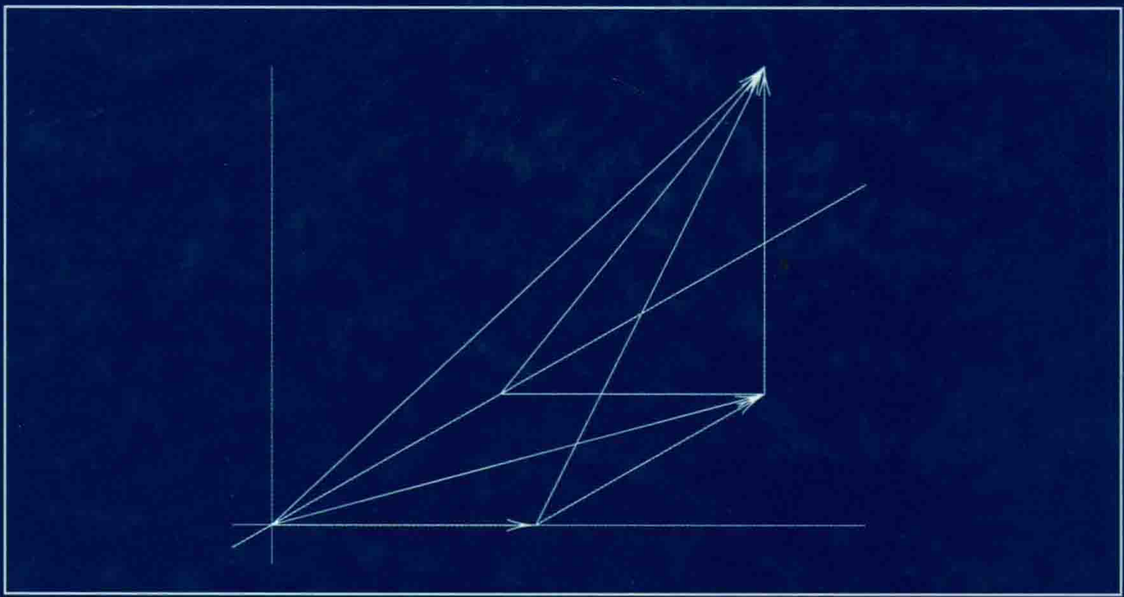
Cambridge Series in Statistical and Probabilistic Mathematics



The Coordinate-Free Approach to Linear Models

Michael J. Wichura

The Coordinate-Free Approach to Linear Models

MICHAEL J. WICHURA

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo

Cambridge University Press
32 Avenue of the Americas, New York, NY 10013-2473, USA

www.cambridge.org
Information on this title: www.cambridge.org/9780521868426

First published 2006

Printed in the United States of America

A catalog record for this publication is available from the British Library.

Library of Congress Cataloging in Publication Data

Wichura, Michael J. (Michael John)
The coordinate-free approach to linear models / Michael J. Wichura.
p. cm. – (Cambridge series on statistical and probabilistic mathematics)
Includes bibliographical references and index.
ISBN-13: 978-0-521-86842-6 (hardback)
ISBN-10: 0-521-86842-4 (hardback)
1. Linear models (Statistics) 2. Analysis of variance. 3. Regression analysis.
4. Analysis of covariance. I. Title. II. Series.
QA279.W53 2006
519.5–dc22 2006004133

ISBN-13 978-0-521-86842-6 hardback
ISBN-10 0-521-86842-4 hardback

The Coordinate-Free Approach to Linear Models

This book is about the coordinate-free, or geometric, approach to the theory of linear models, more precisely, Model I ANOVA and linear regression models with nonrandom predictors in a finite-dimensional setting. This approach is more insightful, more elegant, more direct, and simpler than the more common matrix approach to linear regression, analysis of variance, and analysis of covariance models in statistics. The book discusses the intuition behind and optimal properties of various methods of estimating and testing hypotheses about unknown parameters in the models.

Topics covered include inner product spaces, orthogonal projections, book orthogonal spaces, Tjur experimental designs, basic distribution theory, the geometric version of the Gauss-Markov theorem, optimal and nonoptimal properties of Gauss-Markov, Bayes and shrinkage estimators under the assumption of normality, the optimal properties of F-tests, and the analysis of covariance and missing observations.

