LINEAR MODELS IN STATISTICS

BICENTENNIAL

1807

WILEY

2007

BICENTENNIAL

THE WILEY BICENTENNIAL–KNOWLEDGE FOR GENERATIONS

Each generation has its unique needs and aspirations. When Charles Wiley first opened his small printing shop in lower Manhattan in 1807, it was a generation of boundless potential searching for an identity. And we were there, helping to define a new American literary tradition. Over half a century later, in the midst of the Second Industrial Revolution, it was a generation focused on building the future. Once again, we were there, supplying the critical scientific, technical, and engineering knowledge that helped frame the world. Throughout the 20th Century, and into the new millennium, nations began to reach out beyond their own borders and a new international community was born. Wiley was there, expanding its operations around the world to enable a global exchange of ideas, opinions, and know-how.

For 200 years, Wiley has been an integral part of each generation's journey, enabling the flow of information and understanding necessary to meet their needs and fulfill their aspirations. Today, bold new technologies are changing the way we live and learn. Wiley will be there, providing you the must-have knowledge you need to imagine new worlds, new possibilities, and new opportunities.

Generations come and go, but you can always count on Wiley to provide you the knowledge you need, when and where you need it!

William J. Pesce

Peter Booth Wiley

WILLIAM J. PESCE
PRESIDENT AND CHIEF EXECUTIVE OFFICER

PETER BOOTH WILEY
CHAIRMAN OF THE BOARD

LINEAR MODELS IN STATISTICS

Second Edition

Alvin C. Rencher and G. Bruce Schaalje
Department of Statistics, Brigham Young University, Provo, Utah

WILEY-INTERSCIENCE
A JOHN WILEY & SONS, INC., PUBLICATION

Published by John Wiley & Sons, Inc., Hoboken, New Jersey
Published simultaneously in Canada

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in variety of electronic formats. Some content that appears in print may not be available in electronic formats. For more information about Wiley products, visit our web site at www.wiley.com.

Wiley Bicentennial Logo: Richard J. Pacifico

Library of Congress Cataloging-in-Publication Data:
Rencher, Alvin C., 1934-
Linear models in statistics/Alvin C. Rencher, G. Bruce Schaalje. – 2nd ed.
p. cm.
Includes bibliographical references.
ISBN 978-0-471-75498-5 (cloth)
1. Linear models (Statistics) I. Schaalje, G. Bruce. II. Title.
QA276.R425 2007
519.5′35–dc22

2007024268

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

CONTENTS

Preface xiii

1 Introduction 1

1.1 Simple Linear Regression Model 1
1.2 Multiple Linear Regression Model 2
1.3 Analysis-of-Variance Models 3

2 Matrix Algebra 5

2.1 Matrix and Vector Notation 5
2.1.1 Matrices, Vectors, and Scalars 5
2.1.2 Matrix Equality 6
2.1.3 Transpose 7
2.1.4 Matrices of Special Form 7
2.2 Operations 9
2.2.1 Sum of Two Matrices or Two Vectors 9
2.2.2 Product of a Scalar and a Matrix 10
2.2.3 Product of Two Matrices or Two Vectors 10
2.2.4 Hadamard Product of Two Matrices or Two Vectors 16
2.3 Partitioned Matrices 16
2.4 Rank 19
2.5 Inverse 21
2.6 Positive Definite Matrices 24
2.7 Systems of Equations 28
2.8 Generalized Inverse 32
2.8.1 Definition and Properties 33
2.8.2 Generalized Inverses and Systems of Equations 36
2.9 Determinants 37
2.10 Orthogonal Vectors and Matrices 41
2.11 Trace 44
2.12 Eigenvalues and Eigenvectors 46
2.12.1 Definition 46
2.12.2 Functions of a Matrix 49

2.12.3 Products 50
2.12.4 Symmetric Matrices 51
2.12.5 Positive Definite and Semidefinite Matrices 53
2.13 Idempotent Matrices 54
2.14 Vector and Matrix Calculus 56
2.14.1 Derivatives of Functions of Vectors and Matrices 56
2.14.2 Derivatives Involving Inverse Matrices and Determinants 58
2.14.3 Maximization or Minimization of a Function of a Vector 60

