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 kilograms 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 $x = 0.9$ (kg s/m^2).

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

$$g(x) \quad E(Y|x) \quad \text{for} \quad x \subset D$$

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 $g(x) \quad A + Bx + Cx^2$ or $g(x) = A2^x + B$ or $g(x) = A \log x + B$, and, so on, 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 $g(x)$ 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 1.1.1.

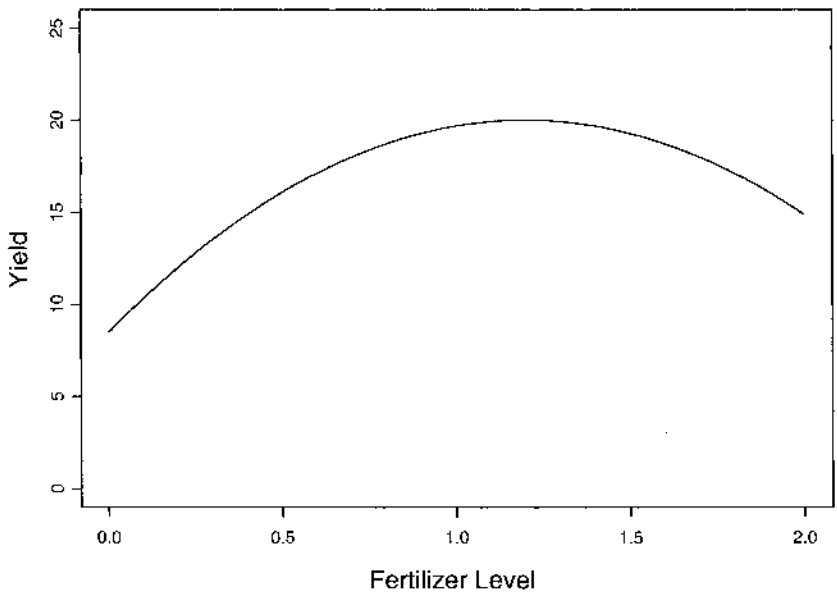

Figure 1.1.1: Regression of yield on fertilizer level

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 - 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 $\beta_0, \beta_1, \beta_2, \beta_3$ 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

Table 1.1.1: Height Data

Individual	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

complex model might suppose

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

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 the 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/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 column 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 Section 1.2, emphasizing geometric aspects.

1.2 Vectors, Inner Products, Lengths

Let Ω (omega) 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 $\mathbf{y}$. While we could easily reform $\mathbf{y}$ into a column vector, it is often convenient to preserve the form of $\mathbf{y}$. The term "n-tuple" means that the n elements of a vector $\mathbf{y} \in \Omega$ are ordered.

Omega 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 $\mathbf{y}$ 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 $i = 1, \ldots, n$.

Omega 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 $h(\mathbf{x}, \mathbf{y})$ (often called the **dot** product) is usually written simply as $(\mathbf{x}, \mathbf{y})$. 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:

$$\begin{aligned} (\mathbf{x}, \mathbf{y}) &= (\mathbf{y}, \mathbf{x}) \\ (a\mathbf{x}, \mathbf{y}) &= a(\mathbf{x}, \mathbf{y}) \\ (\mathbf{x}_1 + \mathbf{x}_2, \mathbf{y}) &= (\mathbf{x}_1, \mathbf{y}) + (\mathbf{x}_2, \mathbf{y}) \end{aligned}$$

for all vectors, and real numbers a.

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

The *distance* between vectors $\mathbf{x}$ and $\mathbf{y}$ is the length of $\mathbf{x} - \mathbf{y}$. Vectors $\mathbf{x}$ and $\mathbf{y}$ 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

$$\mathbf{x} = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 0 & 0 \\ 3 & 0 & \\ & 0 & \end{bmatrix} \quad \text{and} \quad \mathbf{y} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 3 & 0 \\ 0 & 5 & \\ & -1 & \end{bmatrix}$$

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

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. (The author was several years younger than Pythagorus, but knew him well.)

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

□

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, $\hat{y}$ 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 1.2.1). 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 with 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 $\mathbf{x} = 0$, any b will do. Otherwise, $b = (\mathbf{y}, \mathbf{x})/\|\mathbf{x}\|^2$. Thus,

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

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.

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

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

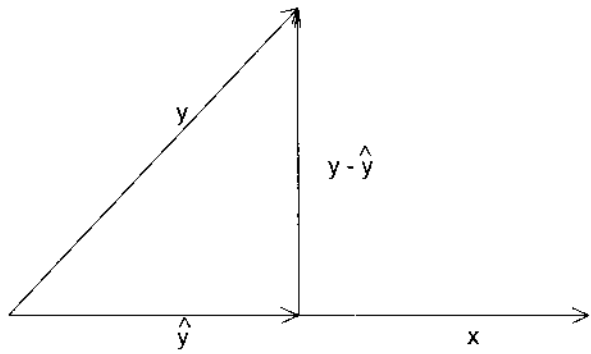

Figure 1.2.1: Projection of $\mathbf{y}$ onto $\mathbf{x}$

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 $\|\mathbf{y} - \hat{\mathbf{y}}\| = 0$, so that, equivalently, $\mathbf{y}$ 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 $b = (\mathbf{y}, \mathbf{I}_A)/\|\mathbf{I}_A\|^2 = \left(\sum_{i \subset A} y_i\right)/N(A)$, 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 four-component row vectors, $\mathbf{y} = (3, 7, 8, 13)$, and A is the indicator Indicator vector 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 five-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 $\mathbf{x} \subset V$ 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 $\mathcal{L}(\mathbf{x}_1, \ldots, \mathbf{x}_k)$.

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 four-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 $V_2 = \mathcal{L}(\mathbf{x}_1, \mathbf{x}_2)$, $V_3 = \mathcal{L}(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)$ and $V_4 = \mathcal{L}(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4)$.*

(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 $V = \mathcal{L}(\mathbf{C}_1, \mathbf{C}_2, \mathbf{C}_3)$.*

(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.*

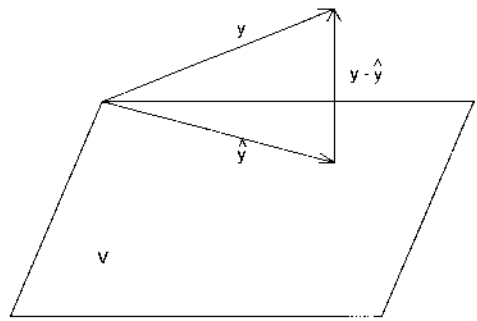

Figure 1.3.1: Projection of $\mathbf{y}$ onto the Subspace V

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 the following chapters.

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 $(\mathbf{y} \quad \hat{\mathbf{y}}) \perp V$ is equivalent to $(\mathbf{y} \cdot \hat{\mathbf{y}}, \mathbf{x}) = 0$ 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 (See Figure 1.3.1).

If vectors $\mathbf{x}_1, \ldots, \mathbf{x}_k$ span a subspace V then a vector $\mathbf{z} \subset V$ 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

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

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. 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 $V = \mathcal{L}(\mathbf{x})$, 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 six-component space Ω of the problem above, and let $V = \mathcal{L}(\mathbf{C}_1, \mathbf{C}_2, \mathbf{C}_3)$. 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 $\hat{\mathbf{y}}_1 - \hat{\mathbf{y}}_2 = \mathbf{0}$, that is, $\hat{\mathbf{y}}_1 = \hat{\mathbf{y}}_2$.

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 this 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, a 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

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

□

Example 1.3.2. Let

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

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 $(\hat{\mathbf{y}},\mathbf{v}_2)=12$. 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 $\mathcal{L}(\mathbf{v}_1,\mathbf{v}_3)=\mathcal{L}(\mathbf{v}_1,\mathbf{v}_2)=V$, and that $(\mathbf{v}_1,\mathbf{v}_3)\neq 0$. $\hat{\mathbf{y}}_1$ remains the same. $\mathbf{v}_3=2\mathbf{v_1}-\mathbf{v}_2=\begin{pmatrix}0\\3\\3\end{pmatrix}$, $\hat{\mathbf{y}}_3=\frac{6}{18}\mathbf{v}_3\quad\begin{pmatrix}0\\1\\1\end{pmatrix}$, 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 $\mathbf{v}_3$. $\mathbf{y}\quad\begin{pmatrix}3\\4\\4\end{pmatrix}=\begin{pmatrix}4\\4\\2\end{pmatrix}$, which is not orthogonal to V. Therefore, $\hat{\mathbf{y}}_1+\hat{\mathbf{y}}_3$ is not the projection of $\mathbf{y}$ on $V=\mathcal{L}(\mathbf{v}_1,\mathbf{v}_3)$.

Since $(\mathbf{y}\quad\hat{\mathbf{y}})\perp\hat{\mathbf{y}}$, we have, by the Pythagorean Theorem,
$\|\mathbf{y}\|^2=\|(\mathbf{y}\quad\hat{\mathbf{y}})+\hat{\mathbf{y}}\|^2=\|\mathbf{y}\quad\hat{\mathbf{y}}\|^2+\|\hat{\mathbf{y}}\|^2$

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

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 1.3.7.

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 $1\leq i\leq k$ let $V_i=\mathcal{L}(\mathbf{x}_1,\ldots,\mathbf{x}_i)$ so that $V_1\subset V_2\subset\cdots\subset V_k$ are properly nested subspaces. Let

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

Then $\mathbf{v}_1$ and $\mathbf{v}_2$ span V_2 and are orthogonal. Thus $p(\mathbf{x}_3|V_2)=p(\mathbf{x}_3|\mathbf{v}_1)+p(\mathbf{x}_3|\mathbf{v}_2)$ and 1.3.4 $\mathbf{v}_3=\mathbf{x}_3-p(\mathbf{x}_3|V_2)$. Continuing in this way, suppose we have defined $\mathbf{v}_1,\ldots,\mathbf{v}_i$ to be mutually orthogonal vectors spanning V_i. Define $\mathbf{v}_{i+1}\quad\mathbf{x}_{i+1}-p(\mathbf{x}_{i+1}|V_i)$. 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 $i\leq k\quad 1$ 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 $p(\mathbf{y}|\mathbf{v}_j)\quad b_j\mathbf{v}_j$, 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 $\|\hat{\mathbf{y}}\|^2 = \sum_{i=1}^k (\mathbf{y}, \mathbf{v}_i^*)^2$. The function "qr" in R and S-Plus determines $\mathbf{U}$ and textbfR for given $\mathbf{X}$.

Example 1.3.3. Consider R_4, the space of four-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

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

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 $\mathbf{u}_i = \mathbf{v}_i / \|\mathbf{v}_i\|$, 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. The vector $\mathbf{U}$ is expressible in the form $\mathbf{U} = \mathbf{XR}$, where $\mathbf{R}$ has zeros below the diagonal. Since $\mathbf{I} = \mathbf{U}'\mathbf{U} = \mathbf{U}'\mathbf{XR}, \mathbf{R}^{-1} = \mathbf{U}'\mathbf{X}$, and $\mathbf{X} = \mathbf{UR}^{-1}$, 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 θ 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 $1 \leq n_1 < n_2 < n$. 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 $\mathbf{y} \in \Omega$,

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

and

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

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

Thus, the definition of the projection $p(\mathbf{y}|V)$ 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 $V = \mathcal{L}(\mathbf{x}_1, \ldots, \mathbf{x}_k)$ we need $(\mathbf{y}, \mathbf{x}_i) = (\hat{\mathbf{y}}, \mathbf{x}_i)$ for all i. This leads to the so-called *normal equations*:

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

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

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

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

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

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.

In the case that the elements of Ω are not column vectors, we can always rewrite its elements as column vectors, and the matrix $\mathbf{M}$ will remain unchanged. Thus, in the general case $\mathbf{M}$ possesses an inverse if and only if the vectors $\mathbf{x}_1, \ldots, \mathbf{x}_k$ are linearly independent. Of course, even in this case with $\Omega = R_n$, the space of n-component column vectors, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_k)$, being $n \times k$, does not have an inverse unless $n = k$. In applications we always have $n > k$.

In the computation on $\mathbf{M} = \mathbf{X}'\mathbf{X}$ it makes little sense to write $\mathbf{X}$ on its side as $\mathbf{X}'$, then $\mathbf{X}$, and then to carry out the computation as the multiplication of two matrices, unless the computer software being used requires this. The matrix $\mathbf{M}$ is the matrix of inner products, and $\mathbf{U}$ is a vector of inner products, and this viewpoint should be emphasized.

Example 1.3.4. Let $\mathbf{y}$, $\mathbf{v}_1$ and $\mathbf{v}_2$ be as in Example 1.3.3. Let $\mathbf{x}_1 = \mathbf{v}_1$ and $\mathbf{x}_2 = 2\mathbf{v}_1 + \mathbf{v}_2$. Then

$$\mathbf{y} = \begin{bmatrix} 7 \\ 0 \\ 2 \end{bmatrix}, \quad \mathbf{x}_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \quad \mathbf{x}_2 = \begin{bmatrix} 4 \\ 1 \\ 1 \end{bmatrix},$$

and $V = \mathcal{L}(\mathbf{v}_1, \mathbf{v}_2) = \mathcal{L}(\mathbf{x}_1, \mathbf{x}_2)$. We compute

$$\mathbf{M} = \begin{bmatrix} 3 & 6 \\ 6 & 18 \end{bmatrix}, \quad \mathbf{U} = \begin{bmatrix} (\mathbf{x}_1, & \mathbf{y}) \\ (\mathbf{x}_2, & \mathbf{y}) \end{bmatrix} = \begin{bmatrix} 9 \\ 30 \end{bmatrix},$$

$$\mathbf{M}^{-1} = \frac{1}{18}\begin{bmatrix} 18 & -6 \\ -6 & 3 \end{bmatrix} = \frac{1}{6}\begin{bmatrix} 6 & -2 \\ -2 & 1 \end{bmatrix},$$

$$\mathbf{b} = \mathbf{M}^{-1}\mathbf{U} = \frac{1}{6}\begin{bmatrix} 6 \\ 12 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

and $\hat{\mathbf{y}} = p(\mathbf{y}|V) = -\mathbf{x}_1 + 2\mathbf{x}_2 = \begin{bmatrix} 7 \\ 1 \\ 1 \end{bmatrix}$, as before.

It is easy to compute lengths of $\hat{\mathbf{y}}$ and of $\mathbf{y} - \hat{\mathbf{y}}$. First,

$$\|\hat{\mathbf{y}}\|^2 = (\mathbf{y}, \hat{\mathbf{y}}) = \left(\mathbf{y}, \sum_1^k b_j \mathbf{x}_j\right) = \sum b_j(\mathbf{y}, \mathbf{x}_j) = \mathbf{b}'\mathbf{U}.$$

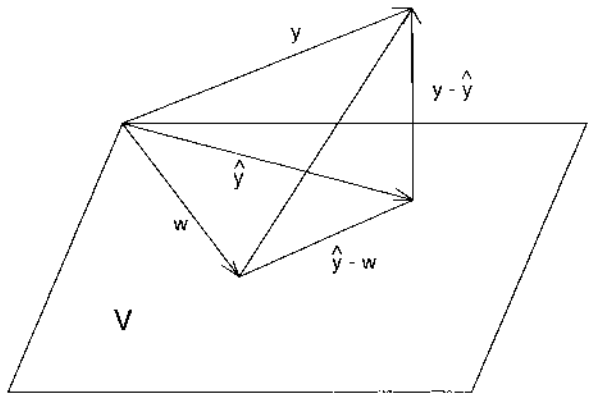

Figure 1.3.2: Projection of $\mathbf{y}$ onto the Subspace V

By the Pythagorean Theorem,

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

For Example 1.3.3, $\|\hat{\mathbf{y}}\|^2 = b_1(\mathbf{y}, \mathbf{x_1}) + b_2(\mathbf{y}, \mathbf{x}_2) = (-1)(9) + 2(30) = 51$.

The projection $\hat{\mathbf{y}} = p(\mathbf{y}|V)$ is the closest vector in V to $\mathbf{y}$, since for any other vector $\mathbf{w} \in V$,

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

by the Pythagorean Theorem, and the facts that $(\hat{\mathbf{y}} - \mathbf{w}) \in V$ and $(\mathbf{y} - \hat{\mathbf{y}}) \perp V$. Thus $\|\mathbf{y} - \mathbf{w}\|^2$ is minimized for $\mathbf{w} \in V$ by taking $\mathbf{w} = \hat{\mathbf{y}}$ (Figure 1.3.2).

For this reason, the vectors $\mathbf{b}$ and $\hat{\mathbf{y}}$ are said to have been obtained by the principle of least squares.

