Linear Statistical Models

JAMES H. STAPLETON
Michigan State University

A Wiley-Interscience Publication
JOHN WILEY & SONS, INC.
New York • Chichester • Brisbane • Toronto • Singapore

This Page Intentionally Left Blank

Linear Statistical Models

THE WILEY SERIES IN PROBABILITY AND STATISTICS

Established by WALTER A. SHEWHART and SAMUEL S. WILKS

A complete list of the titles in this series appears at the end of this volume

Linear Statistical Models

JAMES H. STAPLETON
Michigan State University

A Wiley-Interscience Publication
JOHN WILEY & SONS, INC.
New York • Chichester • Brisbane • Toronto • Singapore

This text is printed on acid-free paper.

Library of Congress Cataloging in Publication Data:

Stapleton, James, 1931–
Linear statistical models / James Stapleton.
p. cm.—(Wiley series in probability and statistics. Probability and statistics section)
"A Wiley–Interscience publication."
ISBN 0-471-57150-4 (acid-free)
1. Linear models (Statistics) I. Title. II. Series.
III. Series: Wiley series in probability and statistics. Probability and statistics.
QA279.S695 1995
519.5'38—dc20 94-39384

To Alicia, who, through all the years, never expressed
a doubt that this would someday be completed,
despite the many doubts of the author.

This Page Intentionally Left Blank

Contents

This Page Intentionally Left Blank

Preface

The first seven chapters of this book were developed over a period of about 20 years for the course Linear Statistical Models at Michigan State University. They were first distributed in longhand (those former students may still be suffering the consequences), then typed using a word processor some eight or nine years ago. The last chapter, on frequency data, is the result of a summer course, offered every three or four years since 1980.

Linear statistical models are mathematical models which are linear in the unknown parameters, and which include a random error term. It is this error term which makes the models statistical. These models lead to the methodology usually called *multiple regression* or *analysis of variance*, and have wide applicability to the physical, biological, and social sciences, to agriculture and business, and to engineering.

The linearity makes it possible to study these models from a vector space point of view. The vectors **Y** of observations are represented as arrays written in a form convenient for intuition, rather than necessarily as column or row vectors. The geometry of these vector spaces has been emphasized because the author has found that the intuition it provides is vital to the understanding of the theory. Pictures of the vectors spaces have been added for their intuitive value. In the author's opinion this geometric viewpoint has not been sufficiently exploited in current textbooks, though it is well understood by those doing research in the field. For a brief discussion of the history of these ideas see Herr (1980).

Bold print is used to denote vectors, as well as linear transformations. The author has found it useful for classroom boardwork to use an arrow notation above the symbol to distinguish vectors, and to encourage students to do the same, at least in the earlier part of the course.

Students studying these notes should have had a one-year course in probability and statistics at the post-calculus level, plus one course on linear algebra. The author has found that most such students can handle the matrix algebra used here, but need the material on inner products and orthogonal projections introduced in Chapter 1.

Chapter 1 provides examples and introduces the linear algebra necessary for later chapters. One section is devoted to a brief history of the early development of least squares theory, much of it written by Stephen Stigler (1986).

Chapter 2 is devoted to methods of study of random vectors. The multivariate normal, chi-square, t and F distributions, central and noncentral, are introduced.

Chapter 3 then discusses the linear model, and presents the basic theory necessary to regression analysis and the analysis of variance, including confidence intervals, the Gauss–Markov Theorem, power, and multiple and partial correlation coefficients. It concludes with a study of a SAS multiple regression printout.

Chapter 4 is devoted to a more detailed study of multiple regression methods, including sections on transformations, analysis of residuals, and on asymptotic theory. The last two sections are devoted to robust methods and to the bootstrap. Much of this methodology has been developed over the last 15 years and is a very active topic of research.

Chapter 5 discusses simultaneous confidence intervals: Bonferroni, Scheffé, Tukey, and Bechhofer.

Chapter 6 turns to the analysis of variance, with two- and three-way analyses of variance. The geometric point of view is emphasized.

Chapter 7 considers some miscellaneous topics, including random component models, nested designs, and partially balanced incomplete block designs.

Chapter 8, the longest, discusses the analysis of frequency, or categorical data. Though these methods differ significantly in the distributional assumptions of the models, it depends strongly on the linear representations, common to the theory of the first seven chapters.

Computations illustrating the theory were done using APL*Plus (Magnugistics, Inc.), S-Plus (Statistical Sciences, Inc.), and SAS (SAS Institute, Inc.). Graphics were done using S-Plus.). To perform simulations, and to produce graphical displays, the author recommends that the reader use a mathematical language which makes it easy to manipulate vectors and matrices.

For the linear models course the author teaches at Michigan State University only Section 2.3, Projections of Random Variables, and Section 3.9, Further Decomposition of Subspaces, are omitted from Chapters 1, 2, and 3. From Chapter 4 only Section 4.1, Linearizing Transformations, and one or two other sections are usually discussed. From Chapter 5 the Bonferroni, Tukey, and Scheffé simultaneous confidence interval methods are covered. From Chapter 6 only the material on the analysis of covariance (Section 6.6) is omitted, though relatively little time is devoted to three-way analysis of variance (Section 6.5). One or two sections of Chapter 7, Miscellaneous Other Models, are usually chosen for discussion. Students are introduced to S-Plus early in the semester, then use it for the remainder of the semester for numerical work.

A course on the analysis of frequency data could be built on Sections 1.1, 1.2, 1.3, 2.1, 2.2, 2.3, 2.4 (if students have not already studied these topics), and, of course, Chapter 8.

The author thanks Virgil Anderson, retired professor from Purdue University, now a statistical consultant, from whom he first learned of the analysis of variance and the design of experiments. He also thanks Professor James Hannan, from whom he first learned of the geometric point of view, and Professors Vaclav Fabian and Dennis Gilliland for many valuable conversations. He is grateful to Sharon Carson and to Loretta Ferguson, who showed much patience as they typed several early versions. Finally, he thanks the students who have tried to read the material, and who found (he hopes) a large percentage of the errors in those versions.

JAMES STAPLETON

This Page Intentionally Left Blank

CHAPTER 1

Linear Algebra, Projections

1.1 INTRODUCTION

Suppose that each element of a population possesses a numerical characteristic x, and another numerical characteristic y. It is often desirable to study the relationship between two such variables x and y in order to better understand how values of x affect y, or to predict y, given the value of x. For example, we may wish to know the effect of amount x of fertilizer per square meter on the yield y of a crop in pounds per square meter. Or we might like to know the relationship between a man's height y and that of his father x.

For each value of the independent variable x, the dependent variable Y may be supposed to have a probability distribution with mean $g(x)$. Thus, for example, $g(0.9)$ is the expected yield of a crop using fertilizer level $x = 0.9$ (kgms/m^2).