3 Random Vectors and Matrices 69

3.1 Introduction 69
3.2 Means, Variances, Covariances, and Correlations 70
3.3 Mean Vectors and Covariance Matrices for Random Vectors 75
3.3.1 Mean Vectors 75
3.3.2 Covariance Matrix 75
3.3.3 Generalized Variance 77
3.3.4 Standardized Distance 77
3.4 Correlation Matrices 77
3.5 Mean Vectors and Covariance Matrices for Partitioned Random Vectors 78
3.6 Linear Functions of Random Vectors 79
3.6.1 Means 80
3.6.2 Variances and Covariances 81

4 Multivariate Normal Distribution 87

4.1 Univariate Normal Density Function 87
4.2 Multivariate Normal Density Function 88
4.3 Moment Generating Functions 90
4.4 Properties of the Multivariate Normal Distribution 92
4.5 Partial Correlation 100

5 Distribution of Quadratic Forms in y 105

5.1 Sums of Squares 105
5.2 Mean and Variance of Quadratic Forms 107
5.3 Noncentral Chi-Square Distribution 112
5.4 Noncentral *F* and *t* Distributions 114
5.4.1 Noncentral *F* Distribution 114
5.4.2 Noncentral *t* Distribution 116
5.5 Distribution of Quadratic Forms 117
5.6 Independence of Linear Forms and Quadratic Forms 119

6 Simple Linear Regression 127

6.1 The Model 127
6.2 Estimation of $\boldsymbol{\beta}_0$, $\boldsymbol{\beta}_1$, and σ^2 128
6.3 Hypothesis Test and Confidence Interval for $\boldsymbol{\beta}_1$ 132
6.4 Coefficient of Determination 133

7 Multiple Regression: Estimation 137

7.1 Introduction 137
7.2 The Model 137
7.3 Estimation of $\boldsymbol{\beta}$ and σ^2 141
7.3.1 Least-Squares Estimator for $\boldsymbol{\beta}$ 145
7.3.2 Properties of the Least-Squares Estimator $\hat{\boldsymbol{\beta}}$ 141
7.3.3 An Estimator for σ^2 149
7.4 Geometry of Least-Squares 151
7.4.1 Parameter Space, Data Space, and Prediction Space 152
7.4.2 Geometric Interpretation of the Multiple Linear Regression Model 153
7.5 The Model in Centered Form 154
7.6 Normal Model 157
7.6.1 Assumptions 157
7.6.2 Maximum Likelihood Estimators for $\boldsymbol{\beta}$ and σ^2 158
7.6.3 Properties of $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ 159
7.7 R^2 in Fixed-x Regression 161
7.8 Generalized Least-Squares: cov($\mathbf{y}$) = $\sigma^2\mathbf{V}$ 164
7.8.1 Estimation of $\boldsymbol{\beta}$ and σ^2 when cov(y) = $\sigma^2\mathbf{V}$ 164
7.8.2 Misspecification of the Error Structure 167
7.9 Model Misspecification 169
7.10 Orthogonalization 174

8 Multiple Regression: Tests of Hypotheses and Confidence Intervals 185

8.1 Test of Overall Regression 185
8.2 Test on a Subset of the $\boldsymbol{\beta}$ Values 189
8.3 F Test in Terms of R^2 196
8.4 The General Linear Hypothesis Tests for H_0: $\mathbf{C}\boldsymbol{\beta} = \mathbf{0}$ and H_0: $\mathbf{C}\boldsymbol{\beta} = \mathbf{t}$ 198
8.4.1 The Test for H_0: $\mathbf{C}\boldsymbol{\beta} = 0$ 198
8.4.2 The Test for H_0: $\mathbf{C}\boldsymbol{\beta} = \boldsymbol{t}$ 203
8.5 Tests on β_j and $\mathbf{a}'\boldsymbol{\beta}$ 204
8.5.1 Testing One β_j or One $\mathbf{a}'\boldsymbol{\beta}$ 204
8.5.2 Testing Several β_j or $\mathbf{a}'_i\boldsymbol{\beta}$ Values 205

