Notes for Statistics 8311
Linear Models

Fall 2006

R. D. Cook, K. Larntz, and S. Weisberg

Contents

Chapter 1

Introduction

Linear models have a dominant role in statistical theory and practice. Most standard statistical methods are special cases of the general linear model, and rely on the corresponding theory for justification.

The goal of this course is to develop the theoretical basis for analyses based on a linear model. We shall be concerned with laying the theoretical foundation for simple as well as complex data sets.

Linear models is one of the oldest topics in the statistics curriculum. The main role of linear models in statistical practice, however, has begun to undergo a fundamental change due in large measure to available computing. Balanced experiments were often required to make analysis possible. This has produced a fundamental change in the way we can think about linear models, as much less stress can be placed on the special cases where computations are easy and more can be placed on general ideas. Topics that might have been standard, such as the recovery of interblock information in an incomplete block experiment, is of much less interest when computers can be used to appropriately maximize functions.

However, standard results are so elegant, and so interesting, that they deserve study in their own right, and for that reason we will study the traditional body of material that makes up linear models, including many standard simple models as well as a general approach.

The goal of these notes is to develop a *coordinate-free approach* to linear models. Coordinates can often sever to make problems unnecessarily complex, and understanding the features of a problems that are not dependent on coordinates is extremely valuable. The problems introduced by parameters are more easily understood given the coordinate-free background.

1.1 Some simple examples

1.1.1 One sample problem

The simplest linear model has data $y_i, i = 1, \ldots, n$ such that each y_i has the same distribution with mean μ and variance $\sigma^2 > 0$. Normality, or other distributional assumptions, are sometimes needed, but will not be used in the first few weeks of the course. In the one-sample problem, the goals are to learn about μ and possibly σ^2.

A *model* for this problem can we obtained by writing:

$$\begin{aligned} y_i &= \mu + (y_i - \mu) \\ &= \mu + \varepsilon_i \end{aligned}$$

where $\varepsilon_i = y_i - \mu$. Each observation is then taken to be the sum of a fixed part, in this case the parameter μ, and a random part ε_i, a random variable with zero mean and variance σ^2. In the spirit of this course, we will collect the responses into a vector $y = (y_1, \ldots, y_n)'$, and the ε_i into $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$. Writing J_n to be a vector of length n of all ones, the one-sample model can be written as

$$y = J_n\mu + \varepsilon$$

We will soon be learning the linear algebraic background to interpret this equation. The vector on the left is any arbitrary vector in n-dimensional space. On the right we have two vectors. ε is also an arbitrary vector in n-dimensional space, while $J_n\mu$ is a vector that is *constrained* to live in a part of n-dimensional space. This will be a characteristic form of (fixed-effect) linear models.

1.1.2 One way layout

Suppose we let y_{ij} be the jth observation in the ith population, $i = 1, \ldots, p; j = 1, \ldots, n_i$ be $n = \sum n_i$ independent observations. We then specify a mean structure:

$$\begin{aligned} \mathrm{E}(y_{ij}|\text{Group} = i) &= \mu_i \\ \mathrm{Var}(y_{ij}|\text{Group} = i) &= \sigma^2 \end{aligned} \tag{1.1}$$

so each group has its own mean but a common variance. This model is (somewhat) more complex because the mean now depends on the index and is therefore conditional. In matrix terms, suppose that $p = 3$, and write $y = (y_{11}, y_{12}, \ldots, y_{pn_p})'$

to be the vector of responses. Then we can write

$$y = \begin{pmatrix} J_{n_1} & 0 & 0 \\ 0 & J_{n_2} & 0 \\ 0 & 0 & J_{n_3} \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix} + \varepsilon \tag{1.2}$$

This is a model is just like the one-sample model except that the description of the fixed part is more complicated. The fixed part is now in a more complex space of dimension p rather than 1.

In such a model we may wish to address several goals:

1. Estimate the cell means μ_i, and obtain estimates of uncertainty.
2. Test hypotheses such as $\mu_1 = \mu_2 = \ldots = \mu_p$, or $\mu_i = \mu_j$ or more generally $\sum \alpha_i \mu_i$ = constant, where the α_i are known numbers.
3. Estimate the index of the largest of the μ_i. A *comparative experiment* is one in which several treatments indexed here from $1, \ldots, p$, are to be compared, and the goal is to decide which is the best one or the best few. This leads to many interestng questions, in particular many questions concerning how to make inferences when faced with mulitple objectives (comparing many treatments).

and so on. This model is linear because it is linear in the unknown location parameters μ_i.

Parameterizaton. A general form of the one-way model given by (1.2), is

$$y = X\beta + \varepsilon$$

where

$$X = \begin{pmatrix} J_{n_1} & 0 & 0 \\ 0 & J_{n_2} & 0 \\ 0 & 0 & J_{n_3} \end{pmatrix}; \qquad \beta = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix}$$

This is in fact a *parametric* or *coordinate* version of a linear model because of the fixed choice of X. In fact, (1.2) is just one of many possible ways of writing the linear model for the one-way classification. If A is any $p \times p$ nonsingular matrix, meaning that there is a matrix A^{-1} such that $AA^{-1} = I$, we can write

$$\begin{aligned} y &= XAA^{-1}\beta + \varepsilon \\ &= (XA)(A^{-1}\beta) + \varepsilon \\ &= X^*\gamma + \varepsilon \end{aligned}$$

which is a completely equivalent form of this linear model, but with parameters γ rather than β. There are several different choices that are commonly used for A (we set $p = 3$ for illustration):

1. The three parameters are the overall mean μ, $\alpha_1 = \mu_1 - \mu$, and $\alpha_2 = \mu_2 - \mu$. This is the "effects" parameterization seen most often.

$$A_1 = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & -1 & -1 \end{pmatrix}$$

2. This sets the parameters to be μ_1, $\mu_2 - \mu_1$ and $\mu_3 - \mu_1$. This parameterization is the default used by R.

$$A_2 = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}$$

3. This is called the Helmert parameterization, and is the default used by S-Plus. It is convenient for computing, but usually not convenient for interpretation.

$$A_3 = \begin{pmatrix} 1 & 1 & 1 \\ 1 & -1 & 1 \\ 1 & 0 & -2 \end{pmatrix}$$

The approach to linear models we use will try to avoid specific parameterization, since it is not relevant to for many important topics.

1.1.3 One-way random effects

The one-way model we have just discussed was a *conditional* model for fixed groups. Suppose that the groups were in fact a random sample from a population of groups. Since all that changes from group to group is the mean, one way to view this problem is to assume that the μ_i are random draws from a population, with mean μ and variance τ^2. The rules for iterated mean and variance can then be applied to get the unconditional model,

$$\begin{aligned} \mathrm{E}(y_{ij} &= \mathrm{E}\left[\mathrm{E}(y_{ij}|\text{Group} = i)\right] \\ &= \mathrm{E}\left[\mu_i\right] \\ &= \mu \end{aligned}$$

and

$$\begin{aligned} \text{Var}(y_{ij} &= \text{E}\left[\text{Var}(y_{ij}|\text{Group} = i)\right] + \text{Var}\left[\text{E}(y_{ij}|\text{Group} = i)\right] \\ &= \text{E}\left[\sigma^2\right] + \text{Var}\left[\mu_i\right] \\ &= \sigma^2 + \tau^2 \end{aligned}$$

Thus, the unconditional model is that $\text{E}(y_{ij}) = \mu$ but $\text{var}(y_{ij}) = \sigma^2 + \tau^2$. In addition, although the y_{ij} are conditionally independent given group, they are unconditionally correlated, since $\text{cov}(y_{ij}, y_{ik}) = \tau^2$. The simpler mean structure for the random effects model is offset by a more complex variance structure.

1.1.4 Simple linear regression

The simple linear regression model is a special case of (1.1), if we take

$$\mu_i = \beta_0 + \beta_1 x_i \tag{1.3}$$

and further assume that the x_i are known, fixed constants. The model can be written as

$$y_{ij} = \beta_0 + \beta_1 x_i + \varepsilon_{ij}, i = 1, \ldots, p; j = 1, \ldots, n_i \tag{1.4}$$

One usually sees this model written as

$$y_k = \beta_0 + \beta_1 x_k + \varepsilon_k, \text{ where } k = 1, \ldots, n = \sum n_i \tag{1.5}$$

losing the identification of observations with a population. For this model we may wish to:

1. Estimate the βs and σ^2.
2. Make tests concerning the βs, in particular of $\beta_1 = 0$.
3. Obtain interval estimates for $\beta_0 + \beta_1 x_i$. This is the *prediction problem*.
4. Examine the assumption that the cell means μ_i are linear in the xs.

and so on. You should have all seen the simple regression model in great detail, especially in Stat 8061 if not elsewhere, and we shall look at regression only as a special case of the general linear model.

Chapter 2

Linear Algebra for Linear Models

Finite dimensional linear algebra is at the foundations of linear model theory. We study this topic only as it provides a basis for this work, not as an end in itself. These notes are very similar to Paul Halmos' superb undergraduate linear algebra textbook, *Finite-Dimensional Vector Spaces*. The book J. Schott (1997), *Matrix Analysis for Statistics*, presents a super set of the material in these notes, and is recommended as a useful reference (but it costs more that $100).

2.1 Basic definitions

Suppose that $V = \{x, y, \ldots\}$ is a set. We write $x, y \in V$.