Michael J. Wichura has 37 years of teaching experience in the Department of Statistics at the University of Chicago. He has served as an associate editor for the *Annals of Probability* and was the database editor for the *Current Index to Statistics* from 1995 to 2000. He is the author of the PiCTeX macros (for drawing pictures in TeX) and the PiCTeX manual and also of the TABLE macros and the TABLE manual.

This series of high-quality upper-division textbooks and expository monographs covers all aspects of stochastic applicable mathematics. The topics range from pure and applied statistics to probability theory, operations research, optimization, and mathematical programming. The books contain clear presentations of new developments in the field and also of the state of the art in classical methods. While emphasizing rigorous treatment of theoretical methods, the books also contain applications and discussions of new techniques made possible by advances in computational practice.

Already Published

1. *Bootstrap Methods and Their Application*, by A. C. Davison and D. V. Hinkley
2. *Markov Chains*, by J. Norris
3. *Asymptotic Statistics*, by A. W. van der Vaart
4. *Wavelet Methods for Time Series Analysis*, by Donald B. Percival and Andrew T. Walden
5. *Bayesian Methods*, by Thomas Leonard and John S. J. Hsu
6. *Empirical Processes in M-Estimation*, by Sara van de Geer
7. *Numerical Methods of Statistics*, by John F. Monahan
8. *A User's Guide to Measure Theoretic Probability*, by David Pollard
9. *The Estimation and Tracking of Frequency*, by B. G. Quinn and E. J. Hannan
10. *Data Analysis and Graphics Using R*, by John Maindonald and John Braun
11. *Statistical Models*, by A. C. Davison
12. *Semiparametric Regression*, by D. Ruppert, M. P. Wand, and R. J. Carroll
13. *Exercises in Probability*, by Loic Chaumont and Marc Yor
14. *Statistical Analysis of Stochastic Processes in Time*, by J. K. Lindsey
15. *Measure Theory and Filtering*, by Lakhdar Aggoun and Robert Elliott
16. *Essentials of Statistical Inference*, by G. A. Young and R. L. Smith
17. *Elements of Distribution Theory*, by Thomas A. Severini
18. *Statistical Mechanics of Disordered Systems*, by Anton Bovier

In memoriam

William H. Kruskal

1919–2005

PREFACE

When I was a graduate student in the mid-1960s, the mathematical theory underlying analysis of variance and regression became clear to me after I read a draft of William Kruskal's monograph on the so-called coordinate-free, or geometric, approach to these subjects. Alas, with Kruskal's demise, this excellent treatise will never be published.

From time to time during the 1970s, 80s, and early 90s, I had the good fortune to teach the coordinate-free approach to linear models, more precisely, to Model I analysis of variance and linear regression with nonrandom predictors. While doing so, I evolved my own set of lecture notes, presented here. With regard to inspiration and content, my debt to Kruskal is clear. However, my notes differ from Kruskal's in many ways. To mention just a few, my notes are intended for a one- rather than three-quarter course. The notes are aimed at statistics graduate students who are already familiar with the basic concepts of linear algebra, such as linear subspaces and linear transformations, and who have already had some exposure to the matricial formulation of the GLM, perhaps through a methodology course, and who are interested in the underlying theory. I have also included Tjur experimental designs and some of the highlights of the optimality theory for estimation and testing in linear models under the assumption of normality, feeling that the elegant setting provided by the coordinate-free approach is a natural one in which to place these jewels of mathematical statistics. As he alluded to in his conversation with Zabell (1994), Kruskal always wished that he could have brought "his book, his potential book, his unborn book" to life. Out of deference to Kruskal, who was my colleague here at the University of Chicago, I have not until now made my notes public.

For motivation, Chapter 1 presents an example illustrating Kruskal's claim in his 1961 Berkeley Symposium paper that "the coordinate-free approach ... permits a simpler, more general, more elegant, and more direct treatment of the general theory ... than do its notational counterparts, the matrix and scalar approaches." I hope that as s/he works through the book, the reader will be more and more convinced that this is indeed the case.

The last section of Chapter 2 reviews the "elementary" concepts from linear algebra which the reader is assumed to know already. The first five sections of that chapter develop the "nonelementary" tools we need, such as (finite-dimensional, real) inner product spaces, orthogonal projections, the spectral theorem for self-adjoint linear transformations, and the representation of linear and bilinear functionals. Section 2.3 uses the notions of book orthogonal subspaces and orthogonal projections, along with the inclusion and intersection of subspaces, to discuss Tjur experimental designs, thereby giving a unified treatment of the algebra underlying the models one usually encounters in a first course in analysis of variance, and much more.