Problem 1.3.3. *Let $\Omega, \mathbf{C}_1, \mathbf{C}_2, \mathbf{C}_3$ be defined as in problem 1.3.2. Let $V = \mathcal{L}(\mathbf{C}_1, \mathbf{C}_2, \mathbf{C}_3)$*

(a) For $\mathbf{y} = \begin{bmatrix} 6 & 11 & 8 \\ 4 & 7 & \\ 2 & & \end{bmatrix}$, *find* $\hat{\mathbf{y}} = p(\mathbf{y}|V), \mathbf{y} - \hat{\mathbf{y}}, \|\mathbf{y}\|^2, \|\hat{\mathbf{y}}\|^2, \|\mathbf{y} - \hat{\mathbf{y}}\|^2$.

(b) Give a general nonmatrix formula for $\hat{\mathbf{y}} = p(\mathbf{y}|V)$ for any $\mathbf{y}$.

Problem 1.3.4. *Let* $\mathbf{x}_1 = (1,1,1,1)'$, $\mathbf{x}_2 = (4,1,3,4)'$, $\mathbf{y} = (1,9,5,5)'$ *(so these are column vectors). Let* $V = \mathcal{L}(\mathbf{x}_1, \mathbf{x}_2)$.

(a) *Find* $\hat{\mathbf{y}} = p(\mathbf{y}|V)$ *and* $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$.

(b) *Find* $\hat{\mathbf{y}}_1 = p(\mathbf{y}|\mathbf{x}_1)$ *and* $\hat{\mathbf{y}}_2 = p(\mathbf{y}|\mathbf{x}_2)$ *and show that* $\hat{\mathbf{y}} \neq \hat{\mathbf{y}}_1 + \hat{\mathbf{y}}_2$.

(c) *Verify that* $\mathbf{e} \perp V$.

(d) *Find* $\|\mathbf{y}\|^2, \|\hat{\mathbf{y}}\|^2, \|\mathbf{y} - \hat{\mathbf{y}}\|^2$, *and verify that the Pythagorean Theorem holds. Compute* $\|\hat{\mathbf{y}}\|^2$ *directly from* $\hat{\mathbf{y}}$ *and also by using the formula* $\|\hat{\mathbf{y}}\|^2 = \mathbf{U}'\mathbf{b}$.

(e) *Use Gram–Schmidt orthogonalization to find four mutually orthogonal non-zero vectors* $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \mathbf{v}_4$, *such that* $V = \mathcal{L}(\mathbf{v}_1, \mathbf{v}_2)$. *Hint: You can choose* $\mathbf{x}_3$ *and* $\mathbf{x}_4$ *arbitrarily, as long as* $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$ *are linearly independent.*

(f) *Express* $\boldsymbol{y}$ *and* $\hat{\mathbf{y}}$ *in terms of the* $\mathbf{v}_i inpart(e)$..

(g) *Let* $\mathbf{w} = (2,8,4,2)'$. *Show that* $\mathbf{w} \in V$ *and verify that* $\|\mathbf{y} - \mathbf{w}\|^2 = \|\mathbf{y} - \hat{\mathbf{y}}\|^2 + \|\hat{\mathbf{y}} - \mathbf{w}\|^2$. *(Why must this equality hold?)*

(h) ***(Important!** Does* $p(\hat{\mathbf{y}}|\mathbf{x}_1) = \hat{\mathbf{y}}_1$? *Is this true for any* $\boldsymbol{y}$? *That is, do we obtain the same vector by (1) first projecting* $\boldsymbol{y}$ *on* V, *then projecting this vector on* $\mathbf{x}_1$ *as by (2) projecting* $\boldsymbol{y}$ *directly on* $\mathbf{x}_1$? *textbfMore generally, if* V *is a subspace, and* V_1 *a subspace of* V, *does* $p(p(\mathbf{y}|V)|V_1) = p(\mathbf{y}|V_1)$? *Hint: Use an argument based entirely on inner products.*

Problem 1.3.5. *Let* $\mathbf{y} = (y_1, \ldots, y_n)', \mathbf{x} = (x_1, \ldots, x_n)', \mathbf{J} = (1, \ldots, 1)'$, *and* $V = \mathcal{L}(\mathbf{J}, \mathbf{x})$.

(a) *Use Gram Schmidt orthogonalization on the vectors* $\boldsymbol{J}$, $\boldsymbol{x}$ *(in this order) to find orthogonal vectors* $\mathbf{J}, \mathbf{x}^*$ *spanning* V. *Express* $\mathbf{x}^*$ *in terms of* $\boldsymbol{J}$ *and* $\boldsymbol{x}$, *then find* b_0, b_1 *such that* $\hat{\mathbf{y}} = b_0\mathbf{J} + b_1\mathbf{x}$. *To simplify the notation, let* $\mathbf{y}^* = \mathbf{y} - p(\mathbf{y}|\mathbf{J}) = \mathbf{y} - \bar{y}\mathbf{J}$,

$$
\begin{aligned}
S_{xy} &= (\mathbf{x}^*, \mathbf{y}^*) = (\mathbf{x}^*, \mathbf{y}) = \sum (x_i - \bar{x})y_i = \sum (x_i - \bar{x})(y_i - \bar{y}) \\
&= \sum x_i y_i - \bar{x}y n, \\
S_{xx} &= (\mathbf{x}^*, \mathbf{x}^*) = \sum (x_i - \bar{x})x_i = \sum x_i^2 - \bar{x}^2 n, \\
S_{yy} &= (\mathbf{y}^*, \mathbf{y}^*) = \sum (y_i - y)^2 = \sum (y_i - \bar{y})y_i.
\end{aligned}
$$

(b) *Suppose* $\hat{\mathbf{y}} = p(\mathbf{y}|V) = a_0\mathbf{J} + a_1\mathbf{x}^*$. *Find formulas for* a_1 *and* a_0 *in terms of* $\bar{y}, S_{xy}$, *and* S_{xx}.

(c) *Express* $\mathbf{x}^*$ *in terms of* $\boldsymbol{J}$ *and* $\boldsymbol{x}$, *and use this to determine formulas for* b_1 *and* b_0 *so that* $\hat{\mathbf{y}} = b_0\mathbf{J} + b_1\mathbf{x}$.

(d) *Express* $\|\hat{\mathbf{y}}\|^2$ *and* $\|\mathbf{y} - \hat{\mathbf{y}}\|^2$ *in terms of* $\bar{y}, S_{xy}, S_{xx}$, *and* S_{yy}.

(e) *Use the formula* $\mathbf{b} = \mathbf{M}^{-1}\mathbf{U}$ *for* $\mathbf{b} = (b_0, b_1)'$ *and verify that they are the same as those found in (c).*

(f) *For* $\mathbf{y} = \begin{pmatrix} 2 \\ 6 \\ 8 \\ 8 \end{pmatrix}, \mathbf{x} = \begin{pmatrix} 0 \\ 1 \\ 2 \\ 3 \end{pmatrix}$ *find* $a_0, a_1, \hat{\mathbf{y}}, b_0, b_1, \|\mathbf{y}\|^2, \|\hat{\mathbf{y}}\|^2, \|\mathbf{y} - \hat{\mathbf{y}}\|^2$. *Verify that* $\|\hat{\mathbf{y}}\|^2 = b_0(\mathbf{y}, \mathbf{J}) + b_1(\mathbf{y}, \mathbf{x})$ *and that* $(\mathbf{y} - \hat{\mathbf{y}}) \perp V$.

(g) *Plot the four* (x_i, y_i) *points and the least-squares line obtained in (f).*

Problem 1.3.6. *Let* $\mathbf{x}_1, \ldots, \mathbf{x}_k$ *be a basis of a subspace* V. *Suppose that* $p(\mathbf{y}|V) = \sum_{j=1}^{k} p(\mathbf{y}|\mathbf{x}_j)$ *for every vector* $\mathbf{y} \in \Omega$. *Prove that* $\mathbf{x}_1, \ldots, \mathbf{x}_k$ *are mutually orthogonal. Hint: Consider the vector* $\mathbf{y} = \mathbf{x}_i$ *for each* i.

Problem 1.3.7. *Consider the collection* $\mathcal{H}$ *of all real-valued functions on the unit interval* $U = [0, 1]$ *having the property* $\int_0^1 f^2(x)\,dx < \infty$. *Define the inner product* $(\mathbf{f}, \mathbf{g}) = \int_0^1 f(x)g(x)\,dx$. *Such an inner product space, with the correct definition of the integral, and a more subtle property called completeness, is called a Hilbert space after the great German mathematician, David Hilbert, of the late nineteenth and early twentieth centuries. The Hilbert space* $\mathcal{H}$ *is not finite dimensional, but our projection theory still applies because we will be interested in projections on finite dimensional subspaces. Consider the function* $h(x) = \sqrt{x}$ *for* $x \in U$. *For each non-negative integer* k *define* $p_k(x) = x^k$. *The functions* h, p_0, p_1, p_2 *determine corresponding points* $\mathbf{h}, \mathbf{p}_0, \mathbf{p}_1, \mathbf{p}_2$ *in* $\mathcal{H}$. *Define* $V_k = \mathcal{L}(\mathbf{p}_0, \mathbf{p}_1, \ldots, \mathbf{p}_k)$, *and* $\hat{\mathbf{h}}_k = p(\mathbf{h}|V_k)$. *The point* $\hat{\mathbf{h}}_k$ *corresponds to a polynomial* $\hat{h}_k$ *of degree* k *on [0, 1]. Though there is a subtle difference between the point functions* $h, p_k, \hat{h}_k$, *and the corresponding points* $\mathbf{h}, \mathbf{p}_k, \hat{\mathbf{h}}_k$ *in* $\mathcal{H}$, *we will ignore this difference. Let* $E_k = \|\mathbf{h} - \mathbf{h}_k\|^2$ *be the measure of error when the function* h_k *is used to approximate* h.

(a) *Find the functions* $\hat{h}_k$ *for* $k = 0, 1, 2$. *Plot* h *and these three functions on the same axes. Hint: The inner products* $(\mathbf{p}_i, \mathbf{p}_j)$ *and* $(\mathbf{p}_i, \mathbf{h})$ *are easy to determine as functions of* i *and* j, *so that the matrices* $\boldsymbol{M}$ *and* $\boldsymbol{U}$ *are easy to determine. If possible use exact arithmetic.*

(b) *Evaluate* E_k *for* $k = 0, 1, 2$.

(c) *Find the Taylor approximation* $\mathbf{h}^*$ *of* $\mathbf{h}$, *using constant, linear, and quadratic terms, and expanding about* $x = 1/2$. *Show that the error* $\|\mathbf{h} - \hat{\mathbf{h}}_2\|^2$ *is smaller than the error* $\|\mathbf{h} - \mathbf{h}^*\|^2$.

(d) *Repeat (a) and (b) for* $h(x) = 1/(1 + x)$. *Hint: Let* $c_k = (\mathbf{h}, \mathbf{p}_k) = \int_0^1 h(x)p_k(x)\,dx$. *Then* $c_k = \int_0^1 x^{k-1}[1 - h(x)]\,dx = (1/k) \cdot c_{k-1}$.

Table 1.4.1: Regression of % of Wormy Fruit on Size of Apple Crop

Tree Number	X 100's of Fruit	Y Percent Wormy	$\hat{Y}$ Est. of $E(Y\|X)$	d_{xy} Dev. from Regr.
1	8	59	56.14	2.86
2	6	58	58.17	-0.17
3	11	56	53.10	2.90
4	22	53	41.96	11.04
5	14	50	50.06	-0.06
6	17	45	47.03	-2.03
7	18	43	46.01	-3.01
8	24	42	39.94	2.06
9	19	39	45.00	−6.00
10	23	38	40.95	-2.95
11	26	30	37.91	−7.91
12	40	27	23.73	3.27
$\sum X = 228$			$\sum Y^2 =$	$25,522$
$\bar{X} = 19$			$(\sum Y)^2/n$	$24,300$
$\sum X^2 = 5,256$			$\sum XY -$	$9,324$
$\sum Y = 540$			$(\sum X)(\sum Y)/n =$	$10,260$
$\bar{Y} = 45$			$(\sum X)^2/n$	$4,332$

1.4 Examples

In this section we discuss four real data examples, formulate them in terms of vector spaces, and carry out some of the computations. At this point we consider only ways of describing observed vectors $\mathbf{y}$ in terms of a few other vectors $\mathbf{x}_1, \ldots, \mathbf{x}_k$.

Example 1.4.1. In their classic book *Statistical Methods for Research Workers*, Snedecor and Cochran (1980, p. 162) present the data of Table 1.4.1 accompanied by this commentary:

> "Regression of injured fruit on crop size. It is rather generally thought that the intensity of the injury by codling moth larvae is greater on apple trees bearing a small crop. Apparently the density of the flying moths is unrelated to the size of the crop on a tree so that the chance of attack for any particular fruit is augmented if there are few fruits in the tree. The data in table 6.5 are adapted from the results of an experiment containing evidence about this phenomenon. The 12 trees were all given a calyx spray of lead arsenate followed by fine cover sprays made up of 3 pounds of managanese arsenate and 1 quart of fish oil per 100 gallons. There is a decided tendency for the percentage of wormy fruits to decrease as the number of apples in the tree increases."

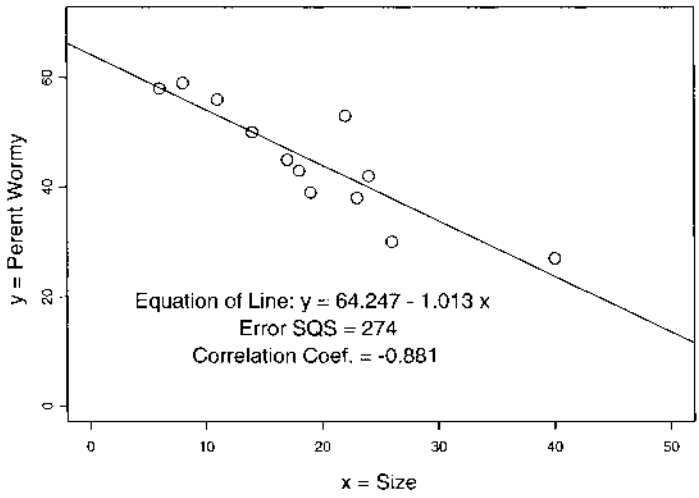

Figure 1.4.1: Regression of percentage of wormy apples on size of apple crop. From *Statistical Methods for Research Workers*, by G. W. Snedecor (1976), Iowa State Press

$$x_i = X_i - \bar{X} \qquad y_i = Y_i - \bar{Y}$$
$$\sum x^2 = 924 \quad \sum y^2 = 1222 \quad \sum xy = -926$$
$$b = \sum xy / \sum x^2 = -936/924 = -1.13 \quad (\% \text{ wormy})$$
$$\hat{Y} = \bar{Y} + b(X - \bar{X}) = 45 - 1.013(X - 19) = 64.247 - 1.103$$
$$\sum d^2_{y.x} = 1.222 - (-936)^2/924 = 273.88$$
$$S^2_{y.x} = \sum d^2_{y.x}/(n-2) = 273.88/10 = 27.388$$

The line on the scatter diagram of Figure 1.4.1 was obtained as follows. Suppose we try to approximate y by a linear function $g(x) = b_0 + b_1 x$. One possible criterion for the choice of the pair (b_0, b_1) is to choose that pair for which

$$Q = Q(b_0, b_1) = \sum_{i=1}^{n} |y_i - (b_0 + b_1 x_i)|^2$$

is minimum. If we define $\mathbf{y}$ and $\mathbf{x}_1$ as 12-component column vectors of y and x values, and $\mathbf{x}_0$ as the 12-component vector of all ones, then

$$Q = \|\mathbf{y} - (b_0 \mathbf{x}_0 + b_1 \mathbf{x}_1)\|^2,$$

so that Q is minimized for $b_0 \mathbf{x}_0 + b_1 \mathbf{x}_1 = \hat{\mathbf{y}}$, the projection of $\mathbf{y}$ onto $\mathcal{L}(\mathbf{x}_0, \mathbf{x}_1)$.

