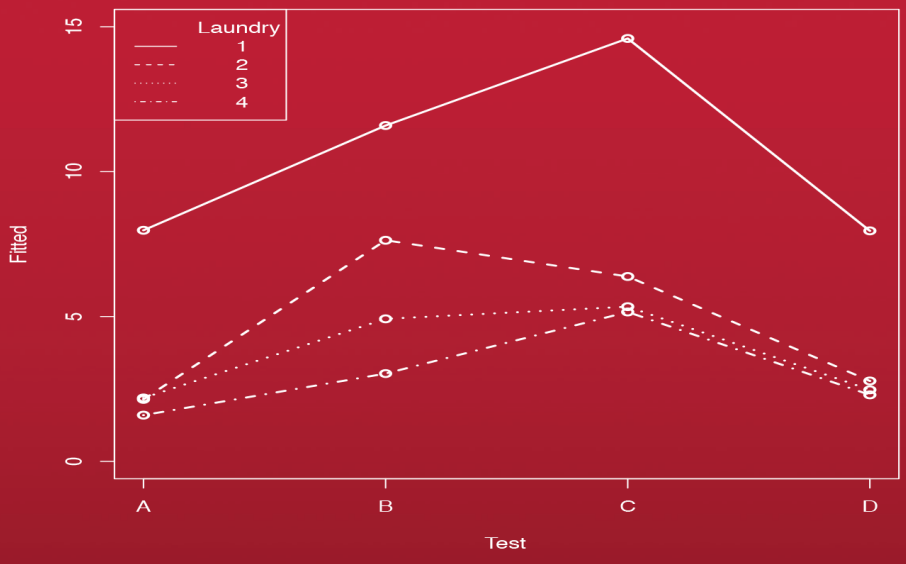
Texts in Statistical Science

Analysis of Variance, Design, and Regression

Linear Modeling for Unbalanced Data

Second Edition



Ronald Christensen

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Analysis of Variance, Design, and Regression

Linear Modeling for Unbalanced Data

Second Edition

CHAPMAN & HALL/CRC
Texts in Statistical Science Series

Statistical Theory: A Concise Introduction
F. Abramovich and Y. Ritov

Practical Multivariate Analysis, Fifth Edition
A. Afifi, S. May, and V.A. Clark

Practical Statistics for Medical Research
D.G. Altman

Interpreting Data: A First Course in Statistics
A.J.B. Anderson

Introduction to Probability with R
K. Baclawski

Linear Algebra and Matrix Analysis for Statistics
S. Banerjee and A. Roy

Mathematical Statistics: Basic Ideas and Selected Topics, Volume I, Second Edition
P. J. Bickel and K. A. Doksum

Mathematical Statistics: Basic Ideas and Selected Topics, Volume II
P. J. Bickel and K. A. Doksum

Analysis of Categorical Data with R
C. R. Bilder and T. M. Loughin

Statistical Methods for SPC and TQM
D. Bissell

Introduction to Probability
J. K. Blitzstein and J. Hwang

Bayesian Methods for Data Analysis, Third Edition
B.P. Carlin and T.A. Louis

Second Edition
R. Caulcutt

The Analysis of Time Series: An Introduction, Sixth Edition
C. Chatfield

Introduction to Multivariate Analysis
C. Chatfield and A.J. Collins

Problem Solving: A Statistician's Guide, Second Edition
C. Chatfield

Statistics for Technology: A Course in Applied Statistics, Third Edition
C. Chatfield

Analysis of Variance, Design, and Regression : Linear Modeling for Unbalanced Data, Second Edition
R. Christensen

Bayesian Ideas and Data Analysis: An Introduction for Scientists and Statisticians
R. Christensen, W. Johnson, A. Branscum, and T.E. Hanson

Modelling Binary Data, Second Edition
D. Collett

Modelling Survival Data in Medical Research, Third Edition
D. Collett

Introduction to Statistical Methods for Clinical Trials
T.D. Cook and D.L. DeMets

Applied Statistics: Principles and Examples
D.R. Cox and E.J. Snell

Multivariate Survival Analysis and Competing Risks
M. Crowder

Statistical Analysis of Reliability Data
M.J. Crowder, A.C. Kimber, T.J. Sweeting, and R.L. Smith

