

Springer Texts in Statistics

Series Editors

G. Allen, Department of Statistics, Houston, TX, USA

R. De Veaux, Department of Mathematics and Statistics, Williams College,
Williamstown, MA, USA

R. Nugent, Department of Statistics, Carnegie Mellon University, Pittsburgh,
PA, USA

Springer Texts in Statistics (STS) includes advanced textbooks from 3rd- to 4th-year undergraduate courses to 1st- to 2nd-year graduate courses. Exercise sets should be included. The series editors are currently Genevera I. Allen, Richard D. De Veaux, and Rebecca Nugent. Stephen Fienberg, George Casella, and Ingram Olkin were editors of the series for many years.

More information about this series at <http://www.springer.com/series/417>

Ronald Christensen

Plane Answers to Complex Questions

The Theory of Linear Models

Fifth Edition

 Springer

Ronald Christensen
Department of Mathematics and Statistics
University of New Mexico
Albuquerque, NM, USA

ISSN 1431-875X
Springer Texts in Statistics
ISBN 978-3-030-32096-6
<https://doi.org/10.1007/978-3-030-32097-3>

ISSN 2197-4136 (electronic)

ISBN 978-3-030-32097-3 (eBook)

4th edition © Springer Science+Business Media, LLC 2011
3rd edition © Springer Science+Business Media New York, 2002
2nd edition © Springer Science+Business Media New York, 1996
1st edition © Springer Science+Business Media New York, 1987
5th edition © Springer Nature Switzerland AG 2020

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

To Dad, Mom, Sharon, Fletch, and Don

Prefaces

Critical assessment of data is the essential task of the educated mind.

Professor Garrett G. Fagan, Pennsylvania State University.

The last words in his audio course *The Emperors of Rome*, The Teaching Company.

Preface to the Fifth Edition

I prepared the fifth edition of *Plane Answers (PA-V)* in conjunction with a new edition of *Advanced Linear Modeling (ALM-III)* (Christensen, 2019). The emphasis in both revisions was to include more material on *Statistical Learning*. *ALM-III* has far more changes in it than *PA-V*. (*ALM-III* is about 50% longer than *ALM-II*.) In *ALM-III*, all but the first three chapters (and Chapter 13) are devoted to dependent data. I regretfully concluded that almost all of the mixed models chapter in *PA* needed to go into *ALM-III*. The one exception is that I moved the discussion of BLUP into Chapter 6 of *PA-V*.

The biggest changes in *PA-V* are listed below:

- Section 1.3 has been restructured to isolate the more difficult parts.
- Section 2.9 is a new section on biased estimation and the variance-bias tradeoff.
- Subsection 3.2.1 is a short new subsection that introduces the importance of small F statistics.
- A new Exercise 3.7b helps establish Fieller's method prior to its application in Exercises 6.9.1, 2, 3.
- Section 4.1 contains some cleaner notation for one-way ANOVA computations.
- Section 5.1 is a new section containing my overall view of common multiple comparison procedures.
- Subsubsection 6.3.3.2 discusses best predictors for loss functions other than squared error.
- Section 6.6 now contains the discussion of BLUP.
- The section on polynomial regression and one-way ANOVA now contains a table of polynomial contrasts.
- Subsection 7.5.3 contains new material on characterizing the interaction space in an unbalanced two-way ANOVA.

- Subsection 9.1.1 introduces ACOVA ideas for models with dependent or heteroscedastic data.
- I thought about just deleting Section 9.3 but opted for attempting to make it more relevant.
- Section 11.2 has some new results on checking whether models qualify as generalized split plot models.
- New Subsection 11.2.3 addresses the analysis of (generalized) split plot designs when there is missing data in the subplots.
- As mentioned, mixed models got moved to *ALM-III* because it fits naturally into *ALM-III*'s emphasis on dependent data.
- Subsection 12.4.1 includes additional discussion of testing for heteroscedasticity.
- Subsection 12.4.2 introduces the Huber–White sandwich estimator.
- I changed the order of the chapters on variable selection and collinearity from previous versions of *PA* in order to smooth the presentation with *ALM-III*.
- The collinearity chapter contains a new Section 13.2 on variance inflation factors.
- The singular value decomposition of a matrix X , Theorem 13.3.1, has been generalized and the relationship between ridge regression and principal component regression explicated.
- Section 13.6 is a new section on different approaches to estimation. While this stands on its own, it also motivates material in *ALM-III*.
- Section 14.1 now discusses information criteria for model selection as well as cost complexity pruning.
- Section 14.2 has examples illustrating issues with larger data sets.
- Section 14.3 contains more discussion of variable selection.
- Section 14.4 introduces *boosting*, *bagging* and the random part of *random forests*. The application of these subjects is closely related to nonparametric regression as discussed in Chapter 1 of *ALM-III*.
- Appendix B has a number of refinements in the results. I also decided to rename “orthogonal matrices” as “orthonormal matrices” because it is a clearly better name.
- Appendix D has a new section on identifiability.