Definition 1.1.1: For each $x \in D$ suppose Y is a random variable with distribution depending on x. Then

$$g(x) = E(Y|x) \qquad \text{for} \quad x \in D$$

is the regression function for Y on x.

Often the domain D will be a subset of the real line, or even the whole real line. However, D could also be a finite set, say $\{1, 2, 3\}$, or a countably infinite set $\{1, 2, \ldots\}$. The experimenter or statistician would like to determine the function g, using sample data consisting of pairs (x_i, y_i) for $i = 1, \ldots, n$. Unfortunately, the number of possible functions $g(x)$ is so large that in order to make headway certain simplifying models for the form of $g(x)$ must be adopted. If it is supposed that $g(x)$ is of the form $g(x) = A + Bx + Cx^2$ or $g(x) = A2^x + B$ or $g(x) = A \log x + B$, etc., then the problem is reduced to one of identifying a few parameters, here labeled as A, B, C. In each of the three forms for $g(x)$ given above, g is linear in these parameters.

In one of the simplest cases we might consider a model for which $g(x) = C + Dx$, where C and D are unknown parameters. The problem of estimating

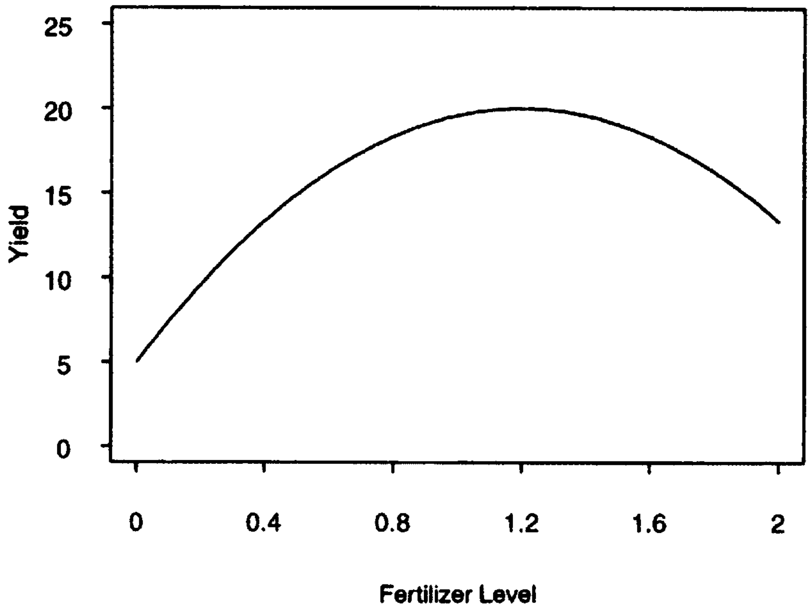

FIGURE 1.1 Regression of yield on fertilizer level.

$g(x)$ then becomes the simpler one of estimating the two parameters C and D. This model may not be a good approximation of the true regression function, and, if possible, should be checked for validity. The crop yield as a function of fertilizer level may well have the form in Figure 1.1.

The regression function g would be better approximated by a second degree polynomial $g(x) = A + Bx + Cx^2$. However, if attention is confined to the 0.7 to 1.3 range, the regression function is approximately linear, and the simplifying model $g(x) = C + Dx$, called the simple linear regression model, may be used.

In attempting to understand the relationship between a person's height Y and the heights of his/her father (x_1) and mother (x_2) and the person's sex (x_3), we might suppose

$$E(Y|x_1, x_2, x_3) = g(x_1, x_2, x_3) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3, \qquad (1.1.1)$$

where x_3 is 1 for males, 0 for females, and β_0, β_1, β_2, β_3 are unknown parameters. Thus a brother would be expected to be β_3 taller than his sister. Again, this model, called a *multiple regression model*, can only be an approximation of the true regression function, valid over a limited range of values of x_1, x_2. A more complex model might suppose

$$g(x_1, x_2, x_3) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1^2 + \beta_5 x_2^2 + \beta_6 x_1 x_2.$$

Table 1.1.1 Height Data

Indiv.	Y	x_1	x_2	x_3
1	68.5	70	62	1
2	72.5	73	66	1
3	70.0	68	67	1
4	71.0	72	64	1
5	65.0	66	60	1
6	64.5	71	63	0
7	67.5	74	68	0
8	61.5	65	65	0
9	63.5	70	64	0
10	63.5	69	65	0

This model is nonlinear in (x_1, x_2, x_3), but linear in the β's. It is the linearity in the β's which makes this model a *linear* statistical model.

Consider the model (1.1.1), and suppose we have data of Table 1.1.1 on (Y, x_1, x_2, x_3) for 10 individuals. These data were collected in a class taught by the author. Perhaps the student can collect similar data in his or her class and compare results.

The statistical problem is to determine estimates $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$ so that the resulting function $\hat{g}(x_1, x_2, x_3) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$ is in some sense a good approximation of $g(x_1, x_2, x_3)$. For this purpose it is convenient to write the model in vector form:

$$E(\mathbf{Y}) = \beta_0 \mathbf{x}_0 + \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + \beta_3 \mathbf{x}_3,$$

where $\mathbf{x}_0$ is the vector of all ones, and $\mathbf{y}$ and $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$ are the column vectors in Table 1.1.1.

This formulation of the model suggests that linear algebra may be an important tool in the analysis of linear statistical models. We will therefore review such material in the next section, emphasizing geometric aspects.

1.2 VECTORS, INNER PRODUCTS, LENGTHS

Let Ω be the collection of all n-tuples of real numbers for a positive integer n. In applications Ω will be the sample space of all possible values of the observation vector $\mathbf{y}$. Though Ω will be in one-to-one correspondence to Euclidean n-space, it will be convenient to consider elements of Ω as arrays all of the same configuration, not necessarily column or row vectors. For example, in application to what is usually called one-way analysis of variance, we might have 3, 4 and 2 observations on three different levels of some treatment effect. Then we might take

$$\mathbf{y} = \begin{bmatrix} y_{11} & y_{12} & y_{13} \\ y_{21} & y_{22} & y_{23} \\ y_{31} & y_{32} & \\ & y_{42} & \end{bmatrix}$$

and Ω the collection of all such y. While we could easily reform y into a column vector, it is often convenient to preserve the form of y. The term "n-tuple" means that the elements of a vector $\mathbf{y} \in \Omega$ are ordered. A vector y may be considered to be a real-valued function on $\{1, \ldots, n\}$.

Ω becomes a linear space if we define $a\mathbf{y}$ for any $\mathbf{y} \in \Omega$ and any real number a to be the element of Ω given by multiplying each component of Ω by a, and if for any two elements $\mathbf{y}_1, \mathbf{y}_2 \in \Omega$ we define $\mathbf{y}_1 + \mathbf{y}_2$ to be the vector in Ω whose ith component is the sum of the ith components of $\mathbf{y}_1$ and $\mathbf{y}_2$, for $i = 1, \ldots, n$.