An Introduction to Generalized Linear Models, Third Edition
A.J. Dobson and A.G. Barnett

Nonlinear Time Series: Theory, Methods, and Applications with R Examples
R. Douc, E. Moulines, and D.S. Stoffer

Introduction to Optimization Methods and Their Applications in Statistics
B.S. Everitt

Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models
J.J. Faraway

Linear Models with R, Second Edition
J.J. Faraway

A Course in Large Sample Theory
T.S. Ferguson

Multivariate Statistics: A Practical Approach
B. Flury and H. Riedwyl

Readings in Decision Analysis
S. French

Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference, Second Edition
D. Gamerman and H.F. Lopes

Bayesian Data Analysis, Third Edition
A. Gelman, J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari, and D.B. Rubin

Multivariate Analysis of Variance and Repeated Measures: A Practical Approach for Behavioural Scientists
D.J. Hand and C.C. Taylor

Practical Longitudinal Data Analysis
D.J. Hand and M. Crowder

Logistic Regression Models
J.M. Hilbe

Richly Parameterized Linear Models: Additive, Time Series, and Spatial Models Using Random Effects
J.S. Hodges

Statistics for Epidemiology
N.P. Jewell

Stochastic Processes: An Introduction, Second Edition
P.W. Jones and P. Smith

The Theory of Linear Models
B. Jørgensen

Principles of Uncertainty
J.B. Kadane

Graphics for Statistics and Data Analysis with R
K.J. Keen

Mathematical Statistics
K. Knight

Introduction to Multivariate Analysis: Linear and Nonlinear Modeling
S. Konishi

Nonparametric Methods in Statistics with SAS Applications
O. Korosteleva

Modeling and Analysis of Stochastic Systems, Second Edition
V.G. Kulkarni

Exercises and Solutions in Biostatistical Theory
L.L. Kupper, B.H. Neelon, and S.M. O'Brien

Exercises and Solutions in Statistical Theory
L.L. Kupper, B.H. Neelon, and S.M. O'Brien

Design and Analysis of Experiments with R
J. Lawson

Design and Analysis of Experiments with SAS
J. Lawson

A Course in Categorical Data Analysis
T. Leonard

Statistics for Accountants
S. Letchford

Introduction to the Theory of Statistical Inference
H. Liero and S. Zwanzig

Statistical Theory, Fourth Edition
B.W. Lindgren

Stationary Stochastic Processes: Theory and Applications
G. Lindgren

Statistics for Finance
E. Lindström, H. Madsen, and J. N. Nielsen

The BUGS Book: A Practical Introduction to Bayesian Analysis
D. Lunn, C. Jackson, N. Best, A. Thomas, and D. Spiegelhalter

Introduction to General and Generalized Linear Models
H. Madsen and P. Thyregod

Time Series Analysis
H. Madsen

Pólya Urn Models
H. Mahmoud

Randomization, Bootstrap and Monte Carlo Methods in Biology, Third Edition
B.F.J. Manly

Introduction to Randomized Controlled Clinical Trials, Second Edition
J.N.S. Matthews

Statistical Rethinking: A Bayesian Course with Examples in R and Stan
R. McElreath

Statistical Methods in Agriculture and Experimental Biology, Second Edition
R. Mead, R.N. Curnow, and A.M. Hasted

Statistics in Engineering: A Practical Approach
A.V. Metcalfe

Statistical Inference: An Integrated Approach, Second Edition
H. S. Migon, D. Gamerman, and F. Louzada

Beyond ANOVA: Basics of Applied Statistics
R.G. Miller, Jr.