A big part of the effort in producing *PA-V* was just cleaning the text. After 30 years you would think, by now, I would be happy with it.

While *PA* is a book on Linear Model Theory, Christensen (2015) illustrates the use of most of the theory presented in this book. There are a number of related topics discussed on my website <http://www.stat.unm.edu/~fletcher/> in various places. These include computer code for the applications book as well as for *ALM-III*, cf. <http://www.stat.unm.edu/~fletcher/Rcode.pdf> and <http://www.stat.unm.edu/~fletcher/R-ALMIII.pdf>.

I have quite assiduously avoided doing asymptotic theory in *PA*, and that remains true in *PA-V*. There are many sources that discuss asymptotics for linear model theory. The appendix to Christensen and Lin (2015) uses a number of the most important results.

I would like to thank Fletcher Christensen and Joe Cavanaugh both of whom I have used as a sounding board for years on linear model issues. I thank Mohammad Hattab for numerous suggestions. Since the last edition of the book, Steve Fienberg and Ingram Olkin have both died. They were the subject matter editors for Springer when *PA* was first published in 1987. I sent the book to a lot of publishers and Steve was the only person who took seriously the efforts of an assistant professor from Montana State University. (Steve had been one of my professors at the University of Minnesota.) He recommended it to Ingram who both liked it a lot and gave me a large number of suggestions (virtually all of which remain in the book). I owe both of them a great debt!

Please note that while most of this book's examples, exercises, and figures draw from real data and are cited accordingly, a few of them are based on simulated data.

Some people think that *Plane Answers* is an example of the old maxim, "If all you have is a hammer, everything looks like a nail." I prefer to think that if you have a good enough hammer, almost everything actually is a nail.

Ronald Christensen
Albuquerque, New Mexico, 2018

Preface to the Fourth Edition

As with the prefaces to the second and third editions, this focuses on changes to the previous edition. The preface to the first edition discusses the core of the book.

Two substantial changes have occurred in Chapter 3. Subsection 3.3.2 uses a simplified method of finding the reduced model and includes some additional discussion of applications. In testing the generalized least squares models of Section 3.8, even though the data may not be independent or homoscedastic, there are conditions under which the standard F statistic (based on those assumptions) still has the standard F distribution under the reduced model. Section 3.8 contains a new subsection examining such conditions.

The major change in the fourth edition has been a more extensive discussion of best prediction and associated ideas of R^2 in Sections 6.3 and 6.4. It also includes a nice result that justifies traditional uses of residual plots. One portion of the new material is viewing best predictors (best linear predictors) as perpendicular projections of the dependent random variable y into the space of random variables that are (linear) functions of the predictor variables x . A new subsection on inner products and perpendicular projections for more general spaces facilitates the discussion. While these ideas were not new to me, their inclusion here was inspired by deLaubenfels (2006).

Section 9.1 has an improved discussion of least squares estimation in ACOVA models. A new Section 9.5 examines Milliken and Graybill's generalization of Tukey's one degree of freedom for nonadditivity test.

A new Section 10.5 considers estimable parameters that can be known with certainty when $C(X) \not\subset C(V)$ in a general Gauss–Markov model. It also contains a relatively simple way to estimate estimable parameters that are not known with certainty. The nastier parts in Sections 10.1–10.4 are those that provide sufficient generality to allow $C(X) \not\subset C(V)$. The approach of Section 10.5 seems more appealing.

In Sections 12.4 and 12.6, the point is now made that ML and REML methods can also be viewed as method of moments or estimating equations procedures.

The biggest change in Chapter 13 is a new title. The plots have been improved and extended. At the end of Section 13.6, some additional references are given on case deletions for correlated data as well as an efficient way of computing case deletion diagnostics for correlated data.

