Springer Texts in Statistics

Advisors:

George Casella Stephen Fienberg Ingram Olkin

Springer Science+Business Media, LLC

Springer Texts in Statistics

Alfred: Elements of Statistics for the Life and Social Sciences
Berger: An Introduction to Probability and Stochastic Processes
Bilodeau and Brenner: Theory of Multivariate Statistics
Blom: Probability and Statistics: Theory and Applications
Brockwell and Davis: Introduction to Times Series and Forecasting,
Second Edition
Carmona: Statistical Analysis of Financial Data in S-Plus
Chow and Teicher: Probability Theory: Independence, Interchangeability,
Martingales, Third Edition
Christensen: Advanced Linear Modeling: Multivariate, Time Series, and
Spatial Data; Nonparametric Regression and Response Surface
Maximization, Second Edition
Christensen: Log-Linear Models and Logistic Regression, Second Edition
Christensen: Plane Answers to Complex Questions: The Theory of Linear
Models, Third Edition
Creighton: A First Course in Probability Models and Statistical Inference
Davis: Statistical Methods for the Analysis of Repeated Measurements
Dean and Voss: Design and Analysis of Experiments
du Toit, Steyn, and Stumpf: Graphical Exploratory Data Analysis
Durrett: Essentials of Stochastic Processes
Edwards: Introduction to Graphical Modelling, Second Edition
Finkelstein and Levin: Statistics for Lawyers
Flury: A First Course in Multivariate Statistics
Heiberger and Holland: Statistical Analysis and Data Display: An Intermediate
Course with Examples in S-PLUS, R, and SAS
Jobson: Applied Multivariate Data Analysis, Volume I: Regression and
Experimental Design
Jobson: Applied Multivariate Data Analysis, Volume II: Categorical and
Multivariate Methods
Kalbfleisch: Probability and Statistical Inference, Volume I: Probability,
Second Edition
Kalbfleisch: Probability and Statistical Inference, Volume II: Statistical Inference,
Second Edition
Karr: Probability
Keyfitz: Applied Mathematical Demography, Second Edition
Kiefer: Introduction to Statistical Inference
Kokoska and Nevison: Statistical Tables and Formulae
Kulkarni: Modeling, Analysis, Design, and Control of Stochastic Systems
Lange: Applied Probability
Lange: Optimization
Lehmann: Elements of Large-Sample Theory

(continued after index)

Ronald Christensen

Plane Answers to Complex Questions

The Theory of Linear Models

Third Edition

With 34 Illustrations

Springer

Ronald Christensen
Department of Mathematics and Statistics
University of New Mexico
Albuquerque, NM 87131
USA

Editorial Board

George Casella
Department of Statistics
University of Florida
Gainesville, FL 32611-8545
USA

Stephen Fienberg
Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213-3890
USA

Ingram Olkin
Department of Statistics
Stanford University
Stanford, CA 94305
USA

Library of Congress Cataloging-in-Publication Data

Christensen, Ronald, 1951–

Plane answers to complex questions : the theory of linear models / Ronald
Christensen. — 3rd ed.

p. cm. — (Springer texts in statistics)

Includes bibliographical references and index.

ISBN 978-1-4419-2971-6 ISBN 978-0-387-21544-0 (eBook)

DOI 10.1007/978-0-387-21544-0

I. Linear models (Statistics) 2. Analysis of variance. I. Title. II. Series.
QA279 .C48 2002

519.5'35—dc21

2001053055

Printed on acid-free paper.

© 2002, 1996, 1987 Springer Science+Business Media New York

Originally published by Springer-Verlag New York, Inc. in 2002

All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher Springer Science+Business Media, LLC **except for** brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden.