Chapter 3 develops basic distribution theory for random vectors taking values in inner product spaces — the first- and second-moment structures of such vectors, and the key fact that if one splits a spherical normal random vector up into its components in mutually orthogonal subspaces, then those components are independent and have themselves spherical normal distributions within their respective subspaces.

The geometric version of the Gauss-Markov theorem is discussed in Chapter 4, from the point of view of both estimating linear functionals of the unknown mean vector and estimating the mean vector itself. These results are based just on assumptions about the first and second moment structures of the data vector. For an especially nice example of how the geometric viewpoint is more insightful than the matricial one, be sure to work the "four-penny problem" in Exercise 4.2.17.

Estimation under the assumption of normality is taken up in Chapter 5. In some respects, Gauss-Markov estimators are optimal; for example, they are minimum variance unbiased. However, in other respects they are not optimal. Indeed, Bayesian considerations lead naturally to the James-Stein shrinkage type estimators, which can significantly outperform GMEs in terms of mean square error.

Again under the assumption of normality, F-testing of null hypotheses and the related issue of interval estimation are taken up in Chapter 6. Chapters 7 and 8 deal with the analysis of covariance and missing observations, respectively.

The book is written at the level of Halmos's *Finite-Dimensional Vector Spaces* (but Halmos is not a prerequisite). Thus the reader will on the one hand need to be comfortable with the yoga of definitions, theorems, and proofs, but on the other hand be comforted by knowing that the abstract ideas will be illustrated by concrete examples and presented with (what I hope are) some insightful comments. To get a feeling for the coordinate-free approach before embarking on a serious study of this book, you might find it helpful to first read one or more of the brief elementary nontechnical expositions of the subject that have appeared in *The American Statistician*, for example, Herr (1980), Bryant (1984), or Saville and Wood (1986). From the perspective of mathematical statistics, there are some very elegant results, and some notable surprises, connected with the optimality theory for Gauss-Markov estimation and F-testing under the assumption of normality. In the sections that deal with these matters — in particular Sections 5.4, 5.5, 6.3, 6.4, and 6.6 — the mathematics is somewhat harder than elsewhere, corresponding to the greater depth of the theory.

Each of the chapters following the Introduction contains numerous exercises, along with a problem set that develops some topic complementing the material in that chapter. Altogether there are about 200 exercises. Most of them are easy but, I hope, instructive. I typically devote most of the class time to having students present solutions to the exercises. Some exercises foreshadow what is to come, by covering a special case of material that will be presented in full generality later. Moreover, the assertions of some exercises are appealed to later in the text. If you are working through the book on your own, you should at least read over each exercise, even if you do not work things out. The problem sets, several of which are based on journal articles, are harder than the exercises and require a sustained effort for their completion.

The students in my courses have typically worked through the whole book in one quarter, but that is admittedly a brisk pace. One semester would be less demanding. For a short course, you could concentrate on the parts of the book that flesh out the outline of the coordinate-free viewpoint that Kruskal set out in Section 2 of his aforementioned Berkeley Symposium paper. That would involve: Chapter 1, for motivation; Sections 2.1, 2.2, 2.5, and 2.7 for notation and basic results from linear algebra; Sections 3.1–3.8 for distribution theory; Sections 4.1–4.7 for the properties of Gauss-Markov estimators; Sections 5.1 and 5.2 for estimation under the assumption of normality; and Sections 6.1, 6.2, and 6.5 for hypothesis testing and interval estimation under normality.

Various graduate students, in particular Neal Thomas and Nathaniel Schenker, have made many comments that have greatly improved this book. My thanks go to all of them, and also to David Van Dyke and Peter Meyer for suggesting how easy/hard each of the exercises is. Thanks are also due to Mitzi Nakatsuka for her help in converting the notes to TEX, and to Persi Diaconis for his advice and encouragement.

Michael J. Wichura
University of Chicago

CONTENTS

CHAPTER 1

INTRODUCTION

In this chapter we introduce and contrast the matricial and geometric formulations of the so-called general linear model and introduce some notational conventions.