8.6 Confidence Intervals and Prediction Intervals 209
8.6.1 Confidence Region for $\boldsymbol{\beta}$ 209
8.6.2 Confidence Interval for β_j 210
8.6.3 Confidence Interval for $\mathbf{a}'\boldsymbol{\beta}$ 211
8.6.4 Confidence Interval for $E(y)$ 211
8.6.5 Prediction Interval for a Future Observation 213
8.6.6 Confidence Interval for σ^2 215
8.6.7 Simultaneous Intervals 215
8.7 Likelihood Ratio Tests 217

9 Multiple Regression: Model Validation and Diagnostics 227

9.1 Residuals 227
9.2 The Hat Matrix 230
9.3 Outliers 232
9.4 Influential Observations and Leverage 235

10 Multiple Regression: Random *x*'s 243

10.1 Multivariate Normal Regression Model 244
10.2 Estimation and Testing in Multivariate Normal Regression 245
10.3 Standardized Regression Coefficents 249
10.4 R^2 in Multivariate Normal Regression 254
10.5 Tests and Confidence Intervals for R^2 258
10.6 Effect of Each Variable on R^2 262
10.7 Prediction for Multivariate Normal or Nonnormal Data 265
10.8 Sample Partial Correlations 266

11 Multiple Regression: Bayesian Inference 277

11.1 Elements of Bayesian Statistical Inference 277
11.2 A Bayesian Multiple Linear Regression Model 279
11.2.1 A Bayesian Multiple Regression Model with a Conjugate Prior 280
11.2.2 Marginal Posterior Density of $\boldsymbol{\beta}$ 282
11.2.3 Marginal Posterior Densities of τ and σ^2 284
11.3 Inference in Bayesian Multiple Linear Regression 285
11.3.1 Bayesian Point and Interval Estimates of Regression Coefficients 285
11.3.2 Hypothesis Tests for Regression Coefficients in Bayesian Inference 286
11.3.3 Special Cases of Inference in Bayesian Multiple Regression Models 286
11.3.4 Bayesian Point and Interval Estimation of σ^2 287

11.4 Bayesian Inference through Markov Chain Monte Carlo Simulation 288
11.5 Posterior Predictive Inference 290

12 Analysis-of-Variance Models 295

12.1 Non-Full-Rank Models 295
12.1.1 One-Way Model 295
12.1.2 Two-Way Model 299
12.2 Estimation 301
12.2.1 Estimation of $\boldsymbol{\beta}$ 302
12.2.2 Estimable Functions of $\boldsymbol{\beta}$ 305
12.3 Estimators 309
12.3.1 Estimators of $\boldsymbol{\lambda}'\boldsymbol{\beta}$ 309
12.3.2 Estimation of σ^2 313
12.3.3 Normal Model 314
12.4 Geometry of Least-Squares in the Overparameterized Model 316
12.5 Reparameterization 318
12.6 Side Conditions 320
12.7 Testing Hypotheses 323
12.7.1 Testable Hypotheses 323
12.7.2 Full-Reduced-Model Approach 324
12.7.3 General Linear Hypothesis 326
12.8 An Illustration of Estimation and Testing 329
12.8.1 Estimable Functions 330
12.8.2 Testing a Hypothesis 331
12.8.3 Orthogonality of Columns of $\mathbf{X}$ 333

13 One-Way Analysis-of-Variance: Balanced Case 339

13.1 The One-Way Model 339
13.2 Estimable Functions 340
13.3 Estimation of Parameters 341
13.3.1 Solving the Normal Equations 341
13.3.2 An Estimator for σ^2 343
13.4 Testing the Hypothesis H_0: $\mu_1 = \mu_2 = \cdots = \mu_k$ 344
13.4.1 Full–Reduced-Model Approach 344
13.4.2 General Linear Hypothesis 348
13.5 Expected Mean Squares 351
13.5.1 Full-Reduced-Model Approach 352
13.5.2 General Linear Hypothesis 354

