CHAPTER 20

The Geometry of Least Squares

Comment: Philosophies on teaching the geometry of least squares vary. An early introduction is certainly possible, as demonstrated by the successful presentation by Box, Hunter, and Hunter (1978, e.g., pp. 179, 197–201) for specific types of designs. For general regression situations, more general explanations are required. Our view is that, while regression can be taught perfectly well without the geometry, understanding of the geometry provides much better understanding of, for example, the difficulties associated with singular or nearly singular regressions, and of the difficulties associated with interpretations of the R^2 statistic. For an advanced understanding of least squares, knowledge of the geometry is essential.

20.1. THE BASIC GEOMETRY

We want to fit, by least squares, the model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{20.1.1}$$

where $\mathbf{Y}$ and $\boldsymbol{\epsilon}$ are both $n \times 1$, $\mathbf{X}$ is $n \times p$, and $\boldsymbol{\beta}$ is $p \times 1$. Consider a Euclidean (i.e., "ordinary") space of n dimensions, call it S. The n numbers $Y_1, Y_2, \ldots, Y_n$ within $\mathbf{Y}$ define a point Y in this space. They also define a vector (a line with length and direction) usually represented by the line joining the origin O, $(0, 0, 0, \ldots, 0)$, to the point Y. (In fact, any parallel line of the same length can also be regarded as the vector $\mathbf{Y}$, but most of the time we think about the vector from O.) The columns of $\mathbf{X}$ also define vectors in the n-dimensional space. Let us assume for the present that all p of the $\mathbf{X}$ columns are linearly independent, that is, none of them can be represented as a linear combination of any of the others. This implies that $\mathbf{X}'\mathbf{X}$ is nonsingular. Then the p columns of $\mathbf{X}$ define a subspace (we call it the estimation space) of S of p ($<n$) dimensions. Consider

$$\mathbf{X}\boldsymbol{\beta} = [\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_{p-1}] \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} = \beta_0\mathbf{x}_0 + \beta_1\mathbf{x}_1 + \cdots + \beta_{p-1}\mathbf{x}_{p-1}, \tag{20.1.2}$$

where each $\mathbf{x}_i$ is an $n \times 1$ vector and the β_i are scalars. (Usually $\mathbf{x}_0 = \mathbf{1}$.) This defines a vector formed as a linear combination of the $\mathbf{x}_i$, and so $\mathbf{X}\boldsymbol{\beta}$ is a vector in the estimation space. Precisely where it is depends on the values chosen for the β_i. We can now draw

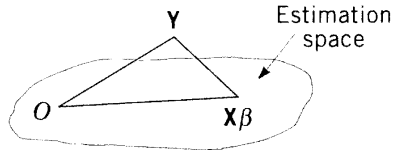

Figure 20.1. A general point $\mathbf{X}\boldsymbol{\beta}$ in the estimation space.

Figure 20.1. We see that the points marked O, $\mathbf{Y}$, $\mathbf{X}\boldsymbol{\beta}$ form a triangle in n-space. In general, the angles will be determined by the values of the β_i, given $\mathbf{Y}$ and $\mathbf{X}$. The three sides of the triangle are the vectors $\mathbf{Y}$, $\mathbf{X}\boldsymbol{\beta}$, and $\boldsymbol{\epsilon} = \mathbf{Y} - \mathbf{X}\boldsymbol{\beta}$. So the model (20.1.1) simply says that $\mathbf{Y}$ can be split up into two vectors, one of which, $\mathbf{X}\boldsymbol{\beta}$, is completely in the estimation space and one of which, $(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$, is partially not, in general. When we estimate $\boldsymbol{\beta}$ by least squares, the solution $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ is the one that minimizes the sum of squares function

$$S(\boldsymbol{\beta}) = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}). \tag{20.1.3}$$

At this point, we need to know the fact that, if $\mathbf{z}$ is any vector, $\mathbf{z}'\mathbf{z}$ is the squared length of that vector. So the sum of squares function $S(\boldsymbol{\beta})$ is just the squared length of the vector joining $\mathbf{Y}$ and $\mathbf{X}\boldsymbol{\beta}$ in Figure 20.1. When is this a minimum? It is when $\boldsymbol{\beta}$ is set equal to $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. It follows that the vector $\mathbf{Xb} = \hat{\mathbf{Y}}$ joins the origin O to the foot of the perpendicular from $\mathbf{Y}$ to the estimation space. See Figure 20.2. Compare it with Figure 20.1. When $\boldsymbol{\beta}$ takes the special value $\mathbf{b}$, the least squares value, the triangle of Figure 20.1 becomes a right-angled one as in Figure 20.2. If that is true, the vectors $\hat{\mathbf{Y}} = \mathbf{Xb}$ and $\mathbf{e} = \mathbf{Y} - \mathbf{Xb}$ should be orthogonal. Two vectors $\mathbf{p}$ and $\mathbf{q}$ are orthogonal if $\mathbf{p}'\mathbf{q} = 0 = \mathbf{q}'\mathbf{p}$. Here,

$$\begin{aligned} 0 = (\mathbf{Xb})'\mathbf{e} &= \mathbf{b}'\mathbf{X}'\mathbf{e} \\ &= \mathbf{b}'\mathbf{X}'(\mathbf{Y} - \mathbf{Xb}) \\ &= \mathbf{b}'(\mathbf{X}'\mathbf{Y} - \mathbf{X}'\mathbf{Xb}). \end{aligned} \tag{20.1.4}$$

Thus the orthogonality requirement in (20.1.4) implies either that $\mathbf{b} = \mathbf{0}$, in which case the vector $\mathbf{Y}$ is orthogonal to the estimation space and $\hat{\mathbf{Y}} = \mathbf{0}$, or that the normal equations $\mathbf{X}'\mathbf{Xb} = \mathbf{X}'\mathbf{y}$ hold. The space in which $\mathbf{e} = \mathbf{Y} - \mathbf{Xb}$ lies is called the error space, and it has $(n - p)$ dimensions. The estimation space of p dimensions and the error space together constitute S. So the least squares fitting of the regression model splits the space S up into two orthogonal spaces; every vector in the estimation space is orthogonal to every vector in the error space.

Exercise 1. Let $\mathbf{Y} = (3.1, 2.3, 5.4)'$, $\mathbf{x}_0 = (1, 1, 1)'$ and $\mathbf{x}_1 = (2, 1, 3)'$, so that we are

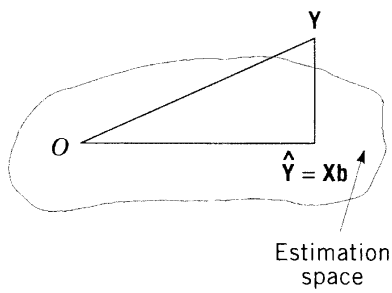

Figure 20.2. The right-angled triangle of vectors $\mathbf{Y}$, $\mathbf{Xb}$, and $\mathbf{e} = \mathbf{Y} - \mathbf{Xb}$.

going to fit the straight line $Y = \beta_0 + \beta_1 x_1 + \epsilon$ via least squares. First, do the regression fit without worrying about the geometry. Show that $\hat{Y} = 0.50x_0 + 1.55x_1$. Then show that $\hat{\mathbf{Y}} = \mathbf{Xb} = (3.60, 2.05, 5.15)'$ and $\mathbf{e} = (-0.50, 0.25, 0.25)'$; confirm that these two vectors are orthogonal and furthermore that $\mathbf{e}$ is orthogonal to any vector of the form $\mathbf{X}\boldsymbol{\beta} = \beta_0\mathbf{x}_0 + \beta_1\mathbf{x}_1$, for all $\boldsymbol{\beta}$, because $\mathbf{e}$ is orthogonal to both (in general, all) the individual columns of $\mathbf{X}$. Write the coordinates of the various points on a diagram like Figure 20.2.