Table 1.4.2: Student Height Data

X	y	$\hat{y}$	e
1 70 62 1	68.5	68.66	−0.16
1 73 66 1	72.5	72.32	0.18
1 68 67 1	70.0	69.87	0.13
1 72 64 1	71.0	70.78	0.22
1 66 60 1	65.0	65.37	−0.37
1 71 63 0	64.5	63.85	0.65
1 74 68 0	67.5	67.99	0.49
1 65 65 0	61.5	61.29	0.29
1 70 64 0	63.5	63.74	0.25
1 69 65 0	63.5	63.63	0.13

Thus, for $\mathbf{X} = (\mathbf{x}_0, \mathbf{x}_1), \mathbf{M} = \mathbf{X}'\mathbf{X}, \mathbf{U} = \mathbf{X}'\mathbf{y}$,

$$\mathbf{b} = \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} = \mathbf{M}^{-1}\mathbf{U}.$$

The matrix $\mathbf{X}$ is the 12×2 matrix whose first column elements are all ones, and whose second column is the column labeled X in Table 1.4.1. The column vector $\mathbf{y}$ was labeled Y by Snedecor. $\hat{\mathbf{y}}$ and $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$ were labeled $\hat{Y}$ and $d_{y.x}$.

$$\mathbf{M} = \begin{bmatrix} 12 & 228 \\ 228 & 5256 \end{bmatrix}, \quad \mathbf{M}^{-1} = \begin{bmatrix} 0.474\,030 & -0.020\,563 \\ -0.020\,563 & 0.001\,082 \end{bmatrix}, \quad \mathbf{U} = \begin{bmatrix} 540 \\ 9324 \end{bmatrix}$$

$$\mathbf{b} = M^{-1}U = \begin{bmatrix} 64.247 \\ -1.013 \end{bmatrix} \quad \|\mathbf{y}\|^2 = 25,522 \quad \|\hat{\mathbf{y}}\|^2 = 25,248 \quad \|\mathbf{y} - \hat{\mathbf{y}}\|^2 = 274$$

Notice that $\|\mathbf{y}\|^2 = \|\hat{\mathbf{y}}\|^2 + \|\mathbf{y} - \hat{\mathbf{y}}\|^2$, as should be the case, by the Pythagorean Theorem. Simple computations verify that $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$ is orthogonal to $\mathbf{x}_0$ and $\mathbf{x}_1$, that is, $\sum e_i = 0$ and $\sum e_i x_i = 0$.

Here we have chosen to use the more general matrix formulas in order to determine b_0 and b_1 even though nonmatrix formulas were developed in Problem 1.2.3. A complete discussion of the simple linear regression model will be included later.

Example 1.4.2. Now consider the height data of Table 1.1. Let us try to approximate the 10-component vector $\mathbf{y}$ with a vector $\hat{\mathbf{y}}$ contained in $\mathcal{L}(\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)$, where $\mathbf{x}_0$ is the 10-component column vector of ones and $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ are as given in Table 1.1.1. The approximation vectors are given in Table 1.4.2.

$$\mathbf{M} = \begin{bmatrix} 10 & 698 & 644 & 5 \\ 698 & 48,796 & 44,977 & 349 \\ 644 & 44,977 & 41,524 & 319 \\ 5 & 349 & 319 & 5 \end{bmatrix}$$

$$\mathbf{M}^{-1} = \begin{bmatrix} 10,927,530 & -55,341 & -108,380 & -150,056 \\ -55,341 & 1,629 & 898 & 1,077 \\ 108,380 & 898 & 2,631 & 3,158 \\ 150,056 & 1,077 & 3,158 & 43,789 \end{bmatrix} 10^{-5}$$

$$\mathbf{U} = \mathbf{X}'\mathbf{y} = \begin{bmatrix} 667.5 \\ 46,648.0 \\ 43,008.5 \\ 347.0 \end{bmatrix} \quad \mathbf{b} = \begin{bmatrix} -7.702 \\ 0.585 \\ 0.477 \\ 5.872 \end{bmatrix} \quad \sum e_i^2 = \|\mathbf{e}\|^2 = 0.08575$$

The height y seems to be predicted very nicely by x_1 (father's height), x_2 (mother's height), and x_3 (sex). We must be cautious, however, in interpreting such an analysis based on 10 observations with 4 independent variables. Predictions of heights for other people, based on the coefficients determined for these data, should not be expected to be as good.

Table 1.4.3: Numbers of Mice Inoculated for Three Strains

Days to Death	9D	11C	DSC1	Total
2	6	1	3	10
3	4	3	5	12
4	9	3	5	17
5	8	6	8	22
6	3	6	19	28
7	1	14	23	38
8		11	22	33
9		4	14	18
10		6	14	20
11		2	7	9
12		3	8	11
13		1	4	5
14			1	1
Total	31	60	133	224
$\sum X$	125	442	1037	1604
$\sum X^2$	561	3602	8961	13,124

Example 1.4.3. (From Snedecor, 1967, p. 278):

> EXAMPLE 10.12.1 The numbers of days survived by mice inoculated with three strains of typhoid organisms are summarized in the following frequency distributions. Thus, with strains 9D, 6 mice survived for 2 days, 4 mice for 3 days, and so on. We have $n_1 = 31, n_2 = 60, n_3 = 133, N = 224$. The purpose of the analysis is to estimate and compare the mean numbers of days to death for the three strains.

Since the variance for strain 9D looks much smaller than for the other strains, it seems wise to calculate s_i^2 separately for each strain, rather than use a pooled s^2 from the analysis of variance.

The calculations are given under Table 1.4.3. Again from (1967) consider the variable days to death for three strains of typhoid organism. Let $\mathbf{y}$ be the table with three columns, having the days to death for 31 mice on 9D in column 1, for 60 mice on 11C in column 2, and 133 mice on DSC1 in column 3. Thus $\mathbf{y}$ has 224 components. Let y_{ij} be the jth component in the ith column of $\mathbf{y}$. Let $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ be the indicators of columns 1, 2, 3. The best approximation to $\mathbf{y}$ by vectors in $\mathcal{L}(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3) = V$ in the least-squares sense is

$$\hat{\mathbf{y}} = p(\mathbf{y}|V) = \sum_{i=1}^{3} p(\mathbf{y}|\mathbf{x}_i) = \sum_{i=1}^{3} y_i \mathbf{x}_i$$

The second equality follows by the orthogonality of $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$. The symbol y_i denotes the mean of the values of y in the ith column. Thus $\hat{\mathbf{y}}$ is the array with 31 $\bar{y}_1$'s in column 1, 60 $\bar{y}_2$'s in column 2, 133$\bar{y}_3$'s in column 3. Easy computation (remembering, for example, that 4 occurs nine times in column 1) shows that

$$\sum_j Y_{1j} = 125, \quad \sum_j Y_{2j} = 442, \quad \text{and} \quad \sum_j Y_{3j} = 1,037.$$

We find $\bar{y}_1 = 4.032, \bar{y}_2 = 7.367, \bar{y}_3 = 7.797$, and the error sum of squares $\|\mathbf{e}\|^2 = \sum_{ij}(y_{ij} - y_i)^2 = 1,278.42, \|\hat{\mathbf{y}}\|^2 = \sum_i n_i y_i^2 = 11,845.58$, and $\|\mathbf{y}\|^2 = \sum_{ij} y_{ij}^2 = 13,124$.

Example 1.4.4. The following data were given in a problem in Dixon and Massey (1957, p. 185):

> The drained weight in ounces of frozen apricots was measured for various types of syrups and various concentrations of syrup. The original weights of the apricots were the same. Differences in drained weights would be attributable to differences in concentrations or type of syrups.

		Syrup Composition				
		All Sucrose	2/3 Sucrose 1/3 Corn Syrup	1/3 Sucrose 2/3 Corn Syrup	All Corn Syrup	y_i.
Conc. of Syrup	30	28.80	28.21	29.28	29.12	28.853
	40	29.12	28.64	29.12	30.24	29.280
	50	29.76	30.40	29.12	28.32	29.400
	$y_{.j}$	29.227	29.083	29.173	29.227	

$\bar{\mathbf{y}}_{..} = 29.178$

Let $\mathbf{y}$ be the 3×4 matrix of drained weights. Let us approximate $\mathbf{y}$ by a linear combination of indicator vectors for rows and columns. Define $\mathbf{R}_i$ to be the indicator of row i and $\mathbf{C}_j$ to be the indicator of column j. Thus, for example,

$$\mathbf{R}_2 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \quad \text{and} \quad \mathbf{C}_3 = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

Take $V = \mathcal{L}(\mathbf{R}_1, \mathbf{R}_2, \mathbf{R}_3, \mathbf{C}_1, \ldots, \mathbf{C}_4)$. Define $\mathbf{x}_0$ to be the 3×4 vector of all ones. Then $\mathbf{x}_0 = \sum_i \mathbf{R}_i = \sum_j \mathbf{C}_j$. Let $\bar{y}_{i\cdot}, \bar{y}_{\cdot j}$ and $\bar{y}_{\cdot\cdot}$ be the means of the ith row, the jth column, and all the y_{ij}, respectively. It is not difficult to show that V has dimension $4 + 3 - 1 = 6$, and that $\hat{\mathbf{y}} = \hat{\mathbf{y}}_0 + \hat{\mathbf{y}}_\mathbf{R} + \hat{\mathbf{y}}_\mathbf{C}$, where

$$\hat{\mathbf{y}}_0 = p(\mathbf{y}|\mathbf{x}_0) = \bar{y}_{\cdot\cdot}\mathbf{x}_0 = \begin{bmatrix} 29.178 & 29.178 & 29.178 & 29.178 \\ 29.178 & 29.178 & 29.178 & 29.178 \\ 29.178 & 29.178 & 29.178 & 29.178 \end{bmatrix},$$

$$\hat{\mathbf{y}}_\mathbf{R} = \sum_i (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})\mathbf{R}_i$$

$$\begin{bmatrix} -0.325 & -0.325 & -0.325 & -0.325 \\ 0.102 & 0.102 & 0.102 & 0.102 \\ 0.222 & 0.222 & 0.222 & 0.222 \end{bmatrix},$$

$$\hat{\mathbf{y}}_\mathbf{C} = \sum_i (\bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot})\mathbf{C}_j$$

$$= \begin{bmatrix} 0.049 & 0.095 & -0.005 & 0.049 \\ 0.049 & -0.095 & -0.005 & 0.049 \\ 0.049 & 0.095 & 0.005 & 0.049 \end{bmatrix}.$$

Notice that $\hat{\mathbf{y}}_0, \hat{\mathbf{y}}_\mathbf{R}$, and $\hat{\mathbf{y}}_\mathbf{C}$ are orthogonal and that the ij element of $\hat{\mathbf{y}}$ is $\hat{y}_{ij} = \bar{y}_{\cdot\cdot} + (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}) + (\bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot})$. Therefore

$$\hat{\mathbf{y}} = \begin{bmatrix} 28.902 & 28.758 & 28.848 & 28.902 \\ 29.329 & 29.186 & 29.276 & 29.329 \\ 29.449 & 29.306 & 29.396 & 29.449 \end{bmatrix},$$

$$\mathbf{e} = \begin{bmatrix} -0.102 & -0.548 & 0.432 & 0.218 \\ 0.209 & 0.546 & 0.156 & 0.911 \\ 0.311 & 1.094 & 0.276 & -1.129 \end{bmatrix}$$

Further computation gives

$$
\begin{aligned}
\|\mathbf{y}\|^2 &= \sum_{ij} y_{ij}^2 = 10,221 \\
\|\hat{\mathbf{y}}\|^2 &= \|\hat{\mathbf{y}}_0\|^2 + \|\mathbf{y_R}\|^2 + \|\hat{\mathbf{y}}_\mathbf{C}\|^2 \\
&= \bar{y}_{..}^2(12) + 4\sum_i (\bar{y}_{i.} - \bar{y}_{..})^2 + 3\sum_j (\bar{y}_{.j} - \bar{y}_{..})^2 \\
&= 10,215.92 + 0.66 + 0.04 = 10,216.62 \\
\|\mathbf{e}\|^2 &= \|\mathbf{y} - \hat{\mathbf{y}}\|^2 = 4.38
\end{aligned}
$$

showing again that the Pythagorean Theorem holds.

Later, after we formulate probability models, and discuss their properties, we will be able to draw further conclusions about the contributions of concentration and composition to variation in drainage weight.

1.5 Some History

In his scholarly and fascinating history of the development of statistics before 1900, Stephen Stigler (1986) begins his first chapter, entitled "Least Squares and the Combination of Observations," with the following:

> The method of least-squares was the dominant theme – the *leitmotif* of nineteenth-century statistics. In several respects it was to statistics what the calculus had been to mathematics a century earlier. "Proofs" of the method gave direction to the development of statistical theory, handbooks explaining its use guided the application of the higher methods, and disputes on the priority of its discovery signaled the intellectual community's recognition of the method's value. Like the calculus of mathematics, this "calculus of observations" did not spring into existence without antecedents, and the exploration of its subtleties and potential took over a century. Throughout much of this time statistical methods were referred to as "the combination of observations". This phrase captures a key ingredient of the method of least squares and describes a concept whose evolution paced the method's development. The method itself first appeared in print in 1805.

Stigler refers to Adrien-Marie Legendre (1752-1833), who in 1805 wrote an eight page book *Nouvelles méthodes pour le détermination des orbites des comètes* (New methods for the determination of the orbit of the planets), with a nine page appendix, "Sur la méthode des maindres quarres" (On the method of least squares). Legendre began the appendix with a statement of his objective; here is Stigler's translation:

> In most investigations where the object is to deduce the most accurate possible results from observational measurements, we are led to a system of equations of the form
>
> $$E = a + bx + cy + fz + \cdots$$
>
> in which $a, b, c, f, \ldots$ are known coefficients, varying from one equation to the other, and $x, y, z, \ldots$ are known quantities, to be determined by the condition that each value of E is reduced either to zero, or to a very small quantity (Legendre, 1805).

In today's notation we might make the substitutions $E = -\varepsilon_i, -a = Y_i, b = x_{1i}, x = \beta_1, c = x_{2i}, y = \beta_2$, etc., and write the model as $-a = bx + cy + \cdots - E$ or $Y_i = \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \varepsilon_i$ or even as $\mathbf{Y} = \beta_1 \mathbf{x}_1 + \cdots + \beta_k \mathbf{x}_k + \varepsilon = \mathbf{X}\beta + \varepsilon$.

Again in Stigler's translation, Legendre wrote

> Of all the principles that can be proposed for this purpose, I think there is none more general, more exact, or more easy to apply, than that which we have used in this work; it consists of making the sum of squares of the errors a minimum. By this method, a kind of equilibrium is established among the errors which, since it prevents the extremes from dominating, is appropriate for revealing the state of the system which most nearly approaches the truth.

Legendre gave an example using data from the 1795 survey of the French meridian arc, in which there were $n = 5$ observations and $k = 3$ unknown parameters.

Though Carl Friedrich Gauss claimed in 1809 that he had used the method of least squares as early as 1795, it seems clear from published writings that Legendre should be given credit for the first development of least squares.

The statistical problem solved by Legendre had been faced earlier by astronomer Johann Tobias Mayer (1723-1762), mathematician Leonhard Euler (1707-1783) and scientist and mathematician Pierre-Simon Laplace (1749-1827) in considering astronomic data. We will illustrate their earlier solutions on some data concerning the motion of Saturn studied by Laplace in 1787. Table 1.5.1 is taken from Stigler's book.

Using Legendre's notation, these eighteenth century scientists considered the problem of solving the "equations"

$$E_i = a_i + w + b_i x + c_i y + d_i z \quad (i = 1, \ldots, 24) \tag{1.5.1}$$

given by setting the E_i's all equal to zero. Observations were made on 24 occasions when Saturn, the moon, and earth were aligned over 200 years. The dependent variable a_i was the difference between the observed longitude of Saturn and that predicted by Laplace's theory. The measurements b_i, c_i, d_i were simple functions of observations made on the orbit of Saturn at those times.

They knew (or would have known) that those 24 equations in four unknowns (w, x, y, z) had no single solutions and that therefore all the E_i's could not be made zero. Mayer's idea was to reduce his collection of equations to a number equal to the number of unknowns by adding across equations. In Mayer's case he had 27 equations with three unknowns, so he grouped the 27 equations into three groups of 9 each, and simply added coefficients to get 3 equations in three unknowns. As applied to the data of the Table 1.5.1 we could add the first 6, next 6, etc. to get 4 equations in four unknowns. Mayer chose the subset of equations to add according to the sizes of the coefficients, grouping large a_i's together, the next smallest together, and so on.

Euler had available observations on Saturn and Jupiter for the years 1582-1745 ($n = 75$) and had k 6 unknowns. He did not combine observations as did Mayer but instead tried to solve for his unknowns by using some periodicity of the coefficients to reduce the number of unknowns and by considering small sets of observations, trying to verify solutions on other small sets. He was largely unsuccessful, and wrote (Stigler's translation)

> Now, from these equations we can conclude nothing; and the reason, perhaps, is that I have tried to satisfy several observations exactly, whereas I should have only satisfied then approximately; and this error has then multiplied itself.