Definition 2.1 (Vector Space) *The set V is a* vector space *if all elements of V satisfy the following addition and scalar multiplication axioms:*

Axiom 2.1 (Addition) *Suppose there is a binary operator "$+$" that acts on elements of V such that $x + y \in V$, and*

1. $x + y = y + x$ *(commutative)*
2. $x + (y + z) = (x + y) + z$ *(associative)*
3. *There exists a unique vector* zero, $0 \in V$ *such that* $0 + x = x + 0 = x$ *for all* $x \in V$.
4. *For all $x \in V$, there exists a unique vector $(-x) \in V$. such that $x + (-x) = 0$.*

Axiom 2.2 (Scalar multiplication) *Let $\alpha, \beta, \ldots$ be real numbers, and let $x, y, \ldots \in V$ be vectors. Then:*

1. $\alpha x \in V$ *(αx exists and is well defined)*
2. $\alpha(\beta x) = (\alpha\beta)x$ *(associative law)*
3. $\alpha(x + y) = \alpha x + \alpha y$ *(distributive law)*
4. $(\alpha + \beta)x = \alpha x + \beta x$ *(distributive law)*
5. *There exists a scalar* 1 *such that* $1x = x$ *for all* $x \in V$. 1 *is unique.*

Remark 2.1 *A vector space is* closed *under both addition and under scalar multiplication.*

Thus, a set V is a vector space if and and only if for all $x, y \in V$ and scalars α, β, we have $\alpha x + \beta y \in V$.

Several of the usual properties of vectors can be deduced from the axioms, including:

1. $0x = 0$, where 0 is a scalar.
2. $(-\alpha)x = -(\alpha x)$
3. $\alpha(0) = 0, 0 \in V$
4. $\alpha x = 0 \Rightarrow \alpha = 0$ or $x = 0$

The symbol "0" describes an element both in V and a scalar. This should cause only the minimum of confusion since in context exactly which meaning for the symbol is intended should be clear.

Example. Suppose that $x' = (\alpha_1, \ldots \alpha_n)$ is an n-tuple of real numbers, $\alpha_j \in \Re$. By convention, all vectors are column vectors, so the transpose is required to display x in a row. This space is called $\Re^n$, and is the basic space of interest in linear models. Let $y' = (\beta_1, \ldots, \beta_n) \in \Re^n$. Then scalar multiplication for any scalar γ is defined by

$$\gamma x = \gamma \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix} = \begin{pmatrix} \gamma\alpha_1 \\ \vdots \\ \gamma\alpha_n \end{pmatrix}$$

and addition is defined by

$$x + y = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix} + \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_n \end{pmatrix} = \begin{pmatrix} \alpha_1 + \beta_1 \\ \vdots \\ \alpha_n + \beta_n \end{pmatrix}$$

With these definitions, $\Re^n$ is a vector space because it satisfies all the axioms.

Consider the following five sets:

$$\begin{aligned} S_0 &= \{(a, a, a)', a \in \Re\} \\ S_1 &= \{(a, 0, a)', a \in \Re\} \\ S_2 &= \{(a, b, a + b)', (a, b) \in \Re\} \\ S_3 &= \{(a, a, a)', a \in \Re^+\} \\ S_4 &= \{(1, 1, 1)' + (a, b, a + b)', (a, b) \in \Re\} \end{aligned}$$

The sets S_0, S_1 and S_2 are vector spaces, but S_3 and S_4 are not. The set S_0 was encountered in the discussion of the one sample problem in Chapter 1.

Example. A real polynomial of degree n is defined by $x = \sum_{i=0}^{n} \alpha_i t^i$, for real numbers $(\alpha_0, \ldots, \alpha_n)$. The space of all such polynomials is called $\mathcal{P}_{n+1}$. If $y = \sum_{i=0}^{n} \beta_i t^i$, then $x + y = \sum_{i=0}^{n} (\alpha_i + \beta_i) t^i$. One can easily show that the axioms are satisfied, and $\mathcal{P}_{n+1}$ is a vector space.

We next turn to the question of relationships between elements of a vector space V, in particular examining linear relationships.

Definition 2.2 (Linear dependence) *A set of vectors $C = \{x_1, \ldots, x_n\}$ is called* linearly dependent *if there exists scalars $\{\alpha_1, \ldots, \alpha_n\}$ not all equal to 0 such that*

$$\sum_{i=1}^{n} \alpha_i x_i = 0.$$

If $\sum \alpha_i x_i = 0 \Rightarrow \alpha_i = 0$ for all i, then C is linearly independent.

The concept of linear dependence is fundamental to the study of linear models. Here are some consequences of this definition.

1. If $0 \in C$, then C is linearly dependent by setting all the scalars to zero except for the scalar associated with the vector 0, which can be arbitrary.

2. If C is linearly independent, then for any $C_1 \subset C$, C_1 is linearly independent.

3. If C_1 is linearly dependent, then is C linearly dependent?

In the vector space $\Re^n$, suppose that e_i is the vector with a "1" for its i-th element, and all other elements equal to zero. Then the set $\{e_1, \ldots, e_n\}$ must be a linearly independent set. If $x \in \Re^n(\alpha_1, \ldots, \alpha_n)'$, then

$$x = \alpha_1 x_1 + \cdots + \alpha_n x_n$$

and every vector in $\Re^n$ is a unique linear combination of the e_i, with uniqueness following from linear independence. The set $\{e_1, \ldots, e_n\}$ is called the *canonical basis* or *standard basis* from $\Re^n$.

Here is a modestly more complicated example. Suppose that

$$S_2 = \{(a, b, a + b)', a, b \in \Re\}$$

let $x_1 = (3, 0, 3)'$ and $x_2 = (0, 4, 4)'$, which are two vectors in S_2. Consider any other vector in S_2, say $x_3 = (a, b, a+b)' = (a/3)x_1+(b/4)x_2$, so then (x_1, x_2, x_3) is a linearly dependent set. All the vectors in S_2 are of length three, but the linearly independent set has only two vectors.

Now consider the set $C = \{x_1, x_2, x_3\}$ given by

$$x_1 = \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}, x_2 = \begin{pmatrix} 16 \\ 12 \\ 3 \end{pmatrix}, x_3 = \begin{pmatrix} 0 \\ 28 \\ 3 \end{pmatrix} \tag{2.1}$$

Each of the $x_i \in \Re^3 = \{x = (\alpha_1, \alpha_2, \alpha_3)' | \alpha_i \in \Re\}$. The set $C = \{x_1, x_2, x_3\}$ is a linearly dependent set because $16x_1 - x_2 + x_3 = 0$.

Suppose $Y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$, $\beta_i \neq 0, i = 1, 2, 3$, where the xs are given by (2.1). What is the meaning of linear dependence for this linear model? Since, for the example given above, $x_2 - x_3 = 16x_1$, or $x_1 = (x_2 - x_3)/16$, by substituting for x_1 we write:

$$\begin{aligned} Y &= \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon \\ &= \beta_1 \left[\frac{x_2 - x_3}{16}\right] + \beta_2 x_2 + \beta_3 x_3 + \varepsilon \\ &= \left[\frac{\beta_1}{16} + \beta_2\right] x_2 + \left[-\frac{\beta_1}{16} + \beta_3\right] x_3 + \varepsilon \\ &= \gamma_1 x_2 + \gamma_2 x_3 + \varepsilon \end{aligned}$$

This result suggests that the γ_j, the parameters in the "reduced" mean function, are uninterpretable, because the value of the parameter depends on β_1, a quantity that cannot be estimated.

Here is another example:

$$C = \left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \right\} = \{e_1, e_2, e_3\}$$

This is the *canonical basis* for $\Re^3$. It is a linearly independent set.

Suppose we have an arbitrary collection, $C = \{x_1, \ldots, x_n\}$. Then the set

$$\{x_1, x_2, \ldots, x_n, \sum_{i=1}^{n} \alpha_i x_i\}$$

is always linearly dependent.

Theorem 2.1 *A collection $C = \{x_1, x_2, \ldots x_n\}$ is linearly dependent if and only if there exists $\alpha_1, \ldots \alpha_n$, and an index $k \leq n$ such that*

$$x_k = \sum_{i \neq k}^{n} \alpha_i x_i$$

Proof. Assume $x_k = \sum_{i \neq k}^{n} \alpha_i x_i$. Then:

$$0 = x_k + (-x_k) = \sum_{i \neq k}^{n} \alpha_i x_i + (-1)x_k$$

$$= \sum_{i=1}^{n} \alpha_i x_i$$

with $\alpha_k = -1$, and hence the x_i are linearly dependent.

Next, assume that C is linearly dependent. Then $0 = \sum_{i=1}^{n} \beta_i x_i$ for some $\{\beta_1, \ldots, \beta_n\}$. Since $\beta_k \neq 0$ for some index k, then

$$0 = \sum_{i \neq k}^{n} \beta_i x_i + \beta_k x_k$$

$$\frac{1}{\beta_k} 0 = \sum_{i \neq k}^{n} \frac{\beta_i}{\beta_k} x_i + x_k$$

$$x_k = \sum_{i \neq k}^{n} \alpha_i x_i, \text{ where } \alpha_i = -\frac{\beta_i}{\beta_k}$$

and the theorem is proved.