Ω becomes an inner product space if for each $\mathbf{x}, \mathbf{y} \in \Omega$ we define the function

$$h(\mathbf{x}, \mathbf{y}) = \sum_1^n x_i y_i,$$

where $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$. If Ω is the collection of n-dimensional column vectors then $h(\mathbf{x}, \mathbf{y}) = \mathbf{x}'\mathbf{y}$, in matrix notation. The inner product $h(\mathbf{x}, \mathbf{y})$ is usually written simply as $(\mathbf{x}, \mathbf{y})$, and we will use this notation. The inner product is often called the dot product, written in the form $\mathbf{x} \cdot \mathbf{y}$. Since there is a small danger of confusion with the pair $(\mathbf{x}, \mathbf{y})$, we will use bold parentheses to emphasize that we mean the inner product. Since bold symbols are not easily indicated on a chalkboard or in student notes, it is important that the meaning will almost always be clear from the context. The inner product has the properties:

$$(\mathbf{x}, \mathbf{y}) = (\mathbf{y}, \mathbf{x})$$

$$(a\mathbf{x}, \mathbf{y}) = a(\mathbf{x}, \mathbf{y})$$

$$(\mathbf{x}_1 + \mathbf{x}_2, \mathbf{y}) = (\mathbf{x}_1, \mathbf{y}) + (\mathbf{x}_2, \mathbf{y})$$

for all vectors, and real numbers a.

We define $\|\mathbf{x}\|^2 = (\mathbf{x}, \mathbf{x})$ and call $\|\mathbf{x}\|$ the (Euclidean) *length of* x. Thus $\mathbf{x} = (3, 4, 12)$ has length 13.

The *distance* between vectors x and y is the length of $\mathbf{x} - \mathbf{y}$. Vectors x and y are said to be *orthogonal* if $(\mathbf{x}, \mathbf{y}) = 0$. We write $\mathbf{x} \perp \mathbf{y}$.

For example, if the sample space is the collection of arrays mentioned above, then

$$\mathbf{x} = \left[\begin{array}{ccc} 1 & 0 & 0 \\ 2 & 0 & 0 \\ 3 & 0 & \\ & 0 & \end{array}\right] \quad \text{and} \quad \mathbf{y} = \left[\begin{array}{ccc} 0 & 1 & 0 \\ 0 & 3 & 0 \\ 0 & 5 & \\ & -1 & \end{array}\right]$$

are orthogonal, with squared lengths 14 and 36. For Ω the collection of 3-tuples, $(2, 3, 1) \perp (-1, 1, -1)$.

The following theorem is perhaps the most important of the entire book. We credit it to Pythagorus (sixth century B.C.), though he would not, of course, have recognized it in this form.

Pythagorean Theorem: Let $\mathbf{v}_1, \ldots, \mathbf{v}_k$ be mutually orthogonal vectors in Ω. Then

$$\left\|\sum_1^k \mathbf{v}_i\right\|^2 = \sum_1^k \|\mathbf{v}_i\|^2$$

Proof:

$$\left\|\sum_1^k \mathbf{v}_i\right\|^2 = \left(\sum_{i=1}^k \mathbf{v}_i, \sum_{j=1}^k \mathbf{v}_j\right) = \sum_{i=1}^k \sum_{j=1}^k (\mathbf{v}_i, \mathbf{v}_j) = \sum_{i=1}^k (\mathbf{v}_i, \mathbf{v}_i)$$

$$= \sum_{i=1}^k \|\mathbf{v}_i\|^2. \qquad \square$$

Definition 1.2.1: The *projection* of a vector $\mathbf{y}$ on a vector $\mathbf{x}$ is the vector $\hat{\mathbf{y}}$ such that

1. $\hat{\mathbf{y}} = b\mathbf{x}$ for some constant b
2. $(\mathbf{y} - \hat{\mathbf{y}}) \perp \mathbf{x}$ (equivalently, $(\hat{\mathbf{y}}, \mathbf{x}) = (\mathbf{y}, \mathbf{x})$)

Equivalently, $\hat{\mathbf{y}}$ is the projection of $\mathbf{y}$ on the subspace of all vectors of the form $a\mathbf{x}$, the subspace spanned by $\mathbf{x}$ (Figure 1.2). To be more precise, these properties define othogonal projection. We will use the word projection to mean orthogonal projection. We write $p(\mathbf{y}|\mathbf{x})$ to denote this projection. Students should not confuse this will conditional probability.

Let us try to find the constant b. We need $(\hat{\mathbf{y}}, \mathbf{x}) = (b\mathbf{x}, \mathbf{x}) = b(\mathbf{x}, \mathbf{x}) = (\mathbf{y}, \mathbf{x})$. Hence, if $\mathbf{x} = \mathbf{0}$, any b will do. Otherwise, $b = (\mathbf{y}, \mathbf{x})/\|\mathbf{x}\|^2$. Thus,

$$\hat{\mathbf{y}} = \begin{cases} \mathbf{0} & \text{for} \quad \mathbf{x} = 0 \\ [(\mathbf{y}, \mathbf{x})/\|\mathbf{x}\|^2]\mathbf{x}, & \text{otherwise} \end{cases}$$

Here $\mathbf{0}$ is the vector of all zeros. Note that if $\mathbf{x}$ is replaced by a multiple $a\mathbf{x}$ of $\mathbf{x}$, for $a \neq 0$ then $\hat{\mathbf{y}}$ remains the same though the coefficient b is replaced by b/a.

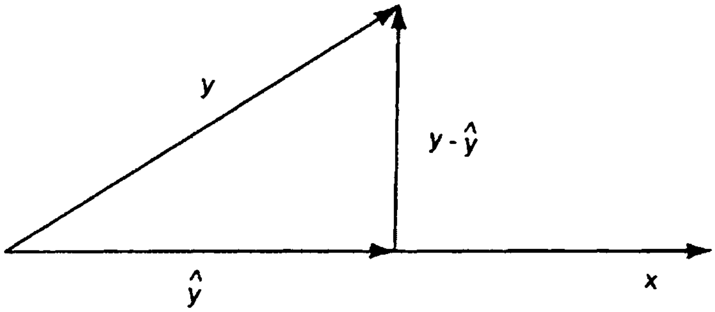

FIGURE 1.2

Example 1.2.1: Let $\mathbf{x} = \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix}$, $\mathbf{y} = \begin{pmatrix} 1 \\ -6 \\ 5 \end{pmatrix}$. Then $(\mathbf{x}, \mathbf{y}) = 18$, $\|\mathbf{x}\|^2 = 6$, $b = 18/6 = 3$, $\hat{\mathbf{y}} = 3\mathbf{x} = \begin{pmatrix} 3 \\ -6 \\ 3 \end{pmatrix}$, $\mathbf{y} - \hat{\mathbf{y}} = \begin{pmatrix} -2 \\ 0 \\ 2 \end{pmatrix} \perp \mathbf{x}$.