20.2. PYTHAGORAS AND ANALYSIS OF VARIANCE

Every analysis of variance table arising from a regression is an application of Pythagoras's Theorem about the "square of the hypotenuse of a right-angled triangle equals the sum of squares of the other two sides." Usually, repeated applications of Pythagoras's result are needed. Look again at Figure 20.2. It implies that

$$\mathbf{Y}'\mathbf{Y} = \hat{\mathbf{Y}}'\hat{\mathbf{Y}} + (\mathbf{Y} - \mathbf{Xb})'(\mathbf{Y} - \mathbf{Xb}) \tag{20.2.1}$$

or

Total sum of squares = Sum of squares due to regression + Residual sum of squares.

The corresponding degrees of freedom equation

$$n = p + (n - p) \tag{20.2.2}$$

corresponds to the dimensional split of S into two orthogonal spaces of dimensions p and $(n - p)$, respectively. The orthogonality of $\hat{\mathbf{Y}}$ and $\mathbf{e}$ is essential for a clear split-up of the total sum of squares. Where a split-up is required of a particular sum of squares that does not have a natural orthogonal split-up, orthogonality must be introduced to achieve it.

Exercise 2. Show that for the data in Exercise 1, the two equations (20.2.1) and (20.2.2) correspond to

$$44.06 = 43.685 + 0.375,$$
$$3 = 2 + 1.$$

Further Split-up of a Regression Sum of Squares

If the $\mathbf{X}_i$ vectors that define the estimation space happen to be all orthogonal to one another, the regression sum of squares can immediately be split down into orthogonal pieces. Suppose, for example, that $\mathbf{X} = (\mathbf{1}, \mathbf{x})$ and that $\mathbf{1}'\mathbf{x} = 0 = \mathbf{x}'\mathbf{1}$. The model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ is of a straight line, Y versus x. Figure 20.3 shows the estimation space defined by the two *orthogonal* vectors $\mathbf{1}$ and $\mathbf{x}$.

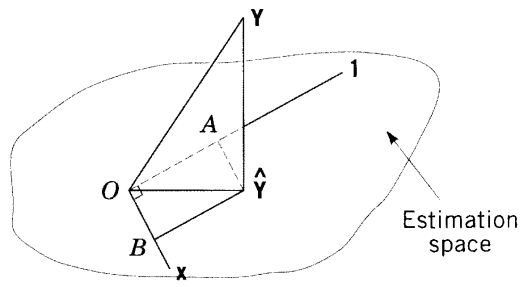

Figure 20.3. Orthogonal vectors in the model function.

Because these vectors are orthogonal, perpendiculars from $\hat{\mathbf{Y}}$ to these lines form a (right-angled) rectangle, so that

$$O\hat{Y}^2 = OA^2 + OB^2 \tag{20.2.3}$$

(or $OA^2 + A\hat{Y}^2$, etc.). The division is unique because of the orthogonality of the base vectors **1**, **x**; this result extends in the obvious way to any number of orthogonal base vectors, in which case the rectangle of Figure 20.3 becomes a rectangular block in the p-dimensional estimation space.

Exercise 3. For the least squares regression problem where

$$\mathbf{Y}' = (6, 9, 17, 18, 26) \quad \text{and} \quad \mathbf{X}' = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ -2 & -1 & 0 & 1 & 2 \end{bmatrix}$$

show that the normal equations $\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{Y}$ split up into two separate pieces and that A and B in Figure 20.3 are the two "individual $\hat{Y}$'s" that come from looking at each piece separately. Hence evaluate Eqs. (20.2.1) and (20.2.2) for these data (1406 = 1395.3 + 10.7, and 4 = 2 + 2).

An Orthogonal Breakup of the Normal Equations

Let us redraw Figure 20.3 with some additional detail; see Figure 20.4, in which lines OY and $O\hat{Y}$ are omitted but YA and YB have been drawn. In other respects the diagrams are intended to be identical. The points marked O, A, $\hat{Y}$, and B form a rectangle, and thus $O\hat{Y}^2 = OA^2 + OB^2$, as mentioned. Also, YA is perpendicular to OA, and YB is perpendicular to OB. Thus OA is "the $\hat{\mathbf{Y}}$ for the regression of $\mathbf{Y}$ on **1** alone" and OB is "the $\hat{\mathbf{Y}}$ for the regression of $\mathbf{Y}$ on **x** alone." This happens only because (here) **1** and **x** are orthogonal. The result also extends to any number of $\mathbf{X}$ vectors provided they are mutually orthogonal.

Exercise 4. For the data of Exercise 3, show that $OA^2 = 1155.2$, $OB^2 = 240.1$, and their sum is $O\hat{Y}^2 = 1395.3$. Also show that OA is the vector $(15.2, 15.2, 15.2, 15.2, 15.2)'$, OB is $(-9.8, -4.9, 0, 4.9, 9.8)'$, and $O\hat{Y}$ is the sum of these orthogonal vectors.

Orthogonalizing the Vectors of X in General

It is frequently desired to break down a regression sum of squares into components even when the vectors of $\mathbf{X}$ are not mutually orthogonal. Such a breakdown can

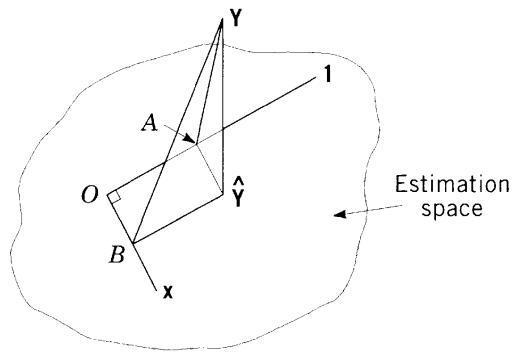

Figure 20.4. Orthogonal breakup of the normal equations when model vectors are orthogonal.

provide a set of sequential sums of squares, each being an extra sum of squares to the entries above it in the sequence. Geometrically, we need to perform a succession of orthogonalization procedures to do this. Figure 20.5 shows the idea for two vectors $\mathbf{1} = \mathbf{x}_0$ and $\mathbf{x}$, for which $\mathbf{1}'\mathbf{x} \neq 0$. We first pick one vector as a starter (we shall pick $\mathbf{1}$, but either would work) and then construct $\mathbf{x}_{\cdot 0}$, the portion of $\mathbf{x}$ (the second vector) that is orthogonal to $\mathbf{1}$. How? We use the standard result of regression that residuals are orthogonal to fitted values. We fit $\mathbf{x}$ (thinking of it as a "$\mathbf{Y}$") on to $\mathbf{1}$ and take residuals, giving $\hat{\mathbf{x}} = b\mathbf{1}$, where $b = "(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}" = (\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}'\mathbf{x} = \bar{x}$. Thus $\mathbf{x}_{\cdot 0} = \mathbf{x} - \bar{x}\mathbf{1}$. It is easy to check that $\mathbf{x}'_{\cdot 0}\mathbf{1} = 0$ so that these two vectors are orthogonal. Now the original regression equation was of the form

$$\hat{\mathbf{Y}} = b_0\mathbf{1} + b_1\mathbf{x}, \tag{20.2.4}$$

where

$$\begin{pmatrix} b_0 \\ b_1 \end{pmatrix} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

The orthogonalized regression equation is

$$\begin{aligned} \hat{\mathbf{Y}} &= \bar{Y}\mathbf{1} + b_1(\mathbf{x} - \bar{x}\mathbf{1}) \\ &= \bar{Y}\mathbf{1} + b_1\mathbf{x}_{\cdot 0}. \end{aligned} \tag{20.2.5}$$

We have used the same symbol b_1 in both equations (20.2.4) and (20.2.5) because the values will be identical. What we have in (20.2.4) and (20.2.5) are two different descriptions of the *same* vector $O\hat{Y}$ in Figure 20.5. This vector, $\hat{\mathbf{Y}}$, can be represented either as the sum of the nonorthogonal vectors OA^* and OB^*, or as the sum of the orthogonal vectors OA and OB. Either description is valid but only the second permits us to split up the sum of squares $O\hat{Y}^2$ as $OA^2 + OB^2$ via the Pythagoras result. Note that $\mathbf{Y}$ and $\hat{\mathbf{Y}}$ are unaffected by what we have done. All we have done is alter the description of $\hat{\mathbf{Y}}$ in terms of a linear combination of vectors in the space. Exercises 5 and 6 explore the fact that we can choose either vector first.

