Springer Texts in Statistics

For other titles published in this series, go to
www.springer.com/series/417

Ronald Christensen

Plane Answers to Complex Questions

The Theory of Linear Models

Fourth edition

Springer

Ronald Christensen
Department of Mathematics and Statistics
MSC01 1115
1 University of New Mexico
Albuquerque, New Mexico, 87131-0001
USA
fletcher@stat.unm.edu

ISSN 1431-875X
ISBN 978-1-4419-9815-6 e-ISBN 978-1-4419-9816-3
DOI 10.1007/978-1-4419-9816-3
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2011928555

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

To Dad, Mom, Sharon, Fletch, and Don

Preface

Preface to the Fourth Edition

"Critical assessment of data is the 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.

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.

Acknowledgments

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 loathe 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

Contents