Theorem 1.2.1: Among all multiples $a\mathbf{x}$ of $\mathbf{x}$, the projection $\hat{\mathbf{y}}$ of $\mathbf{y}$ on $\mathbf{x}$ is the closest vector to $\mathbf{y}$.

***Proof*:** Since $(\mathbf{y} - \hat{\mathbf{y}}) \perp (\hat{\mathbf{y}} - a\mathbf{x})$ and $(\mathbf{y} - a\mathbf{x}) = (\mathbf{y} - \hat{\mathbf{y}}) + (\hat{\mathbf{y}} - a\mathbf{x})$, it follows that

$$\|\mathbf{y} - a\mathbf{x}\|^2 = \|\mathbf{y} - \hat{\mathbf{y}}\|^2 + \|\hat{\mathbf{y}} - a\mathbf{x}\|^2.$$

This is obviously minimum for $a\mathbf{x} = \hat{\mathbf{y}}$. □

Since $\hat{\mathbf{y}} \perp (\mathbf{y} - \hat{\mathbf{y}})$ and $\mathbf{y} = \hat{\mathbf{y}} + (\mathbf{y} - \hat{\mathbf{y}})$, the Pythagorean Theorem implies that $\|\mathbf{y}\|^2 = \|\hat{\mathbf{y}}\|^2 + \|\mathbf{y} - \hat{\mathbf{y}}\|^2$. Since $\|\hat{\mathbf{y}}\|^2 = b^2\|\mathbf{x}\|^2 = (\mathbf{y}, \mathbf{x})^2/\|\mathbf{x}\|^2$, this implies that $\|\mathbf{y}\|^2 \geq (\mathbf{y}, \mathbf{x})^2/\|\mathbf{x}\|^2$, with equality if and only if $\|\mathbf{y} - \hat{\mathbf{y}}\| = 0$, i.e., $\mathbf{y}$ is a multiple of $\mathbf{x}$. This is the famous *Cauchy-Schwarz Inequality*, usually written as $(\mathbf{y}, \mathbf{x})^2 \leq \|\mathbf{y}\|^2\|\mathbf{x}\|^2$. The inequality is best understood as the result of the equality implied by the Pythagorean Theorem.

Definition 1.2.2: Let A be a subset of the indices of the components of a vector space Ω. The indicator of A is the vector $\mathbf{I}_A \in \Omega$, with components which are 1 for indices in A, and 0 otherwise.

The *projection* $\hat{\mathbf{y}}_A$ of $\mathbf{y}$ on the vector $\mathbf{I}_A$ is therefore $b\mathbf{I}_A$ for $b = (\mathbf{y}, \mathbf{I}_A)/\|\mathbf{I}_A\|^2 = \left(\sum_{i \in A} y_i\right)/N(A)$, where $N(A)$ is the number of indices in A. Thus, $b = \bar{y}_A$, the mean of the y-values with components in A. For example, if Ω is the space of 4-component row vectors, $\mathbf{y} = (3, 7, 8, 13)$, and A is the indicator of the second and fourth components, $p(\mathbf{y}|\mathbf{I}_A) = (0, 10, 0, 10)$.

Problem 1.2.1: Let Ω be the collection of all 5-tuples of the form $\mathbf{y} = \begin{pmatrix} y_{11} & y_{21} & \\ y_{12} & y_{22} & y_{31} \end{pmatrix}$. Let $\mathbf{x} = \begin{pmatrix} 1 & 0 & \\ 2 & 1 & 3 \end{pmatrix}$, $\mathbf{y} = \begin{pmatrix} 5 & 1 & \\ 9 & 4 & 11 \end{pmatrix}$.

(a) Find $(\mathbf{x}, \mathbf{y})$, $\|\mathbf{x}\|^2$, $\|\mathbf{y}\|^2$, $\hat{\mathbf{y}} = p(\mathbf{y}|\mathbf{x})$, and $\mathbf{y} - \hat{\mathbf{y}}$. Show that $\mathbf{x} \perp (\mathbf{y} - \hat{\mathbf{y}})$, and $\|\mathbf{y}\|^2 = \|\hat{\mathbf{y}}\|^2 + \|\mathbf{y} - \hat{\mathbf{y}}\|^2$.

(b) Let $\mathbf{w} = \begin{pmatrix} -2 & 1 & \\ 0 & 2 & 0 \end{pmatrix}$ and $\mathbf{z} = 3\mathbf{x} + 2\mathbf{w}$. Show that $(\mathbf{w}, \mathbf{x}) = 0$ and that $\|\mathbf{z}\|^2 = 9\|\mathbf{x}\|^2 + 4\|\mathbf{w}\|^2$. (Why must this be true?)

(c) Let $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$ be the indicators of the first, second and third columns. Find $p(\mathbf{y}|\mathbf{x}_i)$ for $i = 1, 2, 3$.

Problem 1.2.2: Is projection a linear transformation in the sense that $p(c\mathbf{y}|\mathbf{x}) = cp(\mathbf{y}|\mathbf{x})$ for any real number c? Prove or disprove. What is the relationship between $p(\mathbf{y}|\mathbf{x})$ and $p(\mathbf{y}|c\mathbf{x})$ for $c \neq 0$?

Problem 1.2.3: Let $\|\mathbf{x}\|^2 > 0$. Use calculus to prove that $\|\mathbf{y} - b\mathbf{x}\|^2$ is minimum for $b = (\mathbf{y}, \mathbf{x})/\|\mathbf{x}\|^2$.

Problem 1.2.4: Prove the converse of the Pythagorean Theorem. That is, $\|\mathbf{x} + \mathbf{y}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2$ implies that $\mathbf{x} \perp \mathbf{y}$.

Problem 1.2.5: Sketch a picture and prove the parallelogram law:

$$\|\mathbf{x} + \mathbf{y}\|^2 + \|\mathbf{x} - \mathbf{y}\|^2 = 2(\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2)$$

1.3 SUBSPACES, PROJECTIONS

We begin the discussion of subspaces and projections with a number of definitions of great importance to our subsequent discussion of linear models. Almost all of the definitions and the theorems which follow are usually included in a first course in matrix or linear algebra. Such courses do not always include discussion of orthogonal projection, so this material may be new to the student.

Definition 1.3.1: A *subspace* of Ω is a subset of Ω which is closed under addition and scalar multiplication.

That is, $V \subset \Omega$ is a subspace if for every $\mathbf{x} \in V$ and every scalar a, $a\mathbf{x} \in V$ and if for every $\mathbf{v}_1, \mathbf{v}_2 \in V$, $\mathbf{v}_1 + \mathbf{v}_2 \in V$.