Thus, the most prolific of mathematicians, perhaps the greatest of analysts, failed even to proceed as far as Mayer.

In 1787 Laplace, eulogized by Poisson in 1827 as "the Newton of France" (Stigler, 1986, p. 31), and perhaps the greatest contributor to probability and statistics before 1900, considered the Saturn data of Table 1.5.1. Laplace reduced the 24 equations in four unknowns to 4 equations. The first new equation was the sum of all equations. The second was the difference between the sum of the first 12 and the sum of the second 12. The third was the sum of equations 3, 4, 10, 11,17, 18, 23, 24 minus the sum of equations 1, 7, 14, 20, the fourth was the sum of equations 2, 8, 9, 15, 16, 21, 22 minus the sum of equations 5, 6, 13, 19. Stigler describes some of Laplace's motivation, which now seems quite valid: Laplace obtained his jth equation by multiplying the original ith equation by a constant k_{ij} and then adding over i. His jth equation was therefore

$$0 = \sum_i k_{ij} a_i + x \sum_i k_{ij} b_i \; | \; y \sum_i k_{ij} c_i + z \sum_i k_{ij} d_i \qquad (1.5.2)$$

Laplace's k_{ij} were all 1, -1 or 0. Mayer's had all been 0 or 1. Legendre showed that the method of least squares leads to taking $k_{i1} = 1, k_{i2} = b_i, k_{i3} = c_i, k_{i4}$ d_i.

In general, with modern notation, the four simultaneous equations of (1.5.2) may be written as

$$\mathbf{K}'\mathbf{X}\mathbf{b} = \mathbf{K}'\mathbf{y}, \tag{1.5.3}$$

where $\mathbf{X}$ and $\mathbf{y}$ are respectively 4 by 2 and 4 by one matrices. Thus,

$$\mathbf{b} = (\mathbf{K}'\mathbf{X})^{-1}\mathbf{K}'\mathbf{y}, \tag{1.5.4}$$

whenever the inverse exists.

The column in Table 1.5.1 "Halley Residual" had been derived by Edmund Halley in 1676 using a different theory. Details are omitted.

In 1809 Gauss showed the connections among normally distributed errors, most probable parameter values (maximum likelihood estimates) and least squares. In 1810 Laplace published his central limit theorem and argued that this could justify the assumption of normally distributed errors, hence least squares. Laplace showed in 1811 that, at least, asymptotically, least squares estimators are normally distributed, and they are less variable than other linear estimators, i.e ., solutions of (1.5.1). Normality of the errors was not needed.

In 1823 Gauss showed that the asymptotic argument was unnecessary, that the variability of the solutions to (1.5.1) could be studied algebraically, and that least squares estimators had least variability. We will make this precise in Sections 1.3 and 1.4 with a discussion of the famous Gauss-Markov Theorem. The least squares theory and applications developed by Legendre, Gauss and Laplace were widely published. Stigler cites a compilation by Mansfield Merriman in 1877 of "writings related to the method of least squares", including 70 titles between 1805 and 1834, and 179 between 1835 and 1864.

Problem 1.5.1. *Let* $\mathbf{x}_1 = (x_{11}, x_{12}, x_{13}, x_{14}) = (1, 1, 1, 0)$, $\mathbf{x}_2 = (x_{21}, x_{22}, x_{23}, x_{24}) = (0, 1, 1, 1)$. *and* $\mathbf{X}$ *the 4 x 2 matrix with these vectors as columns. Let* $\mathbf{y} = (y_1, y_2, y_3, y_4)' = (3, 6, 4, 2)'$. *Suppose we wish to find* b_1 *and* b_2 *such that* $\hat{y}_i = b_1 x_{1i} + b_2 x_{2i}$ *is "close to"* y_i *for each i. That is, we want* $r_i = y_i - \hat{y}_i$ *to be zero or small for each i.*

(a) Use the method suggested by the technique Mayer used. That is, reduce the problem to the solution of two linear equations in 2 unknowns by summing the first two, and also by summing the last two equations. Find the matrix K of equation (1.5.1), the resulting vectors $\mathbf{b}$ *and* $\hat{\mathbf{y}} = \mathbf{X}\mathbf{b}$.

(b) Repeat (a) by choosing $\mathbf{K}$ *in a manner similar to the method of Laplace, determining one equation by summing over all four, the other by taking the difference between the sum of the first two and the sum of the last two.*

(c) Repeat (a) for the case that $\mathbf{K} = \mathbf{X}$.

(d) Evaluate the error sum of squares $\|\mathbf{y} - \hat{\mathbf{y}}\|^2$ *for each of (a), (b), (c). Are you surprised that the error sum of squares is smallest for (c)?*

Table 1.5.1: Laplace's Saturn Data.*

Eq. no.	Year (i)	$-a_i$	b_i	c_i	d_i	Laplace Residual	Halley Residual	L.S. Residual
1	1591	1′11.9″	−158.0	+0.22041	−0.97541	+1′33″	−0′54″	+1′36″
2	1598	3′32.7″	−151.78	+0.99974	−0.02278	−0.07	+0.37	−0.05
3	1660	5′12.0″	−89.67	+0.79735	+0.60352	−1.36	+2.58	−1.21
4	1664	3′56.7″	−85.54	−0.04241	+0.99910	−0.35	+3.20	−0.29
5	1667	3′31.7″	−82.45	−0.57924	+0.81516	−0.21	−3.50	−0.33
6	1672	3′32.8″	−77.28	−0.98890	−0.14858	−0.58	−3.25	−1.06
7	1679	3′9.9″	−70.01	−0.12591	−0.99204	−0.14	−1.57	−0.08
8	1687	4′49.2″	−62.79	−0.99476	+0.10222	−1.09	−4.54	−0.52
9	1690	3′26.8″	−59.66	+0.72246	+0.69141	+0.25	−7.59	−0.29
10	1694	2′4.9″	−55.52	−0.07303	+0.99733	+1.29	−9.00	+1.23
11	1697	2′37.4″	−52.43	−0.66945	+0.74285	+0.25	−9.35	+0.22
12	1701	2′41.2″	−48.29	−0.99902	−0.04435	+0.01	−8.00	−0.07
13	1731	3′31.4″	−18.27	−0.98712	−0.15998	−0.47	−4.50	−0.53
14	1738	4′9.5″	−11.01	−0.13759	−0.99049	−1.02	−7.49	−0.56
15	1746	4′58.3″	−3.75	+0.99348	+0.11401	−1.07	−4.21	−0.50
16	1749	4′3.8″	−0.65	+0.71410	+0.70004	−0.12	−8.38	+0.03
17	1753	1′58.2″	+3.48	−0.08518	+0.99637	+1.54	−13.39	+1.41
18	1756	1′35.2″	+6.58	−0.67859	+0.73452	+1.37	−17.27	−1.35
19	1760	3′14.0″	+10.72	−0.99838	−0.05691	−0.23	−22.17	−0.29
20	1767	1′40.2″	+17.98	−0.03403	−0.99942	+1.29	−13.12	+1.34
21	1775	3′46.0″	+25.23	+0.99994	+0.01065	+0.19	−2.12	−0.26
22	1778	4′32.9″	+28.33	+0.78255	+0.62559	−0.34	+1.21	−0.19
23	1782	4′4.4″	+32.46	+0.01794	+0.99984	−0.23	−5.18	−0.15
24	1785	4′17.6″	+35.56	−0.59930	+0.80053	−0.56	−12.07	−0.57

Residuals are fitted values minus observed values. *Source*: Laplace (1788). Reprinted with permission from *The History of Statistics: The Measurement of Uncertainty before 1900* by Stephen M. Stigler, Cambridge, MA: The Belknap Press of Harvard University Press.

1.6 Projection Operators

The purpose of this section is to study the transformation $P_V : \mathbf{y} \to \hat{\mathbf{y}}$ which transforms a vector $\mathbf{y} \in \Omega$ into its projection $\hat{\mathbf{y}}$ on a subspace V.

In applications a vector $\mathbf{y}$ will be observed. The model under consideration will specify that $\mathbf{y} = \theta + \epsilon$, for $\theta \subset V$, a known subspace of Ω, with ϵ a random vector, both θ and ϵ unknown. We will usually estimate θ by the projection of $\mathbf{y}$ onto V. We should therefore understand the properties of this projection as well as possible.

The transformation $P : \mathbf{y} \to p(\mathbf{y}|V)$ for a subspace V is linear, since $p(\alpha\mathbf{y}|V) = \alpha p(\mathbf{y}|V)$ and $p(\mathbf{y}_1 + \mathbf{y}_2|V) = p(\mathbf{y}_1|V) + p(\mathbf{y}_2|V)$. (The student should check this.)

Since $\hat{\mathbf{y}} = p(\mathbf{y}|V)$ implies that $p(\hat{\mathbf{y}}|V) = \hat{\mathbf{y}}$, the projection operator P is idempotent, i.e., $P^2 = P$. In addition, P is self-adjoint, since for each $\mathbf{x}, \mathbf{y} \in \Omega$, $(P\mathbf{x}, \mathbf{y}) = (P\mathbf{x}, P\mathbf{y}) = (\mathbf{x}, P\mathbf{y})$.

If Ω is the space of n-component column vectors, this means P may be represented as a symmetric matrix, a *projection matrix*. Thus, for this case the projection operator onto V is an $n \times n$ matrix $\mathbf{P}_V$ such that

$$\mathbf{P}_V' = \mathbf{P}_V \quad \text{and} \quad \mathbf{P}_V^2 = \mathbf{P}_V.$$

For $V = \mathcal{L}(\mathbf{x}_1, \ldots, \mathbf{x}_k)$ with $\mathbf{x}_1, \ldots, \mathbf{x}_k$ linearly independent column vectors, we have

$$p(\mathbf{y}|V) = \mathbf{X}\mathbf{b} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y},$$

where $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_k)$, so that

$$\mathbf{P}_V = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'.$$

It is easy to check that $\mathbf{P}_V$ is symmetric and idempotent.

Example 1.6.1. For simplicity we will refer to a projection operator as a *projection*.

(1) $\mathbf{P} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$ = projection onto the linear subspace of vectors $\begin{bmatrix} y_1 \\ y_2 \\ 0 \end{bmatrix}$ spanned by $\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$ and $\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$.

(2) $\mathbf{P} = \frac{1}{n}\mathbf{J}_n\mathbf{J}_n'$ = projection onto $\mathbf{J}_n$, the column vector of n 1's. Then $\mathbf{P}\mathbf{x} = x\mathbf{J}_n$, where $x = (\mathbf{x}, \mathbf{J}_n)/\|\mathbf{J}_n\|^2 = (\sum x_i)/n$.

(3) $\mathbf{P} = I_n - \frac{1}{n}\mathbf{J}_n\mathbf{J}_n'$ projection onto the subspace of column vectors whose components add to zero, i.e., are orthogonal to $\mathbf{J}_n$. $\mathbf{P}$ adjusts $\mathbf{y}$ by subtracting y from all components. $\mathbf{P}\mathbf{y}$ is the vector of deviations $y_i - y$.

(4) $\mathbf{P} = \mathbf{v}\mathbf{v}'/\|\mathbf{v}\|^2$ projection onto the one-dimensional subspace $\mathcal{L}(\mathbf{v})$.

(5) $\mathbf{P} = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 1/2 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ = projection onto the subspace spanned by $\begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$ and $\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$. Thus, $\mathbf{P}\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} (y_1 + y_2)/2 \\ (y_1 + y_2)/2 \\ y_3 \end{bmatrix}$.

Problem 1.6.1. *Show that for* $\underset{n\times k}{\mathbf{W}} = \underset{n\times k}{\mathbf{X}}\,\underset{k\times k}{\mathbf{B}}$ *with* $\boldsymbol{B}$ *nonsingular,* $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ *remains unchanged if* $\boldsymbol{X}$ *is replaced by* $\boldsymbol{W}$*. Thus,* $\boldsymbol{P}$ *is a function of the subspace spanned by the columns of* $\boldsymbol{X}$*, not of the particular basis chosen for this subspace.*

Theorem 1.6.1. *Let* A *be a linear operator on* Ω *which is idempotent and self-adjoint. Then* A *is the projection operator onto the range of* A*.*

Proof. We must show that for all $\mathbf{y} \in \Omega$, and $\mathbf{x} \subset R$ Range of A, $(A\mathbf{y}, \mathbf{x}) = (\mathbf{y}, \mathbf{x})$. If $\mathbf{x} \subset R$ then $\mathbf{x}$ $A\mathbf{z}$ for some $\mathbf{z} \in \Omega$. But $(A\mathbf{y}, \mathbf{x}) = (\mathbf{y}, A\mathbf{x})$ by self-adjointness (symmetry) and $A\mathbf{x} = AA\mathbf{z} = A\mathbf{z} = \mathbf{x}$ because A is idempotent. □

Problem 1.6.2. *Prove that the projection operator onto* $V^{\perp}$*, the collection of vectors in* Ω *orthogonal to* V*, is* $I - P_V$*. (*I *is the identity transformation.)*

Subspace $V_0 \subset V$: Let V be a subspace of Ω and let V_0 be a subspace of V. Let P and P_0 be the corresponding projection operators. Then

$$(1) \quad PP_0 = P_0 \quad \text{and} \quad (2) \quad P_0P = P_0.$$

Equivalently, if $\hat{\mathbf{y}} = p(\mathbf{y}|V)$ and $\hat{\mathbf{y}}_0 = P(\mathbf{y}|V_0)$ then (1) $p(\hat{\mathbf{y}}_0|V) = \hat{\mathbf{y}}_0$ and (2) $p(\hat{\mathbf{y}}|V_0) = \hat{\mathbf{y}}_0$. It is easy to check these equalities by merely noting in (1) that $\hat{\mathbf{y}}_0 \in V$ and $(\mathbf{v}, \hat{\mathbf{y}}_0) = (\mathbf{v}, \mathbf{y})$ for all $\mathbf{v} \in V_0$, and in (2) that $\hat{\mathbf{y}}_0 \subset V_0$ and $(\mathbf{v}, \hat{\mathbf{y}}_0) = (\mathbf{v}, \hat{\mathbf{y}})$ for all $\mathbf{v} \in V_0$.