13.6 Contrasts 357

13.6.1 Hypothesis Test for a Contrast 357
13.6.2 Orthogonal Contrasts 358
13.6.3 Orthogonal Polynomial Contrasts 363

14 Two-Way Analysis-of-Variance: Balanced Case 377

14.1 The Two-Way Model 377
14.2 Estimable Functions 378
14.3 Estimators of $\boldsymbol{\lambda}'\boldsymbol{\beta}$ and σ^2 382

14.3.1 Solving the Normal Equations and Estimating $\boldsymbol{\lambda}'\boldsymbol{\beta}$ 382
14.3.2 An Estimator for σ^2 384

14.4 Testing Hypotheses 385

14.4.1 Test for Interaction 385
14.4.2 Tests for Main Effects 395

14.5 Expected Mean Squares 403

14.5.1 Sums-of-Squares Approach 403
14.5.2 Quadratic Form Approach 405

15 Analysis-of-Variance: The Cell Means Model for Unbalanced Data 413

15.1 Introduction 413
15.2 One-Way Model 415

15.2.1 Estimation and Testing 415
15.2.2 Contrasts 417

15.3 Two-Way Model 421

15.3.1 Unconstrained Model 421
15.3.2 Constrained Model 428

15.4 Two-Way Model with Empty Cells 432

16 Analysis-of-Covariance 443

16.1 Introduction 443
16.2 Estimation and Testing 444

16.2.1 The Analysis-of-Covariance Model 444
16.2.2 Estimation 446
16.2.3 Testing Hypotheses 448

16.3 One-Way Model with One Covariate 449

16.3.1 The Model 449
16.3.2 Estimation 449
16.3.3 Testing Hypotheses 450

16.4 Two-Way Model with One Covariate 457
16.4.1 Tests for Main Effects and Interactions 458
16.4.2 Test for Slope 462
16.4.3 Test for Homogeneity of Slopes 463
16.5 One-Way Model with Multiple Covariates 464
16.5.1 The Model 464
16.5.2 Estimation 465
16.5.3 Testing Hypotheses 468
16.6 Analysis-of-Covariance with Unbalanced Models 473

17 Linear Mixed Models 479

17.1 Introduction 479
17.2 The Linear Mixed Model 479
17.3 Examples 481
17.4 Estimation of Variance Components 486
17.5 Inference for $\boldsymbol{\beta}$ 490
17.5.1 An Estimator for $\boldsymbol{\beta}$ 490
17.5.2 Large-Sample Inference for Estimable Functions of $\boldsymbol{\beta}$ 491
17.5.3 Small-Sample Inference for Estimable Functions of $\boldsymbol{\beta}$ 491
17.6 Inference for the $\mathbf{a}_i$ Terms 497
17.7 Residual Diagnostics 501

18 Additional Models 507

18.1 Nonlinear Regression 507
18.2 Logistic Regression 508
18.3 Loglinear Models 511
18.4 Poisson Regression 512
18.5 Generalized Linear Models 513

Appendix A Answers and Hints to the Problems 517

References 653

Index 663

PREFACE

In the second edition, we have added chapters on Bayesian inference in linear models (Chapter 11) and linear mixed models (Chapter 17), and have upgraded the material in all other chapters. Our continuing objective has been to introduce the theory of linear models in a clear but rigorous format.

In spite of the availability of highly innovative tools in statistics, the main tool of the applied statistician remains the linear model. The linear model involves the simplest and seemingly most restrictive statistical properties: independence, normality, constancy of variance, and linearity. However, the model and the statistical methods associated with it are surprisingly versatile and robust. More importantly, mastery of the linear model is a prerequisite to work with advanced statistical tools because most advanced tools are generalizations of the linear model. The linear model is thus central to the training of any statistician, applied or theoretical.

This book develops the basic theory of linear models for regression, analysis-of-variance, analysis–of–covariance, and linear mixed models. Chapter 18 briefly introduces logistic regression, generalized linear models, and nonlinear models. Applications are illustrated by examples and problems using real data. This combination of theory and applications will prepare the reader to further explore the literature and to more correctly interpret the output from a linear models computer package.