1. Orientation

Recall the classical framework of the *general linear model (GLM)*. One is given an n-dimensional random vector $\boldsymbol{Y}^{n\times 1} = (Y_1 \ldots, Y_n)^T$, perhaps multivariate normally distributed, with covariance matrix $(\mathrm{Cov}(Y_i, Y_j))^{n\times n} = \sigma^2 \boldsymbol{I}^{n\times n}$ and mean vector $\boldsymbol{\mu}^{n\times 1} = E(\boldsymbol{Y}) = (EY_1, \ldots, EY_n)^T$ of the form

$$\boldsymbol{\mu} = \boldsymbol{X\beta},$$

where $\boldsymbol{X}^{n\times p}$ is known and σ^2 and $\boldsymbol{\beta}^{p\times 1} = (\beta_1, \ldots, \beta_p)^T$ are unknown; in addition, the β_i's may be subject to linear constraints $\boldsymbol{R\beta} = 0$, where $\boldsymbol{R}^{c\times p}$ is known. $\boldsymbol{X}$ is called the *design*, or *regression, matrix*, and $\boldsymbol{\beta}$ is called the *parameter vector.*

1.1 Example. In the classical *two-sample problem*, one has

$$\boldsymbol{X}^T = \Big(\underbrace{\begin{matrix} 1 & 1 & \ldots & 1 \\ 0 & 0 & \ldots & 0 \end{matrix}}_{n_1 \text{ times}} \; \underbrace{\begin{matrix} 0 & 0 & \ldots & 0 \\ 1 & 1 & \ldots & 1 \end{matrix}}_{n_2 \text{ times}}\Big) \quad \text{and} \quad \boldsymbol{\beta} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix},$$

that is,

$$E(Y_i) = \begin{cases} \mu_1, & \text{if } 1 \le i \le n_1, \\ \mu_2, & \text{if } n_1 < i \le n_1 + n_2 = n. \end{cases}$$

•

1.2 Example. In *simple linear regression*, one has

$$\boldsymbol{X}^T = \begin{pmatrix} 1 & 1 & \ldots & 1 \\ x_1 & x_2 & \ldots & x_n \end{pmatrix} \quad \text{and} \quad \boldsymbol{\beta} = \begin{pmatrix} a \\ b \end{pmatrix},$$

that is,

$$E(Y_i) = a + bx_i \quad \text{for } i = 1, \ldots, n.$$

•

Typical problems are the estimation of linear combinations of the β_i's, testing that some such linear combinations are 0 (or some other prescribed value), and the estimation of σ^2.

1.3 Example. In the two-sample problem, one is often interested in estimating the difference $\mu_2 - \mu_1$ or in testing the null hypothesis that $\mu_1 = \mu_2$. •

1.4 Example. In simple linear regression, one seeks estimates of the intercept a and slope b and may want to test, for example, the hypothesis that $b = 0$ or the hypothesis that $b = 1$. •

If you have had some prior statistical training, you may well have already encountered the resolution of these problems. You may know, for example, that provided $\boldsymbol{X}$ is of full rank and no linear constraints are imposed on $\boldsymbol{\beta}$, the *best (minimum variance) linear unbiased estimator (BLUE)* of $\sum_{1\le i\le p} c_i\beta_i$ is $\sum_{1\le i\le p} c_i\hat{\beta}_i$, where

$$(\hat{\beta}_1, \ldots, \hat{\beta}_p)^T = \boldsymbol{C}\boldsymbol{X}^T\boldsymbol{Y}, \quad \text{with} \quad \boldsymbol{C} = \boldsymbol{A}^{-1}, \quad \boldsymbol{A} = \boldsymbol{X}^T\boldsymbol{X};$$

this is called the *Gauss-Markov theorem.*

In this book we will be studying the GLM from a geometric point of view, using linear algebra in place of matrix algebra. Although we will not reach any conclusions that could not be obtained solely by matrix techniques, the basic ideas will emerge more clearly. With the added intuitive feeling and mathematical insight this provides, one will be better able to understand old results and formulate and prove new ones.

From a geometric perspective, the GLM may be described as follows, using some terms that will be defined in subsequent chapters. One is given a *random vector* Y taking values in some given *inner product space* $(V, \langle\cdot,\cdot\rangle)$. It is assumed that Y has a *weakly spherical covariance operator* and the *mean* μ of Y lies in a given *manifold* M of V; for purposes of testing, it is further assumed that Y is *normally distributed.* One desires to estimate μ (or *linear functionals* of μ) and to test hypotheses such as $\mu \in M_0$, where M_0 is a given *submanifold* of M. The Gauss-Markov theorem says that the BLUE of the linear functional $\psi(\mu)$ is $\psi(\hat{\mu})$, where $\hat{\mu}$ is the *orthogonal projection* of Y onto M. As we will see, this geometric description of the problem encompasses the matricial formulation of the GLM not only as it is set out above (take, for example, $V = \mathbb{R}^n$, $\langle\cdot,\cdot\rangle$ = dot-product, $Y = \boldsymbol{Y}$, $\mu = \boldsymbol{\mu}$, and M = the subspace of $\mathbb{R}^n$ spanned by the columns of the design matrix $\boldsymbol{X}$), but also in cases where $\boldsymbol{X}$ is of less than full rank and/or linear constraints are imposed on the β_i's.