Example 1.6.2. We will use Gram–Schmidt S-Plus and orthogonalization to determine an orthogonal basis for the column space V of the matrix W. S-Plus print-out follows.

```
> gramschmidt   # A function defined by the author.

function(W)
{ n = dim(W)[1]
k = dim(W)[2]
B = matrix(W[, 1])
for(j in 2:k) {
v = matrix(W[, j])
B = cbind(B, v - B %*% matrix((t(B) %*% v)/diag(t(B) %*% B)))
}
C = apply(B*B,2,sum)  # Vector of squared lengths of
# the columns of B.
```

```
H = diag(1/C^0.5)
K = B %*% H
list(B, K)  # The result is a list. The collection of columns
# of B is an orthogonal basis for V. The columns of K
# are the same as those of B, except that they have been
# normalized to have length one.
}
```

```
> X
     [,1] [,2] [,3] [,4]
[1,]    1    1    1    6
[2,]    1    2    3    2
[3,]    1    3    1    1
[4,]    1    4    2    0
[5,]    1    5    5    4
```

```
> W = cbind(X,c(2,4,1,0,5)) #Add a column to X so that
#it has rank 5.
```

```
> G = gramschmidt(W)
```

```
> G1 = B[[1]]; G2 = G[[2]]
```

```
> B1 #The collection of columns of G1 is an
#orthogonal basis for the column space of W.
```

```
     [,1] [,2] [,3]   [,4]    [,5]
[1,]    1   -2  0.0  2.200 -0.2641
[2,]    1   -1  1.3 -2.438  0.0704
[3,]    1    0 -1.4 -0.267  0.8098
[4,]    1    1 -1.1 -0.952 -0.7746
[5,]    1    2  1.2  1.457  0.1584
```

```
> G2 #The columns of G2 have length one,
#so G2 is an orthonormal matrix.
```

```
      [,1]   [,2]   [,3]    [,4]    [,5]
[1,] 0.447 -0.632  0.000  0.5904 -0.2268
[2,] 0.447 -0.316  0.518 -0.6543  0.0605
[3,] 0.447  0.000 -0.558 -0.0716  0.6955
[4,] 0.447  0.316 -0.438 -0.2556 -0.6653
[5,] 0.447  0.632  0.478  0.3910  0.1361
```

```
> t(G1)%*%G1    #Inner product matrix. Off-diagonal
#terms are very close to zero.
```

```
             [,1]       [,2]       [,3]       [,4]       [,5]
[1,]  5.00e+000  0.00e+000  6.66e-016 -5.55e-016 -5.55e-016
[2,]  0.00e+000  1.00e+001  6.66e-016 -9.99e-016 -1.33e-015
[3,]  6.66e-016  6.66e-016  6.30e+000 -6.22e-016 -1.57e-015
[4,] -5.55e-016 -9.99e-016 -6.22e-016  1.39e+001  2.91e-015
[5,] -5.55e-016 -1.33e-015 -1.57e-015  2.91e-015  1.36e+000

> t(G2)%*%G2    #Very close to the identity matrix.
             [,1]       [,2]       [,3]       [,4]       [,5]
[1,]  1.00e+000  2.71e-020  1.24e-016 -8.69e-017 -2.17e-016
[2,]  2.71e-020  1.00e+000  7.02e-017 -5.27e-017 -3.64e-016
[3,]  1.24e-016  7.02e-017  1.00e+000 -8.01e-017 -5.25e-016
[4,] -8.69e-017 -5.27e-017 -8.01e-017  1.00e+000  6.92e-016
[5,] -2.17e-016 -3.64e-016 -5.25e-016  6.92e-016  1.00e+000

> y = X %*% matrix(c(6,5,4,3))   #A vector in the column
 #space of X.
>y
     [,1]
[1,]   33
[2,]   34
[3,]   28
[4,]   34
[5,]   63

> P3 = G2[,1:3]%*% t(G2[,1:3])   #Projection onto the space V3
#spanned by the first 3 columns of X (and of G1 and G2).

> P3
             [,1]    [,2]    [,3]       [,4]    [,5]
[1,]  6.00e-001  0.4000  0.2000 -2.04e-017 -0.2000
[2,]  4.00e-001  0.5683 -0.0889 -1.27e-001  0.2476
[3,]  2.00e-001 -0.0889  0.5111  4.44e-001 -0.0667
[4,] -2.04e-017 -0.1270  0.4444  4.92e-001  0.1905
[5,] -2.00e-001  0.2476 -0.0667  1.90e-001  0.8286

>  sum(diag(P3))
[1] 3      #Since P3 is projection onto V3 and dim(V3) = 3.

> P3%*%P3  #Showing that P3 is idempotent. It is symmetric.

             [,1]    [,2]    [,3]       [,4]    [,5]
[1,]  6.0e-001  0.4000  0.2000 -5.10e-017 -0.2000
[2,]  4.0e-001  0.5683 -0.0889 -1.27e-001  0.2476
[3,]  2.0e-001 -0.0889  0.5111  4.44e-001 -0.0667
[4,] -5.1e-017 -0.1270  0.4444  4.92e-001  0.1905
```

```
[5,] -2.0e-001  0.2476 -0.0667  1.90e-001  0.8286

> yhat = P3 %*% y   #Projection of y onto V3.

> cbind(y,yhat)
     [,1] [,2]
[1,]   33 26.4
[2,]   34 41.3
[3,]   28 28.8
[4,]   34 36.9
[5,]   63 58.6

> t(X[,1:3]) %*%(y-yhat)   #Showing that (y - yhat) is
           [,1]   #orthogonal to V3.
[1,]  7.11e-015
[2,]  7.11e-015
[3,] -1.42e-014

> M = t(X[,1:3]) %*% X[,1:3] #Inner product matrix

> M  #for the first 3 columns of X.
     [,1] [,2] [,3]
[1,]    5   15   12
[2,]   15   55   43
[3,]   12   43   40

> MI = solve(M)   #MI is the inverse of M.

> X [,1:3] %*% MI %*% t(X[,1:3])    #Same as P3.
           [,1]    [,2]    [,3]       [,4]    [,5]
[1,]  6.00e-001  0.4000  0.2000  1.26e-015 -0.2000
[2,]  4.00e-001  0.5683 -0.0889 -1.27e-001  0.2476
[3,]  2.00e-001 -0.0889  0.5111  4.44e-001 -0.0667
[4,] -3.89e-016 -0.1270  0.4444  4.92e-001  0.1905
[5,] -2.00e-001  0.2476 -0.0667  1.90e-001  0.8286
```

Direct Sums In regression analysis and, in particular, in the analysis of variance, it will often be possible to decompose the space Ω or a subspace V into smaller subspaces, and therefore to increase understanding of the variation in the observed variable. If these smaller subspaces are mutually orthogonal, simple computational formulas and useful intepretations often result.

For any linear model it will be convenient to decompose Ω into the subspace V, and the error space $V^{\perp}$, so that every observation vector $\mathbf{y}$ is the sum of a vector in V and a vector in $V^{\perp}$.

In Example 1.4.4 V may be decomposed into the spaces $V_0 = \mathcal{L}(\mathbf{x}_0)$, $V_R =$

$\left\{\sum_1^3 a_i\mathbf{R}_i | \sum_1^3 \mathbf{a}_i = 0\right\}$, $V_C = \left\{\sum_1^4 b_j\mathbf{C}_j | \sum b_j = 0\right\}$, so that every vector in V is the sum of its projections onto these three orthogonal subspaces. It follows that every vector $\mathbf{y}$ in Ω is the sum of four orthogonal vectors, each being the projection of $\mathbf{y}$ onto one of the four orthogonal subspaces $V_0, V_R, V_C, V^\perp$. These subspaces were chosen for their simplicity. As will be seen in later chapters, Chapters 3 and 6 in particular, the decomposition of V into orthogonal subspaces, each of a relatively simple structure, provides increased understanding of the variation in the components of $\mathbf{y}$.

Definition 1.6.1. Subspaces $V_1, \ldots, V_k$ of Ω are linearly independent if $\mathbf{x}_i \in V_i$ for $i = 1, \ldots, k$ and $\sum_{i=1}^k \mathbf{x}_i = \mathbf{0}$ implies that $\mathbf{x}_i = \mathbf{0}$ for $i = 1, \ldots, k$.

Let $\mathcal{M}_{ij}$ denote the property: $V_i \cap V_j = \{\mathbf{0}\}$. For $i \neq j$ linear independence of V_i and V_j is equivalent to $\mathcal{M}_{ij}$, so that linear independence of $V_1, \ldots, V_k$ implies $\mathcal{M}$: [$\mathcal{M}_{ij}$ for all $i \neq j$]. However, $\mathcal{M}$ does not imply linear independence of $V_1, \ldots, V_k$. Students are asked to prove these statements in Problem 1.6.12. Thus, linear independence of subspaces is analogous to independence of events. Pairwise independence does not imply independence of more than two events.

Definition 1.6.2. Let $V_1, \ldots, V_k$ be subspaces of Ω. Then

$$V = \left\{\mathbf{x} | \mathbf{x} = \sum_1^k \mathbf{x}_i, \mathbf{x}_i \in V_i, i = 1, \ldots, k\right\}$$

is called the *sum* of $V_1, \ldots, V_k$, and is denoted by

$$V = V_1 + V_2 + \cdots + V_k.$$

If these subspaces are linearly independent we will write

$$V = V_1 \oplus V_2 \oplus \cdots \oplus V_k.$$

We then say that V is the *direct sum* of $V_1, \ldots, V_k$. The use of the $\oplus$ symbol rather than the + symbol implies that the corresponding subspaces are linearly independent.

Theorem 1.6.2. *The representation* $\mathbf{x} = \sum_1^k \mathbf{x}_i$ *for* $\mathbf{x}_i \in V_i$ *of elements* $\mathbf{x} \in V = V_1 + V_2 + \cdots + V_k$ *is unique if and only if the subspaces* $V_1, \ldots, V_k$ *are linearly independent.*

Proof. Suppose that these subspaces are linearly independent. Let $\mathbf{x} = \sum_1^k \mathbf{x}_i = \sum_1^k \mathbf{w}_i$ for $\mathbf{x}_i, \mathbf{w}_i \in V_i, i = 1, \ldots, k$. Then $\sum_{i=1}^k (\mathbf{x}_i - \mathbf{w}_i) = \mathbf{0}$ implying, by the linear independence of the V_i, that $\mathbf{x}_i - \mathbf{w}_i = \mathbf{0}$ for each i.

Suppose that the representation is unique, let $\mathbf{v}_i \in V_i$ for $i = 1, \ldots, k$, and let $\sum_{j=1}^k \mathbf{v}_i = \mathbf{0}$. Since $\mathbf{0} \in V_i$ for each i, and $\mathbf{0} = \mathbf{0} + \cdots + \mathbf{0}$, it follows that $\mathbf{v}_i = \mathbf{0}$ for each i, implying the independence of $V_1, \ldots, V_k$.

□

Theorem 1.6.3. *If $\{\mathbf{v}_{ij}|j = 1,\ldots,n_i\}$ is a basis for V_i for $i = 1,\ldots,k$ and $V_1,\ldots,V_k$ are linearly independent, then $\{\mathbf{v}_{ij}|j = 1,\ldots,n_i, i = 1,\ldots,k\}$ is a basis for $V = V_1 \oplus \cdots \oplus V_k$.*

Proof. For any $\mathbf{x} = \sum_1^k x_i$ for $\mathbf{x}_i \in V_i$, suppose $\mathbf{x}_i = \sum_1^{n_i} b_{ij}\mathbf{v}_{ij}$. Thus, $\mathbf{x} = \sum_{ij} b_{ij}\mathbf{v}_{ij}$, so the $\mathbf{v}_{ij}$ span V. It is enough then to show that the $\mathbf{v}_{ij}$ are linearly independent. Suppose $\sum_{ij} c_{ij}\mathbf{v}_{ij} = \mathbf{0}$ for some c_{ij}'s. By the independence of $V_1,\ldots,V_k$, $\sum_j c_{ij}\mathbf{v}_{ij} = \mathbf{0}$ for each i. The independence of $\mathbf{v}_{i1},\ldots,\mathbf{v}_{in_i}$ then implies $c_{ij} = 0$ for all j and i.

□

Corollary: If $V = V_1 \oplus V_2 \oplus \cdots \oplus V_k$ then

$$\dim(V) = \dim(V_1) + \cdots + \dim(V_k).$$

Definition 1.6.3. For any subspace V of Ω, the collection of all vectors in Ω which are orthogonal to V is called the *orthogonal complement* of V. This orthogonal complement will be denoted by $V^\perp$, read "vee-perp".

It is easy to verify that $V^\perp$ is a subspace, and that $P_{V^\perp} = I - P_V$. Since $V^\perp \cap V = \{\mathbf{0}\}$, $V^\perp$ and V are linearly independent.

Theorem 1.6.4. *Let V_1 and V_2 be subspaces of Ω. Then*

$$(V_1 + V_2)^\perp = V_1^\perp \cap V_2^\perp \quad \textit{and} \quad (V_1 \cap V_2)^\perp = V_1^\perp + V_2^\perp.$$

Proof. We prove only the first equality. The second is proved similarly. Suppose $\mathbf{v} \in (V_1 + V_2)^\perp$. Then for each element $\mathbf{x} \in V_1 + V_2$, it follows that $\mathbf{v} \perp \mathbf{x}$. In particular, $\mathbf{v} \perp \mathbf{x}_1$ for each $\mathbf{x}_1 \in V_1$ and $\mathbf{v} \perp \mathbf{x}_2$, for each $\mathbf{x}_2 \in V_2$. Thus $\mathbf{v} \in V_1^\perp \cap V_2^\perp$ and $(V_1 + V_2)^\perp \subset V_1^\perp \cap V_2^\perp$.

If $\mathbf{v} \in V_1^\perp \cap V_2^\perp$, then $\mathbf{v} \perp \mathbf{x}_1, \mathbf{v} \perp \mathbf{x}_2$ for all $\mathbf{x}_1 \in V_1, \mathbf{x}_2 \in V_2$. It follows that $\mathbf{v} \perp (b_1\mathbf{x}_1 + b_2\mathbf{x}_2)$ for all scalars b_1, b_2, and all $\mathbf{x}_1 \in V_1, \mathbf{x}_2 \in V_2$, hence that $\mathbf{v} \in (V_1 + V_2)^\perp$. Thus, $(V_1 + V_2)^\perp \supset V_1^\perp \cap V_2^\perp$.

□

Theorem 1.6.4 is the linear space version of DeMorgan's Laws for sets:

$$(A \cup B)^c = A^c \cap B^c \quad \text{and} \quad (A \cap B)^c = A^c \cup B^c.$$

Theorem 1.6.5. *For any subspace V and any $\mathbf{x} \in \Omega$, there exist unique elements $\mathbf{x}_1, \mathbf{x}_2$ such that $\mathbf{x} = \mathbf{x}_1 + \mathbf{x}_2, \mathbf{x}_1 = p(\mathbf{x}|V)$ and $\mathbf{x}_2 = p(\mathbf{x}|V^\perp)$.*

Proof. For existence take $\mathbf{x}_1 = p(\mathbf{x}|V), \mathbf{x}_2 = \mathbf{x} - \mathbf{x}_1$. Uniqueness follows from the linear independence of $V^\perp$ and V.

□

Example 1.6.3. Let Ω be the space of 4-component row vectors. Let $\mathbf{x}_1 = (1,1,1,1), \mathbf{x}_2 = (1,1,0,0), \mathbf{x}_3 = (1,0,1,0), V_1 = \mathcal{L}(\mathbf{x}_1,\mathbf{x}_2), V_2 = \mathcal{L}(\mathbf{x}_3)$. Then V_1

and V_2 are linearly independent, so that $V = V_1 \oplus V_2$ $\{(a, b+c, a+b, a+c, a)|a,b,c \in R_1\}$ has dimension 3.

$$V_1^\perp = \{(a,-a,b,-b)|a,b \subset R_1\}$$
$$V_2^\perp = \{(a,b,-a,c)|a,b,c \subset R_1\}$$
$$V^\perp = \{(a,-a,-a,a)|a \in R_1\}$$

so that

$$V^\perp = V_1^\perp \cup V_2^\perp$$

In general, $P_V = P_{V_1} + P_{V_2}$ only if V_1 and V_2 are orthogonal. They are *not* orthogonal in this example. Verify this by projecting $\mathbf{y}$ $(11,4,3,8)$ onto each of V_1, V_2, and V.

Theorem 1.6.6. *Let V be a subspace of Ω and let V_0 be a proper subspace of V. Let $V_1 = V_0^\perp \cap V$. Then (1) V_0 and V_1 are mutually orthogonal subspaces, (2) $V = V_0 \oplus V_1$, and (3) $P_{V_1} = P_V - P_{V_0}$.*

Proof. Part (1) is obvious. To prove (2) let $\mathbf{y} \in V$, and let $\hat{\mathbf{y}}_0 = p(\mathbf{y}|V_0)$. Then $\mathbf{y} = \hat{\mathbf{y}}_0 + (\mathbf{y} - \hat{\mathbf{y}}_0), \hat{\mathbf{y}}_0 \subset V_0, \mathbf{y} - \hat{\mathbf{y}}_0 \in V \cap V_0^\perp$. Thus $V \subset V_0 \oplus V_1$. Since $V \supset V_0$ and $V \supset V_1, V \supset V_0 \oplus V_1$, implying that $V = V_0 \oplus V_1$.

To prove (3) note that, since $V_1 \perp V_0, p(\mathbf{y}|V) = p(\mathbf{y}|V_0) + p(\mathbf{y}|V_1)$ for all $\mathbf{y}$. Thus $P_V = P_{V_0} + P_{V_1}$ and $P_{V_1} = P_V - P_{V_0}$. □

In fact, this theorem shows that Ω may be decomposed into three mutually orthogonal subspaces $V_0, V_0^- \cap V$, and $V^\perp$, whose direct sum is Ω.

Problem 1.6.3. *Let Ω be Euclidean 4-space (column vectors). Let*

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \quad \mathbf{x}_2 = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{x}_3 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 0 \end{bmatrix}$$

and let $V_0 = \mathcal{L}(\mathbf{x}_1)$ for $\mathbf{x}_1 = 3\mathbf{x}_3 - 2\mathbf{x}_2, V = \mathcal{L}(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)$. Find P_{V_0}, P_V and P_{V_1} for $V_1 = V_0^\perp \cap V$. For $\mathbf{y} = \begin{bmatrix} 0 \\ 2 \\ 14 \\ 1 \end{bmatrix}$ find $p(\mathbf{y}|V_0), p(\mathbf{y}|V_1), p(\mathbf{y}|V)$.

