14

*VECTOR GEOMETRY

The star means that strictly speaking, instructors may skip this chapter and later starred sections that are also geometrically oriented. However, many students find a picture worth a thousand words. So those who find that geometry makes a difficult subject easier to follow, are encouraged to study this starred material.

14-1 THE GEOMETRIC INTERPRETATION OF VECTORS

(a) Introduction

Assuming the reader is familiar with vector algebra, we develop its corresponding geometric interpretation in this chapter; this is then used to reinterpret regression and correlation theory. Readers with matrix algebra will have simultaneously taken varying amounts of geometry; hence, some may be able to pick up this argument at a midway point. But for the sake of those who have very little background, this geometry is developed from first principles. For simplicity, we begin by showing vectors in only two or three dimensions. However, interpretations in any number of dimensions are equally valid; thus, we can drop explicit reference to the dimension of the space later on.

Consider the vector

$$\mathbf{x} = (x_1, x_2, \ldots, x_n) \tag{14-1}$$

For example,

$$\mathbf{x} = (2, 4, 3) \tag{14-2}$$

which may be plotted as a point in three dimensions (Figure 14-1). Sometimes it is more convenient to represent it as an arrow from the origin to the point. If a vector is designated as an arrow, it may be shifted, provided its length and direction are maintained—that is, it may be shifted in a

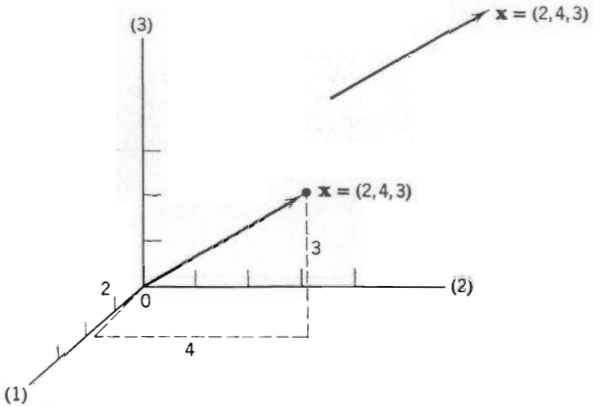


FIGURE 14-1 A three-dimensional vector. This vector is the direction and distance defined by moving two units in the first direction, four units in the second direction, and three units in the third direction.

parallel way. But if a vector is designated as a point only, then of course this point may *not* be shifted.

The simple algebraic manipulations of vectors are set out in Table 14-1, along with the corresponding geometric interpretation. In addition, each geometric operation is detailed in Figures 14-2 to 14-4.

In review, in Figure 14-5 we see that the sum $(\mathbf{x} + \mathbf{y})$ is one diagonal of the parallelogram formed from $\mathbf{x}$ and $\mathbf{y}$, while the difference $(\mathbf{x} - \mathbf{y})$ is the other diagonal.

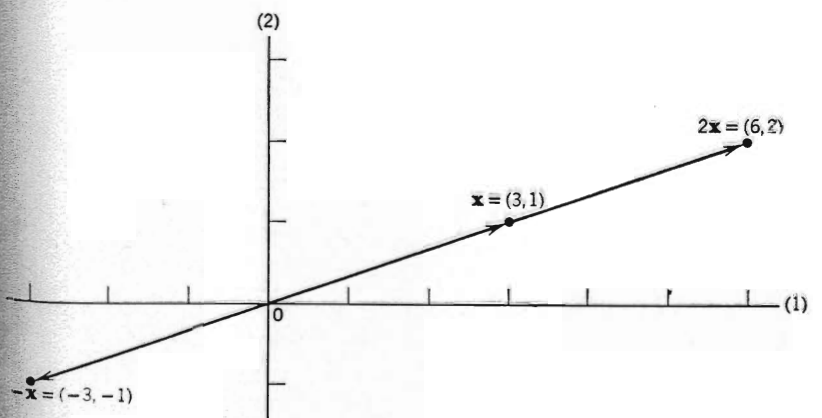


FIGURE 14-2 Scalar multiplication

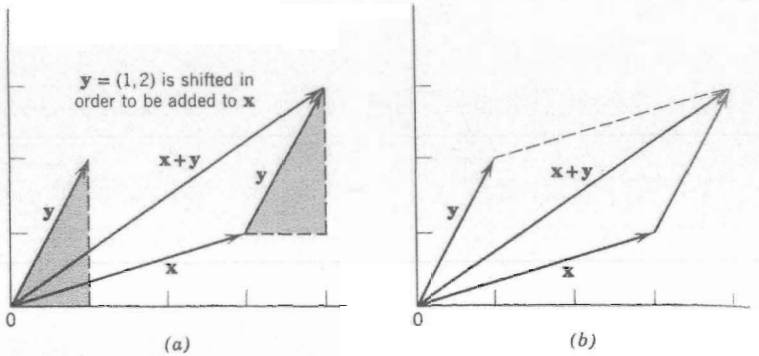


FIGURE 14-3 Vector addition, which in (*b*) is seen to be equivalent to constructing a diagonal of the parallelogram defined by x and y.

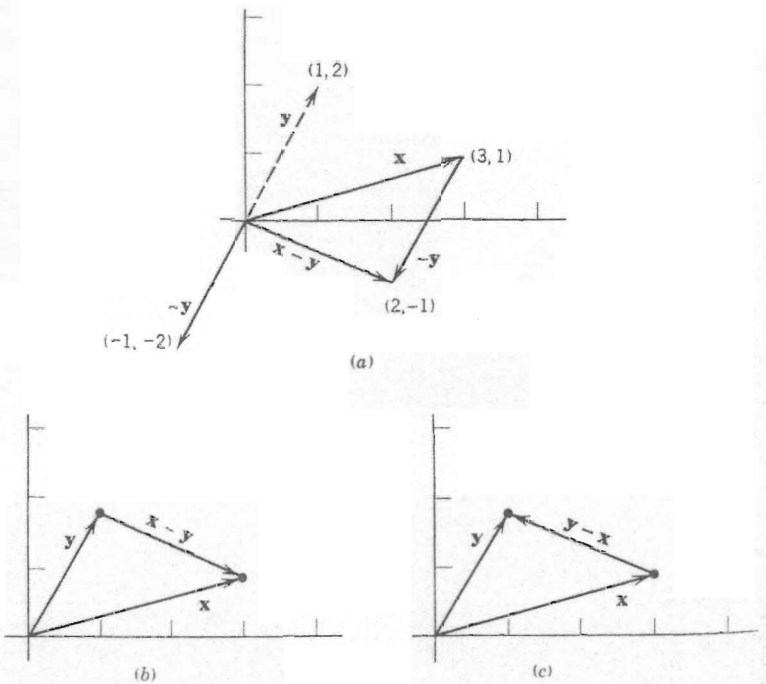


FIGURE 14-4 Vector subtraction (x − y), which in (*b*) is seen to be equivalent to moving from point y to point x. (*c*) The reader can confirm that (y − x) is similarly obtained by moving from point x to point y.

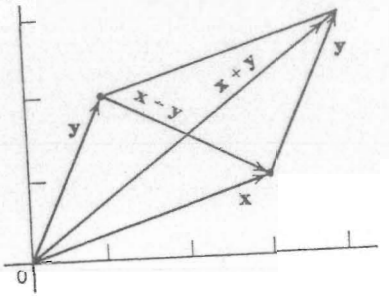


FIGURE 14-5 Vector addition and subtraction compared. Addition is the diagonal obtained by shifting the arrow y to follow the arrow x; subtraction is the diagonal obtained by moving from the point y to the point x.

TABLE 14-1 Comparison of the Algebra and Geometry of Vectors

	Since each manipulation is defined algebraically in this way:	*It follows that it has this geometric interpretation:*
Scalar multiplication by a positive constant	$2(3, 1) = (6, 2)$	Changes length (Figure 14-2)
Scalar multiplication by -1	$-1(3, 1) = (-3, -1)$	Changes direction (Figure 14-2)
Addition	$(3, 1) + (1, 2) = (4, 3)$	Shifts the arrow y to follow the arrow x (Figure 14-3); this is seen to yield the diagonal of the parallelogram constructed from x and y
Subtraction	$(3, 1) - (1, 2) = (2, -1)$	Is equivalent to summing $x + (-y)$, that is, shifting the arrow $(-y)$ to follow the arrow x in Figure 14-4*a*. This is also seen to be the arrow obtained in Figure 14-4*b* by moving from the *point* y to the *point* x

(b) Dot Product

(i) Definition and Properties. The dot product (also called inner product or scalar product) of two vectors is defined as a simple kind of matrix multiplication:

$$\mathbf{x} \cdot \mathbf{y} = \boxed{x_1\ x_2\ \cdots\ x_n} \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix} = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n \qquad (14\text{-}3)$$

For example,

$$(3, 1, -1) \cdot (2, -3, 0) = 3$$

The dot product, of course, obeys all the rules of matrix multiplication; for example,

$$\mathbf{x} \cdot (\mathbf{y} + \mathbf{z}) = \mathbf{x} \cdot \mathbf{y} + \mathbf{x} \cdot \mathbf{z} \quad \text{(distributive law)} \qquad (14\text{-}4)$$

$$\mathbf{x} \cdot (c\mathbf{y}) = (c\mathbf{x}) \cdot \mathbf{y} = c(\mathbf{x} \cdot \mathbf{y}) \qquad (14\text{-}5)$$

But, it also satisfies in addition:

$$\mathbf{x} \cdot \mathbf{y} = \mathbf{y} \cdot \mathbf{x} \quad \text{(commutative law)} \qquad (14\text{-}6)$$

(ii) Length. A special case is the dot product of a vector with itself:

$$\mathbf{x} \cdot \mathbf{x} = x_1^2 + x_2^2 + \cdots + x_n^2 \qquad (14\text{-}7)$$

This is called $\|\mathbf{x}\|^2$. In two dimensions we recognize it as the squared length of the vector, according to the theorem of Pythagoras in Figure 14-6*a*. For example, the vector $\mathbf{x} = (3, 1)$ has squared length

$$\mathbf{x} \cdot \mathbf{x} = 3^2 + 1^2 = 10$$

Thus, its length is $\sqrt{10} = 3.16$.

It is easy to also confirm in three dimensions that $\|\mathbf{x}\|^2$ is the squared length of the vector. For example, in Figure 14-6*b*, first apply the Pythagorean theorem to the horizontal ΔABC, obtaining $x_1^2 + x_2^2$ as the squared length of AC. Then apply the Pythagorean theorem again to the vertical ΔACD, confirming that the squared length of the vector AD is

$$(x_1^2 + x_2^2) + x_3^2 = \|\mathbf{x}\|^2 \qquad (14\text{-}8)$$

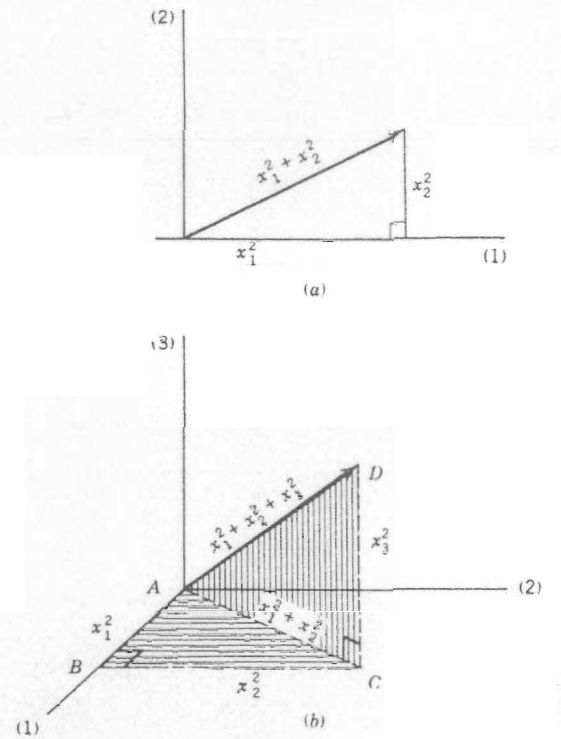


FIGURE 14-6 Squared lengths of vectors, related by the theorem of Pythagoras. (*a*) In two dimensions. (*b*) In three dimensions.