The old Chapter 14 has been divided into two chapters, the first on variable selection and the second on collinearity and alternatives to least squares estimation. Chapter 15 includes a new section on penalized estimation that discusses both ridge and lasso estimation and their relation to Bayesian inference. There is also a new section on orthogonal distance regression that finds a regression line by minimizing orthogonal distances, as opposed to least squares, which minimizes vertical distances.

Appendix D now contains a short proof of the claim: If the random vectors x and y are independent, then any vector-valued functions of them, say $g(x)$ and $h(y)$, are also independent.

Another significant change is that I wanted to focus on Fisherian inference, rather than the previous blend of Fisherian and Neyman–Pearson inference. In the interests of continuity and conformity, the differences are soft-pedaled in most of the book. They arise notably in new comments made after presenting the traditional (one-sided) F test in Section 3.2 and in a new Subsection 5.6.1 on multiple comparisons. The Fisherian viewpoint is expanded in Appendix F, which is where it primarily occurred in the previous edition. But the change is most obvious in Appendix E. In all previous editions, Appendix E existed just in case readers did not already know the material. While I still expect most readers to know the “how to” of Appendix E, I no longer expect most to be familiar with the “why” presented there.

Other minor changes are too numerous to mention and, of course, I have corrected all of the typographic errors that have come to my attention. Comments by Jarrett Barber led me to clean up Definition 2.1.1 on identifiability.

My thanks to Fletcher Christensen for general advice and for constructing Figures 10.1 and 10.2. (Little enough to do for putting a roof over his head all those years. :-)

Ronald Christensen
Albuquerque, New Mexico, 2010

Preface to the Third Edition

The third edition of *Plane Answers* includes fundamental changes in how some aspects of the theory are handled. Chapter 1 includes a new section that introduces generalized linear models. Primarily, this provides a definition so as to allow comments on how aspects of linear model theory extend to generalized linear models.

For years, I have been unhappy with the concept of estimability. Just because you cannot get a linear unbiased estimate of something does not mean you cannot estimate it. For example, it is obvious how to estimate the ratio of two contrasts in an ANOVA, just estimate each one and take their ratio. The real issue is that if the model matrix X is not of full rank, the parameters are not identifiable. Section 2.1 now introduces the concept of identifiability and treats estimability as a special case of identifiability. This change also resulted in some minor changes in Section 2.2.

In the second edition, Appendix F presented an alternative approach to dealing with linear parametric constraints. In this edition, I have used the new approach in Section 3.3. I think that both the new approach and the old approach have virtues, so I have left a fair amount of the old approach intact.

Chapter 8 contains a new section with a theoretical discussion of models for factorial treatment structures and the introduction of special models for homologous factors. This is closely related to the changes in Section 3.3.

In Chapter 9, reliance on the normal equations has been eliminated from the discussion of estimation in ACOVA models—something I should have done years ago! In the previous editions, Exercise 9.3 has indicated that Section 9.1 should be done with projection operators, not normal equations. I have finally changed it. (Now Exercise 9.3 is to redo Section 9.1 with normal equations.)

Appendix F now discusses the meaning of small F statistics. These can occur because of model lack of fit that exists in an unsuspected location. They can also occur when the mean structure of the model is fine but the covariance structure has been misspecified.

In addition, there are various smaller changes including the correction of typographical errors. Among these are very brief introductions to nonparametric regression and generalized additive models, as well as Bayesian justifications for the mixed model equations and classical ridge regression. I will let you discover the other changes for yourself.

Ronald Christensen
Albuquerque, New Mexico, 2001

Preface to the Second Edition

The second edition of *Plane Answers* has many additions and a couple of deletions. New material includes additional illustrative examples in Appendices A and B and Chapters 2 and 3, as well as discussions of Bayesian estimation, near replicate lack of fit tests, testing the independence assumption, testing variance components, the interblock analysis for balanced incomplete block designs, nonestimable constraints, analysis of unreplicated experiments using normal plots, tensors, and properties of Kronecker products and Vec operators. The book contains an improved discussion of the relation between ANOVA and regression, and an improved presentation of general Gauss–Markov models. The primary material that has been deleted are the discussions of weighted means and of log-linear models. The material on log-linear models was included in Christensen (1997), so it became redundant here. Generally, I have tried to clean up the presentation of ideas wherever it seemed obscure to me.