Definition 1.3.2: Let $\mathbf{x}_1, \ldots, \mathbf{x}_k$ be k vectors in an n-dimensional vector space. The subspace *spanned* by $\mathbf{x}_1, \ldots, \mathbf{x}_k$ is the collection of all vectors

$$\mathbf{y} = b_1\mathbf{x}_1 + \cdots + b_k\mathbf{x}_k$$

for all real numbers $b_1, \ldots, b_k$. We denote this subspace by $\mathscr{L}(\mathbf{x}_1, \ldots, \mathbf{x}_k)$.

Definition 1.3.3: Vectors $\mathbf{x}_1, \ldots, \mathbf{x}_k$ are *linearly independent* if $\sum_1^k b_i\mathbf{x}_i = \mathbf{0}$ implies $b_i = 0$ for $i = 1, \ldots, k$.

Definition 1.3.4: A *basis* for a subspace V of Ω is a set of linearly independent vectors which span V.

The proofs of Theorems 1.3.1 and 1.3.2 are omitted. Readers are referred to any introductory book on linear algebra.

Theorem 1.3.1: Every basis for a subspace V on Ω has the same number of elements.

Definition 1.3.5: The dimension of a subspace V of Ω is the number of elements in each basis.

Theorem 1.3.2: Let $\mathbf{v}_1, \ldots, \mathbf{v}_k$ be linearly independent vectors in a subspace V of dimension d. Then $d \geq k$.

Comment: Theorem 1.3.2 implies that if $\dim(V) = d$ then any collection of $d + 1$ or more vectors in V must be linearly dependent. In particular, any collection of $n + 1$ vectors in the n-component space Ω are linearly dependent.

Definition 1.3.6: A vector $\mathbf{y}$ is *orthogonal* to a subspace V of Ω if $\mathbf{y}$ is orthogonal to all vectors in V. We write $\mathbf{y} \perp V$.

Problem 1.3.1: Let Ω be the space of all 4-component row vectors. Let $\mathbf{x}_1 = (1, 1, 1, 1)$, $\mathbf{x}_2 = (1, 1, 0, 0)$, $\mathbf{x}_3 = (1, 0, 1, 0)$, $\mathbf{x}_4 = (7, 4, 9, 6)$. Let $V_2 = \mathscr{L}(\mathbf{x}_1, \mathbf{x}_2)$, $V_3 = \mathscr{L}(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)$ and $V_4 = \mathscr{L}(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4)$.

(a) Find the dimensions of V_2 and V_3.

(b) Find bases for V_2 and V_3 which contain vectors with as many zeros as possible.

(c) Give a vector $\mathbf{z} \neq \mathbf{0}$ which is orthogonal to all vectors in V_3.

(d) Since $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$, $\mathbf{z}$ are linearly independent, $\mathbf{x}_4$ is expressible in the form $\sum_1^3 b_i\mathbf{x}_i + c\mathbf{z}$. Show that $c = 0$ and hence that $\mathbf{x}_4 \in V_3$, by determining $(\mathbf{x}_4, \mathbf{z})$. What is $\dim(V_4)$?

(e) Give a simple verbal description of V_3.

Problem 1.3.2: Consider the space Ω of arrays $\begin{bmatrix} y_{11} & y_{21} & y_{31} \\ y_{12} & y_{22} & \\ y_{13} & & \end{bmatrix}$ and define $\mathbf{C}_1, \mathbf{C}_2, \mathbf{C}_3$ to be the indicators of the columns. Let $V = \mathscr{L}(\mathbf{C}_1, \mathbf{C}_2, \mathbf{C}_3)$.

(a) What properties must $\mathbf{y}$ satisfy in order that $\mathbf{y} \in V$? In order that $\mathbf{y} \perp V$?

(b) Find a vector $\mathbf{y}$ which is orthogonal to V.

The following definition is perhaps the most important in the entire book. It serves as the foundation of all the least squares theory to be discussed in Chapters 1, 2, and 3.

Definition 1.3.7: The projection of a vector $\mathbf{y}$ on a subspace V of Ω is the vector $\hat{\mathbf{y}} \in V$ such that $(\mathbf{y} - \hat{\mathbf{y}}) \perp V$. The vector $\mathbf{y} - \hat{\mathbf{y}} = \mathbf{e}$ will be called the *residual* vector for $\mathbf{y}$ relative to V.

Comment: The condition $(\mathbf{y} - \hat{\mathbf{y}}) \perp V$ is equivalent to $(\mathbf{y} - \hat{\mathbf{y}}, \mathbf{x}) = 0$ for all $\mathbf{x} \in V$. Therefore, in seeking the projection $\hat{\mathbf{y}}$ of $\mathbf{y}$ on a subspace V we seek a vector $\hat{\mathbf{y}}$ in V which has the same inner products as $\mathbf{y}$ with all vectors in V (Figure 1.3).

If vectors $\mathbf{x}_1, \ldots, \mathbf{x}_k$ span a subspace V then a vector $\mathbf{z} \in V$ is the projection of $\mathbf{y}$ on V if $(\mathbf{z}, \mathbf{x}_i) = (\mathbf{y}, \mathbf{x}_i)$ for all i, since for any vector $\mathbf{x} = \sum_{j=1}^{k} b_j \mathbf{x}_j \in V$, this implies that

$$(\mathbf{z}, \mathbf{x}) = \sum b_j(\mathbf{z}, \mathbf{x}_j) = (\mathbf{y}, \sum b_j \mathbf{x}_j) = (\mathbf{y}, \mathbf{x}).$$

It is tempting to attempt to compute the projection $\hat{\mathbf{y}}$ of $\mathbf{y}$ on V by simply summing the projections $\hat{\mathbf{y}}_i = p(\mathbf{y}|\mathbf{x}_i)$. As we shall see, this is only possible in some very special cases.

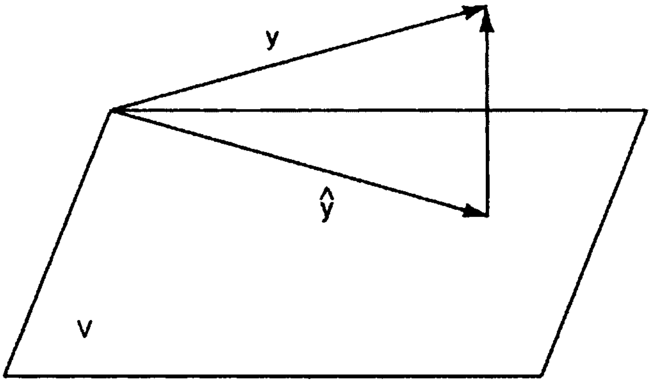

FIGURE 1.3

At this point we have not established the legitimacy of Definition 1.3.7. Does such a vector $\hat{\mathbf{y}}$ always exist and, if so, is it unique? We do know that the projection onto a one-dimensional subspace, say onto $V = \mathscr{L}(\mathbf{x})$, for $\mathbf{x} \neq \mathbf{0}$, does exist and is unique. In fact

$$\hat{\mathbf{y}} = [(\mathbf{y}, \mathbf{x})/\|\mathbf{x}\|^2]\mathbf{x} \quad \text{if} \quad \mathbf{x} \neq \mathbf{0}.$$