As an example, the squared length of the vector $\mathbf{x} = (2, 4, 3)$ is

$$\|\mathbf{x}\|^2 = \mathbf{x} \cdot \mathbf{x} = 2^2 + 4^2 + 3^2 = 29 \qquad (14\text{-}9)$$

Thus, its length is $\sqrt{29} = 5.39$.

$\|\mathbf{x}\|$ has turned out to be length wherever it is physically meaningful (in 1, 2, or 3 dimensions). We will use a little mathematical imagination and call $\|\mathbf{x}\|$ the *length* (or *norm*) in any number of dimensions.

To review:

$$\|\mathbf{x}\|^2 = \mathbf{x} \cdot \mathbf{x} \qquad (14\text{-}10)$$

$$= x_1^2 + x_2^2 \cdots + x_n^2 = \text{squared length} \qquad (14\text{-}11)$$

while

$$\|\mathbf{x}\| = \text{length}$$

One of the most frequently used facts about length is that, if c is positive,

$$\|c\mathbf{x}\| = c\|\mathbf{x}\| \tag{14-12a}$$

This is obvious from Figure 14-2, and may be proved more rigorously in n dimensions.[1] When c is negative, equation (14-12a) cannot be correct because the right side is negative while the left side is positive. By taking the absolute value (magnitude) of c, we may write a generally correct form of (14-12a):

$$\|c\mathbf{x}\| = |c|\,\|\mathbf{x}\| \tag{14-12b}$$

(iii) Perpendicularity. Also called *orthogonality*, and symbolized by $\perp$, this is easily expressed in terms of vector length. From Figure 14-7, it is evident that $\mathbf{x} \perp \mathbf{y}$ iff the length of $(\mathbf{x} + \mathbf{y})$ equals the length of $(\mathbf{x} - \mathbf{y})$, that is, iff

$$\|\mathbf{x} + \mathbf{y}\|^2 = \|\mathbf{x} - \mathbf{y}\|^2$$
$$(\mathbf{x} + \mathbf{y}) \cdot (\mathbf{x} + \mathbf{y}) = (\mathbf{x} - \mathbf{y}) \cdot (\mathbf{x} - \mathbf{y})$$
$$\mathbf{x} \cdot \mathbf{x} + 2\mathbf{x} \cdot \mathbf{y} + \mathbf{y} \cdot \mathbf{y} = \mathbf{x} \cdot \mathbf{x} - 2\mathbf{x} \cdot \mathbf{y} + \mathbf{y} \cdot \mathbf{y}$$
$$4\mathbf{x} \cdot \mathbf{y} = 0$$
$$\mathbf{x} \cdot \mathbf{y} = 0$$

$$\boxed{\mathbf{x} \perp \mathbf{y} \text{ iff } \mathbf{x} \cdot \mathbf{y} = 0} \tag{14-13}$$

That is, two vectors are perpendicular if and only if their dot product is zero.

(c) Subspaces

(i) Generation of Subspaces. In Figure 14-8, we show that when a fixed vector $\mathbf{x}_1$ is multiplied by every possible scalar c_1, a straight line is generated running through $\mathbf{x}_1$ and the origin. Each vector $c_1\mathbf{x}_1$ may be

[1] *Proof of* (14-12a). Since $c\mathbf{x} = (cx_1, cx_2, \ldots)$, from definition (14-11),

$$\|c\mathbf{x}\|^2 = (cx_1)^2 + (cx_2)^2 + \cdots$$
$$= c^2(x_1^2 + x_2^2 + \cdots) = c^2\|\mathbf{x}\|^2$$

Thus

$$\|c\mathbf{x}\| = c\|\mathbf{x}\|$$

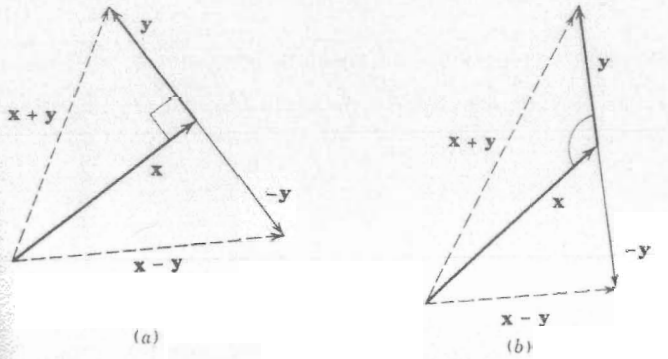


FIGURE 14-7 $\mathbf{x} \perp \mathbf{y}$ iff $\|\mathbf{x} + \mathbf{y}\| = \|\mathbf{x} - \mathbf{y}\|$. (Note that the diagrams are valid in any number of dimensions.) (*a*) $\mathbf{x} \perp \mathbf{y}$. (*b*) $\mathbf{x}$ not $\perp \mathbf{y}$.

represented as an arrow or a point, but the picture is less cluttered if we simply use a point. In summary we write

$$L: \quad c_1\mathbf{x}_1 \qquad -\infty < c_1 < \infty \tag{14-14}$$

In Figure 14-9, we increase dimension by one. In this figure, we use for the first time two conventions about arrowheads. First, arrows within the plane have a light arrowhead, and arrows outside the plane have a dark arrowhead. Second, arrowheads are shown as cones so that when the arrow is pointing away from the reader, the circular base of the cone can be seen. Finally, we represent the plane as a slab, although mathematically speaking,

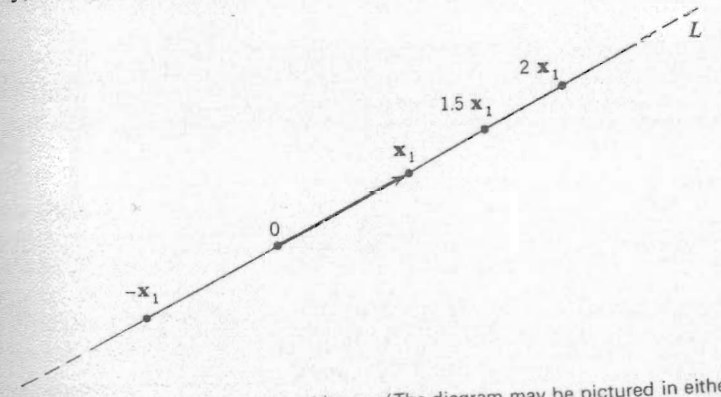


FIGURE 14-8 The line L generated by $\mathbf{x}_1$. (The diagram may be pictured in either two or three dimensions.) L is $c_1\mathbf{x}_1$, where c_1 takes on all values, so that the line extends to infinity.

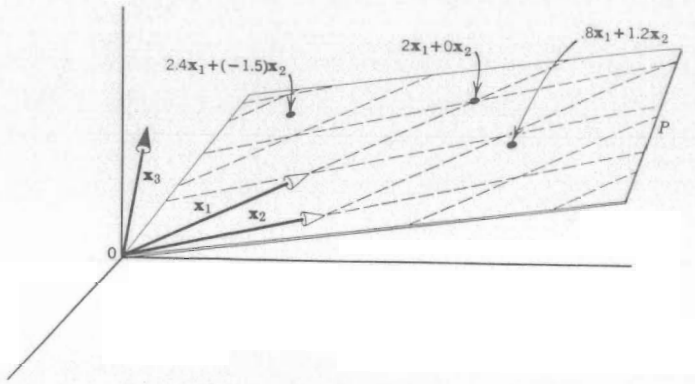


FIGURE 14-9 The plane P generated by $\mathbf{x}_1$ and $\mathbf{x}_2$. P is $c_1\mathbf{x}_1 + c_2\mathbf{x}_2$, where c_1, c_2 take on all values, so that the plane extends to infinity.

it has no thickness. These conventions make it much easier to visualize geometry in n-dimensional space (n-space).

In Figure 14-9, we show the set of points generated by two fixed vectors in 3-space,

$$P: \quad c_1\mathbf{x}_1 + c_2\mathbf{x}_2 \qquad -\infty < c_1, c_2 < \infty \tag{14-15}$$

This is called the *set of all possible linear combinations of* $\mathbf{x}_1$ *and* $\mathbf{x}_2$, and is the plane running through $\mathbf{x}_1, \mathbf{x}_2$, and the origin.[2] Geometrically, we see that we can generate (i.e., get to) any point on this plane P by taking the appropriate linear combination of $\mathbf{x}_1$ and $\mathbf{x}_2$, that is, by appropriately selecting c_1 and c_2 in (14-15); but we cannot generate any point above or below this plane.

To generate the whole 3-space requires a third independent vector, such as $\mathbf{x}_3$, to take us off the plane P. Thus, the whole set of points in this 3-space could be generated by

$$c_1\mathbf{x}_1 + c_2\mathbf{x}_2 + c_3\mathbf{x}_3 \qquad -\infty < c_i < \infty \tag{14-16}$$

This is often stated as: $\mathbf{x}_1, \mathbf{x}_2$, and $\mathbf{x}_3$ generate (or span) this 3-space. Or: $\mathbf{x}_1$, $\mathbf{x}_2$, and $\mathbf{x}_3$ are a *basis* of this 3-space.

This generalizes into n-space; consider the set of points

$$c_1\mathbf{x}_1 + c_2\mathbf{x}_2 + \cdots + c_m\mathbf{x}_m \qquad -\infty < c_i < \infty \tag{14-17}$$

[2] Unless $\mathbf{x}_1, \mathbf{x}_2$, and the origin all lie on a straight line, in which case we can generate only this line. Or worse yet, if $\mathbf{x}_1, \mathbf{x}_2$, and the origin all coincide (i.e., $\mathbf{x}_1 = \mathbf{x}_2 = \mathbf{0}$) then we can generate only one point—the origin. These degenerate cases are called "linear dependence of $\mathbf{x}_1$ and $\mathbf{x}_2$." For simplicity, we will assume throughout this chapter that linear dependence does not occur.

This set of all possible linear combinations of these m fixed vectors is called an m-dimension *subspace*. If $m = 1$, then the subspace is a straight line. If $m = 2$, then the subspace is a plane. If $m > 2$, the subspace is called a hyperplane. Only if $m = n$, and we thus have n linearly independent vectors ($\mathbf{x}_1$, $\mathbf{x}_2 \cdots \mathbf{x}_n$) will we generate all of our n-space. For any vector or point $\mathbf{y}$ in this n-space, a unique set of coefficients $c_1, c_2 \cdots c_n$ can be found such that

$$\mathbf{y} = c_1\mathbf{x}_1 + c_2\mathbf{x}_2 + \cdots + c_n\mathbf{x}_n \tag{14-18}$$

These values $\langle c_1, c_2, \ldots, c_n \rangle$ are called *the coordinates of* $\mathbf{y}$ *with respect to the basis* $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$.

For example, in two-space, the vectors $\mathbf{x}_1 = (1, -1)$ and $\mathbf{x}_2 = (2, 1)$ will generate the whole space. We now use (14-18) to find the coordinates (with respect to this basis $\mathbf{x}_1, \mathbf{x}_2$) of a given vector, say $\mathbf{y} = (4, -1)$; this involves selecting c_1 and c_2 so that

$$c_1\mathbf{x}_1 + c_2\mathbf{x}_2 = \mathbf{y} \tag{14-19}$$

that is[3]

$$c_1\begin{bmatrix}1\\-1\end{bmatrix} + c_2\begin{bmatrix}2\\1\end{bmatrix} = \begin{bmatrix}4\\-1\end{bmatrix} \tag{14-20}$$

that is,

$$c_1 + 2c_2 = 4$$
$$-c_1 + c_2 = -1$$

The algebraic solution to this set of equations is $c_1 = 2, c_2 = 1$. We have expressed $\mathbf{y}$ as a linear combination of $\mathbf{x}_1$ and $\mathbf{x}_2$, with the coefficients $\langle 2, 1 \rangle$ being the coordinates of $\mathbf{y}$ with respect to $\mathbf{x}_1$ and $\mathbf{x}_2$.