2. An illustrative example

To illustrate the differences between the matricial and geometric approaches, we compare the ways in which one establishes the independence of

$$\bar{Y} = \hat{\mu} = \frac{\sum_{1\le i\le n} Y_i}{n} \quad \text{and} \quad s^2 = \hat{\sigma}^2 = \frac{1}{n-1}\sum_{1\le i\le n}(Y_i - \bar{Y})^2$$

in the one-sample problem

$$\boldsymbol{Y}^{n\times 1} \sim N(\mu\boldsymbol{e}, \sigma^2\boldsymbol{I}^{n\times n}) \quad \text{with} \quad \boldsymbol{e} = (1,1,\ldots,1)^T. \tag{2.1}$$

(The vector $\boldsymbol{e}$ is called the *equiangular vector.*)

The classical matrix proof, which uses some facts about multivariate normal distributions, runs like this. Let $\boldsymbol{B}^{n\times n} = (b_{ij})$ be the matrix

$$\begin{pmatrix} \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} & 0 & 0 & \cdots & 0 & 0 \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{-2}{\sqrt{6}} & 0 & \cdots & 0 & 0 \\ \frac{1}{\sqrt{12}} & \frac{1}{\sqrt{12}} & \frac{1}{\sqrt{12}} & \frac{-3}{\sqrt{12}} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ \frac{1}{\sqrt{n(n-1)}} & \frac{1}{\sqrt{n(n-1)}} & \frac{1}{\sqrt{n(n-1)}} & \frac{1}{\sqrt{n(n-1)}} & \cdots & \frac{1}{\sqrt{n(n-1)}} & \frac{-(n-1)}{\sqrt{n(n-1)}} \end{pmatrix}.$$

Note that the rows (and columns) of $\boldsymbol{B}$ are orthonormal $\left(\sum_{1\le k\le n} b_{ik}b_{jk} = \delta_{ij} \equiv \begin{cases} 1, & \text{if } i = j \\ 0, & \text{if } i \ne j \end{cases}\right)$ and that the first row is $\frac{1}{\sqrt{n}}\,\boldsymbol{e}^T$. Set

$$\boldsymbol{Z} = \boldsymbol{B}\boldsymbol{Y}.$$

Then

$$\boldsymbol{Z} \sim N(\boldsymbol{\nu}, \boldsymbol{\Sigma})$$

with

$$\boldsymbol{\nu} = \boldsymbol{B}(\mu\boldsymbol{e}) = \mu\boldsymbol{B}\boldsymbol{e} = (\sqrt{n}\,\mu, 0, \ldots, 0)^T$$

and

$$\boldsymbol{\Sigma} = \boldsymbol{B}(\sigma^2\boldsymbol{I})\boldsymbol{B}^T = \sigma^2\boldsymbol{B}\boldsymbol{B}^T = \sigma^2\boldsymbol{I};$$

that is, $Z_1, Z_2, \ldots, Z_n$ are independent normal random variables, each with variance σ^2, $E(Z_1) = \sqrt{n}\,\mu$, and $E(Z_j) = 0$ for $2 \le j \le n$. Moreover,

$$Z_1 = \frac{1}{\sqrt{n}} \sum_{1\le i\le n} Y_i = \sqrt{n}\,\bar{Y}, \text{ or } \bar{Y} = \frac{Z_1}{\sqrt{n}},$$

while

$$\begin{aligned}(n-1)s^2 &= \sum_{1\le i\le n}(Y_i - \bar{Y})^2 = \sum_{1\le i\le n} Y_i^2 - n\bar{Y}^2 \\ &= \sum_{1\le i\le n} Y_i^2 - Z_1^2 = \sum_{1\le i\le n} Z_i^2 - Z_1^2 = \sum_{2\le i\le n} Z_i^2\end{aligned} \tag{2.2}$$

because

$$\sum_{1\le i\le n} Z_i^2 = \boldsymbol{Z}^T\boldsymbol{Z} = \boldsymbol{Y}^T\boldsymbol{B}^T\boldsymbol{B}\boldsymbol{Y} = \boldsymbol{Y}^T\boldsymbol{Y} = \sum_{1\le i\le n} Y_i^2.$$

This gives the independence of $\bar{Y}$ and s^2, and it is an easy step to get the marginal distributions: $\bar{Y} \sim N(\mu, \sigma^2/n)$ and $(n-1)s^2/\sigma^2 \sim \chi^2_{n-1}$.