Definition 2.3 (Basis of a vector space) *A collection B of vectors in V is a* basis *for V if:*

1. *B is linearly independent*
2. *For all $y \in V$, $\{B, y\}$ is linearly dependent; that is, there exists $\{\alpha_1, \ldots, \alpha_n\}$ such that $y = \sum_{i=1}^{n} \alpha_i x_i$, where $B = \{x_1, \ldots, x_n\}$. Equivalently, then, any $y \in V$ can be written as a linear combination of the elements of B.*

A basis is a fundamental set that generates V. If $V = \Re^3$, here are three of the infinite number of possible bases:

$$B_1 = \left\{ \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \right\}$$

$$B_2 = \left\{ \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ -2 \end{pmatrix} \right\}$$

$$B_3 = \left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \right\}$$

Suppose we select a fixed basis, one of the infinite number of bases. The representation of any $y \in V$ with respect to that basis is unique. Suppose that $C = \{x_1, \ldots, x_n\}$ is a basis, and $y = \sum \alpha_i x_i = \sum \beta_i x_i$. Then: $0 = y - y = \sum(\alpha_i - \beta_i)x_i = \sum \gamma_i x_i \Rightarrow \gamma_i = 0$ for all i (assuming that $y \neq 0$), which is the definition of linear independence. This suggests the following:

Definition 2.4 (Coordinates) *If B is a basis for V and $y = \sum \alpha_i x_i, x_i \in B$, then $\{\alpha_1, \ldots, \alpha_n\}$ are the coordinates of y with respect to the basis B.*

Definition 2.5 (Finite dimensional) *A vector space is* finite dimensional *if there exists a basis with a finite number of vectors.*

Example. $\mathcal{P}_3$ is finite dimensional, since it has basis $B = \{1, t+1, t^2+t+1\}$. If $z \in \mathcal{P}_3 = 4t^2 + 4t + 1 = -5(1) + 0(t+1) + 4(t^2+t+1)$, z has coordinates (5, 0, 4) relative to this basis. If the basis is $A = \{1, t, t^2\}$, the coordinates of z are $(1, 4, 4)$.

Many useful results in linear algebra do not involve coordinates, and these carry over to linear models, so picking a basis may be unnecessary and in fact may be confusing, as it almost surely is in the last example. We will attempt to use the coordinate free approach whenever possible.

Theorem 2.2 (Span of C) *Let C be a set of vectors. Then the set of all possible linear combinations of elements of C is a vector space.*

We call the vector space defined in the last theorem the *span* or *range* of C, $\mathcal{R}(C)$. Suppose $V_1, V_2, \ldots$ are all vector spaces such that if $x_i \in C$ then $x_i \in V_i$ for all i. Then $\mathcal{R}(C) = \cap V_i$, the intersection of all vector spaces containing C. Thus the span of C is the "smallest" vector space that includes all the elements of C.

Theorem 2.3 *Every basis for a finite dimensional vector space V has the same number of elements.*

Proof. Let $B_1 = \{x_1, \ldots, x_n\}$ and $B_2 = \{y_1, \ldots, y_m\}$ be two bases for V. Since B_1 is a basis for V, every element of B_2 can be written as a linear combination of the elements of B_1. Hence, $y_1 \cup B_1$ is linearly dependent. By Theorem 2.1, there is at least one index k such that x_k is a linear combination of the remaining xs and of y_1. Define the set $D_1 = \{y_1\} + B_1 - \{x_k\}$ with k chosen to be the first index that satisfies Theorem 2.1. We show that (1) every vector in V can be written as a linear combination of the elements in D_1, and (2) D_1 is a linearly independent set. Combining these two results, it follows that D_1 is a basis for V. First, B_1 is a basis for V, so any $z \in V$ can be written as

$$z = \sum_{i=1}^{n} \gamma_i x_i = \sum_{i \neq k}^{n} \gamma_i x_i + \gamma_k \times \text{ (lin. comb. } y_1 \text{ and all } x\text{s except } x_k)$$

so any vector z can be written as a linear combination of the elements of D_1, which shows that the span of D_1 is V. Now suppose that D_1 were not linearly independent. Then we must have that some x_j is a linear combination of the other xs and y_1. But y_1 is a linear combination of the xs, and hence x_j must be a linear combination of the xs alone. But the xs are linearly independent, giving a contradiction. Hence, D_1 is a basis for V.

We continue in this manner until all the ys are added to be basis one at a time, giving a sequence D_1, $D_2, \ldots, D_m$. We must have $n \geq m$ or else the last few ys would be linear combinations of the first few. Adding the xs to the ys shows $m \geq n$, proving the result.

Definition 2.6 (Dimension) *The number of elements in a basis B of a vector space V is called the* dimension *of V, written* $\dim(V)$.

Example. The set S_2 is a vector space of dimension two. One possible basis for this space is $\{(1,0,1)', (0,1,1)'\}$. Is $\{e_1, e_2\}$ another basis for this space? Why or why not?

The following are immediate consequences of this definition and the preceding theorem:

1. $\Re^n$ is a vector space of dimension n.
2. Any $n+1$ vectors in $\Re^n$ are linearly dependent.
3. Any set of n linearly independent vectors in $\Re^n$ forms a basis for $\Re^n$.

Theorem 2.4 (Completion of a basis) *If $\{x_1, \ldots, x_k\}$ is a linearly independent set of vectors in V ($\dim(V) = n,\ k < n$), there exists elements $x_{k+1}, \ldots, x_n$ such that $\{x_1, \ldots, x_n\}$ is a basis for V. The set $x_{k+1}, \ldots, x_n$ is not unique.*

Proof. Homework.

Definition 2.7 (Coordinates) *Given a basis $B = \{x_1, \ldots, x_n\}$ for V, any $y \in V$ can be written as*

$$y = \sum \alpha_i x_i$$

uniquely. The vector $(\alpha_1, \ldots, \alpha_n)' \in \Re^n$ is called the coordinates of y relative to the basis B.

Definition 2.8 (Isomorphism) *An* isomorphism *between two vector spaces U and V is a $1 \sim 1$ map that preserves linear relations: for $x, x_1, x_2 \in V$ and $y, y_1, y_2 \in U$, we have $T(x) = y$ and*

$$\begin{aligned} T(\alpha_1 x_1 + \alpha_2 x_2) &= \alpha_1 T(x_1) + \alpha_2 T(x_2) \\ &= \alpha_1 y_1 + \alpha_2 y_2 \end{aligned}$$

The following results follow from the definition of isomorphic spaces.

1. $T(0) = 0$.
2. There is a function $T^{-1} : U \to V$ such that $T^{-1}(T(x)) = x$, for all $x \in V$.

3. $T^{-1}(\beta_1 y_1 + \beta_2 y_2) = \beta_1 T^{-1}(x_1) + \beta_2 T^{-1}(x_2)$

4. $\{x_1, \ldots, x_m\}$ are linearly independent if and only if $\{T(x_1), \ldots, T(x_m)\}$ are linearly independent.

5. Two isomorphic vector spaces have the same dimension.

For example, consider the space $S_2 = \{(a, b, a + b)', (a, b) \in \Re\}$, which can be easily shown to be a vector space of dimension two. Consider the map from $S_2 \rightarrow \Re^2$ defined by $T(x) = T((a, b, a + b)') = (a, b)'$. One can show that the condition of Definition 2.8 is satisfied, and so S_2 is isomorphic to $\Re^2$. This result generalizes as follows:

Theorem 2.5 *Any real n-dimensional vector space is isomorphic to $\Re^n$.*

Proof. Let V be a real n-dimensional vector space. To establish the theorem, we need to construct an isomorphism between V and $\Re^n$. Let $B = \{x_1, \ldots, x_n\}$ be a basis for V. Then for all $y \in V$ there exists unique real coordinates $\{\alpha_1^y, \ldots, \alpha_n^y\}$ such that

$$y = \sum_{i=1}^{n} \alpha_i^y x_i$$

Now, define

$$T(y) = \begin{pmatrix} \alpha_1^y \\ \vdots \\ \alpha_n^y \end{pmatrix} \in \Re^n$$

and

$$\begin{aligned} T(\beta_1 y_1 + \beta_2 y_2) &= T[\beta_1 \textstyle\sum \alpha_i^{y_1} x_i + \beta_2 \sum \alpha_i^{y_2} x_i] \\ &= T[\textstyle\sum (\beta_1 \alpha_i^{y_1} + \beta_2 \alpha_i^{y_2}) x_i] \\ &= \begin{pmatrix} \beta_1 \alpha_1^{y_1} + \beta_2 \alpha_1^{y_2} \\ \vdots \\ \beta_1 \alpha_n^{y_1} + \beta_2 \alpha_n^{y_2} \end{pmatrix} \\ &= \beta_1 T(y_1) + \beta_2 T(y_2) \end{aligned}$$

showing that Definition 2.8 holds. The $1 \sim 1$ property follows from the uniqueness of the αs given the basis.

Example. $\beta_0 + \beta_1 t + \beta_2 t^2 \in \mathcal{P}_2$ is clearer than the equivalent expression of a quadratic polynomial that would be the expression of its coordinates with respect

to a fixed basis. As examples, this polynomial is given by $\begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}$ with respect to the basis $(1, t, t^2)$, or $\begin{pmatrix} \beta_0 - \beta_1 + 2\beta_2 \\ \beta_1 - \beta_2 \\ \beta_2 \end{pmatrix}$ with respect to the basis $1, t+1, t^2+t-1$, or ... We will use bases when convenient, but generally not depend on them.

2.2 Linear Subspaces

Definition 2.9 (Subspace) *A subset $M \subset V$ is a* linear subspace *if for all scalars α, β, and all $x, y \in M$, then $\alpha x + \beta y \in M$.*

Equivalently, M is a linear subspace if for all $x, y \in M$, and all scalars α, $x + y \in M$ and $\alpha x \in M$. Thus M is a subspace if and only if it is closed under vector addition and multiplication.

Example. $\Re^n$. Choose any vector $x_0 \neq 0$, and consider $M = \{\alpha x_0, \alpha \in \Re\}$. Then $M = \mathcal{R}(x_0)$ is a vector space of dimension one. This is a line.