This is seen geometrically in Figure 14-10: The line (subspace) L_1 generated by $\mathbf{x}_1$ is shown, along with the subspace L_2 generated by $\mathbf{x}_2$. Then we complete the parallelogram, confirming that c_1 must be 2, and c_2 must be 1.

In other words, to find the coordinates of $\mathbf{y}$, we project; to find c_1, we project $\mathbf{y}$ onto L_1 in the direction parallel to L_2; or, stated briefly, we project $\mathbf{y}$ *onto* $\mathbf{x}_1$ *along* (parallel to) $\mathbf{x}_2$. Similarly, to find c_2 we project $\mathbf{y}$ onto $\mathbf{x}_2$ along $\mathbf{x}_1$.

The simplest kind of projection occurs when $\mathbf{x}_1 \perp \mathbf{x}_2$; this is called an *orthogonal* projection, as in Figure 14-11.

[3] For convenience, we sometimes write our vectors as columns instead of rows. More formally, we can easily justify (14-20) by transposing (14-19).

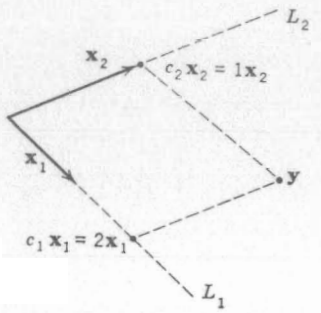


FIGURE 14-10 Finding the coordinates of $\mathbf{y}$ (with respect to $\mathbf{x}_1, \mathbf{x}_2$) geometrically by projection.

(d) Perpendicular Projections and Least Distance

Perpendicular projections are particularly easy to calculate, because the condition for perpendicularity (14-13) is so simple. To work out $\mathbf{y}_1$, the $\perp$ projection of $\mathbf{y}$ onto $\mathbf{x}_1$, consider Figure 14-12, which is valid in any dimension. Of course, since the projection vector $\mathbf{y}_1$ lies on L_1, it is just a (scalar) multiple of $\mathbf{x}_1$; that is,

$$\mathbf{y}_1 = c\mathbf{x}_1 \tag{14-21}$$

with the problem being to determine c. Moreover, we note (Figure 14-5) that $\mathbf{y} - \mathbf{y}_1$ is the vector defined by moving from $\mathbf{y}_1$ to $\mathbf{y}$. But we must keep this perpendicular to $\mathbf{x}_1$, that is, we must find c, so that

$$(\mathbf{y} - \mathbf{y}_1) \perp \mathbf{x}_1 \tag{14-22}$$

Substitute (14-21) into (14-22) and use (14-13):

$$(\mathbf{y} - c\mathbf{x}_1) \cdot \mathbf{x}_1 = 0$$

$$(\mathbf{y} \cdot \mathbf{x}_1) - c(\mathbf{x}_1 \cdot \mathbf{x}_1) = 0$$

$$\boxed{c = \frac{\mathbf{y} \cdot \mathbf{x}_1}{\mathbf{x}_1 \cdot \mathbf{x}_1}} \tag{14-23}$$

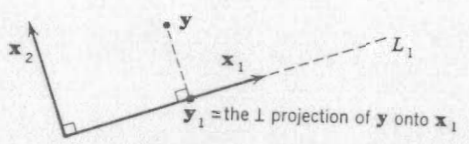


FIGURE 14-11 The orthogonal projection of $\mathbf{y}$ onto $\mathbf{x}_1$.

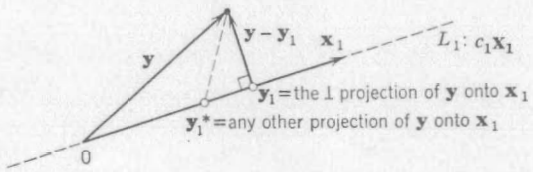


FIGURE 14-12 Orthogonal projection in two dimensions: $\mathbf{y}$ projected onto $\mathbf{x}_1$.

Substituting (14-23) into (14-21) establishes that

$$\boxed{\mathbf{y}_1 = \left(\frac{\mathbf{y} \cdot \mathbf{x}_1}{\mathbf{x}_1 \cdot \mathbf{x}_1}\right)\mathbf{x}_1} \tag{14-24}$$

where $\mathbf{y}_1$ is the $\perp$ projection of $\mathbf{y}$ onto $\mathbf{x}_1$.

The length of this projected vector has a simple formula too:

$$\begin{aligned} \|\mathbf{y}_1\|^2 &= \mathbf{y}_1 \cdot \mathbf{y}_1 \\ &= c^2\mathbf{x}_1 \cdot \mathbf{x}_1 \\ &= \left[\frac{\mathbf{y} \cdot \mathbf{x}_1}{\mathbf{x}_1 \cdot \mathbf{x}_1}\right]^2 \mathbf{x}_1 \cdot \mathbf{x}_1 \end{aligned}$$

$$\boxed{\|\mathbf{y}_1\|^2 = \frac{(\mathbf{y} \cdot \mathbf{x}_1)^2}{\mathbf{x}_1 \cdot \mathbf{x}_1}} \tag{14-25}$$

Hence, the norm or length of $\mathbf{y}_1$ is

$$\|\mathbf{y}_1\| = \frac{|\mathbf{y} \cdot \mathbf{x}_1|}{\|\mathbf{x}_1\|} \tag{14-26}$$

Referring again to Figure 14-12, we see that the *perpendicular* projection $\mathbf{y}_1$ is the point on L_1 closest to $\mathbf{y}$; any nonperpendicular projection, say $\mathbf{y}_1^*$, is farther from $\mathbf{y}$. The proof is simple: The distance $\|\mathbf{y} - \mathbf{y}_1^*\|$ must be greater than the distance $\|\mathbf{y} - \mathbf{y}_1\|$, because the hypotenuse of a right-angled triangle is greater than either side.

This theorem is important enough for regression and correlation theory that it is shown in the three-dimensional case in Figure 14-13.

The perpendicular projection of $\mathbf{y}$ onto the subspace

$$c_1\mathbf{x}_1 + c_2\mathbf{x}_2 + \cdots + c_m\mathbf{x}_m \tag{14-27}$$

is the one point on this subspace closest to $\mathbf{y}$.

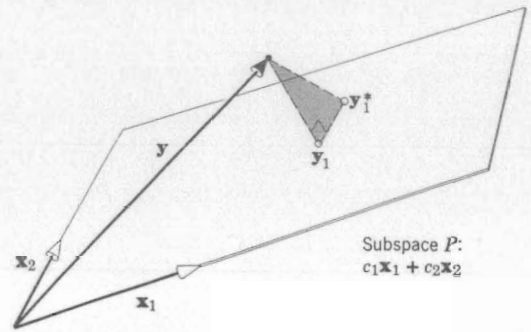


FIGURE 14-13 Orthogonal projection in three dimensions: y projected onto (x_1, x_2) subspace.

(e) Cos θ

We can now obtain a simple formula for the cosine of the angle between any two vectors $\mathbf{x}_1$ and $\mathbf{y}$. Referring to Figure 14-14, we first $\perp$ project $\mathbf{y}$ onto $\mathbf{x}_1$; then, by definition from trigonometry,

$$\cos\theta = \pm\frac{\|\mathbf{y}_1\|}{\|\mathbf{y}\|} \tag{14-28}$$

Moreover, from Figure 14-14, it is clear that the sign of $\cos\theta$ agrees with the sign of the coefficient c in (14-21). Using this equation, we may rewrite (14-28) as

$$\cos\theta = \frac{c\|\mathbf{x}_1\|}{\|\mathbf{y}\|} \tag{14-29}$$

Substituting (14-23)

$$\cos\theta = \frac{\mathbf{y}\cdot\mathbf{x}_1}{\|\mathbf{x}_1\|^2}\frac{\|\mathbf{x}_1\|}{\|\mathbf{y}\|} \tag{14-30}$$

$$\cos\theta = \frac{\mathbf{y}\cdot\mathbf{x}_1}{\|\mathbf{x}_1\|\,\|\mathbf{y}\|} \tag{14-31}$$

To free our notation somewhat, we rename $\mathbf{x}_1$ by $\mathbf{x}$:

$$\boxed{\cos\theta = \frac{\mathbf{x}\cdot\mathbf{y}}{\|\mathbf{x}\|\,\|\mathbf{y}\|}} \tag{14-32}$$

where θ is the angle between $\mathbf{x}$ and $\mathbf{y}$.

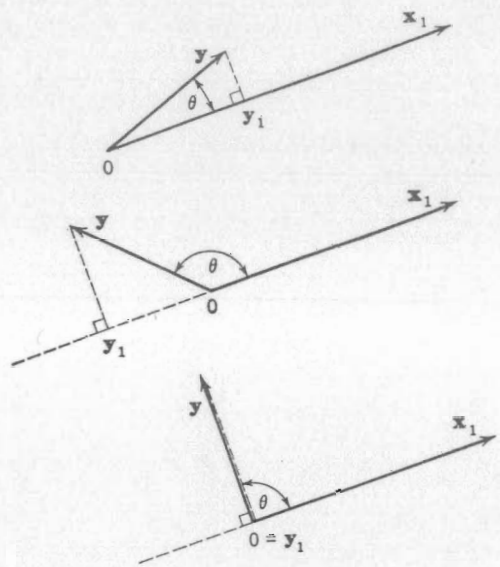


FIGURE 14-14 Cos θ. (*a*) θ acute, $\cos\theta > 0$, $\mathbf{y}_1 = c\mathbf{x}_1$, where $c > 0$. (*b*) θ obtuse, $\cos\theta < 0$; $\mathbf{y}_1 = c\mathbf{x}_1$, where $c < 0$. (*c*) θ is 90° ($\mathbf{y} \perp \mathbf{x}_1$), $\cos\theta = 0$; $\mathbf{y}_1 = c\mathbf{x}_1$, where $c = 0$.

PROBLEMS

14-1 Let $\mathbf{x} = (1, -2)$ and $\mathbf{y} = (3, 1)$. Graph as arrows $\mathbf{x}$ and $\mathbf{y}$ and the following:

(a) $2\mathbf{x}$.
(b) $\mathbf{x} + \mathbf{y}$.
(c) $\mathbf{x} - \mathbf{y}$.
(d) $(\mathbf{x} + \mathbf{y}) + (\mathbf{x} - \mathbf{y})$. Check that this equals $2\mathbf{x}$.
(e) $(\mathbf{x} + \mathbf{y}) - (\mathbf{x} - \mathbf{y})$. Check that this equals $2\mathbf{y}$.
(f) $-3\mathbf{x} + 2\mathbf{y}$.

14-2 Which of the following pairs are orthogonal?

(a) (1, 3) and (−6, 2).
(b) (1, −2) and (1, 2).
(c) (1, 2, −2) and (2, 3, 2).
(d) (1, −2, 1, 0, 1) and (2, 0, 1, −1, −1); call these $\mathbf{x}_1$ and $\mathbf{x}_2$.

14-3 Find c so that $(\mathbf{x}_2 - c\mathbf{x}_1)$ will be perpendicular to $\mathbf{x}_1$ in Problem 14-2(d) above.

14-4 Using the basis $\mathbf{x}_1 = (1, -1), \mathbf{x}_2 = (2, 1)$, find algebraically the coordinates of each of the following points. Then verify your result geometrically (work approximately).

(a) $(-1, -2)$.
(b) $(0, 3)$.
(c) $(-1, 1)$.
(d) $(3, -1)$.

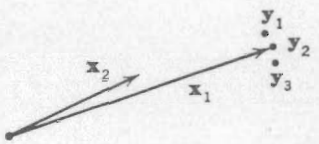
14-5 Find the coordinates of the point $(1.2, .3)$ with respect to

(a) the basis $\mathbf{x}_1 = (.5, .1), \mathbf{x}_2 = (.1, .2)$
(b) the orthogonal basis $\mathbf{x}_1 = (-.4, .2), \mathbf{x}_2 = (.1, .2)$
(c) the orthonormal basis $\mathbf{x}_1 = (-.6, -.8), \mathbf{x}_2 = (-.8, .6)$, where each vector is normalized, that is, of length 1. Which basis is easiest?