Theorem 1.6.7. *Let $V_1, \ldots, V_k$ be mutually orthogonal subspaces of Ω. Let $V = V_1 \oplus \cdots \oplus V_k$. Then $p(\mathbf{y}|V) = \sum_1^k p(\mathbf{y}|V_i)$ for all $\mathbf{y} \subset \Omega$.*

Proof. Let $\hat{\mathbf{y}}_i = p(\mathbf{y}|V_i)$. We must show that for each $\mathbf{x} \subset V, (\mathbf{y}, \mathbf{x}) = \left(\sum_1^k \hat{\mathbf{y}}_i, \mathbf{x}\right)$. Since $\mathbf{x} \in V, \mathbf{x} = \sum_{j=1}^k \mathbf{x}_j$ for some $\mathbf{x}_j \in V_j$ for

$j = 1, \ldots, k$. Thus

$$\left(\sum_1^k \hat{\mathbf{y}}_i, \mathbf{x}\right) = \left(\sum_{i=1}^k \hat{\mathbf{y}}_i, \sum_{i=1}^k \mathbf{x}_j\right) = \sum_{i=1}^k \sum_{i=1}^k (\hat{\mathbf{y}}_i, \mathbf{x}_j)$$
$$= \sum_i (\hat{\mathbf{y}}_i, \mathbf{x}_i) = \sum_i (\mathbf{y}, \mathbf{x}_i) = \left(\mathbf{y}, \sum_{i=1}^k \mathbf{x}_i\right) = (\mathbf{y}, \mathbf{x}).$$

The third equality follows from the orthogonality of the subspaces. The fourth follows from the definition of $\hat{\mathbf{y}}_i$.

□

In the case that $V = \Omega$ we see that $\mathbf{y} = p(\mathbf{y}|V) = \sum_1^k p(\mathbf{y}|V_i)$, and by the Pythagorean Theorem, $\|\mathbf{y}\|^2 = \sum_1^k \|p(\mathbf{y}|V_i)\|^2$. In applying this to the analysis of variance we will frequently make such a decomposition of the squared length of the observation vector $\mathbf{y}$. In fact, the analysis of variance may be viewed as the decomposition of the squared length of a vector into the sum of the squared lengths of several vectors, using the Pythagorean Theorem.

Example 1.6.4. Let Ω be the space of 2×3 matrices. Let $\mathbf{R}_1, \mathbf{R}_2$ be the row indicators and let $\mathbf{C}_1, \mathbf{C}_2, \mathbf{C}_3$, be the column indicators. Let $\mathbf{x}_0 = \sum_i \mathbf{R}_i$ $\sum_j \mathbf{C}_j$ be the matrix of all ones. Define $V_0 = \mathcal{L}(\mathbf{x}_0), V_R = \mathcal{L}(\mathbf{R}_1, \mathbf{R}_2) \cap V_0^\perp, V_C = \mathcal{L}(\mathbf{C}_1, \mathbf{C}_2, \mathbf{C}_3) \cap V_0^\perp$. It is easy to show that $V_R = \{\mathbf{v} | \mathbf{v} = \sum a_i \mathbf{R}_i, a_1 + a_2 = 0\}$ and $V_C = \{\mathbf{v} | \mathbf{v} = \sum b_j \mathbf{C}_j | \sum b_j = 0\}$. For example, $\begin{bmatrix} 2 & -3 & 1 \\ 2 & -3 & 1 \end{bmatrix} \in V_C$. The subspaces V_0, V_R, V_C are linearly independent and mutually orthogonal. Let $V = V_0 \oplus V_R \oplus V_C$. Then $p(\mathbf{y}|V) = \hat{\mathbf{y}}_0 + \hat{\mathbf{y}}_R + \hat{\mathbf{y}}_C$, where $\hat{\mathbf{y}}_0 = p(\mathbf{y}|V_0) = \bar{y}_{..}\mathbf{x}_0, \hat{\mathbf{y}}_R = p(\mathbf{y}|V_R) = \sum_i (\bar{y}_{i.} - \bar{y}_{..})\mathbf{R}_i$, and $\hat{\mathbf{y}}_C = p(\mathbf{y}|V_C) = \sum_i (\bar{y}_{.j} - \bar{y}_{..})\mathbf{C}_j$. Then, since $\Omega = V_0 \oplus V_R \oplus V_C \oplus V^\perp$ is the decomposition of Ω into four mutually orthogonal subspaces, $\mathbf{y} = \hat{\mathbf{y}}_0 + \hat{\mathbf{y}}_R + \hat{\mathbf{y}}_C + \mathbf{e}$, where $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = p(\mathbf{y}|V^\perp)$, and

$$\|\mathbf{y}\|^2 = \|\hat{\mathbf{y}}_0\|^2 + \|\hat{\mathbf{y}}_R\|^2 + \|\hat{\mathbf{y}}_C\|^2 + \|\mathbf{e}\|^2, \quad \|\hat{\mathbf{y}}_0\|^2 = \bar{y}_{..}^2$$
$$\|\hat{\mathbf{y}}_R\|^2 = 3\sum_i (\bar{y}_{i.} - \bar{y}_{..})^2, \quad \|\hat{\mathbf{y}}_C\|^2 = 2\sum_i (\bar{y}_{.j} - \bar{y}_{..})^2$$

Definition 1.6.4. The *null space* of an $m \times n$ matrix $\mathbf{A}$ is the collection of vectors $\mathbf{x} \in R_n$ such that $\mathbf{Ax} = \mathbf{0}$. We denote this null space by $N(\mathbf{A})$. The *column* (or range) *space* of $\mathbf{A}$ is $C(\mathbf{A}) = \{\mathbf{x} | \mathbf{x} = \mathbf{Ab}$ for some $\mathbf{b}\}$.

Theorem 1.6.8. *Let $\mathbf{A}$ be an $m \times n$ matrix. Then*

$$N(\mathbf{A}) = C(\mathbf{A}')^\perp \quad \textit{and} \quad N(\mathbf{A})^\perp = C(\mathbf{A}') \tag{1.6.1}$$

Proof. $\mathbf{w} \in N(\mathbf{A}) \Leftrightarrow \mathbf{w} \perp$ (row space of$\mathbf{A}$) $\Leftrightarrow \mathbf{w} \perp$ (column space of $\mathbf{A}'$) $\Leftrightarrow \mathbf{w} \in C(\mathbf{A}')^\perp$. The second statement of (1.6.1) follows by taking orthogonal complements on both sides.

□

Theorem 1.6.9. *Let $\mathbf{X}$ be an $n \times k$ matrix. Then $C(\mathbf{X}'\mathbf{X}) = C(\mathbf{X}')$.*

Proof. $\mathbf{w} \in C(\mathbf{X}'\mathbf{X})$ implies the existence of $\mathbf{b}$ such that $(\mathbf{X}'\mathbf{X})\mathbf{b} = \mathbf{w} = \mathbf{X}'(\mathbf{X}\mathbf{b})$, which implies $\mathbf{w} \in C(\mathbf{X}')$. Thus $C(\mathbf{X}'\mathbf{X}) \subset C(\mathbf{X}')$. $\mathbf{w} \in C(\mathbf{X}')$ implies that $\mathbf{w} = \mathbf{X}'\mathbf{b}$ for some $\mathbf{b} \in R_n$. Let $\hat{\mathbf{b}} = p(\mathbf{b}|C(\mathbf{X}))$. Then $\mathbf{X}'\hat{\mathbf{b}} = \mathbf{X}'\mathbf{b}$ and, since $\hat{\mathbf{b}} \in C(\mathbf{X})$, there exists $\mathbf{v}$ such that $\mathbf{X}\mathbf{v} = \hat{\mathbf{b}}$. Then $\mathbf{X}'\mathbf{X}\mathbf{v} = \mathbf{X}'\hat{\mathbf{b}} = \mathbf{X}'\mathbf{b} = \mathbf{w}$, so $\mathbf{w} \in C(\mathbf{X}'\mathbf{X})$. Thus $C(\mathbf{X}'\mathbf{X}) \supset C(\mathbf{X}')$. □

It is shown in most introductory courses in linear algebra that the dimensions of the row and column spaces of any matrix $\mathbf{X}$ are equal, and this common dimension is called the *rank* of $\mathbf{X}$. We therefore conclude that $\mathbf{X}, \mathbf{X}', \mathbf{X}'\mathbf{X}$, and $\mathbf{X}\mathbf{X}'$ all have the same rank. In particular, $\mathbf{X}'\mathbf{X}$ ($\mathbf{M}$) has full rank (is nonsingular) if and only if $\mathbf{X}$ has full column rank, i.e., has linearly independent columns.

Problem 1.6.4. *Let Ω R_3. For each subspace give the corresponding projection matrix $\mathbf{P}$. For each verify that $\mathbf{P}$ is idempotent and symmetric.*

(a) $\mathcal{L}(\mathbf{x})$ *for* $\mathbf{x} = (1, 0, -1)'$.

(b) $\mathcal{L}(\mathbf{x}_1, \mathbf{x}_2)$ *for* $\mathbf{x}_1 = (1, 1, 1)', \mathbf{x}_2 = (1, 0, 1)'$.

Problem 1.6.5. *For the subspace $V = \mathcal{L}(\mathbf{J}, \mathbf{x})$ of Problem 1.3.5, what is P_V? (Note that $\mathcal{L}(\mathbf{J}, \mathbf{x}^*) = V$). What is $P_{V^\perp}$? Let $V_0 = \mathcal{L}(\mathbf{J})$ and $V_1 = V \cap V_0^\perp$. What is P_{V_1}?*

Problem 1.6.6. *Let V_1 and V_2 be subspaces of Ω and let $V_0 = V_1 \cap V_2$. Under what conditions does $P_{V_0} = P_{V_1}P_{V_2}$? Always? Never?*

Problem 1.6.7. *Let V_1, V_2, V_3 be subspaces. Does $V_1 \cap (V_2 + V_3) = (V_1 \cap V_2) + (V_1 \cap V_3)$ in general? If not, does this hold if V_2 and V_3 are linearly independent?*

Problem 1.6.8. *(a) For Example 1.6.3 find six mutually orthogonal vectors $\mathbf{v}_i$ for $i = 1, \ldots, 6$ such that*

$$V_0 = \mathcal{L}(\mathbf{v}_1), \quad V_R = \mathcal{L}(\mathbf{v}_2), \quad V_C = \mathcal{L}(\mathbf{v}_3, \mathbf{v}_4), \quad V^\perp = \mathcal{L}(\mathbf{v}_5, \mathbf{v}_6)$$

(b) For $\mathbf{y} = \begin{bmatrix} 12 & 7 & 11 \\ 10 & 1 & 7 \end{bmatrix}$ *find $\hat{\mathbf{y}}_0, \hat{\mathbf{y}}_R, \hat{\mathbf{y}}_C, \hat{\mathbf{y}}, \mathbf{e}$, compute their lengths, and verify that the Pythagorean Theorem holds.*

Problem 1.6.9. *Let* $\mathbf{A} = \begin{bmatrix} 2 & 3 & 7 \\ 1 & 5 & 7 \end{bmatrix}$.

(a) Find a basis for the null space of $\mathbf{A}$ (see Theorem 1.6.8).

(b) Verify Theorem 1.6.9 for $\mathbf{X}$ $\mathbf{A}'$.

Problem 1.6.10. *Let $\mathbf{v}_1, \ldots, \mathbf{v}_n$ be an orthogonal basis for Ω.*

(a) Prove Parseval's Identity: For every $\mathbf{x}, \mathbf{y} \in \Omega$

$$(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} (\mathbf{x}, \mathbf{v}_i)(\mathbf{y}, \mathbf{v}_i)/\|\mathbf{v}_i\|^2.$$

(b) Verify (a) for $\Omega = R^3$, $\mathbf{v}_1 = (1, 1, 1)$, $\mathbf{v}_2 = (1, -1, 0)$, $\mathbf{v}_3 = (1, 1, -2)$, $\mathbf{x} = (3, 5, 8)$, $\mathbf{y} = (2, 1, 4)$.

Problem 1.6.11. *Let* V_1 *and* V_2 *be subspaces of* Ω. *Let* $V = V_1 \oplus V_2$. *Let* P_{V_1}, P_{V_2} *and* P_V *be the corresponding projection operators. Suppose that* $P_V = P_{V_1} + P_{V_2}$. *(This means that* $P_V \mathbf{y} = P_{V_1}\mathbf{y} + P_{V_2}\mathbf{y}$ *for every* $\mathbf{y} \in \Omega$.*) Prove that* $V_1 \perp V_2$. *Hint: Consider* $P_V \mathbf{v}_1$ *for* $\mathbf{v}_1 \in V_1$ *and recall that* $(\mathbf{v}_1 - P_{V_2}\mathbf{v}_1) \perp V_2$.

Problem 1.6.12. *Prove the statements made in the paragraph following Definition 1.6.1. To prove the last statement construct an example.*

Problem 1.6.13. *Let* $V_1, V_2, \ldots, V_k$ *be mutually orthogonal subspaces, none equal to* $\mathcal{L}(\mathbf{0})$. *Prove that they are linearly independent.*

1.7 Eigenvalues and Eigenvectors

In this section we summarize results concerning eigentheory. Though this material will not be heavily used in this course, it will be useful. Most proofs will be omitted.

(1) Let $\mathbf{A}$ be an $n \times n$ matrix. A real number λ and column vectors $\mathbf{v}$ satisfying the equation $\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$ will be called an eigenpair, with λ an eigenvalue, and $\mathbf{v}$ the corresponding eigenvector. The words *characteristic* and *latent* are often used instead of *eigen*. Thus, an eigenvector $\mathbf{v}$ is transformed into a vector whose direction remains the same, but whose length is multiplied by the corresponding eigenvalue λ.

(2) A symmetric matrix $\underset{n\times n}{\mathbf{A}}$ has n real eigenvalues, though these may not all be distinct. Eigenvectors corresponding to different eigenvalues are orthogonal. If there exist k, but not more than k, independent vectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$ corresponding to the same eigenvalue λ, then λ is said to have multiplicity k, and the equation $\det(\lambda\mathbf{I} - \mathbf{A}) = 0$ has root λ of multiplicity k. In this case all vectors in $\mathcal{L}(\mathbf{v}_1, \ldots, \mathbf{v}_k)$ are eigenvectors corresponding to λ, and k such vectors, say $\mathbf{w}_1, \ldots, \mathbf{w}_k$, which are mutually orthogonal, may be chosen.

If such mutually orthogonal eigenvectors are chosen for each different eigenvalue, then the entire collection $\mathbf{u}_1, \ldots, \mathbf{u}_n$ of mutually orthogonal eigenvectors corresponding to eigenvalues $\lambda_1, \ldots, \lambda_n$, where an eigenvalue is repeated k times if its multiplicity is k, span n-space.

Let $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$, the matrix with (ii) element λ_i, off-diagonal terms 0, and $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_n)$. Then $\mathbf{AU} = \mathbf{U}\Lambda$, and if the $\mathbf{u}_i$ are chosen to have length one,

$$\mathbf{U}'\mathbf{U} = \mathbf{I}_n, \quad \mathbf{U}'\mathbf{A}\mathbf{U} = \mathbf{U}'\mathbf{U}\Lambda = \Lambda, \quad \mathbf{A} = \mathbf{U}\Lambda\mathbf{U}'.$$

The representation $\mathbf{A} = \mathbf{U}\Lambda\mathbf{U}'$ is called the *spectral representation* of $\mathbf{A}$.

Recall that the trace of a square matrix $\mathbf{A}$ is the sum of its diagonal elements. It is easy to show that $\text{trace}(\mathbf{BC}) = \text{trace}(\mathbf{CB})$ whenever the matrix product makes sense. It follows therefore that whenever $\mathbf{A}$ has spectral representation $\mathbf{A} = \mathbf{U}\Lambda\mathbf{U}'$, $\text{trace}(\mathbf{A}) = \text{trace}(\Lambda\mathbf{U}'\mathbf{U}) = \text{trace}(\Lambda) = \sum \lambda_i$. Similarly, $\det(\mathbf{A}) = \det(\mathbf{U})\det(\Lambda)\det(\mathbf{U}') = (+1)\det(\Lambda)(+1) = \prod \lambda_i$.

Since, for any $r \times s$ matrix $\mathbf{C} = (\mathbf{c}_1, \ldots, \mathbf{c}_s)$

$$\mathbf{D} = \begin{pmatrix} \mathbf{d}_1 \\ \vdots \\ \mathbf{d}_s \end{pmatrix}, \quad \mathbf{CD} = \sum_{i=1}^{s} \mathbf{c}_i \mathbf{d}_i,$$

we may express $\mathbf{A}$ in the form

$$\mathbf{A} = \mathbf{U}\Lambda\mathbf{U}' = (\lambda_1 \mathbf{u}_1, \ldots, \lambda_n \mathbf{u}_n) \begin{pmatrix} \mathbf{u}_1' \\ \vdots \\ \mathbf{u}_n' \end{pmatrix} = \sum_{1}^{n} \lambda_i \mathbf{u}_i \mathbf{u}_i'.$$

The matrices $\mathbf{u}_i\mathbf{u}_i' = \mathbf{P}_i$ are projections onto the one-dimensional subspaces $\mathcal{L}(\mathbf{u}_i)$. If there are r different eigenvalues with multiplicities $k_1, \ldots, k_r$ then the $\mathbf{P}_i$ corresponding to the same eigenvalue may be summed to get the representation of $\mathbf{A}$,

$$\mathbf{A} = \sum_{1}^{r} \lambda_j \mathbf{P}_j^*,$$

where $\mathbf{P}_j^*$ is the projection onto the k_j-dimensional subspace spanned by the eigenvectors corresponding to λ_j.

