of Quadratic Forms  596
  16.3.1  Non-Central Chi-Squared Distributions  596
  16.3.2  Fundamental Theorems  600
16.4  Other Non-Central Distributions  608
  16.4.1  Non-Central t-Distribution  608
  16.4.2  Non-Central F-Distribution  609
  16.4.3  Extensions  610
Exercises  611

17  Estimation and Inference for Linear Models  615
17.1  Estimation of Parameters  615
  17.1.1  Estimation for the Unconstrained Model  616
  17.1.2  Estimation for the Constrained Model  620
  17.1.3  Best Linear Unbiased Estimation (BLUE) of $\beta$  626
  17.1.4  Partitioned Form of the Model  627
  17.1.5  Reparameterized Models  629
17.2  Tests of Linear Hypotheses on $\beta$  631
  17.2.1  Unconstrained Model  632
  17.2.2  Constrained Model  641
17.3  Confidence Regions and Intervals  643
  17.3.1  Confidence Regions  644
  17.3.2  Confidence Intervals  644
Exercises  646

18  Simultaneous Inference: Tests and Confidence Intervals  653
18.1  Simultaneous Tests  653
  18.1.1  Simultaneous Tests: General Methods  654
xvi Contents

18.1.2 Simultaneous Tests: Cell Means Models 666
18.2 Simultaneous Confidence Intervals 674
  18.2.1 Bonferroni Confidence Intervals 674
  18.2.2 Scheffé Confidence Intervals 674
  18.2.3 Tukey Studentized-Range Intervals 675
Exercises 676

Appendix A: Mathematics 679
  A.I Matrix Algebra 679
  A.II Optimization 694

Appendix B: Statistics 699
  B.I Estimation 699
  B.II Tests of Hypotheses and Confidence Regions 701

Appendix C: Statistical Tables 703

Appendix D: Data Tables 709

References 727
Index 737
Preface to the Second Edition

This edition was prepared in response to users of the first edition. The essential change is that the development of the statistical distribution and inference theory of Part I of the first edition has been moved to the end of the book as Part III in this edition. While this material forms the basis for inference in linear models, it requires fairly intensive study. The result has been that many users of the book spent too much time on that material in a course that was intended to study the methods in linear regression or the analysis of variance. This left too little time for the discussion of the methodology that was the primary intent of the book.

In this edition, Part I contains the material on regression and Part II the material on the analysis of variance. The necessary theoretical concepts are presented as needed to describe the inference that is appropriate for the method being discussed. This is done using intuitive ideas and without proof. References are given to the formal development in Part III. It has been our experience that the theoretical material is best appreciated after having been motivated by the methods in Parts I and II.

In this addition, 63 exercises have been added to the material on regression and 46 exercises have been added to the material on the analysis of variance. The exercises in each chapter have been organized by section number so as to facilitate relating them to the reading material. The AVE method for variance component analysis has been improved by relating it to the AOV method. This simplifies the expressions for the expected values and hence the estimation of individual components. This allows for the development of a general computer program. As in the first edition, data for larger examples are given in Appendix D and these data, as well as those contained in examples and exercises in the text, are available on an ftp file. A solutions manual for the exercises is available from the author.

RONALD R. HOCKING

Ishpeming, Michigan
September 2002
This Page Intentionally Left Blank
Preface to the First Edition

This book presents a thorough treatment of the concepts and methods of linear model analysis and illustrates them with numerical and conceptual examples. A substantial portion of statistical analysis is done in the context of a linear model that describes the data. There is an extensive literature on the methods of analyzing data that is associated with such models. Frequently we encounter lack of agreement on the appropriate analysis, confusion over the application of the existing methods, and improper interpretation of the conclusions. The ready availability of high-speed computers and statistical software encourages the analysis of ever larger and more complex problems while at the same time increasing the likelihood of improper usage. At least part of the problem lies in the gap between the developers of the methodology and the users of those methods. Computer software may fill the gap, but it may not accurately portray the method or it may be improperly interpreted by the user. One of the aims of this book is to bridge that gap by presenting the methods in a conceptually simple way so that the user will more easily understand the application of the methods and be able to assess whether the computer software reflects the method.