14-6 Consider the basis $\mathbf{x}_1 = (1, 0, 2)$ and $\mathbf{x}_2 = (2, -1, 1)$, which generates a plane P in three-space. For each point below, find whether it lies on the plane; if it does, then find its coordinates with respect to $(\mathbf{x}_1, \mathbf{x}_2)$:

(a) $(5, -1, 7)$.
(b) $(4, 0, 1)$.
(c) $(3, -2, 0)$.

14-7 Consider vectors $\mathbf{x}_1, \mathbf{x}_2$ in the two-space: $c_1\mathbf{x}_1 + c_2\mathbf{x}_2$:



(a) Match the point with the correct pair of coordinates with respect to $(\mathbf{x}_1, \mathbf{x}_2)$. Work geometrically, and roughly.

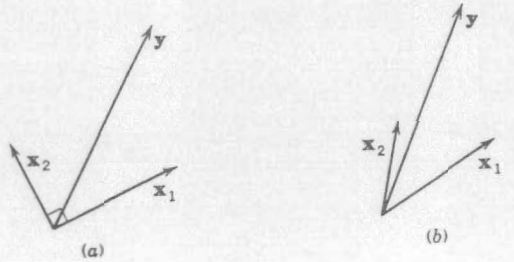
$\mathbf{y}_1$ $\quad \langle 1, 0 \rangle$

$\mathbf{y}_2$ $\quad \langle \frac{1}{2}, 1 \rangle$

$\mathbf{y}_3$ $\quad \langle 1\frac{1}{2}, -1 \rangle$

(b) The three points are close together. Are their coordinates close? Give an intuitive reason why.

14-8



Work geometrically in two-space. In each case, find the $\perp$ projection of $\mathbf{y}$ onto $\mathbf{x}_1$, and $\mathbf{y}$ onto $\mathbf{x}_2$. Call them $\mathbf{y}_1$ and $\mathbf{y}_2$. Under what circumstances does $\mathbf{y} = \mathbf{y}_1 + \mathbf{y}_2$?

14-2 LEAST SQUARES FIT

With this geometry in hand, we now turn to its application to regression. First, consider the problem of fitting a line, as in Chapter 2. To keep the geometry simple, our example in Figure 14-15 consists of only three observed points. The values of x are centered at 0; this can always be achieved by using deviations from the mean (y may or may not be also translated to a zero mean). The mathematical model may be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \tag{14-33}$$

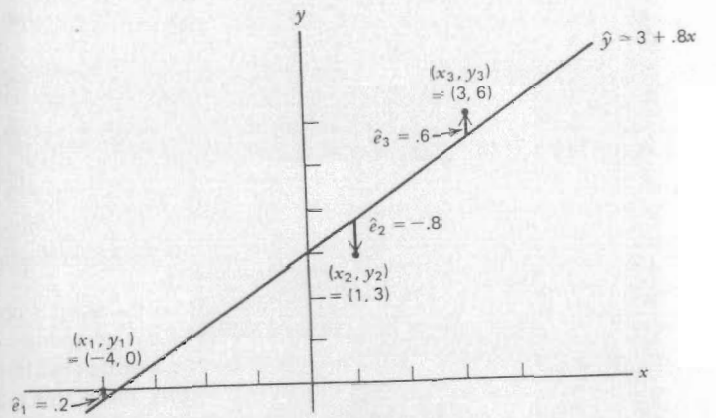


FIGURE 14-15 Regression scatter: points are observations; axes are variables.

The sample values used for estimating $\boldsymbol{\beta}$ are displayed in the form:

$$\mathbf{y} = \mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\mathbf{e}} \tag{14-34}$$

$$\begin{bmatrix} 0 \\ 3 \\ 6 \end{bmatrix} = \begin{bmatrix} 1 & -4 \\ 1 & 1 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} + \begin{bmatrix} \hat{e}_1 \\ \hat{e}_2 \\ \hat{e}_3 \end{bmatrix}$$

or

$$\mathbf{y} = \hat{\mathbf{y}} + \hat{\mathbf{e}} \tag{14-35}$$

observed y = fitted y + residual

where the fitted vector is

$$\hat{\mathbf{y}} = \hat{\beta}_1 \mathbf{x}_1 + \hat{\beta}_2 \mathbf{x}_2$$

$$\hat{\mathbf{y}} = \hat{\beta}_1 \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} + \hat{\beta}_2 \begin{bmatrix} -4 \\ 1 \\ 3 \end{bmatrix} \tag{14-36}$$

Figure 14-16 displays exactly this same information in alternative vector geometry. Whereas in Figure 14-15 each *observation* was plotted as a point in "variable" space (i.e., the space defined by variables on the axes), in Figure 14-16 each *variable* is now plotted as a point or arrow in an "observation" space. Whereas each point in Figure 14-15 was drawn from a *row* of (14-34) [e.g., the first point $(-4, 0)$ was drawn from the first row], each point or vector in Figure 14-16 is a *column* of (14-34).

In our example we also note that $\mathbf{x}_1$ and $\mathbf{x}_2$ are perpendicular. This follows because

$$\mathbf{x}_1 \cdot \mathbf{x}_2 = (1, 1, 1) \cdot (x_1, x_2, x_3) \tag{14-37}$$

$$= x_1 + x_2 + x_3 \tag{14-38}$$

$$= n\bar{x} \tag{14-39}$$

Recall that x was translated so that $\bar{x} = 0$; therefore

$$\mathbf{x}_1 \cdot \mathbf{x}_2 = 0 \tag{14-40}$$

This establishes that $\mathbf{x}_1 \perp \mathbf{x}_2$. This was the motive for translating x onto a zero mean.

Algebraically in (14-36) our problem is to find a fitted value of $\mathbf{y}$ that is a linear combination of $\mathbf{x}_1$ and $\mathbf{x}_2$. Geometrically, in Figure 14-16, this means that we must select a fit somewhere on the plane P generated by $\mathbf{x}_1$ and $\mathbf{x}_2$. If we wish to determine the point or vector on this plane P that best fits, or is closest to the observed $\mathbf{y}$, we should drop a perpendicular from $\mathbf{y}$ onto P. Is

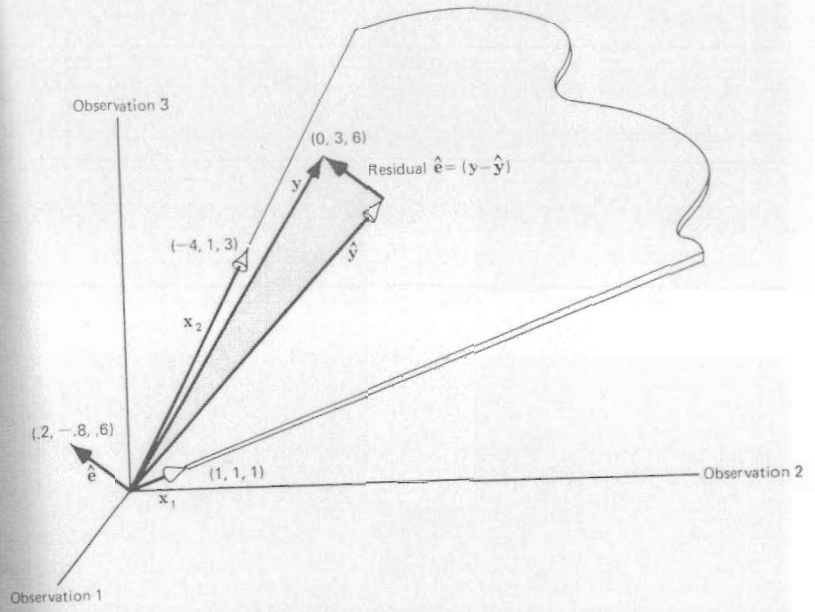


FIGURE 14-16 Same information as in Figure 14-15; but here points (vectors) are variables, axes are observations.

this the least squares solution? The answer is: yes. Recall that least squares involves selecting $\hat{y}_i$ to minimize

$$\sum (y_i - \hat{y}_i)^2 \tag{14-41}$$

In vector notation, according to (14-11), this is:

$$\|\mathbf{y} - \hat{\mathbf{y}}\|^2 \tag{14-42}$$

That is, least squares involves minimizing the squared distance (i.e., minimizing the distance) between $\mathbf{y}$ and $\hat{\mathbf{y}}$, which is accomplished by perpendicular projection according to (14-27).

It is also important to note that the vector of estimated residuals

$$\hat{\mathbf{e}} = (\mathbf{y} - \hat{\mathbf{y}})$$

is perpendicular (orthogonal) to the $(\mathbf{x}_1, \mathbf{x}_2)$ plane; hence $\hat{\mathbf{e}}$ is orthogonal to each of the regressors $\mathbf{x}_1$ and $\mathbf{x}_2$ in the plane. This means

$$\mathbf{x}_1 \cdot \hat{\mathbf{e}} = 0 \quad \text{and} \quad \mathbf{x}_2 \cdot \hat{\mathbf{e}} = 0$$

Finally the equivalence of Figures 14-15 and 14-16 may be confirmed by noting how the estimated residuals $(\hat{e}_1, \hat{e}_2, \hat{e}_3) \simeq (.2, -.8, .6)$ appear in each.

14-3 ORTHOGONAL REGRESSORS

If the regressors are orthogonal, as in Figure 14-16, we can then find simple formulas for $\hat{\beta}_1$ and $\hat{\beta}_2$. Referring to Figure 14-17, for notation, we find that $\hat{\mathbf{y}}$ is just the sum of $\hat{\mathbf{y}}_1$ and $\hat{\mathbf{y}}_2$, the individual $\perp$ projections of $\mathbf{y}$ onto $\mathbf{x}_1$ and $\mathbf{x}_2$, respectively. That is,

$$\hat{\mathbf{y}} = \hat{\mathbf{y}}_1 + \hat{\mathbf{y}}_2 \tag{14-43}$$

From (14-24)

$$\hat{\mathbf{y}} = \left(\frac{\mathbf{y} \cdot \mathbf{x}_1}{\|\mathbf{x}_1\|^2}\right)\mathbf{x}_1 + \left(\frac{\mathbf{y} \cdot \mathbf{x}_2}{\|\mathbf{x}_2\|^2}\right)\mathbf{x}_2 \tag{14-44}$$

Comparing (14-44) with (14-36), we conclude that the coefficient

$$\hat{\beta}_1 = \frac{\mathbf{y} \cdot \mathbf{x}_1}{\|\mathbf{x}_1\|^2} \tag{14-45}$$

Now recall from (14-37) that $\mathbf{x}_1$ is just the unit vector. Thus

$$\mathbf{y} \cdot \mathbf{x}_1 = y_1 + y_2 + y_3, \qquad \text{and in general} = \sum_{i=1}^{n} y_i$$

$$\|\mathbf{x}_1\|^2 = 1 + 1 + 1, \qquad \text{and in general} = n$$

Thus (14-45) reduces to

$$\hat{\beta}_1 = \bar{y} \tag{14-46}$$

(2-13) confirmed

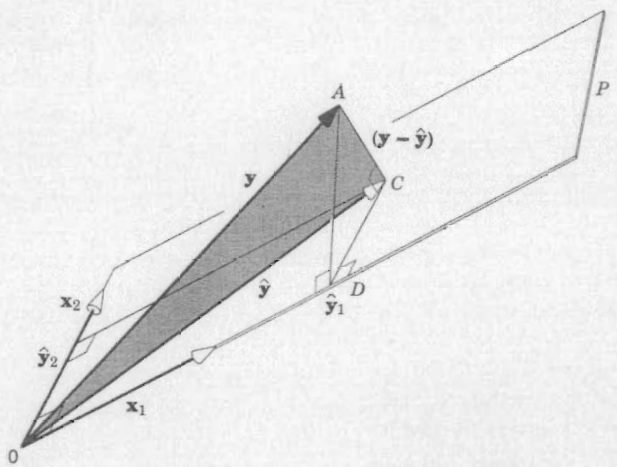


FIGURE 14-17 The projected $\hat{\mathbf{y}}$ is the sum of the individual projections $\hat{\mathbf{y}}_1$ and $\hat{\mathbf{y}}_2$, because $\mathbf{x}_1 \perp \mathbf{x}_2$. Note that $\hat{\mathbf{y}}_1$ (or $\hat{\mathbf{y}}_2$) is that part of $\mathbf{y}$ explained by $\mathbf{x}_1$ (or $\mathbf{x}_2$).