Example. Choose any $x_0, x_1 \in \Re^n$ that are linearly independent vectors, and consider $M = \{\alpha x_0 + \beta x_1 \mid \alpha, \beta \in \Re\}$. Then M is a vector space. What do you suppose its dimension is? This is a plane.

Example. The set $S_2 = \{(a, b, a + b)', (a, b) \in \Re^2\}$ is a subspace of $\Re^3$, with basis $(a, a, 0)', (0, b, b)'$ for any non-zero a and b.

Example. Let C be any set of vectors in V. Then $\mathcal{R}(C) = \{x | x = \sum \alpha_i c_i, c_i \in C, \alpha_i \text{ scalars}\}$ is a vector space contained in V or V itself. $\mathcal{R}(C)$ is a linear subspace of dimension $\leq n$. $\mathcal{R}(C)$ is called a *hyperplane* if it has dimension greater than two.

Theorem 2.6 *If M is a linear subspace, then $0 \in M$.*

Proof. $x \in M \Rightarrow -x \in M \Rightarrow x - x \in M \Rightarrow 0 \in M$. In $\Re^n$, for example, the set of all vector subspaces is the set of all lines, planes and hyperplanes to pass through the origin.

Sometimes in statistical applications it is useful to consider a linear subspace that is shifted or translated from the origin. This can happen, for example, in models that include an intercept. It is therefore helpful to have the following definition of a space that is displaced from the origin.

Definition 2.10 (Flat) *Suppose $M \subset V$ is a linear subspace and $y_0 \in V$. Then a* flat *consists of $\{x + y_0 | x \in M\}$. We will write $y_0 + M$ where M is a subspace to indicate a flat.*

By considering *translations*, flats are equivalent to vector spaces. If Y is a random variable whose domain is the flat $y_0 + M$, then, if y_0 is fixed, $Y - y_0$ has domain M.

Example Set set $S_4 = \{(1, 1, 1)' + z, z \in S_2\}$ is a flat because $0 \notin S_4$.

Example. In $\Re^2$, consider

$$M = \left\{ \alpha \begin{pmatrix} 1 \\ 2 \end{pmatrix} | \alpha \in \Re \right\} \text{ and } y_0 = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$$

Then the flat $y_0 + M$ is given by the set

$$y_0 + M = \left\{ \begin{pmatrix} 2 \\ 2 \end{pmatrix} + \alpha \begin{pmatrix} 1 \\ 2 \end{pmatrix} | \alpha \in \Re \right\}$$

which is just a straight line that does not pass through the origin, but rather through the point $(2, 2)$. The choice of y_0 is not unique and it can be any point $y = y_0 + y_\alpha$, where $y_\alpha = \alpha(1, 2)'$. For example, if $\alpha = -2$, then $y = (0, -2)'$ and if $\alpha = +1$, then $y = (3, 4)'$, and so on. For any y_0 not of this form, we simply get a different flat. This is summarized in the next remark.

Theorem 2.7 *The two spaces*

$$\begin{aligned} F_1 &= \{z | z = y_0 + x, y_0 \in V, x \in M \subset V\} \\ F_2 &= \{z | z = y_1 + x, y_1 \in F_1, x \in M \subset V\} \end{aligned}$$

are the same subspace, so the representation of the flat is not unique.

Definition 2.11 (Sum and intersection of subspaces) *Let H, K be two linear subspaces. Then:*

$$H + K = \{x + y | x \in H, y \in K\}$$

is the sum *of H and K. The intersection of H and K is*

$$H \cap K = \{x | x \in H \text{and } x \in K\}$$

Theorem 2.8 *Both $H + K$ and $H \cap K$ are linear subspaces.*

Proof. Homework

Definition 2.12 (Disjoint subspaces) *Two subspaces are disjoint if* $H \cap K = \{0\}$*, the null vector.*

Theorem 2.9 *If* $H \cap K = \{0\}$*, and* $z \in H + K$*, then the decomposition* $z = x + y$ *with* $x \in H$ *and* $y \in K$ *is unique.*

Proof. Suppose $z = x + y$ and $z = x' + y'$. Then, $x - x' \in H$ and $y - y' \in K$. We must have $x + y = x' + y'$ or $x - x' = y - y'$, which in turn requires that $x - x' = y - y' = 0$, since 0 is the only vector common to H and K. Thus, $x = x'$ and $y = y'$.

Theorem 2.10 *If* $H \cap K = \{0\}$*, then* $\dim(H + K) = \dim(H) + \dim(K)$*. In general,* $\dim(H + K) = \dim(H) + \dim(K) - \dim(H \cap K)$.

Proof. Homework.

Definition 2.13 (Complement of a space) *If* M *and* M^c *are disjoint subspaces of* V *and* $V = M + M^c$*, then* M^c *is called a* complement *of* M.

Remark 2.2 *The complement is not unique. In* $\Re^2$*, a subspace* M *of dimension 1 consists of a line through the origin. A complement of* M *is given by any other line* $M^c \neq \alpha M$ *through the origin, because linear combinations of any two such lines span* $\Re^2$.

2.3 Linear Transformations

Definition 2.14 *A linear transformation* A *on a vector space* V *is a function mapping* $V \to V_1 \subseteq V$ *such that*

$$A(\alpha x + \beta y) = \alpha A(x) + \beta A(y)$$

for all $\alpha, \beta \in \Re$ *and all* $x, y \in V$.

Remark 2.3 *For* $0 \in V$*,* $A(0) = 0$ *for any linear transformation* A.

Examples. If for all $x \in V$,

1. $A(x) = 0$, then A is called the *null transformation*, just written 0.

2. $A(x) = x$, then A is the *identity transformation*. This transformation is generally called I.

3. $A(x) = -x$, then A is a *reflection through origin*.

Definition 2.15 (Sum of linear transformations) *The sum of two linear transformations is defined by:*

$$(A + B)(x) = A(x) + B(x).$$

Remark 2.4 *One can easily show that the set of all linear transformations $A : V \to V$ is itself a vector space $\mathcal{L}(V)$.*

Definition 2.16 (Product of linear transformations) *The product of two linear transformations A and B is defined by: $(AB)(x) = A(B(x))$. Order matters: generally, $AB \neq BA$.*

Here are some easily derived properties of the product:

1. $A0 = 0A = 0$ (here 0 is the null transformation)

2. $AI = IA = A$

3. $A(B + C) = AB + AC$

4. $A(BC) = (AB)C$

We will write $AA = A^2$, and

$$\underbrace{AA \cdots A}_{m} = A^m.$$

Definition 2.17 (Range of a linear transformation) *The* range *of a linear transformation is defined as:*

$$\mathcal{R}(A) = \{y | y = Ax, x \in V\}$$

$\mathcal{R}(A)$ is a linear subspace of V since if $y_1 \in \mathcal{R}(A)$ and $y_2 \in \mathcal{R}(A)$, then there exists x_1 such that $y_1 = A(x_1)$ and x_2 such that $y_2 = A(x_2)$ and $A(\alpha x_1 + \beta x_2) = \alpha A(x_1) + \beta A(x_2) = \alpha y_1 + \beta y_2 \in \mathcal{R}(A)$.

Definition 2.18 (Rank of a transformation) *The rank of a linear transformation $A = \rho(A)$ is the dimension of $\mathcal{R}(A)$.*

Definition 2.19 (Null space) *The null space of a linear transformation A is*

$$N(A) = \{x | x \in V, A(x) = 0\}$$

that is, the set of points x that maps A to zero.

Theorem 2.11 *$N(A)$ is a linear subspace of V.*

Proof. Consider any $x, y \in \mathrm{N}(A)$. Then $A(\alpha x + \beta y) = \alpha A(x) + \beta A(y) = 0$, so $\alpha x + \beta y \in \mathrm{N}(A)$. Since $\mathrm{N}(A)$ is closed under addition and scalar multiplication, it is a linear subspace.

Definition 2.20 *The dimension of $N(A)$ is called $\nu(A)$.*

Example. In $\Re^3$, suppose that

$$x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix},$$

and define

$$A(x) = \begin{pmatrix} 0 \\ x_2 \\ x_3 \end{pmatrix}$$

so $A(x)$ preserves the last two elements of any vector in $\Re^3$, and sets the first element to zero. Here, $\mathcal{R}(A) = \{x|$ first coordinate of $x = 0$, other coordinates arbitrary, $x \in \Re^3\}$, and thus $\rho(A) = 2$. Similarly, $\mathrm{N}(A) = \{x|$ first coordinate of x is arbitrary, other coordinates are $0\}$, which is a subspace of dimension $\nu(A) = 1$.

Theorem 2.12

$$\rho(A) + \nu(A) = n = \dim(V).$$

Proof. Homework.

Example. Suppose $V = \Re^3$, and

$$A \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} x_1 + x_2 \\ x_2 + x_3 \\ x_1 + 2x_2 + x_3 \end{pmatrix} = (x_1 + x_2) \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} + (x_2 + x_3) \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} \quad (2.2)$$

The transformation A maps from $\Re^3$ to the subspace $\mathcal{R}(A) = S_2$. Then N(A) is the set

$$(x_1 + x_2)\begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} + (x_2 + x_3)\begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

or points of the form

$$\alpha \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix} \in N(A)$$

which is seen to be a subspace of dimension one.

Solving linear equations of the form $A(x) = y$ for x is a common problem in linear models. The question of whether or not these equations have a unique solution depends on the null space of A.