The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

9 8 7 6 5 4 3 2

SPIN 11014003

springeronline.com

To Dad, Mom, Sharon, Fletch, and Don

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 non-parametric 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 (1990b), 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

Preface to the Third Edition	vii
Preface to the Second Edition	ix
Preface to the First Edition	xi
1 Introduction	1
1.1 Random Vectors and Matrices	3
1.2 Multivariate Normal Distributions	5
1.3 Distributions of Quadratic Forms	8
1.4 Generalized Linear Models	13
1.5 Additional Exercises	15
2 Estimation	18
2.1 Identifiability and Estimability	19
2.2 Estimation: Least Squares	24
2.3 Estimation: Best Linear Unbiased	28
2.4 Estimation: Maximum Likelihood	30
2.5 Estimation: Minimum Variance Unbiased	30
2.6 Sampling Distributions of Estimates	32
2.7 Generalized Least Squares	33
2.8 Normal Equations	38
2.9 Bayesian Estimation	39
2.9.1 Distribution Theory	43

2.10	Additional Exercises	48
3	Testing Hypotheses	50
3.1	More About Models	51
3.2	Testing Models	53
3.2.1	A Generalized Test Procedure	60
3.3	Testing Linear Parametric Functions	62
3.3.1	A Generalized Test Procedure	71
3.3.2	Testing an Unusual Class of Hypotheses	74
3.4	Testing Single Degrees of Freedom in a Given Subspace . .	76
3.5	Discussion	77
3.6	Breaking a Sum of Squares into Independent Components .	78
3.6.1	Two-Way ANOVA	79
3.6.2	General Theory	81
3.7	Confidence Regions	85
3.8	Tests for Generalized Least Squares Models	86
3.9	Additional Exercises	89
4	One-Way ANOVA	91
4.1	Analysis of Variance	92
4.2	Estimating and Testing Contrasts	100
4.3	Additional Exercises	104
5	Multiple Comparison Techniques	106
5.1	Scheffé's Method	107
5.2	Least Significant Difference Method	112
5.3	Bonferroni Method	113
5.4	Tukey's Method	114
5.5	Multiple Range Tests: Newman-Keuls and Duncan	115
5.6	Summary	117
5.7	Additional Exercises	119
6	Regression Analysis	122
6.1	Simple Linear Regression	123
6.2	Multiple Regression	124
6.2.1	Nonparametric Regression and Generalized Additive Models	129
6.3	Best Linear Prediction: Another Approach to Regression . .	130
6.4	Multiple Correlation and the Coefficient of Determination .	135
6.5	Partial Correlation Coefficients	137
6.6	Testing Lack of Fit	140
6.6.1	The Traditional Test	140
6.6.2	Near Replicate Lack of Fit Tests	142
6.6.3	Partitioning Methods	145
6.7	Polynomial Regression and One-Way Analysis of Variance .	148

6.8	Additional Exercises	152
7	Multifactor Analysis of Variance	156
7.1	Balanced Two-Way ANOVA Without Interaction	156
7.1.1	Contrasts	162
7.2	Balanced Two-Way ANOVA with Interaction	163
7.2.1	Interaction Contrasts	166
7.3	Polynomial Regression and the Balanced Two-Way ANOVA	172
7.4	Two-Way ANOVA with Proportional Numbers	176
7.5	Two-Way ANOVA with Unequal Numbers: General Case .	177
7.6	Three or More Way Analyses	186
7.7	Additional Exercises	193
8	Experimental Design Models	196
8.1	Completely Randomized Designs	197
8.2	Randomized Complete Block Designs: Usual Theory	197
8.3	Latin Square Designs	198
8.4	Factorial Treatment Structures	202
8.5	More on Factorial Treatment Structures	205
8.6	Additional Exercises	207
9	Analysis of Covariance	209
9.1	Estimation of Fixed Effects	211
9.2	Estimation of Error and Tests of Hypotheses	213
9.3	Application: Missing Data	216
9.4	Application: Balanced Incomplete Block Designs	219
9.5	Additional Exercises	228
10	Estimation and Testing in General Gauss–Markov Models	230
10.1	BLUEs with an Arbitrary Covariance Matrix	231
10.2	Geometric Aspects of Estimation	237
10.3	Hypothesis Testing	241
10.4	Least Squares Consistent Estimation	246
11	Split Plot Models	254
11.1	A Cluster Sampling Model	255
11.2	Generalized Split Plot Models	260
11.3	The Split Plot Design	268
11.4	Identifying the Appropriate Error	271
11.5	Exercise: An Unusual Split Plot Analysis	276
12	Mixed Models and Variance Components	279
12.1	Mixed Models	280
12.2	Best Linear Unbiased Prediction	281
12.3	Mixed Model Equations	285
12.4	Variance Component Estimation: Maximum Likelihood . .	287