Similarly, we conclude that

$$\hat{\beta}_2 = \frac{\mathbf{x}_2 \cdot \mathbf{y}}{\|\mathbf{x}_2\|^2} \tag{14-47}$$

and noting that $\mathbf{x}_2$ is, in fact, the variable $\mathbf{x}$ appearing in Chapter 2

$$\hat{\beta}_2 = \frac{\sum x_i y_i}{\sum x_i^2} \tag{14-48}$$

(2-16) confirmed

14-4 ANOVA FOR SIMPLE REGRESSION

If we apply the Pythagorean theorem to Figure 14-17, it turns out to be the ANOVA table in disguise. Specifically, consider ΔADC, reproduced again in Figure 14-18 (noting that the vector $DC = \hat{\mathbf{y}}_2 = \hat{\beta}_2 \mathbf{x}_2$). From the theorem of Pythagoras,

$$\|\mathbf{y} - \hat{\mathbf{y}}_1\|^2 = \|\hat{\beta}_2 \mathbf{x}_2\|^2 + \|\mathbf{y} - \hat{\mathbf{y}}\|^2 \tag{14-49}$$

When we recognize that $\mathbf{x}_2$ here is just the x in Chapter 5, this equation begins to look suspiciously like equation (5-20) expressed in vector notation. This can easily be verified; since

$$\hat{\mathbf{y}}_1 = \hat{\beta}_1 \mathbf{x}_1 \tag{14-50}$$

by (14-46)

$$= \bar{y} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

$$= \begin{bmatrix} \bar{y} \\ \bar{y} \\ \bar{y} \end{bmatrix}$$

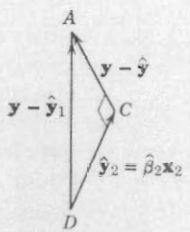


FIGURE 14-18 Pythagorean theorem and ANOVA, for the little triangle of Figure 14-17.

Thus

$$(\mathbf{y} - \hat{\mathbf{y}}_1) = \boxed{\begin{matrix} y_1 - \bar{y} \\ y_2 - \bar{y} \\ y_3 - \bar{y} \end{matrix}} \tag{14-51}$$

Thus fitting $\mathbf{y}$ on the dummy regressor $\mathbf{x}_1$ is just expressing $\mathbf{y}$ in terms of deviations from its mean. The various components of (5-20) are therefore as follows: From (14-51),

$$\|\mathbf{y} - \hat{\mathbf{y}}_1\|^2 = \text{total variation} \tag{14-52}$$

Similarly we can express the right side of (14-49):

$$\begin{aligned} \|\hat{\beta}_2 \mathbf{x}_2\|^2 &= \hat{\beta}_2^2 \|\mathbf{x}_2\|^2 \\ &= \text{explained variation,} \end{aligned} \tag{14-53}$$

and

$$\|\mathbf{y} - \hat{\mathbf{y}}\|^2 = \text{unexplained (residual) variation}$$

Thus, (14-49) becomes

$$\text{Total variation} = \text{explained variation} + \text{unexplained variation} \tag{14-54}$$

(5-20) proved again

More formally, this can be written:

$$\begin{aligned} &\text{total variation after } \mathbf{y} \text{ regressed on } \mathbf{x}_1 \\ &\quad = \text{variation explained by adding regressor } \mathbf{x}_2 \\ &\qquad + \text{variation still left unexplained} \end{aligned} \tag{14-55}$$

14-5 THE STATISTICAL MODEL

In applying statistical tests, we use a mathematical model, that is, a set of assumptions about the parent population of *all* possible outcomes, not just the one outcome we happened to observe. Referring to (14-33), we note that the population consists of all possible observed $\mathbf{y}$ vectors, generated by all possible errors. This is shown schematically in Figure 14-19. If errors are assumed normal, the possible y's we might observe would be spread out in a boundless cloud, thick around $E(\mathbf{y})$, but thinning out in the distance. But to make the geometry manageable, it is necessary to draw an ellipsoid that delimits most of the observed y's, the so-called ellipsoid of concentration. For the independent errors specified in (12-14), the ellipsoid is simply a

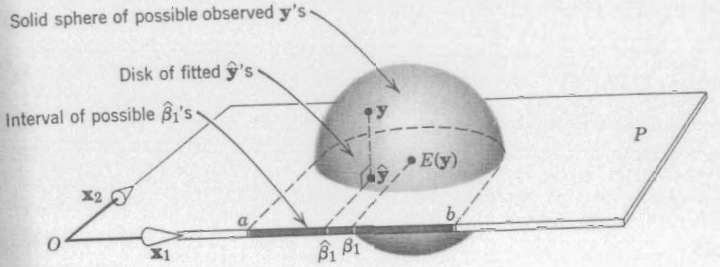


FIGURE 14-19 The distributions of $\mathbf{y}$, $\hat{\mathbf{y}}$ and $\hat{\beta}_1$ (assuming $\mathbf{x}_1 \perp \mathbf{x}_2$).

sphere. This sphere of $\mathbf{y}$ observations (vectors, points) is centered at the mean $E(\mathbf{y})$, which according to (12-13) or (14-33) lies in the plane P generated by $\mathbf{x}_1$ and $\mathbf{x}_2$. Of course, the statistician doesn't know where in this plane $E(\mathbf{y})$ lies; he can only estimate it by observing a sample vector such as the $\mathbf{y}$ vector shown. Note that this observation involves substantial error; that is, $\mathbf{y}$ is quite distant from $E(\mathbf{y})$.

Least squares estimation consists of orthogonally projecting this observed vector $\mathbf{y}$ onto the plane P, the resulting $\hat{\mathbf{y}}$ becoming the estimate of $E(\mathbf{y})$. To derive $\hat{\beta}_1$, the estimate of the true population coefficient β_1, we project $\hat{\mathbf{y}}$ along $\mathbf{x}_2$ onto[4] $\mathbf{x}_1$; similarly, $\hat{\beta}_2$ is derived by projecting $\hat{\mathbf{y}}$ along $\mathbf{x}_1$ onto $\mathbf{x}_2$. We note in this example, that $\hat{\beta}_1$ happened to underestimate β_1, because of the particular error in the observed vector $\mathbf{y}$.

Our more general observations on Figure 14-19 are: $E(\mathbf{y})$ is fixed, while the disk around it, lying in P, represents possible fitted values of $\hat{\mathbf{y}}$, corresponding to the possible observed y's falling in the sphere. *ab* is the projection of this whole disk along $\mathbf{x}_2$ onto $\mathbf{x}_1$. This is the interval of $\hat{\beta}_1$'s around the fixed true β_1. This sampling distribution of $\hat{\beta}_1$ intuitively seems to be unbiased and normal, since the possible observed y's are normally distributed in the sphere centered on $E(\mathbf{y})$; these properties in fact have already been rigorously established in Section 12-4.

---

[4] To be precise, such a projection gives us, for example, $\hat{\beta}_1 \mathbf{x}_1$, rather than the $\hat{\beta}_1$ shown in Figure 14-19. But to keep things simple, we have cheated a little and assumed $\mathbf{x}_1$ is of *unit* length, so that $\hat{\beta}_1 \mathbf{x}_1$ is a vector of length $\hat{\beta}_1$. Thus $\hat{\beta}_1$ may be easily interpreted as the distance along $\mathbf{x}_1$. The target β_1 is similarly interpreted.

But suppose $\mathbf{x}_1$ is not of unit length. (Usually it will not be; indeed in our example, $\mathbf{x}_1$ is the dummy regressor of 1's, so that its length is $\sqrt{n}$.) Then to be precise we must interpret β_1 and β_2 as the coordinates of $E(\mathbf{y})$ on the plane P [i.e., the solution to $E(\mathbf{y}) = \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2$]. Thus β_1 is found by projecting $E(\mathbf{y})$ along $\mathbf{x}_2$ onto $\mathbf{x}_1$, and seeing how many times longer than $\mathbf{x}_1$ this is.

14-6 MULTICOLLINEARITY

Thus far we have only considered two regressors, $\mathbf{x}_1$ and $\mathbf{x}_2$, with $\mathbf{x}_1$ being the unit variable used to estimate the regression intercept $\hat{\beta}_1$, and $\mathbf{x}_2$ being the only bona fide regressor. So long as $\mathbf{x}_2$ is measured as deviations from the mean, $\mathbf{x}_1$ and $\mathbf{x}_2$ must be orthogonal, and Figure 14-19 applies: The projection of $\hat{\mathbf{y}}$ along $\mathbf{x}_2$ onto $\mathbf{x}_1$ is just the $\perp$ projection. Now suppose both $\mathbf{x}_1$ and $\mathbf{x}_2$ are bona fide regressors, in which case they need not be orthogonal. Figure 14-20*a* shows what happens when they are not; the skewed projection of the disk of possible $\hat{\mathbf{y}}$'s along $\mathbf{x}_2$ onto $\mathbf{x}_1$ spreads out the interval of $\hat{\beta}_1$'s.

As the vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ become more nearly collinear, the problem gets worse, as in Figure 14-20*b*; here the interval of $\hat{\beta}_1$'s is dispersed on both sides of the origin. The point estimate $\hat{\beta}_1$ may be positive—but there is now a good chance it may be negative. Moreover, although we see from Figure 14-20*b* that the true β_1 is *not* zero, this is very difficult to establish statistically; usually $H_0(\beta_1 = 0)$ will not be rejected because of the huge standard deviation of $\hat{\beta}_1$.

Although multicollinearity causes a huge spread in $\hat{\beta}_1$, the other attractive properties of $\hat{\beta}_1$ (normality, unbiasedness) are not affected.

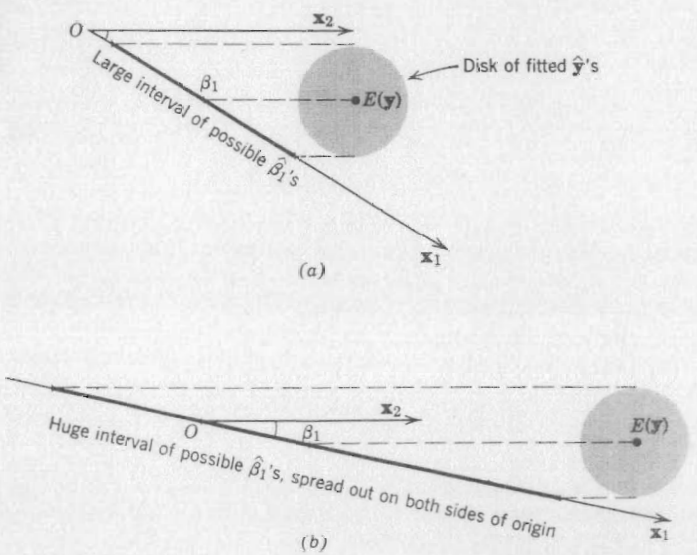


FIGURE 14-20 The plane P from Figure 14-19 is laid flat on the paper, and viewed from above. (*a*) The distributions of $\hat{y}$ and $\hat{\beta}_1$ when $\mathbf{x}_1$ and $\mathbf{x}_2$ are not $\perp$. (*b*) When $\mathbf{x}_1$ and $\mathbf{x}_2$ are nearly collinear.

14-7 CORRELATION AND COS θ

In (14-51) it was established that

$$(\mathbf{y} - \hat{\mathbf{y}}_1) = \begin{bmatrix} y_1 - \bar{y} \\ y_2 - \bar{y} \\ y_3 - \bar{y} \end{bmatrix} \tag{14-56}$$

$$= \text{deviations of } \mathbf{y} \tag{14-57}$$

Throughout the rest of the chapter, we will be interested only in this deviation form for every vector. Also note that since $\mathbf{y}$ as well as $\mathbf{x}$ is expressed in deviation form, according to (2-20) the intercept now disappears, along with the unit regressor used to estimate it. In other words, all x's now become bona fide regressors.