Theorem 2.13 *The equation $Ax = y$ has a solution x_y for each $y \in V$ if and only if $\nu(A) = 0$, or equivalently $A(x) = 0 \Rightarrow x = 0$.*

Proof. Suppose $A(x) = 0 \Rightarrow x = 0$, and then $\nu(A) = 0$. Let $\{x_1, \ldots, x_n\}$ be a basis for V. Then $\{Ax_1, \ldots, Ax_n\}$ are n vectors in V. These vectors are linearly independent (if $\sum \alpha_i Ax_i = 0$, then $A(\sum \alpha_i x_i) = 0$ and $\sum \alpha_i x_i = 0$, which is a contradiction), and, since there are n of them, they form a basis for V. Hence, any $y \in V$ can be written as a linear combination of the columns of Ax, and the coordinates of y with respect to this basis gives the required x.

Suppose that $Ax = y$ has a solution for all $y \in V$. Find n vectors in V, $\{y_1, \ldots, y_n\}$ that are a basis for V and the corresponding vectors $\{x_1, \ldots, x_n\}$. The xs must also be linearly independent because if $\sum \lambda_i x_i = 0$ then

$$\begin{aligned} A(\textstyle\sum \lambda_i x_i) &= \textstyle\sum \lambda_i A(x_i) \\ &= \textstyle\sum \lambda_i y_i \\ &= 0 \end{aligned}$$

which contradicts the fact that the y_i are linearly independent. Then $A(x) = 0$ only if $x = 0$ by linear independence.

Remark 2.5 *If $Ax = y$ has a solution for each $y \in V$ then the solution is unique, since if $y = Ax_1 = Ax_2$ then $A(x_1 - x_2) = 0$, and this is so only if $x_1 - x_2 = 0$ or $x_1 = x_2$.*

Example. The linear transformation defined by (2.2) has null space

$$\mathrm{N}(A) = \mathcal{R}\left(\begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix}\right)$$

and hence the equations $A(x) = y$ will not have a unique solution. If x_0 is a solution to these equations, then so is any vector of the form $x_0 + \alpha(1, -1, 1)'$. This set of solutions forms a flat.

Definition 2.21 (Inverse and Nonsingular) *When $Ax = y$ has a unique solution for all $y \in V$ then the* inverse, *A^{-1} is defined by $A^{-1}y = x$. In this case, A is said to be* nonsingular, *otherwise A is* singular.

Theorem 2.14 *If A is nonsingular then*

1. *$\rho(A) = n$ and $\nu(A) = 0$.*
2. *$A^{-1}A = AA^{-1} = I$.*
3. *If A and B are nonsingular, then AB is nonsingular and $(AB)^{-1} = B^{-1}A^{-1}$.*

Theorem 2.15 *If $\{x_1, \ldots, x_n\}$ is a basis for V and $\{y_1, \ldots, y_n\}$ is also a basis for V then there exists a unique nonsingular linear transformation A such that $Ax_i = y_i$.*

Proof. Homework.

The importance of this proposition in linear models is that one can work with any basis, and transform at the end to any other basis. Again, coordinate systems become irrelevant.

Theorem 2.16 *If B is nonsingular, then $\rho(AB) = \rho(BA) = \rho(A)$.*

Proof. Homework.

Theorem 2.17 *For any linear transformations A and B:*

1. *$\rho(AB) \leq min(\rho(A), \rho(B))$*
2. *$\rho(A + B) \leq \rho(A) + \rho(B)$*

Proof. Homework.

2.4 Projections

Projections are special linear transformations that are extremely useful in linear models. Suppose that V is an n-dimensional vector space, and M and N are subspaces of V such that $M + N = \{z = x + y | x \in M, y =\in N\} = V$ and $M \cap N = \{z | z \in M, z \in N\} = \{0\}$. (Thus, $N = M^c$.)

Definition 2.22 *For all $z \in V$, consider the unique decomposition $z = x+y, x \in M, y \in N$. The transformation $P_{M|N}z = x$ is called the* projection of z *on M along N. Similarly, the linear transformation $P_{N|M}z = y$ is the projection of z on N along M.*

Theorem 2.18 . $P_{N|M} = I - P_{M|N}$.

Because of the above relationship between these two projections, we will define further notation: $Q_{M|N} = I - P_{N|M}$.

Theorem 2.19 *A linear transformation T on V is a projection for some M and N if and only if $T^2z = Tz$ for all $z \in V$.*

Proof. If T is a projection, then for $z = x + y, x \in M, y \in N$,

$$T^2z = TTz = T(Tz) = Tx = x = Tz,$$

since x has no component in N.

Suppose $T^2z = Tz$, for all $z \in V$. Let $N = \mathrm{N}(T) = \{z | Tz = 0\}$, the null space of T, and $M = \{z | Tz = z\}$. We will show $M + N = V$ and $M \cap N = \{0\}$, and hence $T = P_{M|N}$.

1. If $z \in M$, then $Tz = z$. Also, if $z \in N$, then $Tz = 0$, and if $z \in M \cap N, Tz = 0$, so that $M \cap N = \{0\}$.

2. To show $M + N = V$ consider any $z \in V$. Then $z = Tz + (I - T)z$. Let $x = Tz, y = (I - T)z$ and $z = x + y$. Then

   $$Tx = T(Tz) = T^2z = Tz = x$$

   which implies that $x \in M$. Also,

   $$Ty = T(I - T)z = Tz - T^2z = 0$$

   so that $y \in N$. Thus $V = M + N$ and T is precisely the projection on M along N. This completes the proof.

Definition 2.23 (Idempotent) *A linear transformation T is called* idempotent *if* $T^2 = T$.

Consequences of the previous theorems. Let P_1 be the projection of M_1 along N_1, and let P_2 be the projection of M_2 along N_2. Then:

1. $P_1 + P_2$ is a projection if and only if $P_1P_2 = P_2P_1 = 0$. This condition is equivalent to requiring that $M_1 \cap M_2 = \{0\}$.

   To prove this result, multiply to obtain

   $$\begin{aligned}(P_1 + P_2)^2 &= P_1^2 + P_2^2 + P_1P_2 + P_2P_1 \\ &= P_1^2 + P_2^2 + P_1P_2 + P_2P_1\end{aligned}$$

   so we must have that

   $$P_1P_2 + P_2P_1 = 0$$

   Multiply on the left and right by P_1, gives the two equations

   $$\begin{aligned}P_1P_2 + P_1P_2P_1 &= 0 \\ P_1P_2P_1 + P_2P_1 &= 0\end{aligned}$$

   Subtracting these two equations gives $P_1P_2 - P_2P_1 = 0$, and thus $P_1P_2 = P_2P_1 = 0$.

2. $P_1 - P_2$ is a projection if and only if $P_1P_2 = P_2P_1 = P_2$. In this case, we must have $M_2 \subset M_1$.

3. $If P_1P_2 = P_2P_1$, then P_1P_2 is a projection.

2.5 Inner Products

Let V be a finite dimensional vector space.

Definition 2.24 (Inner Product) *A real inner product on V is a function defined on $V \times V \to \Re$, written (x, y), such that, for all $x, y \in V$,*

1. $(x, y) = (y, x)$ *(symmetry)*
2. $(\alpha_1 x_1 + \alpha_2 x_2, y) = \alpha_1(x_1, y) + \alpha_2(x_2, y)$ *(linearity)*

3. *For all* $x \neq 0, (x, x) > 0$ *(nonnegative)*

Example. Suppose x has coordinates $\gamma = (\gamma_1, \ldots, \gamma_n)' \in \Re^n$ and y has coordinates $\lambda = (\lambda_1, \ldots, \lambda_n)' \in \Re^n$, both relative to some fixed basis $\{x_1, \ldots, x_n\}$. The usual inner product is

$$(x, y) = \sum_{i=1}^{n} \gamma_i \lambda_i = \gamma' \lambda \tag{2.3}$$

If $V = \Re^n$, and the basis chosen is the canonical basis $\{e_1, \ldots, e_n\}$, then $\gamma = x$ and $\lambda = y$, and (2.3) corresponds to the usual Euclidean inner product. Another inner product is

$$(x, y)_A = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} \gamma_i \lambda_j \tag{2.4}$$

with the fixed constants a_{ij} selected so that $a_{ij} = a_{ji}$ and $\sum \sum a_{ij} \gamma_i \gamma_j > 0$.

Definition 2.25 (Real inner product space) *A real inner product space* $(V, (\cdot, \cdot))$ *is a pair such that* V *is a real vector space and* $(\cdot, \cdot)$ *is a real inner product defined on* V.

Theorem 2.20 (Cauchy-Schwartz inequality) *For any inner product*

$$| (x, y) | \leq [(x, x)(y, y)]^{1/2}$$

Equality holds if and only if $x = cy$ *for some* $c \in \Re$.

Proof. If $(x, x) = 0$ or $(y, y) = 0$, the result is immediate, since $(x, 0) = 0$, for all $x \in V$, so we can assume $(x, x) > 0$ and $(y, y) > 0$. Let $w = x/(x, x)^{1/2}$ and $z = y/(y, y)^{1/2}$. We show $| (w, z) | \leq 1$:

$$0 \leq (w - z, w - z) = (w, w) - 2(w, z) + (z, z) = 2 - 2(w, z)$$

so that

$$(w, z) \leq 1$$

Similarly,

$$0 \leq (w + z, w + z) = 2 + 2(w, z),$$

so that $(w, z) \geq -1$. Combining these gives $|(w, z)| \leq 1$ as required, To prove the second part, we will have equality if $(w - z, w - z) = 0$ or $(w + z, w + z) = 0$. This will hold only if $w = \pm z$, or $x = \pm((x, x)^{1/2}/(y, y)^{1/2}) \times y$, so equality holds only if $x = cy$.