Much of the work on the second edition was done while on sabbatical at the University of Canterbury in Christchurch, New Zealand. I would particularly like to thank John Deely for arranging my sabbatical. Through their comments and criticisms, four people were particularly helpful in constructing this new edition. I would like to thank Wes Johnson, Snehalata Huzurbazar, Ron Butler, and Vance Berger.

Ronald Christensen
Albuquerque, New Mexico, 1996

Preface to the First Edition

This book was written to rigorously illustrate the practical application of the projective approach to linear models. To some, this may seem contradictory. I contend that it is possible to be both rigorous and illustrative, and that it is possible to use the projective approach in practical applications. Therefore, unlike many other books on linear models, the use of projections and subspaces does not stop after the general theory. They are used wherever I could figure out how to do it. Solving normal equations and using calculus (outside of maximum likelihood theory) are anathema to me. This is because I do not believe that they contribute to the understanding of linear models. I have similar feelings about the use of side conditions. Such topics are mentioned when appropriate and thenceforward avoided like the plague.

On the other side of the coin, I just as strenuously reject teaching linear models with a coordinate free approach. Although Joe Eaton assures me that the issues in complicated problems frequently become clearer when considered free of coordinate systems, my experience is that too many people never make the jump from coordinate free theory back to practical applications. I think that coordinate free theory is better tackled after mastering linear models from some other approach. In particular, I think it would be very easy to pick up the coordinate free approach after learning the material in this book. See Eaton (1983) for an excellent exposition of the coordinate free approach.

By now it should be obvious to the reader that I am not very opinionated on the subject of linear models. In spite of that fact, I have made an effort to identify sections of the book where I express my personal opinions.

Although in recent revisions I have made an effort to cite more of the literature, the book contains comparatively few references. The references are adequate to the needs of the book, but no attempt has been made to survey the literature. This was done for two reasons. First, the book was begun about 10 years ago, right after I finished my Masters degree at the University of Minnesota. At that time, I was not aware of much of the literature. The second reason is that this book emphasizes a particular point of view. A survey of the literature would best be done on the literature's own terms. In writing this, I ended up reinventing a lot of wheels. My apologies to anyone whose work I have overlooked.

Using the Book

This book has been extensively revised, and the last five chapters were written at Montana State University. At Montana State, we require a year of Linear Models for all of our statistics graduate students. In our three-quarter course, I usually end the first quarter with Chapter 4 or in the middle of Chapter 5. At the end of winter quarter, I have finished Chapter 9. I consider the first nine chapters to be the core

material of the book. I go quite slowly because all of our Masters students are required to take the course. For Ph.D. students, I think a one-semester course might be the first nine chapters, and a two-quarter course might have time to add some topics from the remainder of the book.

I view the chapters after 9 as a series of important special topics from which instructors can choose material but which students should have access to even if their course omits them. In our third quarter, I typically cover (at some level) Chapters 11 to 14. The idea behind the special topics is not to provide an exhaustive discussion but rather to give a basic introduction that will also enable readers to move on to more detailed works such as Cook and Weisberg (1982) and Haberman (1974).

Appendices A–E provide required background material. My experience is that the student's greatest stumbling block is linear algebra. I would not dream of teaching out of this book without a thorough review of Appendices A and B.

The main prerequisite for reading this book is a good background in linear algebra. The book also assumes knowledge of mathematical statistics at the level of, say, Lindgren or Hogg and Craig. Although I think a mathematically sophisticated reader could handle this book without having had a course in statistical methods, I think that readers who have had a methods course will get much more out of it.

The exercises in this book are presented in two ways. In the original manuscript, the exercises were incorporated into the text. The original exercises have not been relocated. It has been my practice to assign virtually all of these exercises. At a later date, the editors from Springer-Verlag and I agreed that other instructors might like more options in choosing problems. As a result, a section of additional exercises was added to the end of the first nine chapters and some additional exercises were added to other chapters and appendices. I continue to recommend requiring nearly all of the exercises incorporated in the text. In addition, I think there is much to be learned about linear models by doing, or at least reading, the additional exercises.

Many of the exercises are provided with hints. These are primarily designed so that I can quickly remember how to do them. If they help anyone other than me, so much the better.