If we consider two such deviation vectors, $\mathbf{x}$ and $\mathbf{y}$, it would be interesting to measure how closely they correspond. The standard geometric measure of the closeness of the direction of two vectors is

$$\cos\theta = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \, \|\mathbf{y}\|} \tag{14-58}$$

(14-32) repeated

Writing the dot product explicitly in terms of components,

$$\cos\theta = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2}\sqrt{\sum y_i^2}} \tag{14-59}$$

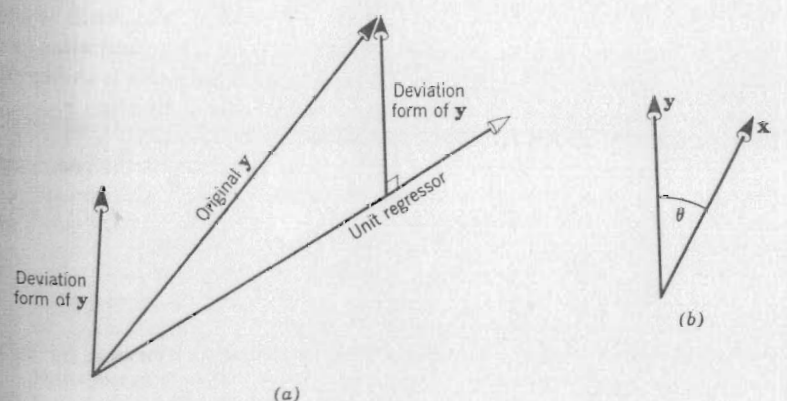


FIGURE 14-21 (*a*) Relation of a vector $\mathbf{y}$ to its derivation form. (*b*) Correlation = $\cos\theta$, where $\mathbf{x}$ and $\mathbf{y}$ are in deviation form.

Since the x_i and y_i values are deviations, we recognize this as the correlation coefficient of (5-6). Thus,

$$\boxed{r = \cos\theta} \tag{14-60}$$

where θ is the angle between the deviation vectors $\mathbf{x}$ and $\mathbf{y}$. In other words, the geometric interpretation of correlation is the closeness of the angle θ. This is shown in Figure 14-21.

Thus, for every geometrical statement about $\cos\theta$, there is an equivalent statistical statement about r. A few such examples are given in Table 14-2. Similarly, the equivalence of the geometry and statistics of regression is given in Table 14-3.

14-8 CORRELATIONS—SIMPLE, MULTIPLE, AND PARTIAL

With all vectors hereafter expressed in deviation form, we see in Figure 14-22 that the correlation r_{YX_2} is just $\cos\theta_2$. To distinguish it from the multiple and partial correlations, r_{YX_2} is sometimes called the simple correlation.

The multiple correlation coefficient R is defined as the simple correlation between $\mathbf{y}$ and $\hat{\mathbf{y}}$—that is, $\cos\lambda$ in Figure 14-22. This provides an index of how well $\mathbf{y}$ can be explained by both regressors[5] $\mathbf{x}_1$ and $\mathbf{x}_2$.

The partial correlation of $\mathbf{y}$ and $\mathbf{x}_2$, designated $r_{YX_2|X_1}$, is the simple correlation of $\mathbf{y}$ and $\mathbf{x}_2$ after the influence of $\mathbf{x}_1$ has been removed from each. The influence of $\mathbf{x}_1$ on $\mathbf{y}$ is the fitted value ($\hat{\mathbf{y}}_1$) when $\mathbf{y}$ is regressed on $\mathbf{x}_1$. When this influence is removed, or subtracted from $\mathbf{y}$, the result is the residual vector ($\mathbf{y} - \hat{\mathbf{y}}_1$). Similarly, $\mathbf{x}_2$ is regressed on $\mathbf{x}_1$ (at A), and when this influence is removed from $\mathbf{x}_2$, the result is the vector AB; this is shifted to CD, forming the angle ϕ_2. Then $\cos\phi_2$ is the partial correlation $r_{YX_2|X_1}$. Similarly, we could show that $\cos\phi_1$ is $r_{YX_1|X_2}$.

In Table 14-4, we extend Table 14-2 to a comparison of the geometry and statistics of multiple and partial correlation.

[5] Specifically,

$$R \equiv r_{y\hat{y}} = \cos\lambda = \frac{\|\hat{\mathbf{y}}\|}{\|\mathbf{y}\|}$$

Bearing in mind that $\mathbf{y}$, like all other variables, is defined as deviations from the mean, $\|\mathbf{y}\|^2$ is its total variation, and we may write

$$R^2 = \frac{\|\hat{\mathbf{y}}\|^2}{\|\mathbf{y}\|^2} = \frac{\text{variation explained by all regressors}}{\text{total variation}} \qquad \text{like (5-51)}$$

TABLE 14-2 Comparison of the Geometrical Interpretation of Cos θ and the Statistical Interpretation of Correlation r (All variables are in deviation form)

Geometry	*Statistics*
$\cos\theta$	r
$-1 \le \cos\theta \le 1$	$-1 \le r \le 1$
$\cos\theta = +1$ iff $\mathbf{x}$ and $\mathbf{y}$ agree perfectly in direction.	$r = +1$ iff x and y move together perfectly
$\cos\theta = -1$ iff $\mathbf{x}$ and $\mathbf{y}$ are in perfectly opposite directions	$r = -1$ iff x and y move together perfectly, but in opposite directions.
$\cos\theta = 0$ iff $\mathbf{x}$ and $\mathbf{y}$ are $\perp$	$r = 0$ iff x and y have no linear relation; iff x and y are uncorrelated

TABLE 14-3 Comparison of the Geometry and Statistics of Regression and ANOVA (All variables are in deviation form).

Geometry	*Statistics*
Squared length of $\mathbf{y}$ Length of $\mathbf{y}$	Variation of y Standard deviation of y (except for the divisor $\sqrt{n-1}$)
$\perp$ projection, yielding $\mathbf{y} - \hat{\mathbf{y}}$ of minimum length	Statistical least squares fit, yielding minimum sum of squared deviations.
Pythagorean theorem: $\|\mathbf{y} - \hat{\mathbf{y}}_1\|^2 = \|\hat{\beta}_2 \mathbf{x}_2\|^2 + \|\mathbf{y} - \hat{\mathbf{y}}\|^2$	ANOVA: Total variation = explained variation + unexplained variation

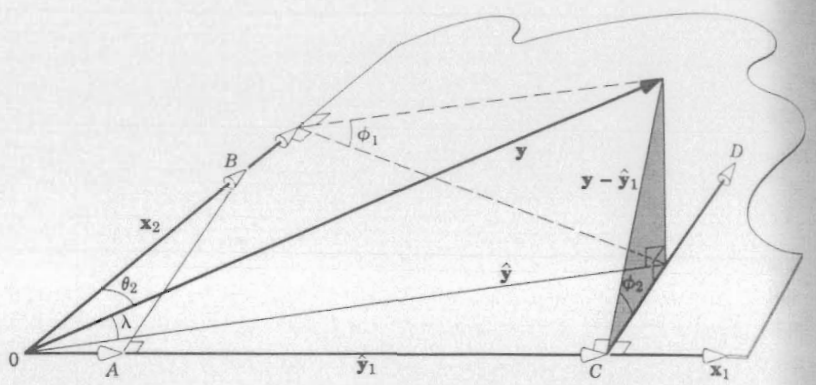


FIGURE 14-22 Multiple correlation coefficient ($R = \cos \lambda$) compared with simple correlation coefficient ($r_{YX_2} = \cos \theta_2$) and partial correlation coefficient ($r_{YX_2|X_1} = \cos \phi_2$).

TABLE 14-4 Comparison of the Geometry and Statistics of Correlations—Simple, Partial, and Multiple (An Extension of Table 14-2. Also refer to Figure 14-22.)

Geometry	*Statistics*
$\cos \theta_2$	Simple correlation r_{YX_2}
$\cos \phi_2$	Partial correlation $r_{YX_2|X_1}$
$\cos \lambda$	Multiple correlation R
$\cos \lambda = 1$ iff $\mathbf{y}$ and $\hat{\mathbf{y}}$ coincide; iff $\mathbf{y}$ lies in the $(\mathbf{x}_1, \mathbf{x}_2)$ subspace.	$R = 1$ iff x_1 and x_2 explain y exactly, leaving no residual.
$\cos \lambda = 0$ iff $\mathbf{y}$ orthogonal to the $(\mathbf{x}_1, \mathbf{x}_2)$ subspace.	$R = 0$ iff $\hat{y} = 0x_1 + 0x_2 = 0$ i.e., x_1 and x_2 do not explain y at all.
$\lvert\cos \theta_2\rvert \leq \lvert\cos \lambda\rvert$	$r_{YX_2} \leq R$
$\lvert\cos \phi_2\rvert \leq \lvert\cos \lambda\rvert$	$r_{YX_2|X_1} \leq R$

14-9 TESTS WHEN THERE ARE k REGRESSORS

(a) ANOVA for Last g Regressors

With a little imagination, in Figure 14-22 we can think of replacing $\mathbf{x}_1$ with a set of regressors $\mathbf{x}_1, \mathbf{x}_2 \cdots \mathbf{x}_{k-g}$, and $\mathbf{x}_2$ with the remaining set of g regressors $\mathbf{x}_{k-g+1} \cdots \mathbf{x}_k$. We now wish to simultaneously test this latter set of g regressors. The only change this causes in the theory of the previous section is that the lines generated by $\mathbf{x}_1$ and $\mathbf{x}_2$ are replaced by subspaces. These subspaces are impossible to draw, so they are still represented in Figure 14-23 by lines. How can we test the null hypothesis that the last g

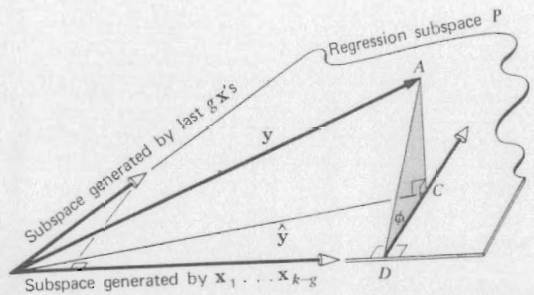


FIGURE 14-23 Multiple regression with k regressors, with the last g being tested.

regressors are all irrelevant, that is, the last g regression coefficients are zero? One method is to apply the Pythagorean theorem to the triangle ADC to obtain the ANOVA identity

$$\|AD\|^2 = \|CD\|^2 + \|AC\|^2$$

$$\begin{pmatrix}\text{unexplained variation after } y \text{ is regressed on } x_1 \cdots x_{k-g} \\ = \text{variation explained by introducing } g \text{ more regressors} \\ + \text{unexplained variation that still remains}\end{pmatrix} \tag{14-61}$$

The two variations on the right side of (14-61) are statistically independent χ^2 variables, with g and $(n - 1 - k)$ d.f. respectively.[6] When we divide by

[6] For proof, see for example, H. Scheffé, 1959. Of course, we are assuming that the true coefficients of the last g regressors are all zero, since this is the null hypothesis being tested.

Our convention in this chapter is to let k represent the number of regressors *excluding* the constant regressor of 1's. This disagrees with Chapter 12 (where k *included* the constant regressor).

these d.f., the variations become variances, and their ratio is

$$F = \frac{\text{additional variance explained by introducing the } g \text{ regressors}}{\text{residual variance}} \tag{14-62}$$

which follows the F distribution and can be used to test the statistical significance of the last g regressors [(14-62) confirms (5-57) noting, that in that earlier chapter, the number of regressors being tested was called r, rather than g.]