This introductory linear models book is designed primarily for a one-semester course for advanced undergraduates or MS students. It includes more material than can be covered in one semester so as to give an instructor a choice of topics and to serve as a reference book for researchers who wish to gain a better understanding of regression and analysis-of-variance. The book would also serve well as a text for PhD classes in which the instructor is looking for a one-semester introduction, and it would be a good supplementary text or reference for a more advanced PhD class for which the students need to review the basics on their own.

Our overriding objective in the preparation of this book has been *clarity of exposition*. We hope that students, instructors, researchers, and practitioners will find this linear models text more comfortable than most. In the final stages of development, we asked students for written comments as they read each day's assignment. They made many suggestions that led to improvements in readability of the book. We are grateful to readers who have notified us of errors and other suggestions for improvements of the text, and we will continue to be very grateful to readers who take the time to do so for this second edition.

Another objective of the book is to tie up loose ends. There are many approaches to teaching regression, for example. Some books present estimation of regression coefficients for fixed x's only, other books use random x's, some use centered models, and others define estimated regression coefficients in terms of variances and covariances or in terms of correlations. Theory for linear models has been presented using both an algebraic and a geometric approach. Many books present classical (frequentist) inference for linear models, while increasingly the Bayesian approach is presented. We have tried to cover all these approaches carefully and to show how they relate to each other. We have attempted to do something similar for various approaches to analysis-of-variance. We believe that this will make the book useful as a reference as well as a textbook. An instructor can choose the approach he or she prefers, and a student or researcher has access to other methods as well.

The book includes a large number of theoretical problems and a smaller number of applied problems using real datasets. The problems, along with the extensive set of answers in Appendix A, extend the book in two significant ways: (1) the theoretical problems and answers fill in nearly all gaps in derivations and proofs and also extend the coverage of material in the text, and (2) the applied problems and answers become additional examples illustrating the theory. As instructors, we find that having answers available for the students saves a great deal of class time and enables us to cover more material and cover it better. The answers would be especially useful to a reader who is engaging this material outside the formal classroom setting.

The mathematical prerequisites for this book are multivariable calculus and matrix algebra. The review of matrix algebra in Chapter 2 is intended to be sufficiently complete so that the reader with no previous experience can master matrix manipulation up to the level required in this book. Statistical prerequisites include some exposure to statistical theory, with coverage of topics such as distributions of random variables, expected values, moment generating functions, and an introduction to estimation and testing hypotheses. These topics are briefly reviewed as each is introduced. One or two statistical methods courses would also be helpful, with coverage of topics such as t tests, regression, and analysis-of-variance.

We have made considerable effort to maintain consistency of notation throughout the book. We have also attempted to employ standard notation as far as possible and to avoid exotic characters that cannot be readily reproduced on the chalkboard. With a few exceptions, we have refrained from the use of abbreviations and mnemonic devices. We often find these annoying in a book or journal article.

Equations are numbered sequentially throughout each chapter; for example, (3.29) indicates the twenty-ninth numbered equation in Chapter 3. Tables and figures are also numbered sequentially throughout each chapter in the form "Table 7.4" or "Figure 3.2." On the other hand, examples and theorems are numbered sequentially within a section, for example, Theorems 2.2a and 2.2b.

The solution of most of the problems with real datasets requires the use of the computer. We have not discussed command files or output of any particular program, because there are so many good packages available. Computations for the numerical examples and numerical problems were done with SAS. The datasets and SAS command files for all the numerical examples and problems in the text are available on the Internet; see Appendix B.

The references list is not intended to be an exhaustive survey of the literature. We have provided original references for some of the basic results in linear models and have also referred the reader to many up-to-date texts and reference books useful for further reading. When citing references in the text, we have used the standard format involving the year of publication. For journal articles, the year alone suffices, for example, Fisher (1921). But for a specific reference in a book, we have included a page number or section, as in Hocking (1996, p. 216).