Acknowledgements

I am a great believer in books. The vast majority of my knowledge about statistics has been obtained by starting at the beginning of a book and reading until I covered what I had set out to learn. I feel both obligated and privileged to thank the authors of the books from which I first learned about linear models: Daniel and Wood, Draper and Smith, Scheffé, and Searle.

In addition, there are a number of people who have substantially influenced particular parts of this book. Their contributions are too diverse to specify, but I should mention that, in several cases, their influence has been entirely by means of their written work. (Moreover, I suspect that in at least one case, the person in

question will be loath to find that his writings have come to such an end as this.) I would like to acknowledge Kit Bingham, Carol Bittinger, Larry Blackwood, Dennis Cook, Somesh Das Gupta, Seymour Geisser, Susan Groshen, Shelby Haberman, David Harville, Cindy Hertzler, Steve Kachman, Kinley Larntz, Dick Lund, Ingram Olkin, S. R. Searle, Anne Torbeyns, Sandy Weisberg, George Zyskind, and all of my students. Three people deserve special recognition for their pains in advising me on the manuscript: Robert Boik, Steve Fienberg, and Wes Johnson.

The typing of the first draft of the manuscript was done by Laura Cranmer and Donna Stickney.

I would like to thank my family: Sharon, Fletch, George, Doris, Gene, and Jim, for their love and support. I would also like to thank my friends from graduate school who helped make those some of the best years of my life.

Finally, there are two people without whom this book would not exist: Frank Martin and Don Berry. Frank because I learned how to think about linear models in a course he taught. This entire book is just an extension of the point of view that I developed in Frank's class. And Don because he was always there ready to help—from teaching my first statistics course to being my thesis adviser and everywhere in between.

Since I have never even met some of these people, it would be most unfair to blame anyone but me for what is contained in the book. (Of course, I will be more than happy to accept any and all praise.) Now that I think about it, there may be one exception to the caveat on blame. If you don't like the diatribe on prediction in Chapter 6, you might save just a smidgen of blame for Seymour (even though he did not see it before publication).

Ronald Christensen
Bozeman, Montana, 1987

References

- Christensen, R. (1997). *Log-linear models and logistic regression* (2nd ed.). New York: Springer.
- Christensen, R. (2015). *Analysis of variance, design, and regression: Linear modeling for unbalanced data* (2nd ed.). Boca Raton: Chapman and Hall/CRC Press.
- Christensen, R. (2019). *Advanced linear modeling III: Statistical learning and dependent data* (3rd ed.). Springer-Verlag, New York.
- Christensen, R. & Lin, Y. (2015). Lack-of-fit tests based on partial sums of residuals. *Communications in Statistics, Theory and Methods*, 44 2862–2880.
- Cook, R. D., & Weisberg, S. (1982). *Residuals and influence in regression*. New York: Chapman and Hall.
- deLaubenfels, R. (2006). The victory of least squares and orthogonality in statistics. *The American Statistician*, 60, 315–321.
- Eaton, M. L. (1983). *Multivariate statistics: a vector space approach*. New York: Wiley. Reprinted in 2007 by IMS Lecture Notes–Monograph Series.
- Haberman, S. J. (1974). *The Analysis of Frequency Data*. University of Chicago Press, Chicago.

Contents

1	Introduction	1
1.1	Random Matrices and Vectors	4
1.2	Multivariate Normal Distributions	7
1.3	Distributions of Quadratic Forms	11
1.3.1	Results for General Covariance Matrices	14
1.4	Generalized Linear Models	16
1.5	Additional Exercises	18
	References	20
2	Estimation	21
2.1	Identifiability and Estimability	22
2.2	Estimation: Least Squares	28
2.3	Estimation: Best Linear Unbiased	33
2.4	Estimation: Maximum Likelihood	34
2.5	Estimation: Minimum Variance Unbiased	35
2.6	Sampling Distributions of Estimates	37
2.7	Generalized Least Squares	38
2.8	Normal Equations	43
2.9	Variance-Bias Tradeoff	44
2.9.1	Estimable Functions	47
2.10	Bayesian Estimation	48
2.10.1	Distribution Theory	52
2.11	Additional Exercises	57
	References	59
3	Testing	61
3.1	More About Models	61
3.2	Testing Models	64
3.2.1	Small Test Statistics	71
3.2.2	A Generalized Test Procedure	72