The topic of linear models has two major sub-categories: linear regression and the analysis of variance. In linear regression, the response is typically assumed to be a function of quantitative variables such as time or temperature. In the analysis of variance, we are concerned with estimating and making inferences about the means of populations and, more generally, about the mean and covariance structure on these populations. Both models use much of the same mathematical theory to develop the analyses, and common features of the analysis have motivated the inclusion of both topics in this book. However, there are sufficient differences to warrant a separation of the discussion.

This book is divided into three parts. Part I contains an introduction to the formulation of linear models, a discussion of the statistical distribution theory for linear and quadratic forms, a presentation of the basic linear model theory, and a review of the concepts and methods of simultaneous inference. Part II addresses linear regression models. The discussion focuses on the development of a model that properly reflects the data and the experimental situation. The problems associated with collinear predictors are discussed and illustrated and remedial solutions are examined. The literature on influential data diagnostics is reviewed and examined. Graphical methods for examining the data and the model are stressed and illustrated. A careful treatment is given to the topic of indicator
variables. The ideas are developed in such a way as to avoid the frequent misuses of this useful concept. Part III is concerned with analysis of variance models. Three chapters are devoted to the discussion of fixed effects models. This methodology is concerned with the analysis of means of populations, and the cell means model is used as the basis for interpretation and hypothesis formulation. The cell means model, introduced by Hocking and Speed (1975) and expanded on by Hocking (1985) has now been commonly accepted in the literature. The discussions further illustrate the utility of this model formulation and describe the relation to the classical effects and interaction model. The model is especially powerful in the analysis of unbalanced data, including the problem of missing cells. The last three chapters of Part III are devoted to the mixed model problem, that is, models with a non-scalar covariance structure. Classical methods of estimation and inference are reviewed and a new approach for estimating variance components, called AVE estimation, is introduced. This method develops the estimates in a way that allows for a diagnostic analysis of the data as well as an assessment of the model and the design. The source of negative estimates of variance components is revealed. The method is especially useful if the objective of the experiment is to identify sources of variability in the process being studied. The method is initially applied to balanced data situations and is then extended to include the unbalanced case. The EM algorithm is used for the extension. In addition to the estimation and diagnostic analysis, the method suggests a general approach for making inferences on both the fixed and the random effects in the model.

This book has been developed over several years of teaching courses in linear regression and the analysis of variance. At Texas A&M University, the material was presented in two semesters to students who have had a minimum of a one-semester course in statistical methods. A first course in mathematical statistics, including the distribution of random variables and an introduction to likelihood methods for inference, is necessary for an appreciation of the results in Part I. Liberal use is made of matrix algebra, and the essential results are summarized in Appendix A. The use of matrix algebra greatly simplifies both the theory and application of the analysis. Most of the results used are contained in an introductory course on matrix algebra. The possible exception is the concept of Kronecker products that greatly simplifies the discussion of analysis of variance. I find it best to introduce such concepts as they arise in the discussion as opposed to using an early segment of the course for a discussion of this material.

The book can be used in several ways. Chapter 1 should be read for a general introduction to the topic and the basic notation and definitions. For an applied course in linear regression, Part II is essentially self-contained if the student is willing to accept the basic results on estimation and inference from least squares analysis. These concepts are motivated for the simple linear regression model in Chapter 5. The focus in these chapters is on the application of the methods to several examples. The graphical methods for examining
influential data and collinear predictors clarify the numerical indicators available in most software packages. Similarly Part III can be used for an applied course in the analysis of variance, although the presentation, using Kronecker products, is more mathematical. The reward for this effort is that we obtain general results for expressing the analysis of a general class of such models. Inference on fixed effects is made clearer by using the cell means model, and the AVE method clarifies the interpretation and estimation of variance components. The chapters in both of these parts contain many exercises. Some of these are conceptual to provide details on the material in the chapters. Many of the exercises lead the student through an analysis of specific examples. A solutions manual is available and professors should contact the author for more information. The data for the larger examples is given in Appendix D and will be available on an ftp file.