What is the nature of the transformation $\boldsymbol{Z} = \boldsymbol{B}\boldsymbol{Y}$? Let $\boldsymbol{b}_1 = \boldsymbol{e}/\sqrt{n}$, $\boldsymbol{b}_2, \ldots, \boldsymbol{b}_n$ denote the transposes of the rows of $\boldsymbol{B}$. The coordinates of $\boldsymbol{Y} = \sum_{1\le j\le n} C_j \boldsymbol{b}_j$ with respect to this new orthonormal basis for $\mathbb{R}^n$ are given by

$$C_i = \boldsymbol{b}_i^T\boldsymbol{Y} = Z_i, \quad i = 1, \ldots, n.$$

The effect of the change of coordinates $\boldsymbol{Y} \to \boldsymbol{Z}$ is to split $\boldsymbol{Y}$ into its components along, and orthogonal to, the equiangular vector $\boldsymbol{e}$.

Now I will show you the geometric proof, which uses some properties of (weakly) spherical normal random vectors taking values in an inner product space $(V, \langle\cdot,\cdot\rangle)$, here ($\mathbb{R}^n$, dot-product). The assumptions imply that $\boldsymbol{Y}$ is spherical normally distributed about its mean $E(\boldsymbol{Y})$ and $E(\boldsymbol{Y})$ lies in the manifold M spanned by $\boldsymbol{e}$. Let P_M denote orthogonal projection onto M and Q_M orthogonal projection onto the orthogonal complement $M^\perp$ of M. Basic distribution theory says that $P_M\boldsymbol{Y}$ and $Q_M\boldsymbol{Y}$ are independent. But

$$P_M\boldsymbol{Y} = \frac{\langle \boldsymbol{e}, \boldsymbol{Y}\rangle}{\langle \boldsymbol{e}, \boldsymbol{e}\rangle}\boldsymbol{e} = \bar{Y}\boldsymbol{e} \tag{2.3}$$

and

$$Q_M\boldsymbol{Y} = \boldsymbol{Y} - P_M\boldsymbol{Y} = \boldsymbol{Y} - \bar{Y}\boldsymbol{e} = (Y_1 - \bar{Y}, \ldots, Y_n - \bar{Y})^T;$$

it follows that $\bar{Y}$ and $(n-1)s^2 = \sum_{1\le i\le n}(Y_i - \bar{Y})^2 = \|Q_M\boldsymbol{Y}\|^2$ are independent. Again it is an easy matter to get the marginal distributions.

To my way of thinking, granted the technical apparatus, the second proof is clearer, being more to the point. The first proof does the same things, but (to the uninitiated) in an obscure manner.

3. Notational conventions

The chapters are organized into sections. Within each section of the current chapter, enumerated items are numbered consecutively in the form

$$(\textit{section\_number.item\_number}).$$

References to items in a different chapter take the expanded form

$$(\textit{chapter\_number.section\_number.item\_number}).$$

For example, (2.4) refers to the 4$^{\text{th}}$ numbered item (which may be an example, exercise, theorem, formula, or whatever) in the 2$^{\text{nd}}$ section of the current chapter, while (6.1.3) refers to the 3$^{\text{rd}}$ numbered item in the 1$^{\text{st}}$ section of the 6$^{\text{th}}$ chapter.

Each exercise is assigned a difficulty level using the syntax

Exercise $[d]$,

where d is an integer in the range 1 to 5 — the larger is d, the harder the exercise. The value of d depends both on the intrinsic difficulty of the exercise and the length of time needed to write up the solution.

To help distinguish between the matricial and geometric points of view, matrices, including row and column vectors, are written in ***italic boldface*** type while linear transformations and elements of abstract vector spaces are written simply in *italic* type. We speak, for example, of the design matrix $\boldsymbol{X}$ but of vectors v and w in an inner product space V.

The end of a proof is marked by a ∎, of an example by a •, of an exercise by a ◇, and of a part of problem set by a ∘.