- 3.3 Testing Linear Parametric Functions 74
 - 3.3.1 A Generalized Test Procedure 83
 - 3.3.2 Testing an Unusual Class of Hypotheses 86
- 3.4 Discussion 88
- 3.5 Testing Single Degrees of Freedom in a Given Subspace 89
- 3.6 Breaking a Sum of Squares into Independent Components 90
 - 3.6.1 General Theory 91
 - 3.6.2 Two-Way ANOVA 95
- 3.7 Confidence Regions 97
- 3.8 Tests for Generalized Least Squares Models 98
 - 3.8.1 Conditions for Simpler Procedures 101
- 3.9 Additional Exercises 103
- References 105
- 4 One-Way ANOVA 107**
 - 4.1 Analysis of Variance 108
 - 4.2 Estimating and Testing Contrasts 116
 - 4.3 Additional Exercises 120
- 5 Multiple Comparison Techniques 123**
 - 5.1 Basic Ideas 124
 - 5.2 Scheffé’s Method 128
 - 5.3 Least Significant Difference Method 133
 - 5.4 Bonferroni Method 135
 - 5.5 Tukey’s Method 135
 - 5.6 Multiple Range Tests: Newman–Keuls and Duncan 137
 - 5.7 Summary 139
 - 5.7.1 Fisher Versus Neyman–Pearson 142
 - 5.8 Additional Exercises 142
 - References 143
- 6 Regression Analysis 145**
 - 6.1 Simple Linear Regression 146
 - 6.2 Multiple Regression 148
 - 6.2.1 Partitioned Model 151
 - 6.2.2 Nonparametric Regression and Generalized Additive Models 153
 - 6.3 General Prediction Theory 155
 - 6.3.1 Discussion 155
 - 6.3.2 General Prediction 156
 - 6.3.3 Best Prediction 156
 - 6.3.4 Best Linear Prediction 160
 - 6.3.5 Inner Products and Orthogonal Projections in General Spaces 164
 - 6.4 Multiple Correlation 165

- 6.4.1 Squared Predictive Correlation 168
- 6.5 Partial Correlation Coefficients 170
- 6.6 Best Linear Unbiased Prediction 172
- 6.7 Testing Lack of Fit 177
 - 6.7.1 The Traditional Test 178
 - 6.7.2 Near Replicate Lack of Fit Tests 180
 - 6.7.3 Partitioning Methods 182
 - 6.7.4 Nonparametric Methods 185
- 6.8 Polynomial Regression and One-Way ANOVA 186
- 6.9 Additional Exercises 191
- References 194
- 7 Multifactor Analysis of Variance 197**
 - 7.1 Balanced Two-Way ANOVA Without Interaction 197
 - 7.1.1 Contrasts 203
 - 7.2 Balanced Two-Way ANOVA with Interaction 204
 - 7.2.1 Interaction Contrasts 207
 - 7.3 Polynomial Regression and the Balanced Two-Way ANOVA 214
 - 7.4 Two-Way ANOVA with Proportional Numbers 217
 - 7.5 Two-Way ANOVA with Unequal Numbers: General Case 219
 - 7.5.1 Without Interaction 219
 - 7.5.2 Interaction 222
 - 7.5.3 Characterizing the Interaction Space 228
 - 7.6 Three or More Way Analyses 230
 - 7.6.1 Balanced Analyses 230
 - 7.6.2 Unbalanced Analyses 232
 - 7.7 Additional Exercises 238
 - References 240
- 8 Experimental Design Models 241**
 - 8.1 Completely Randomized Designs 242
 - 8.2 Randomized Complete Block Designs: Usual Theory 242
 - 8.3 Latin Square Designs 243
 - 8.4 Factorial Treatment Structures 247
 - 8.5 More on Factorial Treatment Structures 250
 - 8.6 Additional Exercises 253
 - References 253
- 9 Analysis of Covariance 255**
 - 9.1 Estimation of Fixed Effects 256
 - 9.1.1 Generalized Least Squares 260
 - 9.2 Estimation of Error and Tests of Hypotheses 261
 - 9.3 Another Adjusted Model and Missing Data 264
 - 9.4 Balanced Incomplete Block Designs 267