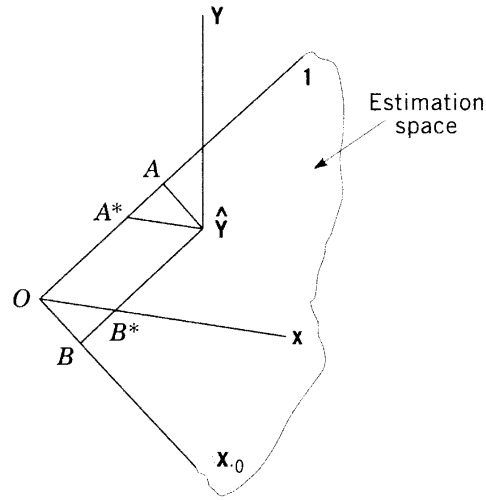

Figure 20.5. Orthogonalizing a second predictor with respect to $\mathbf{1}$.

Exercise 5. For the data of Exercise 1, namely, $\mathbf{Y} = (3.1, 2.3, 5.4)'$, $\mathbf{x}_0 = (1, 1, 1)'$, $\mathbf{x} = (2, 1, 3)'$, and $\hat{\mathbf{Y}} = 0.50\mathbf{1} + 1.55\mathbf{x}$, show that $\mathbf{x}_{\cdot 0} = (0, -1, 1)'$, and that $\hat{\mathbf{Y}} = 3.5\mathbf{1} + 1.55\mathbf{x}_{\cdot 0}$ produces the same fitted values and residuals. Draw a (fairly) accurate diagram of the form of Figure 20.5, working out the lengths of OA^*, OA, OB, OB^*, $O\hat{Y}$, and $Y\hat{Y}$ beforehand to get it (more or less) right. Note that YA is the orthogonal projection of $\mathbf{Y}$ onto $\mathbf{1}$ and YB is the orthogonal projection of $\mathbf{Y}$ on to $\mathbf{x}_{\cdot 0}$. Show that

$$O\hat{Y}^2 = OA^2 + OB^2 \neq OA^{*2} + OB^{*2} \tag{20.2.6}$$

or

$$43.685 = 38.88 + 4.805 \neq 0.125 + 33.635.$$

The quantity $OA^2 = \mathrm{SS}(b_0) = n\overline{Y}^2 = 38.88$, while $OB^2 = \mathrm{SS}(b_1|b_0) = 4.805$. Note that, above, we picked $\mathbf{x}_0 = \mathbf{1}$ first and found $\mathbf{x}_{\cdot 0}$.

Exercise 6. Pick, for the data of Exercise 5, $\mathbf{x}$ first and determine $\mathbf{x}_1$ such that $\mathbf{x}_1'\mathbf{x} = 0$. Let A and B be replaced by perpendiculars from $\hat{\mathbf{Y}}$ to $\mathbf{x}_1$ and $\mathbf{x}$ at C and D, say. Show that $43.685 = O\hat{Y}^2 = OC^2 + OD^2 = OA^2 + OB^2$ but that the split-up is different in the two cases, that is, $OC \neq OA$ and $OD \neq OB$. Specifically, $0.107 = OC^2 \neq OA^2 = 38.88$ and $43.578 = OD^2 \neq OB^2 = 4.805$.

In what we have done above, we see, geometrically, a fact we already know algebraically. Unless all the vectors in $\mathbf{X}$ are mutually orthogonal, the sequential sums of squares will depend on the sequence selected for the entry of the predictor variables. After the entry of the variable selected first, the others are (essentially) successively orthogonalized to all the ones entered before them. We reemphasize a point of great importance in this. $\hat{\mathbf{Y}}$ is unique and fixed no matter what the entry sequence may be. We merely give $\hat{\mathbf{Y}}$ different descriptions according to the vectors we select in the orthogonalized sequence. That is all.

20.3. ANALYSIS OF VARIANCE AND *F*-TEST FOR OVERALL REGRESSION

We reconsider the case of Figure 20.5 where the model is $\mathbf{Y} = \beta_0\mathbf{1} + \beta_1\mathbf{x} + \boldsymbol{\epsilon}$ and $\mathbf{1}'\mathbf{x} \neq 0$. The standard analysis of variance table takes the form of Table 20.1.

Note that the analysis of variance table is simply a Pythagoras split-up of (first)

$$O\hat{Y}^2 = OA^2 + OB^2$$

followed by

$$(OA^2 + OB^2) + Y\hat{Y}^2 = OY^2,$$

and the F-test for $H_0: \beta_1 = 0$ versus $\beta_1 \neq 0$ is simply a comparison of the "length-squared per df of OB" versus the "length-squared per df of $Y\hat{Y}$." Then rejection of

TABLE 20.1. Analysis of Variance Table for a Straight Line Fit

Source	SS	df	MS	F
b_0	$OA^2 = n\overline{Y}^2$	1	—	—
$b_1\|b_0$	$OB^2 = S_{XY}^2/S_{XX}$	1	S_{XY}^2/S_{XX}	$F = S_{XY}^2/(s^2 S_{XX})$ $= \{OB^2/1\}/\{Y\hat{Y}^2/(n-2)\}$
Residual	$Y\hat{Y}^2$ = By subtraction	$n-2$	s^2	
Total	$OY^2 = \mathbf{Y}'\mathbf{Y}$	n		

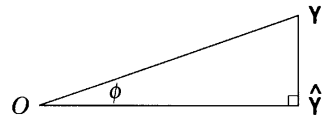

Figure 20.6. Basic geometry of the F-test for no regression at all, not even β_0.

H_0 is caused by larger values of OB^2 rather than smaller, so we are essentially asking: "Is B close to O compared with the size of s?" (do not reject H_0) or "Is B *not* close to O compared with the size of s"? (reject H_0). [The fact that the F-ratio follows the $F(1, n-2)$ distribution when H_0 is true can be established algebraically, but it also has a geometric interpretation, given below.]

An F-test for $H_0: \beta_0 = \beta_1 = 0$ versus H_1: not so, would similarly involve

$$F = \{(OA^2 + OB^2)/2\}/\{Y\hat{Y}^2/(n-2)\}.$$

Note that (see Figure 20.6) this involves a comparison of the "per-df lengths" of $O\hat{Y}^2$ with $Y\hat{Y}^2$. Significant regression will be one in which $O\hat{Y}$ is "large" compared with $Y\hat{Y}$. This implies that the angle ϕ on Figure 20.6 is "small." If we think of OY (the data) as being a fixed axis and $O\hat{Y}$ as one possible position of the fitted vector, which could lie anywhere at the same angle ϕ to OY (in positive or negative direction), we have a geometrical interpretation of the F-test (see Figure 20.7). The tail probability of the F-statistic is the proportional area of the two circular "caps," defined as $\hat{Y}$ rotates around Y, on a sphere of any radius.

What if the test were on nonzero values of β_0 and β_1? Suppose we wished to test $H_0: \beta_0 = \beta_{00}$ and $\beta_1 = \beta_{10}$? Then the point O^* defined by $\mathbf{Y}_0 = \beta_{00}\mathbf{1} + \beta_{10}\mathbf{x}$ would replace O in what we have said above but the other elements of the geometry would essentially be preserved. (See Figure 20.8.) We leave this as an exercise, noting only that in nearly all such problems $\beta_{10} = 0$ anyway, even when $\beta_{00} \neq 0$.