More advanced students may wish to study the remaining chapters in Part I in detail. Most of the statistical and mathematical results are presented in a concise manner in Chapters 2 and 3. The issue of simultaneous inference is presented in Chapter 4 and some of the results are illustrated graphically to reveal the source of confusion that arises when making inferences on several parameter functions in an analyses.

In addition to its structure as a text, this book is designed to serve as a reference for applied statisticians and researchers in linear model methodology. Many of the problems presented remain unsolved, and it is hoped that this book will provide a basis for further research.

RONALD R. HOCKING

Ishpeming, Michigan
February 1996
PART I

REGRESSION MODELS
This Page Intentionally Left Blank
The objective of this chapter is to present a discussion and a formal definition of a general class of linear models. The presentation throughout this book will separate the discussion of regression models and analysis of variance models, though they are mathematically equivalent and subject to much of the same analysis. In this chapter, the focus is on regression models and the definition is illustrated with a variety of examples. Matrix algebra is used to give a compact description of the models and will be used extensively in the chapters to follow. An elementary knowledge of matrix algebra is assumed but a review of the basic concepts and more advanced material to be used in this book is given in Appendix A. This chapter introduces much of the notation and terminology to be used throughout the book.

1.1 BACKGROUND INFORMATION

In the area of applied statistics, a substantial portion of the analyses comes under the heading of linear models. This general heading covers the major areas of regression analysis and the analysis of variance but includes other topics such as time series and multivariate analysis. The first two topics are the primary focus of this book.

The basis for the computational procedure used in the analysis can be traced back to the writings of the French mathematicians Gauss and LeGendre in the early years of the nineteenth century. In their writings they describe the method of least squares for determining a line or plane to give an approximate description for a scatter of points. The method of least squares was given statistical credibility, under the assumption of normally distributed errors, by the likelihood-based methodology pioneered by R. A. Fisher in the first quarter of the twentieth century. Books such as *Applied Regression Analysis* by Draper and Smith (1966), *The Analysis of Variance* by Scheffé (1959) and *An Introduction to Linear Statistical Models* by Graybill (1961) summarize the methodology that was developed in the first half of this century. These methods contributed significantly to developments in all areas of research and are still widely used.
In the early stages of the development of these methods, the applications were limited by the ability to perform the required calculations, specifically the inversion of large matrices. The developments in high-speed computing in the two decades following the Second World War allowed us to consider regression models with many variables and the analysis of variance with many factors. The dramatic developments in computer technology in recent years have prompted new research in linear model methods. Numerical and graphical procedures have greatly expanded our ability to extract information from data. In regression analysis, techniques for identifying and examining the role of unusual observations and for detecting and understanding problems with collinear predictors enable the analyst to obtain a better understanding of the system or process being studied.

In the analysis of variance, much of the confusion caused by the mathematical statement of the model was removed by the revival of what is now called the cell means model. (We write key words and phrases in boldface.) This form of the model was especially useful in resolving questions with regard to unbalanced designs. Increased computer speed and capacity encouraged the application of likelihood-based methods for the estimation of variance components in mixed models. The search for diagnostic methods in the analysis of mixed models led to an alternative approach to describing the model and to a new computational procedure known as AVE which simplifies the computations and provides insight into the sources of variability.

A fundamental problem in all areas of science, and especially in statistics, is the gap between the development of new ideas and their implementation in practice. In the current computing environment, the problem becomes even more serious as new ideas are introduced into readily available statistical packages but not necessarily understood by practitioners. The motivation for the presentation in this book is to bridge the gap between theory and practice. This is done in Parts I and II by providing an intuitive discussion of the theory and then giving a thorough discussion of the techniques necessary for applying the theory to the analysis of data. Parts I and II can be understood by the student with a first course in statistical methods who is willing to accept the fundamental concepts and focus on the applications. The techniques are illustrated by numerical examples from different disciplines with particular emphasis on the ways in which new methods can provide insight into the analysis. Part III contains the necessary mathematical theory for the reader interested in a complete understanding of the basis for the analyses. That part requires an understanding of statistical theory at the level of an introductory course in mathematical statistics. No attempt is made to evaluate existing computer packages since they are constantly being modified and updated. It is hoped that the discussions of methods will allow the reader to be a discriminating user of such packages.
1.2 MATHEMATICAL AND STATISTICAL MODELS