9.5	Testing a Nonlinear Full Model	276
9.6	Additional Exercises	277
	References	279
10	General Gauss–Markov Models	281
10.1	BLUEs with an Arbitrary Covariance Matrix	282
10.2	Geometric Aspects of Estimation	288
10.3	Hypothesis Testing	291
10.4	Least Squares Consistent Estimation	296
10.5	Perfect Estimation and More	304
	References	311
11	Split Plot Models	313
11.1	A Cluster Sampling Model	314
11.2	Generalized Split Plot Models	318
	11.2.1 Estimation and Testing of Estimable Functions	322
	11.2.2 Testing Models	326
	11.2.3 Unbalanced Subplots	328
11.3	The Split Plot Design	328
11.4	Identifying the Appropriate Error	332
	11.4.1 Subsampling	332
	11.4.2 Two-Way ANOVA with Interaction	335
11.5	Exercise: An Unusual Split Plot Analysis	337
	References	338
12	Model Diagnostics	341
12.1	Leverage	344
	12.1.1 Mahalanobis Distances	345
	12.1.2 Diagonal Elements of the Projection Operator	347
	12.1.3 Examples	348
12.2	Checking Normality	354
	12.2.1 Other Applications for Normal Plots	360
12.3	Checking Independence	362
	12.3.1 Serial Correlation	363
12.4	Heteroscedasticity and Lack of Fit	369
	12.4.1 Heteroscedasticity	369
	12.4.2 Huber–White (Robust) Sandwich Estimator	374
	12.4.3 Lack of Fit	377
	12.4.4 Residual Plots	380
12.5	Updating Formulae and Predicted Residuals	380
12.6	Outliers and Influential Observations	384
12.7	Transformations	388
	References	390

- 13 Collinearity and Alternative Estimates 393**
 - 13.1 Defining Collinearity 394
 - 13.2 Tolerance and Variance Inflation Factors 398
 - 13.3 Regression in Canonical Form and on Principal Components 401
 - 13.3.1 Regression in Canonical Form 401
 - 13.3.2 Principal Component Regression 404
 - 13.3.3 Generalized Inverse Regression 404
 - 13.4 Classical Ridge Regression 405
 - 13.4.1 Ridge Applied to Principal Components 408
 - 13.5 More on Mean Squared Error 410
 - 13.6 Robust Estimation and Alternative Distance Measures 410
 - 13.7 Orthogonal Regression 413
 - References 416

- 14 Variable Selection 419**
 - 14.1 All Possible Regressions and Best Subset Regression 421
 - 14.1.1 R^2 421
 - 14.1.2 Adjusted R^2 422
 - 14.1.3 Mallows's C_p 423
 - 14.1.4 Information Criteria: AIC, BIC 425
 - 14.1.5 Cost Complexity Pruning 427
 - 14.2 Stepwise Regression 428
 - 14.2.1 Traditional Forward Selection 428
 - 14.2.2 Backward Elimination 431
 - 14.2.3 Other Methods 433
 - 14.3 Discussion of Traditional Variable Selection Techniques 433
 - 14.3.1 R^2 434
 - 14.3.2 Influential Observations 435
 - 14.3.3 Exploratory Data Analysis 435
 - 14.3.4 Multiplicities 436
 - 14.3.5 Predictive Models 436
 - 14.3.6 Overfitting 437
 - 14.4 Modern Forward Selection: Boosting, Bagging, and Random Forests 437
 - 14.4.1 Boosting 438
 - 14.4.2 Bagging 441
 - 14.4.3 Random Forests 445
 - References 446

Appendix A: Vector Spaces	447
Appendix B: Matrix Results	457
B.1 Basic Ideas	457
B.2 Eigenvalues and Related Results	459
B.3 Projections	463
B.4 Miscellaneous Results	471
B.5 Properties of Kronecker Products and Vec Operators	473
B.6 Tensors	474
B.7 Exercises	476
Appendix C: Some Univariate Distributions	481
Appendix D: Multivariate Distributions	485
D.1 Identifiability	489
Appendix E: Inference for One Parameter	491
E.1 Testing	493
E.2 P Values	495
E.3 Confidence Intervals	496
E.4 Final Comments on Significance Testing	497
Appendix F: Significantly Insignificant Tests	499
F.1 Lack of Fit and Small F Statistics	501
F.2 The Effect of Correlation and Heteroscedasticity on F Statistics ...	503
Appendix G: Randomization Theory Models	509
G.1 Simple Random Sampling	509
G.2 Completely Randomized Designs	511
G.3 Randomized Complete Block Designs	513
References	516
Author Index	517
Subject Index	521