When there is just one regressor to be tested, we set $g = 1$ in (14-62), and so obtain:

$$F = \frac{\text{additional variance explained by the last regressor } \mathbf{x}_k}{\text{unexplained variance}} \tag{14-63}$$

(b) Partial Correlation

Alternatively we could test the last regressor by examining the partial correlation $r_{Yx_k|x_1x_2,\ldots,x_{k-1}}$, which we shall abbreviate to r in this discussion. In Figure 14-24, when $g = 1$, the partial correlation r is cos ϕ (just as in Figure 14-22). Instead of asking [as in (14-63)] whether the squared length of CD in Figure 14-24 is large enough (relative to AC) to reject H_0, why not ask the equivalent question: Is angle ϕ close enough, that is, is r large enough? To

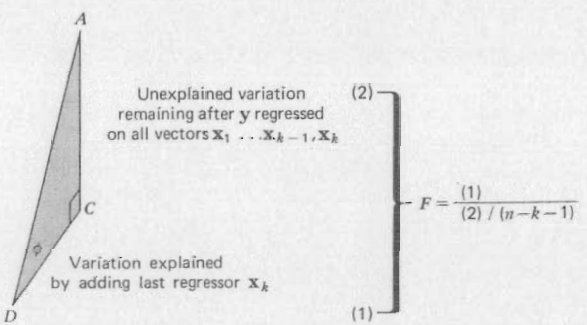


FIGURE 14-24 ANOVA, showing the vectors in triangle ADC in Figure 14-23 when $g = 1$.

do this we simply express F in (14-63) in terms of r, rather than squared lengths, as follows. From (14-63), or Figure 14-24, we may write

$$F = \frac{\|CD\|^2}{\|AC\|^2/(n-k-1)} \tag{14-64}$$

Thus

$$\frac{F}{n-k-1} = \frac{\|CD\|^2}{\|AC\|^2}$$

Dividing the numerator and denominator on the RHS by $\|AD\|^2$ and noting that $r^2 = (\cos\phi)^2 = \|CD\|^2/\|AD\|^2$,

$$\frac{F}{n-k-1} = \frac{\|CD\|^2/\|AD\|^2}{\|AC\|^2/\|AD\|^2} = \frac{r^2}{\frac{\|AD\|^2 - \|CD\|^2}{\|AD\|^2}} = \frac{r^2}{1-r^2} \tag{14-65}$$

Therefore,

$$F = \frac{r^2(n-k-1)}{1-r^2} \tag{14-66}$$

As expected, the closer is ϕ, the greater is r, and the greater is F. Thus (14-66) and (14-63) are seen to be alternative ways of testing the null hypothesis that $\mathbf{y}$ is not related to the last regressor $\mathbf{x}_k$.

Finally, if we take the square root of (14-66), according to (13-44) we then have

$$t = \sqrt{F} = \frac{r\sqrt{n-k-1}}{\sqrt{1-r^2}} \tag{14-67}$$

This is the same t that appeared in our tests in Chapters 2 and 3. Note that (14-67) nicely shows the relation of t, F, and the partial correlation r.

PROBLEMS

14-9 True or false? If false, correct it:

(a) In studying the relation of three variables, if $\mathbf{x}$ and $\mathbf{y}$ are each uncorrelated with $\mathbf{z}$, then $r_{XY|Z} = r_{XY}$, that is, the partial and simple correlations coincide.

(b) If the multiple correlation of $\mathbf{y}$ with $\mathbf{x}_1 \cdots \mathbf{x}_k$ is zero, then the partial correlation of $\mathbf{y}$ with $\mathbf{x}_k$ is also zero, as is the simple correlation of $\mathbf{y}$ with $\mathbf{x}_k$.

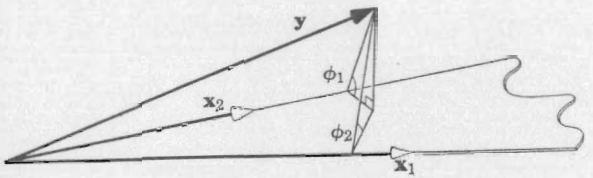
(c) The partial correlation of $\mathbf{y}$ and $\mathbf{x}_k$ is the simple correlation of $\mathbf{y}$ and $\hat{\mathbf{y}}$ after the influence of $\mathbf{x}_k$ has been removed from each.

14-10 Suppose 50 observations were used to regress $\mathbf{y}$ on 3 regressors $\mathbf{x}_1$, $\mathbf{x}_2$, and $\mathbf{x}_3$. If the partial correlation of $\mathbf{y}$ and $\mathbf{x}_3$ was .28, is this statistically significant (discernible) at the 5% level (2-sided)?

14-11 Refer to Figure 14-23, with $g = 1$. Suppose that instead of observing $\mathbf{x}_k$, an economist observed a variable $\mathbf{z}_k$ that was more closely correlated to the previous variables $\mathbf{x}_1 \cdots \mathbf{x}_{k-1}$, yet still generated the same regression subspace P. Suppose further that there is no doubt that the previous variables $\mathbf{x}_1 \cdots \mathbf{x}_{k-1}$ belong in the model. The only question is whether the last variable ($\mathbf{x}_k$ or $\mathbf{z}_k$) belongs.
True or false? If false, correct it, giving a brief reason.

(a) The multiple correlation of $\mathbf{y}$ with $\mathbf{x}_1 \cdots \mathbf{x}_{k-1}, \mathbf{z}_k$ would equal the multiple correlation of $\mathbf{y}$ with $\mathbf{x}_1 \cdots \mathbf{x}_{k-1}, \mathbf{x}_k$.
(b) The partial correlation of $\mathbf{y}$ with $\mathbf{z}_k$ would equal the partial correlation of $\mathbf{y}$ with $\mathbf{x}_k$.

14-12 Referring to Figure 14-22, suppose the regressors $\mathbf{x}_1$ and $\mathbf{x}_2$ were more correlated, so that the picture was like this:



Suppose two economists, using the same sample of 24 observations, made the following two different analyses of the model:

$$\mathbf{y} = \beta_0 + \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + \text{error} \tag{14-68}$$

1. The first economist makes a test of the null hypothesis $\beta_2 = 0$. She calculates the t value, which turns out to be only 1.1 (reflecting a small partial correlation, i.e., a wide angle ϕ_2). At the 5% level, the null hypothesis is not rejected, and so she recommends using the model

$$\mathbf{y} = \beta_0 + \beta_1 \mathbf{x}_1 + \text{error} \tag{14-69}$$

The estimated coefficients turn out to be

$$\mathbf{y} = 1.70 + 1.3\mathbf{x}_1 + \text{residual} \tag{14-70}$$

with the coefficient 1.3 being statistically significant (discernible).

2. In the model (14-68) the second economist makes a test of the null hypothesis $\beta_1 = 0$. In this case, the t value turns out to be only 1.4 (reflecting a small partial correlation, i.e., a wide angle ϕ_1). At the 5% level, the null hypothesis is not rejected, and so he recommends the model

$$\mathbf{y} = \beta_0 + \beta_2 \mathbf{x}_2 + \text{error} \tag{14-71}$$

The estimated coefficients turn out to be

$$\mathbf{y} = 1.70 + 4.7\mathbf{x}_2 + \text{error} \tag{14-72}$$

with the coefficient 4.7 being statistically significant.

(a) Is it possible that the two economists could validly arrive at such different conclusions as (14-70) and (14-72), or can the discrepancy be explained away as a computational error?
(b) Which economist has the better model? Or is there an even better model than these two? If you cannot answer this categorically, list the possible criteria for choosing between models.

*14-10 FORWARD STEPWISE REGRESSION[7]

In this section we will discuss in more detail the stepwise procedure introduced in Section 5-3. Consider a forward stepwise regression[8] of $\mathbf{y}$ on $\mathbf{x}_1$ and $\mathbf{x}_2$, where we start from scratch, adding one regressor at a time; initially suppose that we have specified *a priori* that $\mathbf{x}_1$ will be tested first, and $\mathbf{x}_2$ second.

Before examining this procedure, in Figure 14-25 we show the result of applying the standard multiple regression of $\mathbf{y}$ on $\mathbf{x}_1$ and $\mathbf{x}_2$. (Although our remarks are illustrated for two regressors, they are easily generalized to k regressors.) Although the true coefficients β_1 and β_2 are not shown, they must be kept in mind as the targets; we suppose that β_1 and β_2 are both nonzero, so that $\mathbf{y}$ depends on both $\mathbf{x}_1$ and $\mathbf{x}_2$. The near collinearity between $\mathbf{x}_1$ and $\mathbf{x}_2$ results in large standard errors for the estimators $\hat{\beta}_1$ and $\hat{\beta}_2$; but at least $\hat{\beta}_1$ and $\hat{\beta}_2$ are unbiased, and the residual $\mathbf{y} - \hat{\mathbf{y}}$ is minimized.

[7] This section is starred because of its difficulty. Also, it includes a fallacy that fortunately appears in the literature less often today than in the past.
[8] In practice, the forward procedure is typically used by computer programs in the interest of cost, since alternative stepwise procedures involve fitting regressions of larger dimension. For more detail on alternatives, see Draper and Smith, 1966.

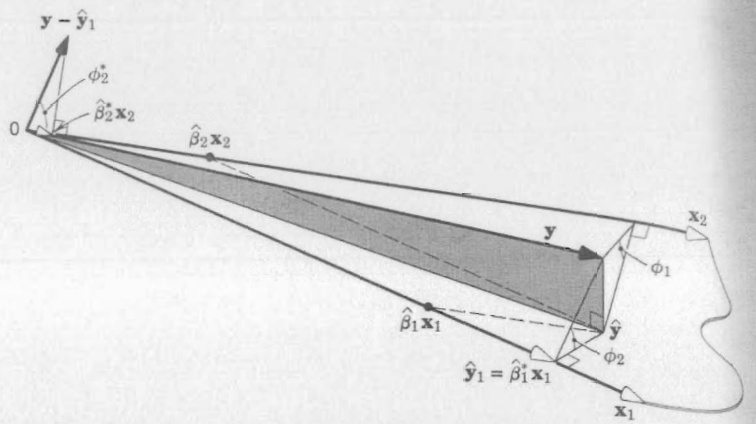


FIGURE 14-25 Problems in stepwise regression.

In stepwise regression, the first step is to regress $\mathbf{y}$ on $\mathbf{x}_1$ alone, yielding the fit $\hat{\mathbf{y}}_1 = \hat{\beta}_1^* \mathbf{x}_1$. Clearly, so long as the regressors $\mathbf{x}_1$ and $\mathbf{x}_2$ are not orthogonal, $\hat{\beta}_1^*$ will be a biased estimate of β_1. (In the case shown in Figure 14-25 it is larger than the unbiased $\hat{\beta}_1$.)

The second step is to consider the second regressor $\mathbf{x}_2$. If we are not careful we may follow the natural temptation, now that $\mathbf{x}_1$ has been "netted out," to regress the rest of $\mathbf{y}$ still left unexplained, that is, the residual $\mathbf{y} - \hat{\mathbf{y}}_1$, on $\mathbf{x}_2$. In Figure 14-25 we shift this residual vector to the origin, and show the resulting estimate $\hat{\beta}_2^*$. Again, as long as $\mathbf{x}_1$ and $\mathbf{x}_2$ are not orthogonal, $\hat{\beta}_2^*$ will be a biased estimator of β_2 (in the case shown it is smaller than the unbiased $\hat{\beta}_2$). Furthermore, a test of significance on $\hat{\beta}_2^*$ would be very weak, being based on $\cos \phi_2^*$, which is nearly zero.[9]

(And there is even more damage: The final residual $\mathbf{y} - \hat{\beta}_1^* \mathbf{x}_1 - \hat{\beta}_2^* \mathbf{x}_2$ will not be as small as in the standard multiple regression.)[10]

We have earlier concluded that with multicollinear regressors, it is difficult in any case to establish statistical significance; here we are using a biased method that makes it even more difficult to establish the significance of regressors that are tested last.

The correct unbiased test of the relationship between $\mathbf{y}$ and $\mathbf{x}_2$ involves a test of $r_{YX_2|X_1}$, or $\cos \phi_2$. But this is the angle between the residual $\mathbf{y} - \hat{\mathbf{y}}_1$ and the vector $\hat{\mathbf{y}} - \hat{\mathbf{y}}_1$, not $\mathbf{x}_2$. If this test leads to the decision to include $\mathbf{x}_2$, then we obtain the correct $\hat{\beta}$ coefficients by a full multiple regression of $\mathbf{y}$ on $\mathbf{x}_1$ and $\mathbf{x}_2$. (In larger models, after each such test, a multiple regression is run on all the $\mathbf{x}$'s that have so far been included.)