A Primer on Linear Models
J.F. Monahan

Applied Stochastic Modelling, Second Edition
B.J.T. Morgan

Elements of Simulation
B.J.T. Morgan

Probability: Methods and Measurement
A. O'Hagan

Introduction to Statistical Limit Theory
A.M. Polansky

Applied Bayesian Forecasting and Time Series Analysis
A. Pole, M. West, and J. Harrison

Statistics in Research and Development, Time Series: Modeling, Computation, and Inference
R. Prado and M. West

Introduction to Statistical Process Control
P. Qiu

Sampling Methodologies with Applications
P.S.R.S. Rao

A First Course in Linear Model Theory
N. Ravishanker and D.K. Dey

Essential Statistics, Fourth Edition
D.A.G. Rees

Stochastic Modeling and Mathematical Statistics: A Text for Statisticians and Quantitative Scientists
F.J. Samaniego

Statistical Methods for Spatial Data Analysis
O. Schabenberger and C.A. Gotway

Bayesian Networks: With Examples in R
M. Scutari and J.-B. Denis

Large Sample Methods in Statistics
P.K. Sen and J. da Motta Singer

Spatio-Temporal Methods in Environmental Epidemiology
G. Shaddick and J.V. Zidek

Decision Analysis: A Bayesian Approach
J.Q. Smith

Analysis of Failure and Survival Data
P. J. Smith

Applied Statistics: Handbook of GENSTAT Analyses
E.J. Snell and H. Simpson

Applied Nonparametric Statistical Methods, Fourth Edition
P. Sprent and N.C. Smeeton

Data Driven Statistical Methods
P. Sprent

Generalized Linear Mixed Models: Modern Concepts, Methods and Applications
W. W. Stroup

Survival Analysis Using S: Analysis of Time-to-Event Data
M. Tableman and J.S. Kim

Applied Categorical and Count Data Analysis
W. Tang, H. He, and X.M. Tu

Elementary Applications of Probability Theory, Second Edition
H.C. Tuckwell

Introduction to Statistical Inference and Its Applications with R
M.W. Trosset

Understanding Advanced Statistical Methods
P.H. Westfall and K.S.S. Henning

Statistical Process Control: Theory and Practice, Third Edition
G.B. Wetherill and D.W. Brown

Generalized Additive Models: An Introduction with R
S. Wood

Epidemiology: Study Design and Data Analysis, Third Edition
M. Woodward

Practical Data Analysis for Designed Experiments
B.S. Yandell

Texts in Statistical Science

Analysis of Variance, Design, and Regression

Linear Modeling for Unbalanced Data

Second Edition

Ronald Christensen

University of New Mexico
Albuquerque, USA

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A CHAPMAN & HALL BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Version Date: 20151221

International Standard Book Number-13: 978-1-4987-7405-5 (eBook - PDF)

To Mark, Karl, and John

It was great fun.

Contents

Preface

Background

Big Data are the future of Statistics. The electronic revolution has increased exponentially our ability to measure things. A century ago, data were hard to come by. Statisticians put a premium on extracting every bit of information that the data contained. Now data are easy to collect; the problem is sorting through them to find meaning. To a large extent, this happens in two ways: doing a crude analysis on a massive amount of data or doing a careful analysis on the moderate amount of data that were isolated from the massive data as being meaningful. It is quite literally impossible to analyze a million data points as carefully as one can analyze a hundred data points, so "crude" is not a pejorative term but rather a fact of life.

The fundamental tools used in analyzing data have been around a long time. It is the emphases and the opportunities that have changed. With thousands of observations, we don't need a perfect statistical analysis to detect a large effect. But with thousands of observations, we might look for subtle effects that we never bothered looking for before, and such an analysis must be done carefully—as must any analysis in which only a small part of the massive data are relevant to the problem at hand. The electronic revolution has also provided us with the opportunity to perform data analysis procedures that were not practical before, but in my experience, the new procedures (often called *machine learning*), are sophisticated applications of fundamental tools.

This book explains some of the fundamental tools and the ideas needed to adapt them to big data. It is not a book that analyzes big data. The book analyzes small data sets carefully but by using tools that 1) can easily be scaled to large data sets or 2) apply to the haphazard way in which small relevant data sets are now constructed. Personally, I believe that it is not safe to apply models to large data sets until you understand their implications for small data. There is also a major emphasis on tools that look for subtle effects (interactions, homologous effects) that are hard to identify.