20.4. THE SINGULAR X'X CASE: AN EXAMPLE

What happens to the geometry in a regression where the problem is singular, that is, $\det(\mathbf{X}'\mathbf{X}) = 0$? The answer is "very little" and the understanding of this answer will reduce one's fear of meeting such problems in practice. Look at Figure 20.9. Although discussed in the framework of three dimensions, the issues discussed are general ones. Suppose that the estimation space is the plane shown, but the $\mathbf{X} = (\mathbf{1}, \mathbf{x}_1, \mathbf{x}_2)$ matrix consists of three vectors. It is obvious that the $\mathbf{X}'\mathbf{X}$ matrix is singular because the three vectors in $\mathbf{X}$ must be dependent. If we suppose that any two of the vectors define the plane (i.e., the three vectors are distinct ones) then the third vector must be a linear combination of the other two. Now it is sometimes mistakenly thought

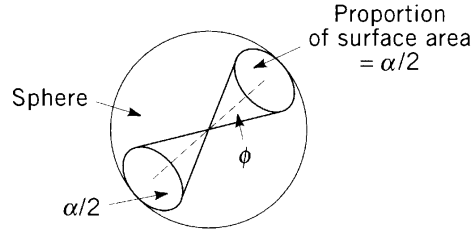

Figure 20.7. A geometrical interpretation of the F-test probability.

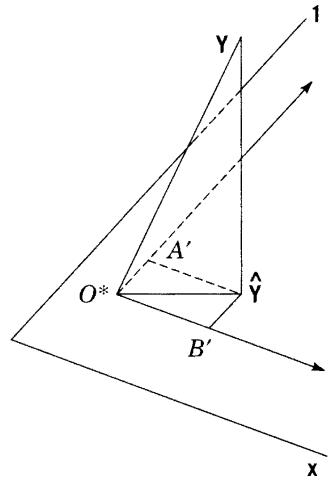

Figure 20.8. Testing that $(\beta_0, \beta_1) = (\beta_{00}, \beta_{10})$.

that the least squares problem has no solution in these circumstances, because $(\mathbf{X}'\mathbf{X})^{-1}$ does not exist. We can see from Figure 20.9 that not only is there a least squares solution but, as always, it is unique. That is, we can drop a perpendicular onto the estimation space at $\hat{Y}$,$\hat{Y}$ is unique, and we have a unique vector $\mathbf{Y} - \hat{\mathbf{Y}}$ orthogonal to the estimation space, and thus orthogonal to all the columns of $\mathbf{X}$. What is not unique is the description of $\hat{\mathbf{Y}}$ in terms of $\mathbf{1}$, $\mathbf{x}_1$, and $\mathbf{x}_2$. Because we have (one, here) too many base vectors, there are an infinite number of ways in which $\hat{\mathbf{Y}}$ can be described. The normal equations exist and can be solved, but the solution for the parameter estimates is not unique. We illustrate this with the smallest possible example, two parameters and two vectors $\mathbf{1}$, $\mathbf{x}$ that are multiples of each other.

Example

Suppose we have n data values at $X = X^*$. (See Figure 20.10 where $n = 5$, although any n can be used.) Consider fitting the line $Y = \beta_0 + \beta_1 X + \epsilon$ by least squares. Clearly there are an infinite number of solutions. *Any* line through the point $(X, Y) = (X^*, \overline{Y})$ will provide a least squares fit. Geometrically the problem is that the two vectors that define the predictor space, $\mathbf{1}$ and $\mathbf{x} = X^*\mathbf{1}$, are multiples of one

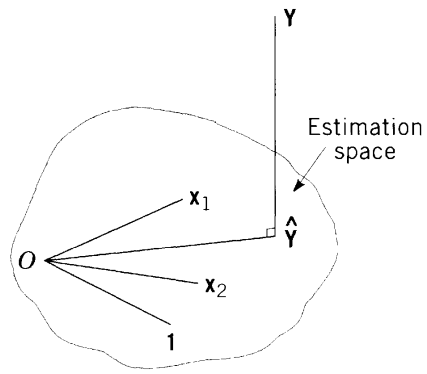

Figure 20.9. A two-dimensional estimation space defined by three vectors, one too many.

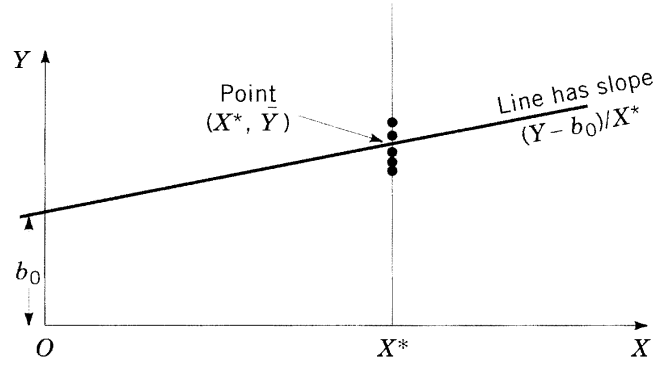

Figure 20.10. Singular regression problem with multiple descriptions for a unique $\hat{\mathbf{Y}}$.

another. (See Figure 20.11.) Thus the foot of the perpendicular from $\mathbf{Y}$ to the predictor space can be described by an infinite number of linear combinations of the vectors $\mathbf{1}$ and $X^*\mathbf{1}$. Note that $\hat{\mathbf{Y}} = \bar{Y}\mathbf{1}$ is unique, however. We now write down the two normal equations $\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{Y}$ as

$$\begin{aligned} nb_0 + nX^*b_1 &= n\bar{Y}, \\ nX^*b_0 + nX^{*2}b_1 &= nX^*\bar{Y}, \end{aligned} \tag{20.4.1}$$

and see immediately they are not independent but are one equation whose solution is $(b_0, b_1 = (\bar{Y} - b_0)/X^*)$ for any b_0. The fitted "line" is thus

$$\hat{Y} = b_0 + b_1X = b_0 + X(\bar{Y} - b_0)/X^*, \tag{20.4.2}$$

representing an infinity of lines with different slopes and intercepts, all passing through the point $(X, Y) = (X^*, \bar{Y})$. Whatever value is assigned to b_0, we have a least squares solution. At $X = X^*$, $\hat{\mathbf{Y}} = \bar{Y}\mathbf{1}$, which is unique whatever value b_0 takes.

20.5 ORTHOGONALIZING IN THE GENERAL REGRESSION CASE

We consider the fit of a general linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ and write $\mathbf{X} = (\mathbf{Z}_1, \mathbf{Z}_2)$, representing any division of $\mathbf{X}$ into two parts that would, in general, not be orthogonal. We divide $\boldsymbol{\beta}' = (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2')$ in a corresponding fashion. Thus we write

$$\mathbf{Y} = \mathbf{Z}_1\boldsymbol{\theta}_1 + \mathbf{Z}_2\boldsymbol{\theta}_2 + \boldsymbol{\epsilon}, \tag{20.5.1}$$

where $\mathbf{Y}$ is n by 1, $\mathbf{Z}_1$ is n by p_1, $\boldsymbol{\theta}_1$ is p_1 by 1, $\mathbf{Z}_2$ is n by p_2, and $\boldsymbol{\theta}_2$ is p_2 by 1. If we fit to just the $\mathbf{Z}_1\boldsymbol{\theta}_1$ part of the model, we get

$$\hat{\mathbf{Y}} = \mathbf{Z}_1(\mathbf{Z}_1'\mathbf{Z}_1)^{-1}\mathbf{Z}_1'\mathbf{Y} = \mathbf{P}_1\mathbf{Y}, \tag{20.5.2}$$

say, where $\mathbf{P}_1$ is the *projection matrix*, that is, the matrix that projects the vector $\mathbf{Y}$

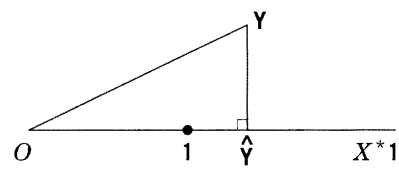

Figure 20.11. The estimation space is overdefined by the vectors $\mathbf{1}$ and $X^*\mathbf{1}$.

down into the estimation space defined by the columns of $\mathbf{Z}_1$. The residual vector is $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I} - \mathbf{P}_1)\mathbf{Y}$. Note that in previous discussions we called $\mathbf{P}_1$ the *hat* matrix because it converts the Y's into the $\hat{Y}$'s. We now rename it to emphasize its geometrical properties.