In summary, we consider several major problems involved in stepwise regression when regressors are not orthogonal. *Even if the correct procedure is used* (i.e., even if the relationship between $\mathbf{y}$ and $\mathbf{x}_2$ is tested by examining $\cos \phi_2$, not $\cos \phi_2^*$) there are two problems:

1. In the initial step of testing whether $\mathbf{x}_1$ should be included, $\hat{\beta}_1^*$ is a biased estimate.
2. Suppose in the initial step we have tested $\mathbf{x}_1$ and decided to include it;[11] then we test the significance of $\mathbf{x}_2$ by examining $\cos \phi_2$ in Figure 14-25. Although the multiple correlation of $\mathbf{y}$ on $\mathbf{x}_1$ and $\mathbf{x}_2$ is high, the partial correlation ($\cos \phi_2$) may be statistically insignificant because of multicollinearity. Then the final regression fit would include only $\mathbf{x}_1$. On the other hand, consider what would happen if we took up the regressors in the other order. In the first step we would include[12] $\mathbf{x}_2$; then, in testing the significance of $\mathbf{x}_1$, we might find the partial correlation ($\cos \phi_1$) statistically insignificant. In this case, the final regression equation would include only $\mathbf{x}_2$. Thus, the variables appearing in our final model may depend on the order in which they are brought into consideration. For this reason, in the absence of prior grounds to justify a prescribed ordering of the variables, a computer program should be selected that will automatically pick up first the regressor that is most highly correlated with $\mathbf{y}$.

Now suppose an incorrect procedure is used (i.e., suppose we have already decided to include $\mathbf{x}_1$, and in then testing $\mathbf{x}_2$ we erroneously examine $\cos \phi_2^*$, rather than $\cos \phi_2$; or equivalently, suppose we erroneously regress $\mathbf{y} - \hat{\mathbf{y}}_1$ on $\mathbf{x}_2$). Then there are two additional problems:

1. $\hat{\beta}_2^*$ will be a severely biased estimate of the effect of the second regressor $\mathbf{x}_2$, and any test based on it will find it very difficult to establish its statistical significance.
2. The test will also be weak because of the excessively large residual.

[9] *Reason.* Since the residual $\mathbf{y} - \hat{\mathbf{y}}_1$ is perpendicular to $\mathbf{x}_1$, and since $\mathbf{x}_1$ and $\mathbf{x}_2$ are nearly parallel, $\mathbf{y} - \hat{\mathbf{y}}_1$ will be nearly perpendicular to $\mathbf{x}_2$, that is, ϕ_2^* will be nearly 90°.

[10] The standard least squares regression coefficients $\hat{\beta}_1$ and $\hat{\beta}_2$ were chosen, by definition, to make $\mathbf{y} - \hat{\beta}_1 \mathbf{x}_1 - \hat{\beta}_2 \mathbf{x}_2$ a minimum, and thus smaller than $\mathbf{y} - \hat{\beta}_1^* \mathbf{x}_1 - \hat{\beta}_2^* \mathbf{x}_2$. (Unless, of course, $\mathbf{x}_1$ and $\mathbf{x}_2$ are orthogonal, in which case the two residuals coincide.)

[11] In the first step of regressing $\mathbf{y}$ on $\mathbf{x}_1$ only, the test of $\mathbf{x}_1$ could be viewed either as examining the biased $\hat{\beta}_1^*$, or alternatively examining $\cos \theta_1$, where θ_1 is the angle between $\mathbf{y}$ and $\mathbf{x}_1$ in Figure 14-25. This close angle leads us to conclude that $\mathbf{x}_1$ is a statistically significant regressor.

[12] Note that this test of $\mathbf{x}_2$ would be statistically significant for the same reason as our initial test of $\mathbf{x}_1$ in footnote 11 above.

In conclusion, if there are clear prior guidelines indicating that a few specific regressors are appropriate, then they should all be used right away in a full multiple regression, rather than tested one at a time with any sort of stepwise approach. If there are no such prior guidelines, but the number of regressors must be kept small to provide a more manageable model, then a stepwise technique may be reasonable. But it must be recognized that this procedure tends to discriminate against regressors tested last, even if correctly applied; and if incorrectly applied, it discriminates even more.

*PROBLEMS

14-13 As in Figure 14-25 suppose $\mathbf{x}_1$ and $\mathbf{x}_2$ are highly collinear. In addition, suppose that the true (as opposed to observed) $\mathbf{y}$ has a perfect positive correlation with $\mathbf{x}_2$, that is, the true model is: $\mathbf{y} = 0\mathbf{x}_1 + \beta_2 \mathbf{x}_2$. Also, suppose that we are lucky enough in our sample to observe $\mathbf{y}$ being perfectly correlated with $\mathbf{x}_2$, that is, $\mathbf{y}$ is perfectly explained by this single regressor. Hence, the fitted standard multiple regression of $\mathbf{y}$ on $\mathbf{x}_1$ and $\mathbf{x}_2$ is

$$\mathbf{y} = 0\mathbf{x}_1 + \hat{\beta}_2 \mathbf{x}_2 \tag{14-77}$$

(a) Show this geometrically.
(b) Is $\hat{\beta}_2$ unbiased?
(c) What is the vector of residuals?

Now, to show how badly a stepwise analysis can go wrong if applied carelessly, suppose an erroneous stepwise procedure is undertaken and in the first step $\mathbf{y}$ is regressed on $\mathbf{x}_1$ as follows:

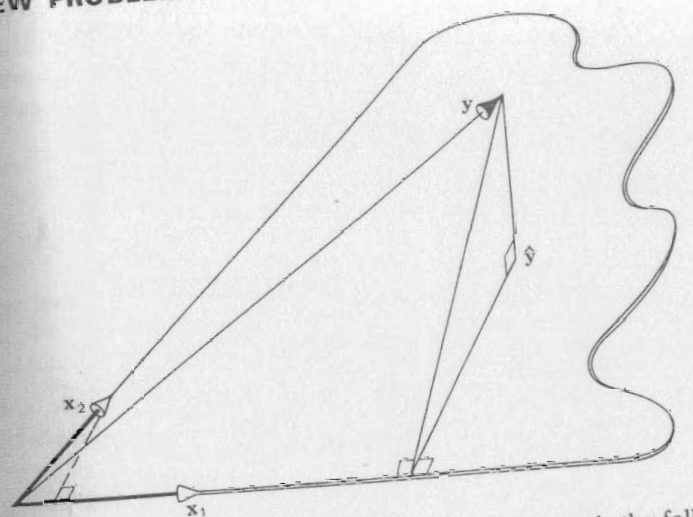
$$\mathbf{y}_1 = \hat{\beta}_1^* \mathbf{x}_1$$

(d) Is $\hat{\beta}_1^*$ biased? Is it possible for us to conclude that $\hat{\beta}_1^*$ is significantly different from zero?
(e) Suppose after erroneously including $\mathbf{x}_1$ as a regressor, we further err by regressing the residual vector $(\mathbf{y} - \hat{\mathbf{y}}_1)$ on $\mathbf{x}_2$. Is the resulting estimate of β_2 biased? Is it possible that we will therefore, reject $\mathbf{x}_2$ as a regressor (even though $\mathbf{x}_2$ in fact perfectly explains $\mathbf{y}$)?
(f) How does the resulting fitted equation compare with (14-77)?
(g) How does the final residual vector resulting from this stepwise procedure compare with the residual vector in (c)?
(h) Could this disastrous result have occurred if, in the first step, we had used a computer routine that introduced the regressor most highly correlated with $\mathbf{y}$ first?

14-14 Suppose an economist is examining the effect of socioeconomic background (x_1) and education (x_2) on income (Y). After explaining Y by regressing it on x_1 only, he states that no further significant explanation of Y can be established using x_2. He concludes that income is related to socioeconomic background, but not education. You are to discuss his paper. What has gone wrong? Illustrate using vector geometry; also point out under what special circumstances this sort of two-step approach could be justified. Are these circumstances present in this case?

REVIEW PROBLEMS

14-15



(a) Make a copy of the diagram above, and mark the following items (draw in any additional vectors you require).
(i) $\hat{\beta}_1$ and $\hat{\beta}_2$, the coefficients of the multiple regression of $\mathbf{y}$ on $\mathbf{x}_1$ and $\mathbf{x}_2$.
(ii) The residual vector.
(iii) The angles ϕ_1 and ϕ_2, where $\cos \phi_1 = r_{Y\tilde{x}_1|x_2}$ and $\cos \phi_2 = r_{Yx_2|x_1}$.
(iv) The angle λ, where $\cos \lambda = R$.
(v) $\hat{\beta}$, the coefficient of simple regression of $\mathbf{y}$ on $\mathbf{x}_1$. Is $\hat{\beta}$ the same as the multiple regression coefficient $\hat{\beta}_1$? Under what conditions would $\hat{\beta} = \hat{\beta}_1$?
(vi) θ_1, where $\cos \theta_1 = r_{Yx_1}$.

(b) Using this diagram, what is the F test for $\beta_2 = 0$? What is the F test for $\beta_1 = 0$?

14-16 An economist fitted the simple regression

$$\mathbf{y} = a + b\mathbf{x} + \hat{\mathbf{e}} \text{ (residual)}$$

The next day she decided that she should include another explanatory variable $\mathbf{z}$, for the same data; she therefore fitted the multiple regression

$$\mathbf{y} = a' + b'\mathbf{x} + c'\mathbf{z} + \hat{\mathbf{e}}'$$

The letters a, b, a', etc., refer to the fitted (OLS) values, not the true (population) values. Under what circumstances will the following be true? (Answer very carefully; for example: never, or always, or usually, except when . . . , or rarely; only when . . .). In each case, explain why.

(a) $b' = b$

(b) $\sum_{i=1}^{n} (\hat{e}_i')^2 \leq \sum_{i=1}^{n} (\hat{e}_i)^2$

(c) b' is statistically significant (discernible) at the 5% level, yet b is not.

(d) b is statistically significant at the 5% level, yet b' is not.

14-17 (a) In a sample of 30 observations, suppose the multiple correlation of $\mathbf{y}$ with 5 variables is .72. Including a sixth variable x_6 increases the multiple correlation to .75.
Test the hypothesis $\beta_6 = 0$ ($\mathbf{x}_6$ is irrelevant) at the 5% level (2-sided).

(b) If $R = .75$ does not achieve statistical significance (discernability), what value of R would?

(c) What is the partial correlation of $\mathbf{y}$ with $\mathbf{x}_6$?

15
OTHER REGRESSION TOPICS

15-1 SPECIFICATION ERROR

How much do estimates err when the model is misspecified? For example, a relationship that is actually nonlinear may be specified to be linear; the problem this raises has already been briefly discussed in Section 2-10. The present section is devoted to another kind of misspecification, the kind that occurs if too many or too few regressors are included in the model (a model is "too long" or "too short").

(a) Too Many Regressors: A Model that is Too Long

The problem of too many regressors may be formulated as follows: Suppose some regressors really don't belong in the model (12-9), that is, their coefficients should have been set equal to zero, *a priori*. If we inadvertently keep these extraneous regressors in the model, that is, estimate their coefficients, will this raise any problems?

In terms of bias, the answer is: no problems. According to (12-31), regression coefficients are unbiased. So the extraneous regressors will have coefficients whose expectation is zero, while the relevant regressors will have coefficients whose expectation is also correct. Thus, no bias is introduced by too many regressors.

However, making the model too long does increase the variance of the estimators. Obviously this is true for the extraneous regressors: Because these β_i are zero, it would be better to specify them as zero exactly, rather than use estimates $\hat{\beta}_i$ that fluctuate around zero. Moreover, the inclusion of irrelevant regressors will also increase the variance for the *relevant* regressors—to the degree that the extraneous regressors produced multicollinearity problems [although we do not prove this, we illustrate it geometrically in part (c)].