Example 1.3.1: Consider the 6-component space Ω of the problem above, and let $V = \mathscr{L}(\mathbf{C}_1, \mathbf{C}_2, \mathbf{C}_3)$. Let $\mathbf{y} = \begin{pmatrix} 6 & 4 & 7 \\ 10 & 8 & \\ 5 & & \end{pmatrix}$. It is easy to show that the vector $\hat{\mathbf{y}} = \sum p(\mathbf{y}|\mathbf{C}_i) = 7\mathbf{C}_1 + 6\mathbf{C}_2 + 7\mathbf{C}_3$ satisfies the conditions for a projection onto V. As will soon be shown the representation of $\hat{\mathbf{y}}$ as the sum of projections on linearly independent vectors spanning the space is possible because $\mathbf{C}_1$, $\mathbf{C}_2$, and $\mathbf{C}_3$ are mutually othogonal.

We will first show uniqueness of the projection. Existence is more difficult. Suppose $\hat{\mathbf{y}}_1$ and $\hat{\mathbf{y}}_2$ are two such projections of $\mathbf{y}$ onto V. Then $\hat{\mathbf{y}}_1 - \hat{\mathbf{y}}_2 \in V$ and $(\hat{\mathbf{y}}_1 - \hat{\mathbf{y}}_2) = (\mathbf{y} - \hat{\mathbf{y}}_2) - (\mathbf{y} - \hat{\mathbf{y}}_1)$ is orthogonal to all vectors in V, in particular to itself. Thus $\|\hat{\mathbf{y}}_1 - \hat{\mathbf{y}}_2\|^2 = (\hat{\mathbf{y}}_1 - \hat{\mathbf{y}}_2, \hat{\mathbf{y}}_1 - \hat{\mathbf{y}}_2) = \mathbf{0}$, implying $\hat{\mathbf{y}}_1 - \hat{\mathbf{y}}_2 = \mathbf{0}$, i.e., $\hat{\mathbf{y}}_1 = \hat{\mathbf{y}}_2$.

We have yet to show that $\hat{\mathbf{y}}$ always exists. In the case that it does exist (we will show that it always exists) we will write $\hat{\mathbf{y}} = p(\mathbf{y}|V)$.

If we are fortunate enough to have an *orthogonal* basis (a basis of mutually orthogonal vectors) for a given subspace V, it is easy to find the projection. Students are warned that that method applies *only* for an orthogonal basis. We will later show that all subspaces possess such orthogonal bases, so that the projection $\hat{\mathbf{y}} = p(\mathbf{y}|V)$ always exists.

Theorem 1.3.3: Let $\mathbf{v}_1, \ldots, \mathbf{v}_k$ be an orthogonal basis for V, subspace of Ω. Then

$$p(\mathbf{y}|V) = \sum_{i=1}^{k} p(\mathbf{y}|\mathbf{v}_i)$$

Proof: Let $\hat{\mathbf{y}}_i = p(\mathbf{y}|\mathbf{v}_i) = b_i\mathbf{v}_i$ for $b_i = (\mathbf{y}, \mathbf{v}_i)/\|\mathbf{v}_i\|^2$. Since $\hat{\mathbf{y}}_i$ is a scalar multiple of $\mathbf{v}_i$, it is orthogonal to $\mathbf{v}_j$ for $j \neq i$. From the comment on the previous page, we need only show that $\sum \hat{\mathbf{y}}_i$ and $\mathbf{y}$, have the same inner product with each $\mathbf{v}_j$, since this implies that they have the same inner product with all $\mathbf{x} \in V$. But

$$\left(\sum_i \hat{\mathbf{y}}_i, \mathbf{v}_j\right) = \sum_i b_i(\mathbf{v}_i, \mathbf{v}_j) = b_j\|\mathbf{v}_j\|^2 = (\mathbf{y}, \mathbf{v}_j). \qquad \square$$

Example 1.3.2: Let

$$\mathbf{y} = \begin{pmatrix} 7 \\ 0 \\ 2 \end{pmatrix}, \qquad \mathbf{v}_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \qquad \mathbf{v}_2 = \begin{pmatrix} 2 \\ -1 \\ -1 \end{pmatrix}, \qquad V = \mathscr{L}(\mathbf{v}_1, \mathbf{v}_2).$$

Then $\mathbf{v}_1 \perp \mathbf{v}_2$ and

$$p(\mathbf{y}\,|\,V) = \hat{\mathbf{y}} = p(\mathbf{y}\,|\,\mathbf{v}_1) + p(\mathbf{y}\,|\,\mathbf{v}_2) = \left(\frac{9}{3}\right)\mathbf{v}_1 + \left(\frac{12}{6}\right)\mathbf{v}_2 = \begin{pmatrix} 3 \\ 3 \\ 3 \end{pmatrix} + \begin{pmatrix} 4 \\ -2 \\ -2 \end{pmatrix} = \begin{pmatrix} 7 \\ 1 \\ 1 \end{pmatrix}.$$

Then $(\mathbf{y}, \mathbf{v}_1) = 9$, $(\mathbf{y}, \mathbf{v}_2) = 12$, $(\hat{\mathbf{y}}, \mathbf{v}_1) = 9$, and $(\hat{\mathbf{y}}, \mathbf{v}_1) = 12$. The residual vector is $\mathbf{y} - \hat{\mathbf{y}} = \begin{pmatrix} 0 \\ -1 \\ 1 \end{pmatrix}$, which is orthogonal to V.

Would this same procedure have worked if we replaced this orthogonal basis $\mathbf{v}_1$, $\mathbf{v}_2$ for V by a nonorthogonal basis? To experiment, let us leave $\mathbf{v}_1$ in the new basis, but replace $\mathbf{v}_2$ by $\mathbf{v}_3 = 2\mathbf{v}_1 - \mathbf{v}_2$. Note that $\mathscr{L}(\mathbf{v}_1, \mathbf{v}_3) = \mathscr{L}(\mathbf{v}_1, \mathbf{v}_2) = V$, and that $(\mathbf{v}_1, \mathbf{v}_2) \neq 0$. $\hat{\mathbf{y}}_1$ remains the same. $\mathbf{v}_3 = 2\mathbf{v}_1 - \mathbf{v}_2 = \begin{pmatrix} 0 \\ 3 \\ 3 \end{pmatrix}$, $\hat{\mathbf{y}}_3 = \frac{6}{18}\mathbf{v}_3 = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}$, and $\hat{\mathbf{y}}_1 + \hat{\mathbf{y}}_3 = \begin{pmatrix} 3 \\ 4 \\ 4 \end{pmatrix}$, which has inner products 11 and 24 with $\mathbf{v}_1$ and $\mathbf{v}_3$. $\mathbf{y} - \begin{pmatrix} 3 \\ 4 \\ 4 \end{pmatrix} = \begin{pmatrix} 4 \\ -4 \\ -2 \end{pmatrix}$, which is not orthogonal to V. Therefore, $\hat{\mathbf{y}}_1 + \hat{\mathbf{y}}_3$ is not the projection of $\mathbf{y}$ on $V = \mathscr{L}(\mathbf{v}_1, \mathbf{v}_3)$.