It is obvious that the fitted $\hat{\mathbf{Y}}$ and the residual $\mathbf{Y} - \hat{\mathbf{Y}}$ are orthogonal because

$$\begin{aligned} \hat{\mathbf{Y}}'\mathbf{e} &= \mathbf{Y}'\mathbf{P}_1'(\mathbf{I} - \mathbf{P}_1)\mathbf{Y} \\ &= \mathbf{Y}'(\mathbf{P}_1' - \mathbf{P}_1'\mathbf{P}_1)\mathbf{Y} \\ &= 0, \end{aligned} \tag{20.5.3}$$

because $\mathbf{P}_1$ is symmetric (so $\mathbf{P}_1' = \mathbf{P}_1$) and idempotent (so $\mathbf{P}_1^2 = \mathbf{P}_1$). This calculation is a repeat of (20.1.4) with different notation. The matrix $\mathbf{Z}_2$, when orthogonalized to $\mathbf{Z}_1$, becomes, by analogy to $\mathbf{Y} - \hat{\mathbf{Y}}$,

$$\begin{aligned} \mathbf{Z}_{2\cdot 1} &= \mathbf{Z}_2 - \hat{\mathbf{Z}}_2 \\ &= (\mathbf{I} - \mathbf{P}_1)\mathbf{Z}_2 \\ &= \mathbf{Z}_2 - \mathbf{Z}_1(\mathbf{Z}_1'\mathbf{Z}_1)^{-1}\mathbf{Z}_1'\mathbf{Z}_2 \\ &= \mathbf{Z}_2 - \mathbf{Z}_1\mathbf{A}. \end{aligned} \tag{20.5.4}$$

$\mathbf{A}$ is usually called the *alias* or *bias* matrix.

Exercise 7. Prove that $\mathbf{Z}_{2\cdot 1}$ and $\mathbf{Z}_1$ are orthogonal matrices.

We can now write the full model in the orthogonalized form

$$\begin{aligned} \mathbf{Y} &= \mathbf{Z}_1\boldsymbol{\theta}_1 + \mathbf{Z}_2\boldsymbol{\theta}_2 + \boldsymbol{\epsilon} \\ &= \mathbf{Z}_1(\boldsymbol{\theta}_1 + \mathbf{A}\boldsymbol{\theta}_2) + (\mathbf{Z}_2 - \mathbf{Z}_1\mathbf{A})\boldsymbol{\theta}_2 + \boldsymbol{\epsilon} \\ &= \mathbf{Z}_1\boldsymbol{\theta} + \mathbf{Z}_{2\cdot 1}\boldsymbol{\theta}_2 + \boldsymbol{\epsilon}. \end{aligned} \tag{20.5.5}$$

We see immediately that $\hat{\boldsymbol{\theta}} = (\mathbf{Z}_1'\mathbf{Z}_1)^{-1}\mathbf{Z}_1'\mathbf{Y}$ estimates *not* just $\boldsymbol{\theta}_1$ but $\boldsymbol{\theta}_1 + \mathbf{A}\boldsymbol{\theta}_2$. Thus $\mathbf{A}\boldsymbol{\theta}_2$ provides the *biases* in the estimates of $\boldsymbol{\theta}_1$ if only the model $\mathbf{Y} = \boldsymbol{\theta}_1\mathbf{Z}_1 + \boldsymbol{\epsilon}$ is fitted.

Exercise 8. Show this by evaluating $E(\hat{\mathbf{Y}}) = E(\mathbf{Z}_1\hat{\boldsymbol{\theta}})$ with $E(\mathbf{Y}) = \boldsymbol{\theta}_1\mathbf{Z}_2$.

For the full regression we have the analysis of variance table of Table 20.2, where $\boldsymbol{\eta} = \mathbf{Z}_1\boldsymbol{\theta}_1 + \mathbf{Z}_2\boldsymbol{\theta}_2$.

Note the following points:

1. The $\hat{\boldsymbol{\theta}}$ obtained from fitting $\mathbf{Y} = \mathbf{Z}_1\boldsymbol{\theta}_1 + \mathbf{Z}_{2\cdot 1}\boldsymbol{\theta}_2 + \boldsymbol{\epsilon}$ is identical to the $\hat{\boldsymbol{\theta}}_1$ obtained by fitting $\mathbf{Y} = \mathbf{Z}_1\boldsymbol{\theta}_1 + \boldsymbol{\epsilon}$.
2. The $\hat{\boldsymbol{\theta}}_2$ obtained from fitting $\mathbf{Y} = \mathbf{Z}_1\boldsymbol{\theta}_1 + \mathbf{Z}_2\boldsymbol{\theta}_2 + \boldsymbol{\epsilon}$ is identical to the $\hat{\boldsymbol{\theta}}_2$ obtained by fitting just $\mathbf{Y} = \mathbf{Z}_{2\cdot 1}\boldsymbol{\theta}_2 + \boldsymbol{\epsilon}$.
3. The values $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ in the analysis of variance table can be thought of as "test

TABLE 20.2. Analysis of Variance Table for the Orthogonalized General Regression

Source	df	SS
Response for $\mathbf{Z}_1$ only	p_1	$(\boldsymbol{\theta}_1 - \hat{\boldsymbol{\theta}}_1)'\mathbf{Z}_1'\mathbf{Z}_1(\boldsymbol{\theta}_1 - \hat{\boldsymbol{\theta}}_1)$
Extra for $\mathbf{Z}_2$	p_2	$(\boldsymbol{\theta}_2 - \hat{\boldsymbol{\theta}}_2)'\mathbf{Z}_{2\cdot 1}'\mathbf{Z}_{2\cdot 1}(\boldsymbol{\theta}_2 - \hat{\boldsymbol{\theta}}_2)$
Residual	$n - p_1 - p_2$	By subtraction
Total	n	$(\mathbf{Y} - \boldsymbol{\eta})'(\mathbf{Y} - \boldsymbol{\eta})$

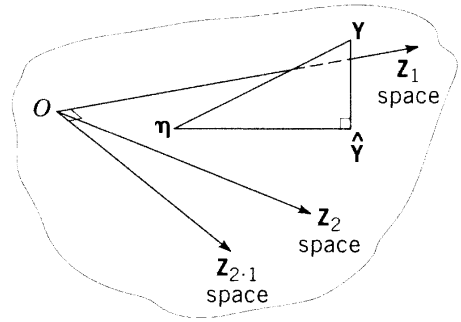

Figure 20.12. $\hat{\mathbf{Y}}$ can be described either as a linear combination of the spaces spanned (defined) by $\mathbf{Z}_1$ and $\mathbf{Z}_2$, which are not orthogonal, or of those spanned by $\mathbf{Z}_1$ and $\mathbf{Z}_{2\cdot1}$, which are orthogonal.

values" and the geometry is modified by a move from the origin to $\boldsymbol{\eta} = \mathbf{Z}_1\boldsymbol{\theta}_1 + \mathbf{Z}_2\boldsymbol{\theta}_2$ (see Figure 20.12).