For our purposes, we describe a mathematical model as a functional relation between variables. In particular, we are interested in models that relate a set of input variables to a set of output variables. It is convenient to think of this situation in a generic form. Thus we think of a response as the output of a process that depends on one or more inputs. This idea is shown in the schematic below. The box indicates a process in which the three inputs are transformed into the single output. For much of this book we will consider a single output but allow for several inputs.

\[
\text{Inputs} \rightarrow \boxed{\text{Process}} \rightarrow \text{Output}
\]

Schematic for mathematical models.

Mathematically, we describe the relation as

\[ y = g(x_1, x_2, x_3), \]  

(1.1)

where \( y \) denotes the output, \( x_1, x_2, \) and \( x_3 \) denote the inputs and \( g(x_1, x_2, x_3) \) denotes the functional relation by which the inputs are converted into the output. We will refer to this as the response function. The concepts are illustrated by the following example.

**Example 1.1.** In the manufacture of particle boards, small wood particles are mixed with an adhesive, formed into sheets of a given thickness, and baked in an oven. The company is interested in the relation of the strength of the boards to the baking temperature. To examine this relation, several boards are produced at each of several temperatures and the strength, \( y \) and the temperature, \( t \), are measured. Based on this information, the analyst wishes to determine the functional relation.

The search for the functional relation in Example 1.1 may be aided by a scatter plot of the temperature-strength pairs. (This is convenient here since we have only one input variable.) The plot might suggest that the relation is nearly a straight line, but that there are departures from that linear relation. Further, we may note that observations at the same temperature do not yield the same strength. These differences could be caused by other factors that influence the strength and have not been held constant in the study. In many cases, these unexplained differences can only be attributed to natural variability in the material or the process that does not have a mathematical explanation. To include such variability we introduce the concept of a statistical model.
particular, we consider the extension of the mathematical model that merely adds a random variable to the input side of equation (1.1). Thus, we write the model as

\[ y = g(x_1, x_2, x_3) + e. \]  (1.2)

Here, \( e \) denotes the added random variable, often called the \textit{error term}. The properties of this random variable will depend on the situation, but it is often assumed to follow a univariate normal distribution with mean zero and variance \( \sigma^2 \). We shall elaborate on this definition in Section 1.3. Implicit in this assumption is the fact that the output, \( y \), can be viewed as a random variable with mean (expected value), \( g(x_1, x_2, x_3) \), and variance \( \sigma^2 \). Thus we may write the deterministic part of the model, using \( E[y] \) to denote the expected value, as

\[ E[y] = g(x_1, x_2, x_3). \]  (1.3)

In most situations, the functional form of the mathematical model may be known apart from the values of certain parameters. Thus, it may be known that the relation is a straight line, but the slope and intercept are not known. Letting \( \beta \) denote a vector of such parameters and \( x \) the vector of inputs, we write the mean function in (1.3) as

\[ E[y] = g(x, \beta). \]  (1.4)

If we add the assumption of independence to the random variables associated with the individual responses, the data for Example 1.1 can be viewed as a random sample from a population with mean given by (1.4) and variance \( \sigma^2 \). In this case, the population is only defined conceptually as the collection of possible boards that could be produced. In other cases, it may be possible to enumerate the population, such as the population of students in the college of engineering at a given university. In that case we might sample the population rather than observe all students or, alternatively, we might view this group of students as a sample from the collection of all possible engineering students at this university. The data could have been collected to develop a model for the relation between the score on a qualifying exam as a function of the student's grade point average.

The concept of an input variable is quite general. For example, when modeling the daily amount of water used by an oil refinery, the inputs may include the size of the refinery, the amount of crude oil processed, the number of cooling towers and the types of products. In general, inputs may be \textit{quantitative}, that is, measurements such as temperature or amount, or they may be \textit{qualitative}, indicating the presence or absence of a factor or the type of product.

The observations may arise as a result of a carefully \textit{designed experiment}. For example, if we wish to assess the effects of temperature on the strength of a product, we could conduct a controlled experiment in which the production