(3) By definition a square matrix $\mathbf{A}$ is positive definite if the quadratic function $Q(\mathbf{x}) = \mathbf{x}'\mathbf{A}\mathbf{x} > 0$ for all $\mathbf{x} \neq \mathbf{0}$. It is non-negative definite if $Q(\mathbf{x}) \geq 0$ for all $\mathbf{x}$.

Example 1.7.1. Let $\mathbf{v}_1 = (1,1,1,1)'$, $\mathbf{v}_2 = (1,-1,0,0)'$, $\mathbf{v}_3 = (1,1,-2,0)'$, $\mathbf{v}_4 = (1,1,1,-3)'$. These $\mathbf{v}_i$ are mutually orthogonal. Let $\mathbf{P}_i$ be projection onto $\mathcal{L}(\mathbf{v}_i)$. Thus,

$$\mathbf{P}_i = \mathbf{v}_i\mathbf{v}_i'/\|\mathbf{v}_i\|^2.$$

Let

$$\mathbf{A} = 8\mathbf{P}_1 + 8\mathbf{P}_2 + 12\mathbf{P}_3 = \begin{bmatrix} 8 & 0 & -2 & 2 \\ 0 & 8 & -2 & 2 \\ -2 & -2 & 10 & 2 \\ 2 & 2 & 2 & 2 \end{bmatrix}.$$

Working backwards from $\mathbf{A}$, the roots of the fourth degree polynomial $\det(\lambda\mathbf{I}_4 - \mathbf{A}) = 0$ are $\lambda = 8, 8, 12, 0$ with corresponding eigenvectors $\mathbf{w}_1$, $\mathbf{w}_2$, $\mathbf{w}_3$, $\mathbf{w}_4$. The vectors $\mathbf{w}_1$, $\mathbf{w}_2$ may be arbitrarily chosen vectors in $\mathcal{L}(\mathbf{v}_1, \mathbf{v}_2)$, the subspace onto which $\mathbf{P}_1 + \mathbf{P}_2$ projects. They may be chosen to be orthogonal, and could be chosen to be $\mathbf{v}_1$ and $\mathbf{v}_2$; $\mathbf{w}_3$ and $\mathbf{w}_4$ are nonzero vectors in $\mathcal{L}(\mathbf{v}_3)$ and $\mathcal{L}(\mathbf{v}_4)$,

respectively. The lengths of eigenvectors are arbitrary. Since one eigenvalue is 0, $\mathbf{A}$ has rank 3. The determinant of $\mathbf{A}$ is the product of its eigenvalues, 0 in this case. The trace of $\mathbf{A}$ is the sum of its eigenvalues, 28 in this example.

Let $\mathbf{u}_i = \mathbf{v}_i/\|\mathbf{v}_i\|$, so these $\mathbf{u}_i$ have length one. Let $\mathbf{U} = (\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3, \mathbf{u}_4)$ and $\Lambda = \text{diag}(8, 8, 12, 0)$. Then $\mathbf{AU} = \mathbf{U}\Lambda$, $\mathbf{U}$ is an orthogonal matrix, and $\mathbf{A} = \mathbf{U}\Lambda\mathbf{U}'$. Here

$$\mathbf{U} = \begin{bmatrix} 1/2 & 1/\sqrt{2} & 1/\sqrt{6} & 1/(2\sqrt{3}) \\ 1/2 & -1/\sqrt{2} & 1/\sqrt{6} & 1/(2\sqrt{3}) \\ 1/2 & 0 & -2/\sqrt{6} & 1/(2\sqrt{3}) \\ 1/2 & 0 & 0 & -\sqrt{3}/2 \end{bmatrix}$$

$$\Lambda = \begin{bmatrix} 8 & 0 & 0 & 0 \\ 0 & 8 & 0 & 0 \\ 0 & 0 & 12 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \qquad \mathbf{AU} = \begin{bmatrix} 4 & 8/\sqrt{2} & 12/\sqrt{6} & 0 \\ 4 & -8/\sqrt{2} & 12/\sqrt{6} & 0 \\ 4 & 0 & -24/\sqrt{6} & 0 \\ 4 & 0 & 0 & 0 \end{bmatrix}$$

$$\mathbf{U}\Lambda\mathbf{U}' = \begin{bmatrix} 8 & 0 & -2 & 2 \\ 0 & 8 & -2 & 2 \\ -2 & 2 & 10 & 2 \\ 2 & 2 & 2 & 2 \end{bmatrix} = \mathbf{A}$$

Consider the quadratic form

$$Q(\mathbf{x}) = \mathbf{x}'\mathbf{A}\mathbf{x} = 8x_1^2 - 4x_1x_3 + 4x_1x_4 + 8x_2^2 - 4x_2x_3 + 4x_2x_4 + 10x_3^2 - 4x_3x_4 + 2x_4^2.$$

Since

$$\mathbf{A} = \sum_{i=1}^{4} \lambda_i \mathbf{P}_i, \quad Q(\mathbf{x}) = \sum_{i=1}^{4} \lambda_i (\mathbf{x}', \mathbf{P}_i\mathbf{x}) = \sum \lambda_i \|\hat{x}_i\|^2,$$

where $\hat{x}_i = \mathbf{P}_i\mathbf{x} = |(\mathbf{v}_i', \mathbf{x})/\|\mathbf{v}_i\|^2|\mathbf{v}_i$, and therefore $\|\hat{x}_i\|^2 = (\mathbf{v}_i', \mathbf{x})^2/\|\mathbf{v}_i\|^2$. Since one eigenvalue is zero, the others positive, $\mathbf{A}$ is nonnegative definite and $Q(\mathbf{x}) \geq 0$ for all $\mathbf{x}$. $\mathbf{A}$ is not positive definite since $Q(\mathbf{v}_4) = 0$.

These computations can be handled easily using R or S-Plus using the function "eigen".

```
> eigen(A)

$values:
[1] 12  8  8  0   #The order is different but the eigenvalues
#and the corresponding eigenvectors are the same.
$vectors:
             [,1]          [,2] [,3]     [,4]
[1,] -4.0825e-001  7.0711e-001 -0.5 -0.28868
[2,] -4.0825e-001 -7.0711e-001 -0.5 -0.28868
[3,]  8.1650e-001 -1.4634e-016 -0.5 -0.28868
[4,] -1.9429e-016  1.2561e-016 -0.5  0.86603
```

Using the representation $\mathbf{A} = \sum_1^n \lambda_i \mathbf{u}_i \mathbf{u}_i'$ above it is easy to show that a square symmetric matrix $\mathbf{A}$ is positive definite if and only if its eigenvalues are all positive, non-negative definite if and only if its eigenvalues are all nonnegative.

If $\mathbf{A}$ is non-negative definite we can write $\Lambda^{1/2} = \text{diag}(\lambda_1^{1/2}, \ldots, \lambda_n^{1/2})$, so $\mathbf{A} = \mathbf{U}\Lambda\mathbf{U}' = \mathbf{U}\Lambda^{1/2}\Lambda^{1/2}\mathbf{U}' = (\mathbf{U}\Lambda^{1/2})(\mathbf{U}\Lambda^{1/2})' = \mathbf{B}\mathbf{B}'$ for $\mathbf{B} = \mathbf{U}\Lambda^{1/2}$. The decomposition $\mathbf{A} = \mathbf{B}\mathbf{B}'$ is quite useful. It is not unique, since if $\mathbf{C}$ is any orthonormal matrix (satisfying $\mathbf{C}\mathbf{C}' = \mathbf{I}$), then $(\mathbf{B}\mathbf{C})(\mathbf{B}\mathbf{C})' = \mathbf{B}\mathbf{C}\mathbf{C}'\mathbf{B}' = \mathbf{B}\mathbf{B}' = \mathbf{A}$.

Letting $\mathbf{C} = \mathbf{U}\Lambda^{1/2}\mathbf{U}' = \sum \lambda_i^{1/2}\mathbf{P}_i$, we get $\mathbf{C}' = \mathbf{C}$, with $\mathbf{A} = \mathbf{C}'\mathbf{C} = \mathbf{C}^2$. The matrix $\mathbf{C}$ is the unique symmetric square root of $\mathbf{A}$.

Letting $\mathbf{y} = \mathbf{U}'\mathbf{x}$ for $\mathbf{U}$ as defined above, we get

$$Q(\mathbf{x}) = \mathbf{x}'\mathbf{A}\mathbf{x} = (\mathbf{U}\mathbf{y})'\mathbf{A}(\mathbf{U}\mathbf{y}) = \mathbf{y}'\mathbf{U}'\mathbf{A}\mathbf{U}\mathbf{y} = \mathbf{y}'\Lambda\mathbf{y} = \sum_1^k \lambda_i y_i^2$$

(4) Let $\mathbf{P}_V$ be the projection operator onto a subspace V of Ω. Then for $\mathbf{x} \in V$, $\mathbf{P}_V\mathbf{x} = \mathbf{x}$ so that all vectors in V are eigenvectors of $\mathbf{P}_V$ with eigenvalues 1. For $\mathbf{x} \in V^\perp$, $\mathbf{P}_V\mathbf{x} = \mathbf{0}$, so that all vectors in $V^\perp$ are eigenvectors of $\mathbf{P}_V$ with eigenvalue 0. The eigenvalue 1 has multiplicity equal to the dimension of V, while the eigenvalue 0 has multiplicity equal to $\dim(V^\perp) = n - \dim(V)$. Since from (2) $\text{trace}(\mathbf{A}) = \sum \lambda_i$, the trace of a projection matrix is the dimension of the subspace onto which it projects.

Partitioned Matrices (Seber, 1977):

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}' & \mathbf{D} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{A}^{-1} + \mathbf{F}\mathbf{E}^{-1}\mathbf{F}' & -\mathbf{F}\mathbf{E}^{-1} \\ -\mathbf{E}^{-1}\mathbf{F}' & \mathbf{E}^{-1} \end{bmatrix}, \quad \text{where} \quad \begin{matrix} \mathbf{E} = \mathbf{D} - \mathbf{B}'\mathbf{A}^{-1}\mathbf{B} \\ \mathbf{F} = \mathbf{A}^{-1}\mathbf{B} \end{matrix}$$

Singular Value Decomposition (Seber, 1977, p. 392): For $\mathbf{X}$ an $n \times k$ matrix of rank $r, n > k > r$, let the r positive eigenvalues of $\mathbf{X}\mathbf{X}'$ be $\sigma_1^2 \geq \sigma_2^2 \geq \cdots \geq \sigma_r^2 > 0$. Let $\mathbf{D}$ be the diagonal matrix with diagonal $(\sigma_1, \ldots, \sigma_r)$. Let the eigenvector of $\mathbf{X}\mathbf{X}'$ corresponding to σ_i^2 be $\mathbf{p}_i$ for each i, $1 \leq i \leq r$, and let $\mathbf{q}_i = \mathbf{X}'\mathbf{p}_i/\sigma_i$. Then $\mathbf{q}_i$ is an eigenvector of $\mathbf{X}'\mathbf{X}$ corresponding to eigenvalue σ_i^2. These vectors $\mathbf{p}_i$ may be chosen to be mutually orthonormal. It follows that the $\mathbf{q}_i$ are also orthogonal. Define

$$\mathbf{P} = (\mathbf{p}_1, \ldots, \mathbf{p}_r), \quad \mathbf{Q} = (\mathbf{q}_1, \ldots, \mathbf{q}_r)' = \mathbf{D}^{-1}\mathbf{P}'\mathbf{X}.$$

Then $\mathbf{X} = \mathbf{P}\mathbf{D}\mathbf{Q} = \sum \sigma_i \mathbf{p}_i \mathbf{q}_i'$. Thus, the linear transformation $\mathbf{X}\mathbf{x} = \mathbf{y}$, taking vectors $\mathbf{x} \in R_k$ into C = column space of $\mathbf{X}$, proceeds as follows. $\mathbf{Q}$ takes a vector $\mathbf{x} \in R_k$ with $(\mathbf{x}, \mathbf{q}_i) = c_i$ into $(c_1, \ldots, c_p)'$. $\mathbf{D}$ then multiplies each c_i by σ_i. $\mathbf{P}(\mathbf{D}\mathbf{Q})\mathbf{x} = \mathbf{X}\mathbf{x}$ is then $\sum_j c_j \sigma_j \mathbf{p}_j$, a vector in the column space of $\mathbf{X}$.

Example 1.7.2.
```
> X = matrix(c(1,1,1,1,1,1,2,3,4,5,1,3,1,2,5,6,
      2,1,0,4),5)
```

```
> X
     [,1] [,2] [,3] [,4]
[1,]    1    1    1    6
[2,]    1    2    3    2
[3,]    1    3    1    1
[4,]    1    4    2    0
[5,]    1    5    5    4

W = svd(X)
> W
$d:
[1] 11.31442  5.00647  1.87972  0.62113    #The sigmas, square
#roots of the eigen values of X'X.

$v:                      #The transpose of Q. The rows of Q form an
#orthogonal basis for the row space of X.
         [,1]       [,2]     [,3]     [,4]
[1,] -0.18514 -0.018964 -0.27068 -0.94451
[2,] -0.60544 -0.531496 -0.52255  0.27910
[3,] -0.53188 -0.245396  0.80148 -0.12050
[4,] -0.56238  0.810514 -0.10634  0.12444

$u: #The matrix P. The columns of P form an orthogonal basis
#for the column space of X.
         [,1]       [,2]     [,3]       [,4]
[1,] -0.41511  0.812394 -0.33505 -0.063254
[2,] -0.36382 -0.039372  0.46601 -0.803283
[3,] -0.27361 -0.209396 -0.60818 -0.166268
[4,] -0.32442 -0.526467 -0.40322 -0.111271
[5,] -0.71778 -0.132103  0.37164  0.557413

> P = W$u
> D = diag(W$d)
 #The 4 by 4 diagonal matrix with d on the diagonal.
> Q = t(W$v)

 t(P)%*%P     #The columns of P define an orthonormal basis for
  #the column space of X.
              [,1]          [,2]          [,3]          [,4]
[1,]  1.0000e+000  1.4583e-016 -2.4788e-016  3.7161e-017
[2,]  1.4583e-016  1.0000e+000  2.3180e-016 -8.0983e-017
[3,] -2.4788e-016  2.3180e-016  1.0000e+000  8.4527e-017
[4,]  3.7161e-017 -8.0983e-017  8.4527e-017  1.0000e+000
```

```
> Q %*%t(Q) #The rows of Q define an orthonormal basis for
#the row space of X.

              [,1]          [,2]          [,3]          [,4]
[1,]  1.0000e+000 -1.4417e-016 -1.3461e-016  4.9019e-017
[2,] -1.4417e-016  1.0000e+000 -1.8065e-016  9.6792e-017
[3,] -1.3461e-016 -1.8065e-016  1.0000e+000 -1.1267e-017
[4,]  4.9019e-017  9.6792e-017 -1.1267e-017  1.0000e+000

> P %*% D %*% Q      #Showing that PDQ = X.
     [,1] [,2] [,3]        [,4]
[1,]    1    1    1  6.000e+000
[2,]    1    2    3  2.000e+000
[3,]    1    3    1  1.000e+000
[4,]    1    4    2 -3.439e-015
[5,]    1    5    5  4.000e+000


> PX = X %*% solve(t(X)%*%X)%*%t(X)
> PX  #The projection matrix onto the column space of X.
          [,1]       [,2]      [,3]      [,4]       [,5]
[1,]  0.948560  0.0137174  0.157750 -0.150892  0.0308642
[2,]  0.013717  0.9963420 -0.042067  0.040238 -0.0082305
[3,]  0.157750 -0.0420668  0.516232  0.462734 -0.0946502
[4,] -0.150892  0.0402378  0.462734  0.557385  0.0905350
[5,]  0.030864 -0.0082305 -0.094650  0.090535  0.9814815

> PX %*% PX    #Verifying that PX is idempotent.
          [,1]       [,2]      [,3]      [,4]       [,5]
[1,]  0.948560  0.0137174  0.157750 -0.150892  0.0308642
[2,]  0.013717  0.9963420 -0.042067  0.040238 -0.0082305
[3,]  0.157750 -0.0420668  0.516232  0.462734 -0.0946502
[4,] -0.150892  0.0402378  0.462734  0.557385  0.0905350
[5,]  0.030864 -0.0082305 -0.094650  0.090535  0.9814815
```

Moore-Penrose or Pseudo-Inverse: The Moore-Penrose inverse or pseudo-inverse of the $n \times k$ matrix $\mathbf{X}$ is the $k \times n$ unique matrix $\mathbf{X}^+$ having the four properties: (1) $\mathbf{X}^+\mathbf{X}\mathbf{X}^+ = \mathbf{X}^+$, (2) $\mathbf{X}\mathbf{X}^+\mathbf{X} = \mathbf{X}$, (3) $\mathbf{X}^+\mathbf{X}$ is symmetric, (4) $\mathbf{X}\mathbf{X}^+$ is symmetric. For any vector $\mathbf{y} \in R_n$, $\mathbf{b} = \mathbf{X}^+\mathbf{y}$ is the unique vector in the row space of $\mathbf{X}$ such that $\mathbf{X}\mathbf{b}$ is the projection of $\mathbf{y}$ on the column space of $\mathbf{X}$. If $\mathbf{X}$ is nonsingular then $\mathbf{X}^+ = \mathbf{X}^{-1}$. The matrix $\mathbf{X}^+\mathbf{X}$ is the projection onto the row space of $\mathbf{X}$. The matrix $\mathbf{X}\mathbf{X}^+$ is the projection onto the column space of $\mathbf{X}$. If $\mathbf{X}$ has full column rank then $\mathbf{X}^+ = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. If V is the column space of $\mathbf{X}$, and $p(\mathbf{y}|V) = \mathbf{X}\hat{\boldsymbol{\beta}}$, then $\hat{\boldsymbol{\beta}} = \mathbf{X}^+\mathbf{y}$.