4. The need for $\mathbf{Z}_2\boldsymbol{\theta}_2$ in the model will be indicated by a large "extra for $\mathbf{Z}_2$" sum of squares.
5. The "extra for $\mathbf{Z}_2$" sum of squares can also be obtained by fitting $\mathbf{Y} = \mathbf{Z}_1\boldsymbol{\theta}_1 + \mathbf{Z}_2\boldsymbol{\theta}_2 + \boldsymbol{\epsilon}$, then fitting $\mathbf{Y} = \mathbf{Z}_1\boldsymbol{\theta}_1 + \boldsymbol{\epsilon}$, and finding the difference between:
   a. The two regression sums of squares.
   b. The two regression sums of squares but with both corrected by $n\overline{Y}^2$.
   c. The two residual sums of squares in reverse order (to give a positive sign result).

More on the geometry of this is given in Chapter 21.

20.6. RANGE SPACE AND NULL SPACE OF A MATRIX M

The range space of a matrix **M**, written $R(\mathbf{M})$, is the space of all vectors defined by the columns of **M**. The dimension of the space is the column rank of **M**, cr(**M**), that is, the number of linearly independent columns of **M**. (Thus the number of columns in **M** may exceed the dimension of the space they define.) The null space of **M**, written $N(\mathbf{M})$, consists of the range space of all vectors **v** such that $\mathbf{Mv} = 0$, that is, all vectors **v** orthogonal to the *rows* of **M**. If we want to define the null space of all vectors orthogonal to the *columns* of **M** we must write $N(\mathbf{M}')$.

Projection Matrices

Let E_n be n-dimensional Euclidean space (i.e., "ordinary" n-dimensional space). Let Ω be a p-dimensional subspace of E_n. Let $\Omega^\perp$ be the rest of E_n, that is, the subspace of E_n orthogonal to Ω. Let $\mathbf{P}_\Omega$ be an n by n projection matrix that projects a general n-dimensional vector **Y** entirely into the space Ω. (We shall write simply **P** in statements involving only the Ω-space.) Then a number of statements can be proved, as follows.

1. Every vector **Y** can be expressed uniquely in the form $\hat{\mathbf{Y}} + \mathbf{e}$, where $\hat{\mathbf{Y}}$ is wholly in Ω (we write $\hat{\mathbf{Y}} \in \Omega$) and $\mathbf{e} \in \Omega^\perp$.
2. If $\hat{\mathbf{Y}} = \mathbf{PY}$, then **P** is unique.
3. **P** can be written as $\mathbf{P} = \mathbf{TT}'$, where the p columns of the n by p matrix **T** form an orthonormal basis (*not* unique) for the Ω-space. **T** is *not* unique and many choices are possible, even though **P** is unique.

[*Note:* A *basis* of Ω is a set of vectors that span the space of Ω, that is, permit every vector of Ω to be expressed as a linear combination of the base vectors. An *orthogonal* basis is one in which all basis vectors are orthogonal to one another. An *orthonormal* basis is an orthogonal basis for which the basis vectors have length (and so squared length) one, that is, $\mathbf{v}'\mathbf{v} = 1$ for all basis vectors $\mathbf{v}$.]

4. $\mathbf{P}$ is symmetric ($\mathbf{P}' = \mathbf{P}$) and idempotent ($\mathbf{P}^2 = \mathbf{P}$).
5. The vectors of $\mathbf{P}$ span the space Ω, that is, $R(\mathbf{P}) = \Omega$.
6. $\mathbf{I} - \mathbf{P}$ is the projection matrix for $\Omega^\perp$, the orthogonal part of E_n *not* in Ω. Thus $R(\mathbf{I} - \mathbf{P}) = \Omega^\perp$.
7. Any symmetric idempotent n by n matrix $\mathbf{P}$ represents an orthogonal projection matrix onto the space spanned by the columns of $\mathbf{P}$, that is, onto $R(\mathbf{P})$.

Statements 1–7 are generally true for any Ω even though we have given them in a notation that fits into the concept of regression. Statements 8 and 9 now make the connection with regression. We think of fitting the model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ by least squares, and $\Omega = R(\mathbf{X})$ will now be defined by the p columns of $\mathbf{X}$ and so will constitute the estimation space for our regression problem.

8. Suppose Ω is a space spanned by the columns of the n by p matrix $\mathbf{X}$. Suppose $(\mathbf{X}'\mathbf{X})^-$ is any generalized inverse of $\mathbf{X}'\mathbf{X}$ (see Appendix 20A). Then $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}'$ is the unique projection matrix for Ω. Note carefully that $\mathbf{X}$ is not unique, nor is $(\mathbf{X}'\mathbf{X})^-$, but $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}'$ is unique and so is $\hat{\mathbf{Y}} = \mathbf{P}\mathbf{Y}$.
9. If $\text{cr}(\mathbf{X}) = p$ so that the columns of $\mathbf{X}$ are linearly independent, and $(\mathbf{X}'\mathbf{X})^{-1}$ exists, then $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and $R(\mathbf{X}) = R(\mathbf{P}) = \Omega$. This is the situation that will typically hold.

For proofs of these statements, see Seber (1977, pp. 394–395).

The practical consequences of these statements are as follows.

Given an estimation space Ω (and so, necessarily, an error space $\Omega^\perp$), any vector $\mathbf{Y}$ can uniquely be regressed into Ω, via a unique projection matrix $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}'$. The facts that, even when $\mathbf{X}$ and $(\mathbf{X}'\mathbf{X})^-$ are not unique, $\mathbf{P}$ is unique and so is $\hat{\mathbf{Y}} = \mathbf{P}\mathbf{Y}$ for a given Ω are completely obvious when thought of geometrically. For a given Ω and $\mathbf{Y}$, there is a unique projection of $\mathbf{Y}$ onto Ω that defines $\hat{\mathbf{Y}}$ uniquely in Ω. It doesn't matter *how* Ω is described [i.e., what $\mathbf{X}$ is chosen, or what choice of $(\mathbf{X}'\mathbf{X})^-$ is made when $\mathbf{X}'\mathbf{X}$ is singular] or *how* $\hat{\mathbf{Y}}$ is described as a function $\mathbf{X}\mathbf{b} = \mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{Y}$ of the b's. The basic triangle of Figure 20.13 joining $\mathbf{0}$, $\mathbf{Y}$, and $\hat{\mathbf{Y}}$ remains fixed forever, given Ω and $\mathbf{Y}$.

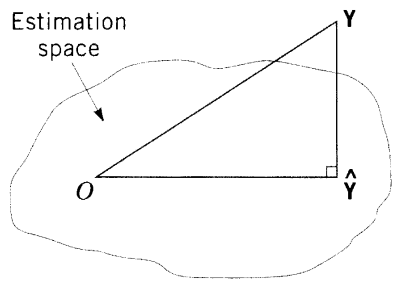

Figure 20.13. The triangle is unique, given Ω and $\mathbf{Y}$.