Definition 2.26 (Cosines) *The* cosine function *is a function from* $V \times V \rightarrow \Re$ *defined by:*

$$\cos(x, y) = \frac{(x, y)}{[(x, x)(y, y)]^{1/2}}$$

provided $\| x \| \neq 0$ *and* $\| y \| \neq 0$.

The Cauchy-Schwartz inequality says $| \cos(x, y) | \leq 1$. The cosine function is invariant under multiplication of x and y by positive scalars, as one might hope.

Example. In $\Re^2$ with the usual inner product,

$$\cos(\theta) = \frac{x_1 y_1 + x_2 y_2}{\sqrt{(x_1^2 + x_2^2)(y_1^2 + y_2^2)}}$$

Definition 2.27 (Norm) *A function* $\| x \|$ *is a* norm *on a vector space* V *if, for any* $x, y \in V$*,*

1. $\| x \| \geq 0$.
2. $\| x \| = 0$ *if and only if* $x = 0$.
3. $\| cx \| = |c| \times \| x \|$ *for any* $c \in \Re$.
4. $\| x + y \| \leq \| x \| + \| y \|$

Definition 2.28 (Distance) *A function* $\delta(x, y)$ *is a* distance *on a vector space* V *if, for any* $x, y \in V$*,*

1. $\delta(x, y) \geq 0$.
2. $\delta(x, y) = 0$ *if and only if* $x = y$.
3. $\delta(x, y) = \delta(y, x)$*, so distance is symmetric.*
4. $\delta(x, z) \leq \delta(x, y) + \delta(y, z)$*, the triangle inequality.*

There are many choices for norms and distance functions. If the coordinates of x relative to a given basis are $\gamma = (\gamma_1, \ldots, \gamma_n)'$, a general class of norms is given by

$$\| x \|_p = \left\{ \sum |\gamma_i|^p \right\}^{1/p}$$

The most familiar member of this family is the *two-norm*, also called the Euclidean norm,

$$\| x \|_2 = (x, x)^{1/2} = \gamma'\gamma$$

Other important members of this class include the *sum norm*,

$$\| x \|_1 = \sum |\gamma_i|$$

and the infinity norm

$$\| x \|_\infty = \max_i |\gamma_i|$$

All these norms satisfy the definition.

2.6 Orthogonality

Let $(V, (\cdot, \cdot))$ be an inner product space.

Definition 2.29 *Vectors $x, y \in (V, (\cdot, \cdot))$ are* orthogonal *if $(x, y) = 0$. We will write this as $x \perp y$.*

Example. $\Re^n$, usual inner product. If $x = \begin{pmatrix} \gamma_1 \\ \vdots \\ \gamma_n \end{pmatrix}, y = \begin{pmatrix} \eta_1 \\ \vdots \\ \eta_n \end{pmatrix}$, then $x \perp y$ if $\sum \gamma_i \eta_i = 0$, or equivalently if $\cos(x, y) = 0$ (the angle between x and y is $\pi/2$).

Example. $\Re^n$. If $x = \begin{pmatrix} \gamma_1 \\ \vdots \\ \gamma_n \end{pmatrix}, y = \begin{pmatrix} \eta_1 \\ \vdots \\ \eta_n \end{pmatrix}$, but the inner product is given by (2.4), then $x \perp y$ if $(x, Ay) = \sum \sum a_{ij} \gamma_i \eta_j = 0$. In the usual Euclidean sense, orthogonality does not imply perpendicular (angle $= \pi/2$) relative to $(\cdot, \cdot)$.

Definition 2.30 (Orthogonal vectors) *The set of vectors $\{x_1, \ldots, x_m\}$ is called orthogonal if $(x_i, x_j) = 0, i \neq j$. The set is called* orthonormal *if in addition $(x_i, x_i) = \| x \|^2 = 1, i = 1, ..., m$.*

Definition 2.31 (Orthogonal Spaces) *Subsets $S_1 \subset V$ and $S_2 \subset V$ are* orthogonal, *$S_1 \perp S_2$ if for all $x \in S_1, y \in S_2,\ x \perp y$. If $S_1 \perp S_2$, then the linear subspaces spanned by S_1 and S_2 are orthogonal.*

Definition 2.32 (Orthogonal flats) *Two flats are orthogonal if their corresponding linear subspaces are orthogonal:* $F_1 = x_1 + M_1; F_2 = x_2 + M_2,$ *then* $M_1 \perp M_2 \Rightarrow F_1 \perp F_2$.

Definition 2.33 (Orthogonal basis) *A basis for* V *is called an* orthogonal basis *if the basis set is orthogonal. The basis is* orthonormal *or* onb *if all elements of an orthogonal basis have unit length.*

Definition 2.34 (Orthogonal complement) *If* C *is a linear subspace in* V*, the* orthogonal complement *of* C*, written* $C^\perp$ *(and read* C perp*), is the set of all vectors in* V *that are orthogonal to all vectors in* C.

We have previously defined the complement C^c of C to be any subspace such that the direct sum $C + C^c = V$. The space $C^\perp$ is a particular C^c with the additional property.

Example. $\Re^2$ with the usual inner product. If C is a line through the origin, $C^\perp$ is a perpendicular line through the origin. One can verify that $C^\perp$ is a linear subspace $(C^\perp)^\perp = C$.

Definition 2.35 *Suppose* $C \subset D \subset V$*, with an inner product* $(\cdot, \cdot)$*. Then we define the orthogonal complement of* C *relative to* D *as:*

$$C^\perp(D) = \{y | y \in D \text{ and } (x, y) = 0, \text{ for all } x \in C\}$$

Theorem 2.21 *Every linear subspace has an orthonormal basis.*

Proof. Many of the important results in linear models are vastly simplified by using an orthonormal basis. The proof of this theorem is constructive: we actually find an orthonormal basis from any arbitrary basis $\{x_1, \ldots, x_n\}$.

Gram Schmidt Orthogonalization. We begin with any $\{x_1, \ldots, x_n\}$ for the subspace M. We will construct an orthonormal basis $\{y_1, \ldots, y_n\}$ with the additional useful property that $\mathcal{R}(\{x_1, \ldots, x_k\}) = \mathcal{R}(\{y_1, \ldots, y_k\})$, $k = 1, ..n$.

1. Let $y_1 = x_1$.

2. We want to find y_2, and linear combination of y_1 and x_2 such that: $(y_1, y_2) = 0$ and $\mathcal{R}(x_1, x_2) = \mathcal{R}(y_1, y_2)$. Set

$$y_2 = x_2 + \alpha_1 y_1$$

so that $(y_1, y_2) = (y_1, x_2 + \alpha_1 y_1) = (y_1, x_2) + \alpha(y_1, y_1) = 0$. Solving for α gives:

$$\alpha_1 = -\frac{(y_1, x_2)}{\| y_1 \|^2}$$

and thus

$$y_2 = x_2 - \frac{(y_1, x_2)}{\| y_1 \|^2} y_1$$

The vectors y_1 and y_2 span the same space as $\mathcal{R}(x_1, x_2)$ because they are not collinear and they are just linear combinations of x_1 and x_2, so they are an orthogonal basis for this space.

3. Continuing with this process, we next find y_3 such that:

$$y_3 \neq 0$$

$$(y_1, y_3) = (y_2, y_3) = 0$$

$$\mathcal{R}(y_1, y_2, y_3) = \mathcal{R}(x_1, x_2, x_3)$$

The reasonable choice is:

$$y_3 = x_3 + \alpha_1 y_1 + \alpha_2 y_2$$

so that

$$\begin{aligned} 0 &= (y_1, y_3) \\ &= (y_1, x_3) + \alpha_1 (y_1, y_1) + \alpha_2 (y_1, y_2) \\ &= (y_1, x_3) + \alpha_1 (y_1, y_1) + 0 \end{aligned}$$

so $\alpha_1 = -(x_3, y_1)/\| y_1 \|^2$. By a similar argument, $\alpha_2 = -(x_3, y_2)/\| y_2 \|^2$. For the general case, take

$$y_k = x_k - \sum_{i=1}^{k-1} \frac{(x_k, y_i)}{\| y_i \|^2} y_i.$$

This yields an orthogonal basis that spans the same space as $\{x_1, \ldots, x_n\}$. If an orthonormal basis is wanted, simply normalize the y_i: $z_i = y_i/\| y_i \|$.

This is the simplest algorithm, but it is numerically deficient, especially if some of the xs are of vastly different lengths, or if any of the cosines between the xs are particularly small. This can lead to cancellation of all significant digits in a computed result. There are many other orthogonalization algorithms.

Modified Gram-Schimdt. A more stable approach to getting an orthonormal basis does the computations in a different order.

1. Start with any basis, say $\{x_1^{(0)}, \ldots, x_n^{(0)}\}$.

2. Let $y_1 = x_1^{(0)} / \parallel x_1^{(0)} \parallel$, so y_1 has unit length. Renormalize and orthogonalize to y_1 the remaining $\{x_2^{(1)}, ..., x_n^{(1)}\}$ via

$$x_i^{(1)} = x_i^{(0)} - (x_i^{(0)}, y_1)y_1, \ i = 2, \ldots, n$$

3. Let $y_2 = x_2^{(1)} / \parallel x_2^{(1)} \parallel$, so y_2 has unit length.

4. Renormalize and orthogonalize to y_1 the remaining $\{x_3^{(2)}, \ldots, x_n^{(2)}\}$ via

$$x_i^{(2)} = x_i^{(1)} - (x_i^{(1)}, y_2)y_2, i = 3, \ldots, n$$

   and continue in the same manner.

The modified Gram Schmidt has the advantage of being computationally more stable, since it renormalizes at each step. Of course there are many other methods of getting an orthonormal basis, particularly the QR method we will learn shortly.