12.5	Maximum Likelihood Estimation for Singular Normal Distributions	291
12.6	Variance Component Estimation: REML	293
12.7	Variance Component Estimation: MINQUE	295
12.8	Variance Component Estimation: MIVQUE	298
12.9	Variance Component Estimation: Henderson's Method 3	298
12.10	Exact F Tests for Variance Components	302
12.10.1	Wald's Test	302
12.10.2	Öfversten's Second Method	304
12.10.3	Comparison of Tests	306
12.11	Recovery of Interblock Information in BIB Designs	309
12.11.1	Estimation	310
12.11.2	Model Testing	313
12.11.3	Contrasts	316
12.11.4	Alternative Inferential Procedures	317
12.11.5	Estimation of Variance Components	318
13	Checking Assumptions, Residuals, and Influential Observations	320
13.1	Leverage	323
13.1.1	Mahalanobis Distances	324
13.1.2	Diagonal Elements of the Projection Operator	326
13.1.3	Examples	327
13.2	Checking Normality	333
13.2.1	Other Applications for Normal Plots	338
13.3	Checking Independence	341
13.3.1	Serial Correlation	342
13.4	Heteroscedasticity and Lack of Fit	348
13.4.1	Heteroscedasticity	348
13.4.2	Lack of Fit	355
13.5	Updating Formulae and Predicted Residuals	359
13.6	Outliers and Influential Observations	363
13.7	Transformations	366
14	Variable Selection and Collinearity	369
14.1	All Possible Regressions and Best Subset Regression	370
14.1.1	R^2	371
14.1.2	Adjusted R^2	371
14.1.3	Mallows's C_p	372
14.2	Stepwise Regression	373
14.2.1	Forward Selection	374
14.2.2	Tolerance	375
14.2.3	Backward Elimination	375
14.2.4	Other Methods	376
14.3	Discussion of Variable Selection Techniques	376

14.4	Defining Collinearity	378
14.5	Regression in Canonical Form and on Principal Components	383
14.5.1	Principal Component Regression	386
14.5.2	Generalized Inverse Regression	386
14.6	Classical Ridge Regression	387
14.7	More on Mean Squared Error	389
Appendix A: Vector Spaces		391
Appendix B: Matrix Results		399
B.1	Basic Ideas	399
B.2	Eigenvalues and Related Results	401
B.3	Projections	406
B.4	Miscellaneous Results	415
B.5	Properties of Kronecker Products and Vec Operators	416
B.6	Tensors	418
B.7	Exercises	420
Appendix C: Some Univariate Distributions		424
Appendix D: Multivariate Distributions		428
Appendix E: Inference for One Parameter		432
E.1	Confidence Intervals	433
E.2	Hypothesis Testing	434
Appendix F: Significantly Insignificant Tests		437
F.1	Lack of Fit and Small F Statistics	438
F.2	The Effect of Correlation and Heteroscedasticity on F Statistics	441
Appendix G: Randomization Theory Models		447
G.1	Simple Random Sampling	448
G.2	Completely Randomized Designs	449
G.3	Randomized Complete Block Designs	451
References		455
Author Index		463
Subject Index		466