20.7. THE ALGEBRA AND GEOMETRY OF PURE ERROR

We consider a general linear regression situation with

$$\begin{aligned} &Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{p-1} X_{p-1} + \epsilon, \\ \text{or} \quad &\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}. \end{aligned} \tag{20.7.1}$$

Our model has p parameters. Some of the X's could be transformations of other X's. Thus, for example, the model could be a polynomial of second order. We assume $\mathbf{X}'\mathbf{X}$ is nonsingular; if it were not, we would redefine $\mathbf{X}$ to ensure that it was. Suppose we have m data sites with $n_1, n_2, n_3, \ldots, n_m$ repeats at these sites, with $n_1 + n_2 + \cdots + n_m = n$. Naturally, $p \leq m$ or we cannot estimate the parameters; preferably $p < m$, or we cannot test for lack of fit. Without loss of generality, some n_j could equal 1, but not all of them, or we would have no pure error with which to test lack of fit. We now define an n by m matrix $\mathbf{X}_e$ of form

$$\mathbf{X}_e = \begin{bmatrix} 1 & & & \\ 1 & & & \\ \cdots & & & \\ 1 & & & \\ \hline & 1 & & \\ & 1 & & \\ & \cdots & & \\ & 1 & & \\ \hline & & \cdots & \\ \hline & & & 1 \\ & & & 1 \\ & & & \cdots \\ & & & 1 \end{bmatrix}, \tag{20.7.2}$$

which is such that the jth column contains only n_j ones and $n - n_j$ zeros, the ones positioned in the row locations $n_1 + n_2 + \cdots + n_{j-1} + 1$ to $n_1 + n_2 + \cdots + n_j$. For a site j with no repeats there will be only a single one in the corresponding column. Consider the model

$$\mathbf{Y} = \mathbf{X}_e\boldsymbol{\mu} + \boldsymbol{\epsilon} \tag{20.7.3}$$

where $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_m)'$. If $p \leq m$, model (20.7.3) is inclusive of model (20.7.1) because we can reexpress the m μ's in terms of p β's, so $R(\mathbf{X})$ is included within $R(\mathbf{X}_e)$. Geometrically the m column vectors of $\mathbf{X}_e$ are linearly combined to the p column vectors of $\mathbf{X}$. This implies that, if we define $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and $\mathbf{P}_e = \mathbf{X}_e(\mathbf{X}_e'\mathbf{X}_e)^{-1}\mathbf{X}_e'$ as the respective projection matrices, the space $R(\mathbf{P}) = R(\mathbf{X})$ lies within the space $R(\mathbf{P}_e) = R(\mathbf{X}_e)$.

Now $\mathbf{X}_e'\mathbf{X}_e$ = diagonal $(n_1, n_2, \ldots, n_m)$, a matrix with terms n_i along the upper-left

to lower-right main diagonal and zeros elsewhere, so that $(\mathbf{X}_e'\mathbf{X}_e)^{-1}$ = diagonal $(n_1^{-1}, n_2^{-1}, \ldots, n_m^{-1})$. It follows (an exercise for the reader here!) that $\mathbf{P}_e$ consists of a matrix with m main-diagonal blocks of sizes $n_1, n_2, \ldots, n_m$. The jth of these has the form

$$\mathbf{B}_j = \begin{bmatrix} n_j^{-1} & n_j^{-1} & \cdots & n_j^{-1} \\ n_j^{-1} & n_j^{-1} & \cdots & n_j^{-1} \\ \vdots & \vdots & & \vdots \\ n_j^{-1} & n_j^{-1} & \cdots & n_j^{-1} \end{bmatrix} = n_j^{-1}\mathbf{1}\mathbf{1}' \tag{20.7.4}$$

where the $\mathbf{1}$ is an n_j by 1 vector.

We now consider the breakup of the residual sum of squares into lack of fit and pure error. We can write the residual vector as

$$\mathbf{Y} - \hat{\mathbf{Y}} = \tilde{\mathbf{Y}} - \hat{\mathbf{Y}} + \mathbf{Y} - \tilde{\mathbf{Y}}, \tag{20.7.5}$$

where $\tilde{\mathbf{Y}} = \mathbf{P}_e\mathbf{Y}$. Recalling that $\hat{\mathbf{Y}} = \mathbf{P}\mathbf{Y}$, we thus have

$$(\mathbf{I} - \mathbf{P})\mathbf{Y} = (\mathbf{P}_e - \mathbf{P})\mathbf{Y} + (\mathbf{I} - \mathbf{P}_e)\mathbf{Y}, \tag{20.7.6}$$

that is,

$$\text{Residual vector} = \text{Lack of fit vector} + \text{Pure error vector}.$$

The pure error vector consists of deviations of the individual observations from their own pure error group averages. If an $n_j = 1$, a zero appears in the appropriate position, of course. Note that

$$(\mathbf{P}_e - \mathbf{P})'(\mathbf{I} - \mathbf{P}_e) = \mathbf{P}_e - \mathbf{P}_e^2 - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{I} - \mathbf{P}_e) = \mathbf{0}. \tag{20.7.7}$$

The first pair of terms cancel each other because $\mathbf{P}_e$ is a projection matrix and so idempotent. In the third term $\mathbf{X}'(\mathbf{I} - \mathbf{P}_e) = \mathbf{0}$ due to the special forms of $\mathbf{X}$ and $\mathbf{I} - \mathbf{P}_e$ when repeats occur. Specifically, the product $\mathbf{X}'(\mathbf{I} - \mathbf{P}_e)$ consists of a series of products, the jth of which takes the form

$$\begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ a & a & a & \cdots & a \\ \vdots & \vdots & \vdots & & \vdots \\ z & z & z & & z \end{bmatrix} \begin{bmatrix} 1 - n_j^{-1} & -n_j^{-1} & \cdots & -n_j^{-1} \\ -n_j^{-1} & 1 - n_j^{-1} & \cdots & -n_j \\ \vdots & \vdots & & \vdots \\ -n_j^{-1} & -n_j^{-1} & \cdots & 1 - n_j^{-1} \end{bmatrix}, \tag{20.7.8}$$

where $(a, \ldots, z)$ are the values of $(X_1, \ldots, X_k)$ in the jth group of repeats; each letter $a, \ldots, z$ occurs n_j times. We could also write this product more compactly as

$$\begin{bmatrix} 1 \\ a \\ \vdots \\ z \end{bmatrix} \mathbf{1}'(\mathbf{I} - n_j^{-1}\mathbf{1}\mathbf{1}') = \mathbf{0} \tag{20.7.9}$$

because $\mathbf{1}'\mathbf{1} = n_j$ for the jth group.

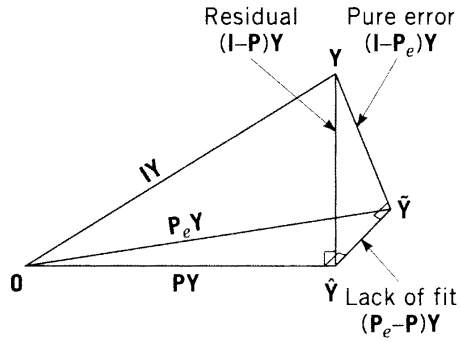

Figure 20.14. The orthogonal breakup of the residual vector into the lack of fit and pure error vectors. All three of these vectors are orthogonal to the estimation space $R(\mathbf{X})$.

The Geometry of Pure Error

From the argument above we see that the geometry of pure error is an orthogonal decomposition of the residual vector $\mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I} - \mathbf{P})\mathbf{Y}$ into orthogonal pieces $(\mathbf{P}_e - \mathbf{P})\mathbf{Y}$ for lack of fit and $(\mathbf{I} - \mathbf{P}_e)\mathbf{Y}$, for pure error, as in (20.7.6). If we define

$$n_e = (n_1 - 1) + (n_2 - 1) + \cdots + (n_m - 1) = n - m \tag{20.7.10}$$

as the pure error degrees of freedom, the corresponding dimensional breakup is

$$n - p = (n - p - n_e) + n_e \tag{20.7.11}$$

for the lack of fit and pure error spaces. As always, the corresponding sums of squares are the squared lengths of the vectors, namely,

$$\mathbf{Y}'(\mathbf{I} - \mathbf{P})\mathbf{Y} = \mathbf{Y}'(\mathbf{P}_e - \mathbf{P})\mathbf{Y} + \mathbf{Y}'(\mathbf{I} - \mathbf{P}_e)\mathbf{Y}, \tag{20.7.12}$$

where we have reduced the matrix powers of the quadratic forms by invoking the idempotency of $\mathbf{I} - \mathbf{P}$, $\mathbf{P}_e - \mathbf{P}$, and $\mathbf{I} - \mathbf{P}_e$. Figure 20.14 illustrates the geometry, and Figure 20.15 shows the dimensions (degrees of freedom) of the various spaces.