2.6.1 Coordinates with respect to an Orthonormal Basis

. Suppose we have an orthonormal basis $\{x_1, \ldots, x_n\}$. What are the coordinates of any $y \in V$ with respect to this basis? We have $y = \sum \lambda_i x_i$. How do we find the λ_i? We can compute the inner product $(y, x_j) = (\sum \lambda_i x_i, x_j) = \lambda_j$, so we can recover the λ_i just be computing inner products. Hence,

$$y = \sum (y, x_i) x_i. \tag{2.5}$$

Also, $\parallel y \parallel^2 = \parallel \sum \lambda_i x_i \parallel^2 = \sum \sum \lambda_i \lambda_j (x_i, x_j) = \sum \lambda_i^2$ so the norm-squared of a vector is the sum of the squared coefficients with respect to an orthonormal basis.

2.6.2 Orthogonal Projections

For any subspace $M \in V$, there are many projections, one for each choice of N such that $M + N = V$. By requiring that M and N be orthogonal we will end up with a unique orthogonal projection with many useful and elegant geometric and statistical properties. Since orthogonality depends on the inner product, orthogonal projections will be unique only up to choice of the inner product.

Recall that for $M \in V$, where $\{V, (\cdot, \cdot)\}$, is a real inner product space, we defined the orthogonal complement as

$$M^{\perp} = \{x | (x, y) = 0, y \in M, x \in V\}$$

so $M^{\perp}$ is the set of all vectors in V that are orthogonal to all vectors in M. Suppose $\dim(M) = m, \dim(V) = n$.

Theorem 2.22

$$M \cap M^{\perp} = \{0\}.$$

Proof. The only vector that satisfies $(x, x) = 0$ is $x = 0$, and this is the only vector in common to these sets.

Theorem 2.23

$$M + M^{\perp} = V \text{ and } (M^{\perp})^{\perp} = M.$$

Proof. Let $\{x_1, \ldots, x_m\}$ be a basis for M, and extend it to $\{x_1, \ldots, x_n\}$ for V. By the Gram-Schmidt method we can without loss of generality assume that $\{x_1, \ldots, x_m\}$ is an orthonormal basis for M and $\{x_1, \ldots, x_n\}$ is an orthonormal basis for V. Thus, $x_{m+1}, \ldots, x_n \in M^{\perp}$ by construction. They are linearly independent, so $\dim(M^{\perp}) \geq n - m$. But $\dim(M + M^{\perp}) \leq n$ since $M + M^{\perp} \subset V$ and $M \cap M^{\perp} = \{0\}$. Thus $\dim(M) + \dim(M^{\perp}) \leq n$ and $\dim(M^{\perp}) \leq n - m$ and therefore $\dim(M^{\perp}) = n - m$. The completion of the x-basis is an orthonormal basis for $M^{\perp}$ and $V = M + M^{\perp}$.

The proof of $(M^{\perp})^{\perp} = M$ starts with the orthonormal basis for $M^{\perp}$ and extends it to V in the same way.

Definition 2.36 *The projection on M along $M^{\perp}$ is called the* orthogonal projection relative to $(\cdot, \cdot)$. *We will usually write $P_{M|M^{\perp}}$ as P_M and $P_{M^{\perp}|M}$ as Q_M.*

Given an orthonormal basis, evaluation of an orthogonal projection is easy. Suppose M is a linear subspace, $\dim(M) = m$. We want to find a specific representation for P_M with the inner product $(\cdot, \cdot)$. Find an orthonormal basis for M, $\{x_1, \ldots, x_n\}$ which we know can be constructed from any basis. Then any $z \in V$ can be written as $z = \sum_{i=1}^{n} \lambda_i x_i$, and so

$$\begin{aligned} z &= \sum_{i=1}^{n} \lambda_i x_i \\ &= \sum_{i=1}^{m} \lambda_i x_i + \sum_{i=m+1}^{n} \lambda_i x_i \\ &= x + y \\ &= P_M z + Q_M z \end{aligned}$$

In addition,

$$\begin{aligned} P_M z &= \sum_{i=1}^{m} (z, x_i) x_i = \sum_{i=1}^{m} \lambda_i x_i \\ \| P_M z \|^2 &= \sum_{i=1}^{m} (z, x_i)^2 = \sum_{i=1}^{m} \lambda_i^2 \\ Q_M z &= \sum_{i=m+1}^{n} (z, x_i) x_i = \sum_{i=m+1}^{n} \lambda_i x_i = z - P_M z \\ &= Iz - P_M z = (I - P_M) z \\ \| Q_M z \|^2 &= \sum_{i=m+1}^{n} (z, x_i)^2 = \sum_{i=m+1}^{n} \lambda_i^2 = \| z \|^2 - \| P_M z \|^2 \end{aligned}$$

By construction $x = P_M z \in M$ and $y = Q_M z \in M^{\perp}$, and this decomposition of z is unique. We don't actually need to have a basis for $M^{\perp}$, as we can compute y from $y = z - x$, and $Iz = P_M z + Q_M z$ gives $Q_M z = I - P_M z$.

Finding the closest vector in a subspace. Consider $z \in V, M \subset V$. We know that there is a unique decomposition $z = x + y,\ x \in M,\ y \in M^{\perp}$. What is the vector $w \in M$ that *minimizes* $\| z - w \|^2$ (as usual, relative to $(\cdot, \cdot)$)? Is it unique? Now

$$\begin{aligned} \| z - w \|^2 &= \| x + y - w \|^2 \\ &= \| x - w + y \|^2 \\ &= \| x - w \|^2 + \| y \|^2 + 2(x - w, y) \end{aligned}$$

but both x and $w \in M$ and $y \in M^{\perp}$, and thus $(x - w, y) = (x, y) + (w, y) = 0 + 0 = 0$, which means that

$$\| z - w \|^2 = \| x - w \|^2 + \| y \|^2$$

To minimize this, only the $\| x - w \|$ term matters, since we are free to choose any $w \in M$ and $x \in M$, simply set $w = x$. The minimum value of the norm is just $\| y \|^2$.

Write $x = P_M z$ and $y = Q_M z$. So, relative to the inner product $(\cdot, \cdot)$, the closest point in M to z is the projection of z onto M along $M^{\perp}, P_M z$, and its distance from M is $\| Q_M z \|$.

Also, the above construction shows that

$$\| z \|^2 = \| P_M z \|^2 + \| Q_M z \|^2.$$

These results are intimately related to linear models, as can be seen from the following picture. In the linear model problem, we may wish to choose $\hat{\mu} \in \mathcal{R}(X) = M$ so that $\| y - \hat{\mu} \|^2$ is minimized. For a given inner product, we know the answer is $\hat{\mu} = P_M Y$, and the minimum value of $\| y - \hat{\mu} \|^2 = \| Q_M Y \|^2$. Of course, the answer depends on the inner product.

2.7 More on Transformations and Projections

In this section we return to the discussion of linear transformations.

Definition 2.37 *The transpose of A, denoted as A' with respect to the inner product $(\cdot, \cdot)$ is defined by the following identity in x and y:*

$$(Ax, y) = (x, A'y), \text{ for all } x, y \in V.$$

From the definition, $(A')' = A$.

Theorem 2.24 *A' is a linear transformation on V.*

To prove this we need the following theorem.

Theorem 2.25 *For A and B linear transformations on $\{V, (\cdot, \cdot)\}$, if $(x, Ay) = (x, By)$, for all $x, y \in V$, then $A = B$.*

Proof. Homework.

Definition 2.38 *A linear transformation is* symmetric *if for all* $x, y \in V$, $(x, Ay) = (Ax, y)$. *We write this as* $A = A'$. *The notion of symmetry depends on the inner product* $(\cdot, \cdot)$.

Definition 2.39 *A symmetric linear transformation is* positive definite *if for all* $x \neq 0 \in V, (x, Ax) > 0$, *and is* positive semi-definite *if* $(x, Ax) \geq 0$ *for all* $x \neq 0 \in V$.

Definition 2.40 *A symmetric linear transformation is* nonsingular *if* $Ax = y$ *has a solution for each* $y \in V$ *or, equivalently,* $Ax = 0 \Rightarrow x = 0$.

Theorem 2.26 *If* A *is symmetric and positive definite, then* A *is nonsingular.*

Proof We need to show that for A positive definite, $Ax = 0 \Rightarrow x = 0$. Now by the Schwartz inequality

$$| (x, Ax) | \leq \| x \| \| Ax \|$$

Suppose $\| Ax \| = 0$ for $x \neq 0$. Then for this x, $| (x, Ax) | = (x, Ax) = 0$ and thus A is not positive definite. By contradiction, then, no such x exists. We here therefore justified in using the two different terms, nonsingular and positive definite as synonyms.

Thus far, all results have been stated for a fixed inner product. One is then led to ask how results change when the inner product changes, or if there is a relationship between results with different inner product. The connection between inner products is provided by the next theorem.

Theorem 2.27 *If* A *is positive definite symmetric with respect to* $(\cdot, \cdot)$, *then* $((x, y)) = (Ax, y)$ *is also an inner product on* V. *Thus one can generate a different inner product for every symmetric positive definite linear transformation* A.

Proof. We need to verify the definition of an inner product.

1. $((x, y)) = (Ax, y) = (x, Ay)$ (symmetry of A)
2. $((\alpha_1 x_1 + \alpha_2 x_2, y)) = (\alpha_1 Ax_1 + \alpha_2 Ax_2, y) = \alpha_1((x, y)) + \alpha_2((x_2, y))$
3. $((x, x)) = (Ax, y) > 0$ if $x \neq 0$ since A is positive definite.