The Moore-Penrose inverse may be used to find solutions to the linear equa-

tion $\mathbf{Xb} = \mathbf{c}$. If this equation has a solution then $\mathbf{c}$ is in the column space of $\mathbf{X}$. That is, there exists some $\mathbf{w}$ such that $\mathbf{Xw} = \mathbf{c}$. Let $\mathbf{b} = \mathbf{X}^+\mathbf{c}$. Then $\mathbf{Xb} = \mathbf{XX}^+\mathbf{Xw} = \mathbf{Xw} = \mathbf{c}$. The general solution to the equation $\mathbf{Xb} = \mathbf{c}$ is given by $\mathbf{b} = \mathbf{X}^+\mathbf{c} + (\mathbf{I}_p \quad \mathbf{X}^+\mathbf{X})\mathbf{d}$, for $\mathbf{d}$ any vector in R_k. Taking $\mathbf{d}$ to be any vector orthogonal to the row space of $\mathbf{X}$, we get the unique solution $\mathbf{X}^+\mathbf{c}$ in the row space of $\mathbf{X}$.

The pseudo-inverse is related to the singular value decomposition of $\mathbf{X}$ in that $\mathbf{X}^+ = \mathbf{Q}'\mathbf{D}^{-1}\mathbf{P}'$.

Example 1.7.3. Let X be as defined in the S-Plus example above. Let W = svd(X), P = W$u , and Q = transpose(W$v) be as defined there. S-Plus print-out:

```
> XP = t(Q) %*% diag(1/W$d) %*% t(P)

> XP     #The Moore-Penrose inverse.
            [,1]      [,2]      [,3]      [,4]      [,5]
[1,]  0.14814815  1.16049  0.345679  0.234568 -0.888889
[2,]  0.00068587 -0.46685  0.131230  0.135345  0.199588
[3,] -0.15089163  0.37357 -0.203932 -0.109282  0.090535
[4,]  0.15843621 -0.17558 -0.019204 -0.068587  0.104938

> XP %*% X    #Projection onto the row space of X.
       #In this case this is the 4 by 4 identity matrix.

              [,1]          [,2]          [,3]          [,4]
[1,]  1.0000e+000  2.7756e-016  1.1102e-016  4.9960e-016
[2,] -7.9472e-017  1.0000e+000 -5.2356e-016 -5.0459e-016
[3,]  2.7756e-017 -1.6653e-016  1.0000e+000  3.3307e-016
[4,] -1.2143e-016 -1.4225e-016 -3.1572e-016  1.0000e+000

> X %*% XP  #Projection onto the columns space of X. See above.
          [,1]       [,2]      [,3]      [,4]       [,5]
[1,]  0.948560  0.0137174  0.157750 -0.150892  0.0308642
[2,]  0.013717  0.9963420 -0.042067  0.040238 -0.0082305
[3,]  0.157750 -0.0420668  0.516232  0.462734 -0.0946502
[4,] -0.150892  0.0402378  0.462734  0.557385  0.0905350
[5,]  0.030864 -0.0082305 -0.094650  0.090535  0.9814815
```

For a full discussion see *Regression and the Moore-Penrose Pseudoinverse* by Arthur Albert (1972).

Triangular Decomposition:

Let $\mathbf{A}$ be a symmetric non-negative definite matrix. There exist an infinite number of $n \times n$ matrices $\mathbf{B}$ such that $\mathbf{BB}' = \mathbf{A}$. Perhaps the easiest such

matrix to find is one of the form (lower triangular)

$$\mathbf{B} = \begin{bmatrix} b_{11} & 0 & \dots & 0 \\ b_{21} & b_{22} & \cdots & 0 \\ \cdot & \cdot & \cdots & 0 \\ \cdot & \cdot & \cdots & 0 \\ \cdot & \cdot & \cdots & 0 \\ b_{n1} & b_{n2} & \cdots & b_{nn} \end{bmatrix}$$

The **Cholesky decomposition** is of the form $\mathbf{A} = \mathbf{C}'\mathbf{C}$, with $\mathbf{C}$ upper-triangluar, so that $\mathbf{C}' = \mathbf{B}$. Elements of B may be found as follows: $b_{11}^2 = a_{11}$, so $b_{11} = \sqrt{a_{11}}$. Then, since $b_{i1}b_{11} = a_{i1}$ we have

$$b_{i1} = a_{i1}/b_{11} \quad \text{for} \quad i = 2, \dots, n.$$

Suppose b_{ij} has already been found for $j = 1, \dots, k-1$ and $i = 1, \dots, n$ for $k \geq 1$. Then we can find b_{ik} inductively. Since, $\sum_{j=1}^{k} b_{kj}b_{kj} = a_{kk}$, it follows that $b_{kk}^2 = a_{kk} - \sum_{j=1}^{k-1} b_{kj}^2$. Then

$$b_{kk} = \left(a_{kk} - \sum_{j=1}^{k-1} b_{kj}^2 \right)^{1/2}.$$

Since $\sum_{j=1}^{k} b_{ij}b_{kj} = a_{ik}$ for $i > k$, it follows that

$$b_{ik} = \left(a_{ik} - \sum_{j=1}^{k-1} b_{ij}b_{kj} \right) / b_{kk} \quad \text{for} \quad i > k.$$

Repeating for each k produces $\mathbf{B}$.

To summarize:

(1) Compute $b_{11} = (a_{11})^{1/2}$, let $b_{i1} = a_{i1}/b_{11}$, and let $k = 2$.

(2) Let $b_{kk} = \left(a_{kk} - \sum_{j=1}^{k-1} b_{kj}^2 \right)^{1/2}$. ($\mathbf{A}$ is non-negative definite if and only if the term in parentheses is non-negative for each k.)

(3) Let $b_{ik} = \left(a_{ik} - \sum_{j=1}^{k-1} b_{ij}b_{kj} \right) \Big/ b_{kk}$ for $i > k$.

(4) Replace k by $k + 1$ and repeat (2) and (3) until $k > n$.

(5) Let $b_{ij} = 0$ for $i < j$.

If any $b_{kk} = 0$ in step (3) then set $b_{ik} = 0$ for $i \geq k$.

All of this can be accomplished using R or S-Plus, using the function "chol". For example, let A be as in Example 1.7.1. More S-Plus:

```
> A = matrix(c(8,0,-2,2,0,8,-2,2,-2,-2,10,2,2,2,2,2),4)

> A
     [,1] [,2] [,3] [,4]
[1,]    8    0   -2    2
[2,]    0    8   -2    2
[3,]   -2   -2   10    2
[4,]    2    2    2    2

> C = chol(A)
Warning messages:
  Choleski decomposition not of full rank in: chol(A)

> B = t(C)

> B
         [,1]     [,2] [,3] [,4]
[1,]  2.82843  0.00000    0    0
[2,]  0.00000  2.82843    0    0
[3,] -0.70711 -0.70711    3    0
[4,]  0.70711  0.70711    1    0

> B%*%t(B)
     [,1] [,2] [,3] [,4]
[1,]    8    0   -2    2
[2,]    0    8   -2    2
[3,]   -2   -2   10    2
[4,]    2    2    2    2
```

Problem 1.7.1. *Let* $\mathbf{A} = \begin{pmatrix} 14 & -2 \\ -2 & 11 \end{pmatrix}$.

(a) Find the eigenvalues λ_1, λ_2 *and corresponding length-one eigenvectors* $\mathbf{u}_1, \mathbf{u}_2$ *for* $\mathbf{A}$.

(b) Define $\mathbf{U}$ *and* Λ *as in Section 1.7 and show that* $\mathbf{A} = \mathbf{U}\Lambda\mathbf{U}'$ *and* $\mathbf{U}\mathbf{U}' = \mathbf{I}_2$.

(c) Give the projections $\mathbf{P}_1^*$ *and* $\mathbf{P}_1^*$ *of Section 1.7 and show that* $\mathbf{A} = \lambda_1\mathbf{P}_1^* + \lambda_2\mathbf{P}_2^*$.

(d) Is $\boldsymbol{A}$ positive definite? Why?

Problem 1.7.2. *What are the eigenvalues and eigenvectors of the projection matrices $\boldsymbol{P}$ of examples 1, 2, 3, 4, 5 of Example 1.6.2?*

Problem 1.7.3. *For $n \times k$ matrix $\boldsymbol{X}$ of rank k, what are the eigenvalues and vectors for $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$? What is $det(\boldsymbol{P})$ if $n > k$? If $n = k$?*

Problem 1.7.4. *Let $n \times n$ matrix $\boldsymbol{A}$ have nonzero eigenvalue λ and corresponding eigenvector $\boldsymbol{v}$. Show that*

(a) $\mathbf{A}^{-1}$ (if it exists) has an eigenvalue λ^{-1}, eigenvector $\boldsymbol{v}$.

(b) $\mathbf{I} - \mathbf{A}$ has an eigenvalue $1 - \lambda$, eigenvector $\boldsymbol{v}$.

(c) For $\mathbf{A} = \mathbf{BC}$, $\boldsymbol{CB}$ has eigenvalue λ, eigenvector $\boldsymbol{Cv}$.

Problem 1.7.5. *Give 2×2 matrices which satisfy the following:*

(a) Positive definite.

(b) non-negative definite, but not positive definite.

(c) Not non-negative definite.

Problem 1.7.6. *Let $\boldsymbol{A}$ be positive definite and let $\mathbf{v} \subset R_n$. Prove that $(\mathbf{A} + \mathbf{vv}')^{-1} = \mathbf{A}^{-1}(\mathbf{I} - c\mathbf{vv}'\mathbf{A}^{-1})$ for $c = 1/(1 + \mathbf{v}'\mathbf{A}^{-1}\mathbf{v})$.*

Problem 1.7.7. *Determine whether the quadratic form $Q(x_1, x_2, x_3) = 2x_1^2 + 2x_2^2 + 11x_3^2 + 16x_1x_2 - 2x_1x_3 - 2x_2x_3$ is non-negative definite. Hint: What is the matrix corresponding to $\boldsymbol{Q}$? One of its eigenvalues is 12.*

Problem 1.7.8. *For $\mathbf{A} = \begin{bmatrix} 5 & -1 \\ -1 & 10 \end{bmatrix}$ find a matrix $\boldsymbol{B}$ such that $\mathbf{A} = \mathbf{BB}'$.*

Problem 1.7.9. *Let $\mathbf{G} = \begin{bmatrix} 2 & 1 & -1 \\ 1 & 3 & 0 \\ 1 & 0 & 4 \end{bmatrix}$.*

(a) Use the formula for the inverse of partitioned matrices for the case that A is a one by one matrix to find $\mathbf{G}^{-1}$.

(b) Use the formula for the inverse of partitioned matrices for the case that A is a two by two matrix to find $\mathbf{G}^{-1}$.

Problem 1.7.10. *Let $\mathbf{X} = \begin{bmatrix} 5 & -1 \\ -1 & 5 \\ 2 & 2 \end{bmatrix}$. Find the singular value decomposition of $\boldsymbol{X}$. Also find the Moore-Penrose inverse $\mathbf{X}^+$ and verify its four defining properties.*

Problem 1.7.11. *Let* $\mathbf{A} = \mathbf{UDV}$ *be the singular value decomposition of* $\boldsymbol{A}$. *Express the following matrices in terms of* $\boldsymbol{U}$, $\boldsymbol{D}$, *and* $\boldsymbol{V}$.

(a) $\mathbf{A}'\mathbf{A}$

(b) $\mathbf{AA}'$

(c) $\mathbf{A}^{-1}$ *(assuming* $\boldsymbol{A}$ *is nonsingular)*

(d) $\mathbf{A}^n \quad \mathbf{AA}\cdots\mathbf{A}$ *(*n *products), assuming* $\boldsymbol{A}$ *is square.*
In the case that the singular values are $\sigma_1 > \sigma_2 \geq \sigma_3 \geq \cdots \geq \sigma_r > 0$, *show that* $\lim_{n\to\infty} \mathbf{A}^n/\sigma_1^n = \mathbf{p_1 p_1'}$.

(e) Projection onto the column space of $\boldsymbol{A}$.

(f) Projection onto the row space of $\boldsymbol{A}$.

(g) What are $\boldsymbol{U}$, $\boldsymbol{D}$, *and* $\boldsymbol{V}$ *for the case that* $\mathbf{A} = \mathbf{a}$ *is an* $n\times 1$ *matrix? What is* $\mathbf{A}^+$?

Problem 1.7.12. *Let* $\boldsymbol{A}$ *be a symmetric* $n \times n$ *matrix of rank one.*

(a) Show that $\boldsymbol{A}$ *can be expressed in the form* $\mathbf{A} = c\mathbf{vv}'$, *for a real number* c, *vector* $\boldsymbol{v}$ *of length one.*

(b) Prove that either $\boldsymbol{A}$ *or* $-\mathbf{A}$ *is non-negative definite .*

(c) Give the spectral decomposition for $\boldsymbol{A}$ *in terms of* c *and* $\boldsymbol{v}$.

Problem 1.7.13. *Let* $\mathbf{G} = \begin{bmatrix} 2 & 1 & 2 & 0 \\ 1 & 3 & 1 & 1 \\ -1 & -2 & 1 & -1 \\ 0 & 1 & 4 & 5 \end{bmatrix}$.

Use the partitioned matrix formula with $\boldsymbol{A}$ *the two by two matrix on the upper-left to show that* $\mathbf{G}^{-1} = (1/8)\begin{bmatrix} 1 & -17 & 27 & -2 \\ 0 & 8 & 8 & 0 \\ 5 & 13 & 23 & 2 \\ -4 & -12 & -20 & 0 \end{bmatrix}$.

Problem 1.7.14. *Let* $\mathbf{A} = \begin{bmatrix} 35 & 7 & 23 & 11 & 31 \\ -7 & 14 & 19 & 10 & -11 \\ -23 & 19 & 33 & 19 & -23 \\ -11 & 10 & 19 & 14 & -7 \\ 31 & -11 & -23 & -7 & 35 \end{bmatrix}$.

(a) Find the eigenvalues and corresponding eigenvectors for A. *Hints: The eigenvalues and eigenvectors consist entirely of integers. Two of them are any multiples of* $(1,1,1,1,1)'$ *and* $(2,-1,\ 2,\ -1,2)'$). A *has rank 3.*

(b) What are the eigenvalues and eigenvectors of A^2 *and, more generally, of* A^n?

(c) What is the determinant of A?

(d) Verify that that trace(Λ) is the sum of the eigenvalues.

(e) Let V be the column space of A. What is the projection matrix P_V onto V? Verify that P_V is idempotent.

(f) What is the Moore-Penrose inverse of A?

Problem 1.7.15. *For X as in example 1.7.2, find X^+, its Moore-Penrose inverse.*