Since $(\mathbf{y} - \hat{\mathbf{y}}) \perp \hat{\mathbf{y}}$, we have, by the Pythagorean Theorem,

$$\|\mathbf{y}\|^2 = \|(\mathbf{y} - \hat{\mathbf{y}}) + \hat{\mathbf{y}}\|^2 = \|\mathbf{y} - \hat{\mathbf{y}}\|^2 + \|\hat{\mathbf{y}}\|^2$$

$$\|\mathbf{y}\|^2 = 53, \qquad \|\hat{\mathbf{y}}\|^2 = \frac{9^2}{3} + \frac{12^3}{6} = 51, \qquad \|\mathbf{y} - \hat{\mathbf{y}}\|^2 = \left\|\begin{pmatrix} 0 \\ -1 \\ 1 \end{pmatrix}\right\|^2 = 2.$$

Warning: We have shown that when $\mathbf{v}_1, \ldots, \mathbf{v}_k$ are mutually orthogonal the projection $\hat{\mathbf{y}}$ of $\mathbf{y}$ on the subspace spanned by $\mathbf{v}_1, \ldots, \mathbf{v}_k$ is $\sum_{j=1}^{k} p(\mathbf{y}|\mathbf{v}_j)$. This is true for all $\mathbf{y}$ *only* if $\mathbf{v}_1, \ldots, \mathbf{v}_k$ are mutually orthogonal. Students are asked to prove the "only" part in Problem 1.3.5.

Every subspace V of Ω of dimension $r > 0$ has an orthogonal basis (actually an infinity of such bases). We will show that such a basis exists by using *Gram–Schmidt orthogonalization.*

Let $\mathbf{x}_1, \ldots, \mathbf{x}_k$ be a basis for a subspace V, a k-dimensional subspace of Ω. For $1 \le i \le k$ let $V_i = \mathscr{L}(\mathbf{x}_1, \ldots, \mathbf{x}_i)$ so that $V_1 \subset V_2 \subset \cdots \subset V_k$ are properly nested subspaces. Let

$$\mathbf{v}_1 = \mathbf{x}_1, \qquad \mathbf{v}_2 = \mathbf{x}_2 - p(\mathbf{x}_2|\mathbf{v}_1).$$

Then $\mathbf{v}_1$ and $\mathbf{v}_2$ span V_2 and are othogonal. Thus $p(\mathbf{x}_3|V_2) = p(\mathbf{x}_3|\mathbf{v}_1) + p(\mathbf{x}_3|\mathbf{v}_2)$ and we can define $\mathbf{v}_3 = \mathbf{x}_3 - p(\mathbf{x}_3|V_2)$. Continuing in this way, suppose we have defined $\mathbf{v}_1, \ldots, \mathbf{v}_i$ to be mutually orthogonal vectors spanning V_i. Define $\mathbf{v}_{i+1} = \mathbf{x}_{i+1} - p(\mathbf{x}_{i+1}|V_i)$. Then $\mathbf{v}_{i+1} \perp V_i$ and hence $\mathbf{v}_1, \ldots, \mathbf{v}_{i+1}$ are mutually orthogonal and span V_{i+1}. Since we can do this for each $i \le k - 1$ we get the orthogonal basis $\mathbf{v}_1, \ldots, \mathbf{v}_k$ for V.

If $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ is an orthogonal basis for a subspace V then, since $\hat{\mathbf{y}} \equiv p(\mathbf{y}|V) = \sum_{j=1}^{k} p(\mathbf{y}|\mathbf{v}_j)$ and $p(\mathbf{y}|\mathbf{v}_j) = b_j\mathbf{v}_j$, with $b_j = [(\mathbf{y}, \mathbf{v}_j)/\|\mathbf{v}_j\|^2]$, it follows by the Pythagorean Theorem that

$$\|\hat{\mathbf{y}}\|^2 = \sum_{j=1}^{k} \|b_j\mathbf{v}_j\|^2 = \sum_{j=1}^{k} b_j^2\|\mathbf{v}_j\|^2 = \sum_{j=1}^{k} (\mathbf{y}, \mathbf{v}_j)^2/\|\mathbf{v}_j\|^2.$$

Of course, the basis $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ can be made into an *orthonormal* basis (all vectors of length one) by dividing each by its own length. If $\{\mathbf{v}_1^*, \ldots, \mathbf{v}_k^*\}$ is such an orthonormal basis then $\hat{\mathbf{y}} = p(\mathbf{y}|V) = \sum_1^k p(\mathbf{y}|\mathbf{v}_i^*) = \sum_1^k (\mathbf{y}, \mathbf{v}_i^*)\mathbf{v}_i^*$ and $\|\hat{\mathbf{y}}\|^2 = \sum_{i=1}^{k} (\mathbf{y}, \mathbf{v}_i^*)^2$.

Example 1.3.3: Consider R_4, the space of 4-component column vectors. Let us apply Gram–Schmidt orthogonalization to the columns of $\mathbf{X} = \begin{bmatrix} 1 & 1 & 4 & 8 \\ 1 & 1 & 0 & 10 \\ 1 & 5 & 12 & 0 \\ 1 & 5 & 8 & 10 \end{bmatrix}$, a matrix chosen carefully by the author to keep the arithmetic simple. Let the four columns be $\mathbf{x}_1, \ldots, \mathbf{x}_4$. Define $\mathbf{v}_1 = \mathbf{x}_1$. Let

$$\mathbf{v}_2 = \mathbf{x}_2 - \frac{12}{4}\mathbf{v}_1 = \begin{bmatrix} -2 \\ -2 \\ 2 \\ 2 \end{bmatrix}, \qquad \mathbf{v}_3 = \mathbf{x}_3 - \left[\frac{24}{4}\mathbf{v}_1 + \frac{32}{16}\mathbf{v}_2\right] = \begin{bmatrix} 2 \\ -2 \\ 2 \\ -2 \end{bmatrix},$$

and

$$\mathbf{v}_4 = \mathbf{x}_4 - \left[\frac{28}{4}\mathbf{v}_1 + \frac{(-16)}{16}\mathbf{v}_2 + \frac{(-24)}{16}\mathbf{v}_3\right] = \begin{bmatrix} 2 \\ -2 \\ -2 \\ 2 \end{bmatrix}.$$