The fundamental tools examined here are linear structures for modeling data; specifically, how to incorporate specific ideas about the structure of the data into the model for the data. Most of the book is devoted to adapting linear structures (regression, analysis of variance, analysis of covariance) to examine measurement (continuous) data. But the exact same methods apply to either-or (Yes/No, binomial) data, count (Poisson, multinomial) data, and time-to-event (survival analysis, reliability) data. The book also places strong emphasis on foundational issues, e.g., the meaning of significance tests and the interval estimates associated with them; the difference between prediction and causation; and the role of randomization.

The platform for this presentation is the revision of a book I published in 1996, *Analysis of Variance, Design, and Regression: Applied Statistical Methods*. Within a year, I knew that the book was not what I thought needed to be taught in the 21st century, cf., Christensen (2000). This book, *Analysis of Variance, Design, and Regression: Linear Modeling of Unbalanced Data*, shares with the earlier book lots of the title, much of the data, and even some of the text, but the book is radically different. The original book focused greatly on balanced analysis of variance. This book focuses on modeling unbalanced data. As such, it generalizes much of the work in the previous book. The more general methods presented here agree with the earlier methods for balanced data. Another advantage of taking a modeling approach to unbalanced data is that by making the effort to treat unbalanced analysis of variance, one can easily handle a wide range of models for nonnormal data, because the same fundamental methods apply. To that end, I have included new chapters on logistic regression,

log-linear models, and time-to-event data. These are placed near the end of the book, not because they are less important, but because the real subject of the book is modeling with linear structures and the methods for measurement data carry over almost immediately.

In early versions of this edition I made extensive comparisons between the methods used here and the balanced ANOVA methods used in the 1996 book. In particular, I emphasized how the newer methods continue to give the same results as the earlier methods when applied to balanced data. While I have toned that down, comparisons still exist. In such comparisons, I do not repeat the details of the balanced analysis given in the earlier book. CRC Press/Chapman & Hall have been kind enough to let me place a version of the 1996 book on my website so that readers can explore the comparisons in detail. Another good thing about having the old book up is that it contains a chapter on confounding and fractional replications in 2^n factorials. I regret having to drop that chapter, but the discussion is based on contrasts for balanced ANOVA and did not really fit the theme of the current edition.

When I was in high school, my two favorite subjects were math and history. On a whim, I made the good choice to major in Math for my BA. I mention my interest in history to apologize (primarily in the same sense that C.S. Lewis was a Christian "apologist") for using so much old data. Unless you are trying to convince 18-year-olds that Statistics is sexy, I don't think the age of the data should matter.

I need to thank Adam Branscum, my coauthor on Christensen et al. (2010). Adam wrote the first drafts of Chapter 7 and Appendix C of that book. Adam's work on Chapter 7 definitely influenced this work and Adam's work on Appendix C is what got me programming in R. This is also a good time to thank the people who have most influenced my career: Wes Johnson, Ed Bedrick, Don Berry, Frank Martin, and the late, great Seymour Geisser. My colleague Yan Lu taught out of a prepublication version of the book, and, with her students, pointed out a number of issues. Generally, the first person whose opinions and help I sought was my son Fletcher.

After the effort to complete this book, I'm feeling as unbalanced as the data being analyzed.

Specifics

I think of the book as something to use in the traditional Master's level year-long course on regression and analysis of variance. If one needed to actually separate the material into a regression course and an ANOVA course, the regression material is in Chapters 6–11 and 20–23. Chapters 12–19 are traditionally viewed as ANOVA. But I much prefer to use both regression and ANOVA ideas when examining the generalized linear models of Chapters 20–22. Well-prepared students could begin with Chapter 3 and skip to Chapter 6. By well-prepared, I tautologically mean students who are already familiar with Chapters 1, 2, 4, and 5.

For less well-prepared students, obviously I would start at the beginning and deemphasize the more difficult topics. This is what I have done when teaching data analysis to upper division Statistics students and graduate students from other fields. I have tried to isolate more difficult material into clearly delineated (sub)sections. In the first semester of such a course, I would skip the end of Chapter 8, include the beginning of Chapter 12, and let time and student interest determine how much of Chapters 9, 10, and 13 to cover. But the book wasn't written to be a text for such a course; it is written to address unbalanced multi-factor ANOVA.