We can now apply these ideas to characterize orthogonal projections.

Theorem 2.28 *P, a linear transformation on $\{V, (\cdot,\cdot)\}$ is an* orthogonal projection *if and only if $P = P^2 = P'$.*

Before beginning the proof, we recall that P is a projection if $P = P^2$, so it is the final imposition of symmetry that makes P an orthogonal projection. Since the notion of symmetry depends on the inner product, so does the notion of orthogonal projection.

Proof

1. Suppose the linear transformation P is an orthogonal projection onto $M \subset V$. Consider $z \in V, z = x + y, x \in M, y \in M^{\perp}$, so that $Pz = x$. Now for any $w \in V$, we will have that $((I-P)w, Pz) = 0$ because $(I-P)w \in M^{\perp}$ and $Pz \in M$. We can therefore write:

$$\begin{aligned}(w, Pz) &= (Pw + (I-P)w, Pz)\\ &= (Pw, Pz) + ((I-P)w, Pz)\\ &= (Pw, Pz)\\ &= (Pw, Pz) + (Pw, (I-P)z)\\ &= (Pw, Pz + (I-P)z)\\ &= (Pw, z)\end{aligned}$$

   and thus $P = P'$. $P = P^2$ because P is a projection.

2. Now suppose $P = P^2 = P'$. P is thus a projection by idempotency, and it projects on $\mathcal{R}(P)$ along $\mathrm{N}(P)$. An orthogonal projection requires that $\mathrm{N}(P) = \mathcal{R}(P)^{\perp}$, so we must show that

$$[\mathcal{R}(P)]^{\perp} = \mathrm{N}(P)$$

   First, take $x \in \mathcal{R}(P)$ and $y \in \mathrm{N}(P)$, then since $P = P^2$, $Px = x$ and $Py = 0$, and

$$(x, y) = (Px, y) = (x, Py) = (x, 0) = 0$$

   This shows that

$$y \in [\mathcal{R}(P)]^{\perp}, \text{ and } \mathrm{N}(P) \subset [\mathcal{R}(P)]^{\perp}$$

   Suppose

$$y \in [\mathcal{R}(P)]^{\perp} \text{ but } y \notin \mathrm{N}(P).$$

   Thus, $Py \neq 0$. If $x \neq 0$ and $x \in \mathcal{R}(P), 0 = (x, y) = (Px, y) = (x, Py)$. Since $Py \in \mathcal{R}(P)$, substitute Py for x and get $(Py, Py) = 0$. Thus $y \in \mathrm{N}(P)$.

Theorem 2.29 *If P is an orthogonal projection, then*

1. $(Px, x) \geq 0$
2. $\| Px \|^2 \leq \| x \|^2$.

Proof.

1. $(Px, x) = (P(Px), x) = (Px, Px) \geq 0$
2. $\| x \|^2 = \| (I - P + P)x \|^2 = \| (I - P)x + Px \|^2 = \| (I - P)x \|^2 + \| Px \|^2 + 2(Px, (I - P)x)$. Since P is an orthogonal projection, this last inner product can be written $(x, P'(I - P)x) = (x, P(I - P)x) = 0$, and thus
$$\| x \|^2 = \| (I - P)x \|^2 + \| Px \|^2 \geq \| Px \|^2$$
because $\| (I - P)x \|^2 \geq 0$.

Definition 2.41 *Two orthogonal projections A and B are called* orthogonal*, and written $A \perp B$, if $\mathcal{R}(A) \perp \mathcal{R}(B)$.*

Theorem 2.30 *Let A and B be orthogonal projections. Then $A \perp B$ if and only if $AB = 0$.*

Proof

1. If $A \perp B$, then $\mathcal{R}(A) \perp \mathcal{R}(B)$ since for $x \in \mathcal{R}(A), y \in R(B), (x, y) = 0$ by orthogonality of the spaces. Thus, for all $w, z \in V, 0 = (Aw, Bz) = (w, ABz)$ only if $AB = 0$.
2. If $AB = 0$, then $A(Bx) = 0$ for all $x \in V$ which means $\mathcal{R}(B)$ is contained in $\mathrm{N}(A) = [R(A)]^{\perp}$, so $R(A) \perp R(B)$.

Theorem 2.31 *Let $A_1, \ldots, A_k$ be orthogonal projections and let $A = \sum_{i=1}^{k} A_i$. A is an orthogonal projection if and only if $A_i A_j = 0, i \neq j$.*

Proof. If $A_i A_j = 0, i \neq j$, easy calculations show $A = A^2$ and $A = A'$. If A is an orthogonal projection, then we must show that $A_i(A_j)(x) = 0, \ i \neq j$. Consider

any fixed vector $x \in \mathcal{R}(A_j)$, so $A_j x = x$ and $\| A_j x \|^2 = \| x \|^2$.

$$\begin{aligned}
\| x \|^2 &\geq \| Ax \|^2 \\
&= (Ax, x) \\
&= \sum_{i=1}^{k} (A_i x, x) \\
&= \sum_{i=1}^{k} \| A_i x \|^2 \\
&= \sum_{i=1}^{k} \| A_i A_j x \|^2 \\
&= \| A_j x \|^2 + \sum_{i \neq j}^{k} \| A_i A_j x \|^2 \\
&= \| x \|^2 + \sum_{i \neq j}^{k} \| A_i A_j x \|^2
\end{aligned}$$

This shows that $A_i A_j x = 0$ for $j \neq i$ for all $x \in \mathcal{R}(A)$. Now for any $z \notin \mathcal{R}(A)$, we can write $z = x + y$, with $x \in \mathcal{R}(A)$ and $y \in$ the orthogonal complement of $\mathcal{R}(A) = \mathcal{R}(I - A)$. Then if $x \in \mathcal{R}(A_j)$, $A_i A_j z = A_i A_j x = 0$, and so the same proof will apply to any $z \in V$.

Let $\{M_1, \ldots, M_k\}$ be orthogonal subspaces such that $M = M_1 + M_2 + M_3 + \ldots + M_k$. Let $\{P_1, \ldots, P_k\}$ be the respective orthogonal projections onto the Ms. Then $P = \sum_{i=1}^{k} P_i$ is an orthogonal projection onto M and each $x \in M$ can be uniquely represented as $x = x_1 + \ldots + x_k$ with $x_i \in M_i$. This is an important result for fitting linear models, where each M_k can be identified with an "effect" of interest, and Y can be uniquely decomposed into components along each M.

For any $x \in M$,

$$Px = \sum_{i=1}^{k} P_i x = \sum_{i=1}^{k} P_i x_i = \sum_{i=1}^{k} x_i = x$$

and

$$\begin{aligned}
\| Px \|^2 &= \| \textstyle\sum_{i=1}^{k} P x_i \|^2 \\
&= (\sum_{i=1}^{k} P_i x, \sum_{i=1}^{k} P_i x)
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=1}^{k}\sum_{j=1}^{k}(P_i x, P_j x) \\
&= \sum_{i=1}^{k}(P_i x, P_i x) \\
&= \sum_{i=1}^{k} \| P_i x \|^2
\end{aligned}
$$

2.8 Orthogonal transformations

Definition 2.42 *A linear transformation A is called an* orthogonal transformation *if $\| Ax \| = \| x \|$, for all $x \in V$.*

An orthogonal transformation is nonsingular since if $Ax = 0$, then $x = 0$ (if not, then $(x, x) \neq 0$).

Theorem 2.32 *A is* orthogonal *if and only if $(Ax, Ay) = (x, y)$, for all $x, y \in V$.*

Proof Assume A is orthogonal. From a homework problem,

$$(x, y) = (1/4)[\| x + y \|^2 - \| x - y \|^2]$$

so

$$
\begin{aligned}
(Ax, Ay) &= (1/4)[\| Ax + Ay \|^2 - \| Ax - Ay \|^2] \\
&= (1/4)[\| A(x + y) \|^2 - \| A(x - y) \|^2] \\
&= (1/4)[\| x + y \|^2 - \| x - y \|^2] = (x, y)
\end{aligned}
$$

where the last results follow because A is orthogonal.

Now assume $(Ax, Ay) = (x, y)$ for all $x, y \in V$. Then $\| Ax \|^2 = (Ax, Ax) = (x, x) = \| x \|^2$ which by definition says that A is orthogonal.

The direct implication of this theorem is that (1) orthogonal transformations preserve lengths, and (2) orthogonal transformations preserve angles, that is, cosines, or equivalently inner products, are preserved. The proposition says A is orthogonal if and only if $(Ax, Ay) = (x, y)$, or $(x, y) = (x, A'Ay)$, so $(A')A = I$ (from a homework problem).

Theorem 2.33 *If $\{x_1, \ldots, x_n\}$ is an orthonormal basis for V and A is orthogonal, then $\{Ax_1, \ldots, Ax_n\}$ is an orthonormal basis for V.*

Proof $(Ax_i, Ax_j) = \delta_{ij}$, where $\delta_{ij} = 1$ if $i = j$ and $= 0$ otherwise.

Theorem 2.34 *Let* $\{x_1, \ldots, x_n\}$ *and* $\{y_1, \ldots, y_n\}$ *be orthonormal bases for* V. *The linear transformation defined by* $Ax_i = y_i$ *is orthogonal.*

Proof. For $x \in V, x = \sum \lambda_i x_i$ and $\| x \|^2 = \sum \lambda_i^2$. Then

$$\| Ax \|^2 = \| A \sum \lambda_i x_i \|^2 = \| \sum \lambda_i A x_i \|^2 = \| \sum \lambda_i y_i \|^2 = \sum \lambda_i^2 = \| x \|^2.$$

and thus A is orthogonal.