We can multiply these $\mathbf{v}_i$ by arbitrary constants to simplify them without losing their orthogonality. For example, we can define $\mathbf{u}_i = \mathbf{v}_i/\|\mathbf{v}_i\|^2$, so that $\mathbf{u}_1$, $\mathbf{u}_2$, $\mathbf{u}_3$, $\mathbf{u}_4$ are unit length orthogonal vectors spanning Ω. Then $\mathbf{U} = (\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3, \mathbf{u}_4)$ is an orthogonal matrix. $\mathbf{U}$ is expressible in the form $\mathbf{U} = \mathbf{X}\mathbf{R}$, where $\mathbf{R}$ has zeros below the diagonal. Since $\mathbf{I} = \mathbf{U}'\mathbf{U} = \mathbf{U}'\mathbf{X}\mathbf{R}$, $\mathbf{R}^{-1} = \mathbf{U}'\mathbf{X}$, and $\mathbf{X} = \mathbf{U}\mathbf{R}^{-1}$, where $\mathbf{R}^{-1}$ has zeros below the diagonal (see Section 1.7).

As we consider linear models we will often begin with a model which supposes that $\mathbf{Y}$ has expectation $\boldsymbol{\theta}$ which lies in a subspace V_2, and will wish to decide whether this vector lies in a smaller subspace V_1. The orthogonal bases provided by the following theorem will be useful in the development of convenient formulas and in the investigation of the distributional properties of estimators.

Theorem 1.3.4: Let $V_1 \subset V_2 \subset \Omega$ be subspaces of Ω of dimensions $1 \le n_1 < n_2 < n$. Then there exist mutually orthogonal vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ such that $\mathbf{v}_1, \ldots, \mathbf{v}_{n_i}$ span V_i, $i = 1, 2$.

***Proof*:** Let $\{\mathbf{x}_1, \ldots, \mathbf{x}_{n_1}\}$ be a basis for V_1. Then by Gram–Schmidt orthogonalization there exists an orthogonal basis $\{\mathbf{v}_1, \ldots, \mathbf{v}_{n_1}\}$ for V_1. Let $\mathbf{x}_{n_1+1}, \ldots, \mathbf{x}_{n_2}$ be chosen consecutively from V_2 so that $\mathbf{v}_1, \ldots, \mathbf{v}_{n_1}, \mathbf{x}_{n_1+1}, \ldots, \mathbf{x}_{n_2}$ are linearly independent. (If this could not be done, V_2 would have dimension less than n_2.) Then applying Gram–Schmidt orthogonalization to $\mathbf{x}_{n_1+1}, \ldots, \mathbf{x}_{n_2}$ we have an orthogonal basis for V_2. Repeating this for V_2 replaced by Ω and $\mathbf{v}_1, \ldots, \mathbf{v}_{n_1}$ by $\mathbf{v}_1, \ldots, \mathbf{v}_{n_2}$ we get the theorem. □

For a nested sequence of subspaces we can repeat this theorem consecutively to get Theorem 1.3.5.

Theorem 1.3.5: Let $V_1 \subset V_2 \subset \cdots \subset V_k \subset \Omega = V_{k+1}$ be subspaces of Ω of dimensions $1 \leq n_1 < n_2 < \cdots < n_k < n = n_{k+1}$. Then there exists an orthogonal basis $\mathbf{v}_1, \ldots, \mathbf{v}_n$ for Ω such that $\mathbf{v}_1, \ldots, \mathbf{v}_{n_i}$ is a basis for V_i for $i = 1, \ldots, k + 1$.

We can therefore write for any $\mathbf{y} \in \Omega$,

$$p(\mathbf{y} \mid V_i) = \sum_{j=1}^{n_i} \frac{(\mathbf{y}, \mathbf{v}_j)}{\|\mathbf{v}_j\|^2} \mathbf{v}_j \qquad \text{for} \quad i = 1, \ldots, k + 1,$$

and

$$\|p(\mathbf{y} \mid V_i)\|^2 = \sum_{j=1}^{n_i} \frac{(\mathbf{y}, \mathbf{v}_j)^2}{\|\mathbf{v}_j\|^2} \qquad \text{for} \quad i = 1, \ldots, k + 1.$$

The $\mathbf{v}_j$ can be chosen to have length one, so these last formulas simplify still further.

Thus, the definition of the projection $p(\mathbf{y} \mid V)$ has been justified. Fortunately, it is not necessary to find an orthogonal basis in order to find the projection in the general case that the basis vectors $(\mathbf{x}_1, \ldots, \mathbf{x}_k)$ are not orthogonal. The Gram–Schmidt method is useful in the development of nonmatrix formulas for regression coefficients.

In order for $\hat{\mathbf{y}} = b_1\mathbf{x}_1 + \cdots + b_k\mathbf{x}_k$ to be the projection of $\mathbf{y}$ on $V = \mathscr{L}(\mathbf{x}_1, \ldots, \mathbf{x}_k)$ we need $(\mathbf{y}, \mathbf{x}_i) = (\hat{\mathbf{y}}, \mathbf{x}_i)$ for all i. This leads to the so-called *normal equations*:

$$(\hat{\mathbf{y}}, \mathbf{x}_i) = \sum_1^k b_j(\mathbf{x}_j, \mathbf{x}_i) = (\mathbf{y}, \mathbf{x}_i) \qquad \text{for} \quad i = 1, \ldots, k$$

It is convenient to write these k simultaneous linear equations in matrix form:

$$\underset{k \times k}{\mathbf{M}} \; \underset{k \times 1}{\mathbf{b}} = \mathbf{U},$$

where $\mathbf{M}$ is the matrix of inner products among the $\mathbf{x}_j$ vectors, $\mathbf{b}$ is the column vector of b_j's, and $\mathbf{U}$ is the $k \times 1$ column vector of inner products of $\mathbf{y}$ with the $\mathbf{x}_j$. If Ω is taken to be the space of n-component column vectors, then we can write $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_k)$, and we get $\mathbf{M} = \mathbf{X}'\mathbf{X}$, $\mathbf{U} = \mathbf{X}'\mathbf{y}$, so the normal equations are:

$$\mathbf{M}\mathbf{b} = (\mathbf{X}'\mathbf{X})\mathbf{b} = \mathbf{X}'\mathbf{y} = \mathbf{U}$$

Of course, if $\mathbf{M} = ((\mathbf{x}_i, \mathbf{x}_j))$ has an inverse we will have an explicit solution

$$\mathbf{b} = \mathbf{M}^{-1}\mathbf{U}$$

of the normal equations. It will be shown in Section 1.6 that $\mathbf{M}$ has rank k if and only if $\mathbf{x}_1, \ldots, \mathbf{x}_k$ are linearly independent. Thus $\mathbf{b} = \mathbf{M}^{-1}\mathbf{U}$ if and only if $\mathbf{x}_1, \ldots, \mathbf{x}_k$ are linearly independent.