The book requires very little pre-knowledge of math, just algebra, but does require that one not be afraid of math. It does not perform calculus, but it discusses that integrals provide areas under curves and, in an appendix, gives the integral formulae for means and variances. It largely avoids matrix algebra but presents enough of it to enable the matrix approach to linear models to be introduced. For a regression-ANOVA course, I would supplement the material after Chapter 11 with occasional matrix arguments. Any material described as a regression approach to an ANOVA problem lends itself to matrix discussion.

Although the book starts at the beginning mathematically, it is not for the intellectually unsophisticated. By Chapter 2 it discusses the impreciseness of our concepts of populations and how

the deletion of outliers must change those concepts. Chapter 2 also discusses the "murky" transformation from a probability interval to a confidence interval and the differences between significance testing, Neyman–Pearson hypothesis testing, and Bayesian methods. Because a lot of these ideas are subtle, and because people learn best from specifics to generalities rather than the other way around, Chapter 3 reiterates much of Chapter 2 but for general linear models. Most of the remainder of the book can be viewed as the application of Chapter 3 to specific data structures. Well-prepared students could start with Chapter 3 despite occasional references made to results in the first two chapters.

Chapter 4 considers two-sample data. Perhaps its most unique feature is, contrary to what seems popular in introductory Statistics these days, the argument that testing equality of means for two independent samples provides much less information when the variances are different than when they are the same.

Chapter 5 exists because I believe that if you teach one- and two-sample continuous data problems, you have a duty to present their discrete data analogs. Having gone that far, it seemed silly to avoid analogs to one-way ANOVA. I do not find the one-way ANOVA F test for equal group means to be all that useful. Contrasts contain more interesting information. The last two sections of Chapter 5 contain, respectively, discrete data analogs to one-way ANOVA and a method of extracting information similar to contrasts.

Chapters 6, 7, and 8 provide tools for exploring the relationship between a single dependent variable and a single measurement (continuous) predictor. A key aspect of the discussion is that the methods in Chapters 7 and 8 extend readily to more general linear models, i.e., those involving categorial and/or multiple predictors. The title of Chapter 8 arises from my personal research interest in testing lack of fit for linear models and the recognition of its relationship to nonparametric regression.

Chapters 9, 10, and 11 examine features associated with multiple regression. Of particular note are new sections on modeling interaction through generalized additive models and on lasso regression. I consider these important concepts for serious students of Statistics. The last of these chapters is where the book's use of matrices is focused. The discussion of principal component regression is located here, not because the discussion uses matrices, but because the discussion requires matrix knowledge to understand.

The rest of the book involves categorical predictor variables. In particular, *the material after Chapter 13 is the primary reason for writing this edition*. The first edition focused on multifactor balanced data and looking at contrasts, not only in main effects but contrasts within two- and three-factor interactions. This edition covers the same material for unbalanced data.

Chapters 12 and 13 cover one-way analysis of variance (ANOVA) models and multiple comparisons but with an emphasis on the ideas needed when examining multiple categorical predictors. Chapter 12 involves one categorical predictor much like Chapter 6 involved one continuous predictor.

Chapter 14 examines the use of two categorial predictors, i.e., two-way ANOVA. It also introduces the concept of homologous factors. Chapter 15 looks at models with one continuous and one categorical factor, analysis of covariance. Chapter 16 considers models with three categorical predictors.

Chapters 17 and 18 introduce the main ideas of experimental design. Chapter 17 introduces a wide variety of standard designs and concepts of design. Chapter 18 introduces the key idea of defining treatments with factorial structure. The unusual aspect of these chapters is that the analyses presented apply when data are missing from the original design.

Chapter 19 introduces the analysis of dependent data. The primary emphasis is on the analysis of split-plot models. A short discussion is also given of multivariate analysis. Both of these methods require groups of observations that are independent of other groups but that are dependent within the groups. Both methods require balance within the groups but the groups themselves can be unbalanced. Subsection 19.2.1 even introduces a method for dealing with unbalance within groups.

It seems to have become popular to treat fixed and random effects models as merely two options