Our selection of topics is intended to prepare the reader for a better understanding of applications and for further reading in topics such as mixed models, generalized linear models, and Bayesian models. Following a brief introduction in Chapter 1, Chapter 2 contains a careful review of all aspects of matrix algebra needed to read the book. Chapters 3, 4, and 5 cover properties of random vectors, matrices, and quadratic forms. Chapters 6, 7, and 8 cover simple and multiple linear regression, including estimation and testing hypotheses and consequences of misspecification of the model. Chapter 9 provides diagnostics for model validation and detection of influential observations. Chapter 10 treats multiple regression with random x's. Chapter 11 covers Bayesian multiple linear regression models along with Bayesian inferences based on those models. Chapter 12 covers the basic theory of analysis-of-variance models, including estimability and testability for the overparameterized model, reparameterization, and the imposition of side conditions. Chapters 13 and 14 cover balanced one-way and two-way analysis-of-variance models using an overparameterized model. Chapter 15 covers unbalanced analysis-of-variance models using a cell means model, including a section on dealing with empty cells in two-way analysis-of-variance. Chapter 16 covers analysis of covariance models. Chapter 17 covers the basic theory of linear mixed models, including residual maximum likelihood estimation of variance components, approximate small-sample inferences for fixed effects, best linear unbiased prediction of random effects, and residual analysis. Chapter 18 introduces additional topics such as nonlinear regression, logistic regression, loglinear models, Poisson regression, and generalized linear models.

In our class for first-year master's-level students, we cover most of the material in Chapters 2–5, 7–8, 10–12, and 17. Many other sequences are possible. For example, a thorough one-semester regression and analysis-of-variance course could cover Chapters 1–10, and 12–15.

Al's introduction to linear models came in classes taught by Dale Richards and Rolf Bargmann. He also learned much from the books by Graybill, Scheffé, and Rao. Al expresses thanks to the following for reading the first edition manuscript and making many valuable suggestions: David Turner, John Walker, Joel Reynolds, and Gale Rex Bryce. Al thanks the following students at Brigham Young University (BYU) who helped with computations, graphics, and typing of the first edition: David Fillmore, Candace Baker, Scott Curtis, Douglas Burton, David Dahl, Brenda Price, Eric Hintze, James Liechty, and Joy Willbur. The students in Al's Linear Models class went through the manuscript carefully and spotted many typographical errors and passages that needed additional clarification.

Bruce's education in linear models came in classes taught by Mel Carter, Del Scott, Doug Martin, Peter Bloomfield, and Francis Giesbrecht, and influential short courses taught by John Nelder and Russ Wolfinger.

We thank Bruce's Linear Models classes of 2006 and 2007 for going through the book and new chapters. They made valuable suggestions for improvement of the text. We thank Paul Martin and James Hattaway for invaluable help with LaTex. The Department of Statistics, Brigham Young University provided financial support and encouragement throughout the project.

Second Edition

For the second edition we added Chapter 11 on Bayesian inference in linear models (including Gibbs sampling) and Chapter 17 on linear mixed models.

We also added a section in Chapter 2 on vector and matrix calculus, adding several new theorems and covering the Lagrange multiplier method. In Chapter 4, we presented a new proof of the conditional distribution of a subvector of a multivariate normal vector. In Chapter 5, we provided proofs of the moment generating function and variance of a quadratic form of a multivariate normal vector. The section on the geometry of least squares was completely rewritten in Chapter 7, and a section on the geometry of least squares in the overparameterized linear model was added to Chapter 12. Chapter 8 was revised to provide more motivation for hypothesis testing and simultaneous inference. A new section was added to Chapter 15 dealing with two-way analysis-of-variance when there are empty cells. This material is not available in any other textbook that we are aware of.

This book would not have been possible without the patience, support, and encouragement of Al's wife LaRue and Bruce's wife Lois. Both have helped and supported us in more ways than they know. This book is dedicated to them.

ALVIN C. RENCHER AND G. BRUCE SCHAALJE

Department of Statistics
Brigham Young University
Provo, Utah