APPENDIX 20A. GENERALIZED INVERSES M⁻

(Sometimes these are also called pseudo-inverses, as well as other terms.) Suppose, first, that $\mathbf{M}$ is a nonsingular square matrix. Then $\mathbf{M}^- = \mathbf{M}^{-1}$. If $\mathbf{M}$ is singular, $\mathbf{M}^-$ is any matrix for which

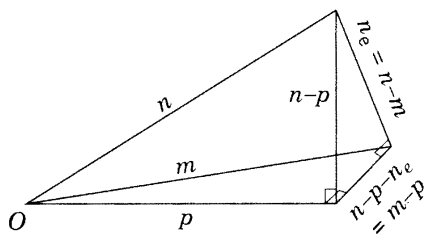

Figure 20.15. The corresponding degrees of freedom of the residual vector breakup are shown as the dimensions of the subspaces in which the various vectors lie. For all subspaces to exist, we must have $n > m > p$, that is, there must be fewer sites than observations but more sites than parameters.

$$\mathbf{M}\mathbf{M}^{-}\mathbf{M} = \mathbf{M}. \tag{20A.1}$$

This concept of a generalized inverse is an interesting extension of the idea of the inverse matrix to singular matrices. (Although it *is* interesting, it can usually be circumvented in practical regression situations!) A generalized inverse satisfying $\mathbf{M}\mathbf{M}^{-}\mathbf{M} = \mathbf{M}$ always exists and is not unique.

(*Note:* It is also possible to define $\mathbf{M}^{-}$ in the same way when $\mathbf{M}$ is not square. For regression purposes, we do not need this extension.)

Moore–Penrose Inverse

A unique definition (the so-called Moore–Penrose inverse) can be obtained by insisting on $\mathbf{M}^{-}$ satisfying three more conditions, namely,

$$\begin{aligned} \mathbf{M}^{-}\mathbf{M}\mathbf{M}^{-} &= \mathbf{M}^{-}, \\ (\mathbf{M}\mathbf{M}^{-})' &= \mathbf{M}\mathbf{M}^{-}, \\ (\mathbf{M}^{-}\mathbf{M})' &= \mathbf{M}^{-}\mathbf{M}. \end{aligned} \tag{20A.2}$$

Some writers write $\mathbf{M}^{+}$ for the Moore–Penrose inverse.

Getting a Generalized Inverse

We assume $\mathbf{M}$ is square and, for our regression applications, is of the symmetric form $\mathbf{X}'\mathbf{X}$. Let $\mathbf{M}$ be p by p and have rank (row or column rank) $r < p$. Probably the easiest method for getting an $\mathbf{M}^{-}$ is the following.

A Method for Getting $\mathbf{M}^{-}$

Examine $\mathbf{M}$ to find an r by r submatrix that has full rank, that is, is nonsingular. If it occupies the upper-left corner, then

$$\mathbf{M} = \begin{bmatrix} \mathbf{M}_{11} & \mathbf{M}_{12} \\ \mathbf{M}_{21} & \mathbf{M}_{22} \end{bmatrix}, \tag{20A.3}$$

where $\mathbf{M}_{11}$ is nonsingular, so that $\mathbf{M}_{11}^{-1}$ exists. Then

$$\mathbf{M}^{-} = \begin{bmatrix} \mathbf{M}_{11}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \tag{20A.4}$$

will be a generalized inverse for $\mathbf{M}$.

Proof: Form the product $\mathbf{M}\mathbf{M}^{-}\mathbf{M}$ to get

$$\begin{bmatrix} \mathbf{M}_{11} & \mathbf{M}_{12} \\ \mathbf{M}_{21} & \mathbf{M}_{21}\mathbf{M}_{11}^{-1}\mathbf{M}_{12} \end{bmatrix}, \tag{20A.5}$$

which has three submatrices correct. Because of the implicit assumption that the $(p - r)$ later columns of $\mathbf{M}$ depend on the first r, there exists an r by $(p - r)$ matrix $\mathbf{Q}$, say, such that

$$\mathbf{M}_{11}\mathbf{Q} = \mathbf{M}_{12} \quad \text{and} \quad \mathbf{M}_{21}\mathbf{Q} = \mathbf{M}_{22}. \tag{20A.6}$$

Solving for $\mathbf{Q}$ in the first of these gives $\mathbf{Q} = \mathbf{M}_{11}^{-1}\mathbf{M}_{12}$ whence $\mathbf{M}_{21}\mathbf{M}_{11}^{-1}\mathbf{M}_{12} = \mathbf{M}_{21}\mathbf{Q} = \mathbf{M}_{22}$. The method works in exactly the same way *wherever* the elements of the nonsingular ($\mathbf{M}_{11}$) matrix are located. It is inverted where it stands and zeros occupy all other places.

Our regression application is the following. If $\mathbf{X}'\mathbf{X}$ is singular and $(\mathbf{X}'\mathbf{X})^-$ is any generalized inverse, the normal equations $\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{Y}$ are satisfied by $\mathbf{b} = (\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{Y}$. See Seber (1977, pp. 76 and 391). Note that although different choices of $(\mathbf{X}'\mathbf{X})^-$ produce different $\mathbf{b}$ estimates, $\hat{\mathbf{Y}} = \mathbf{X}\mathbf{b}$ is invariant due to the geometry.

Example

Consider the fitting of a straight line $Y = \beta_0 + \beta_1 X + \epsilon$ to n data points all at the same location $X = X^*$. The least squares solution is any line through the point $(\bar{X}, \bar{Y}) = (X^*, \bar{Y})$, because a unique solution is obtained only when there are two or more X-sites, and here we have only one. The general solution is thus of the form

$$\hat{Y} = b_0 + (\bar{Y} - b_0)(X/X^*) \tag{20A.7}$$

for any choice of b_0. [Note that when $X = X^*$, $\hat{Y} = \bar{Y}$ so that $\hat{\mathbf{Y}} = (\bar{Y}, \bar{Y}, \ldots, \bar{Y})'$ and is unique, as we know it must be from the geometry.] For this problem, the normal equations are

$$\begin{bmatrix} n & nX^* \\ nX^* & nX^{*2} \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \begin{bmatrix} \Sigma Y_i \\ X^* \Sigma Y_i \end{bmatrix} \tag{20A.8}$$

and do not have a unique solution. We now look at what is achieved by specific choices of $(\mathbf{X}'\mathbf{X})^-$, when evaluating $\mathbf{b} = (\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{Y}$.

Choice 1. Let

$$(\mathbf{X}'\mathbf{X})^- = \begin{bmatrix} n^{-1} & 0 \\ 0 & 0 \end{bmatrix}. \tag{20A.9}$$

Then $b_0 = \bar{Y}$ and $b_1 = 0$. We obtain a horizontal straight line through $(X^*, \bar{Y})$.

Choice 2. Let

$$(\mathbf{X}'\mathbf{X})^- = \begin{bmatrix} 0 & 0 \\ 0 & (nX^{*2})^{-1} \end{bmatrix}. \tag{20A.10}$$

Then $b_0 = 0$, $b_1 = \bar{Y}/X^*$. We have a straight line joining the origin to the point $(X^*, \bar{Y})$.

Choice 3. Let

$$(\mathbf{X}'\mathbf{X})^- = \begin{bmatrix} 0 & (nX^*)^{-1} \\ 0 & 0 \end{bmatrix}. \tag{20A.11}$$

Then $b_0 = \bar{Y}$ and $b_1 = 0$, which is the same solution as Choice 1.

Choice 4. Let

$$(\mathbf{X}'\mathbf{X})^- = \begin{bmatrix} 0 & 0 \\ (nX^*)^{-1} & 0 \end{bmatrix}. \tag{20A.12}$$

Then $b_0 = 0$, $b_1 = \bar{Y}/X^*$, the same solution as Choice 2.

We note two features from this (somewhat limited) example:

1. Any specific choice of $(\mathbf{X}'\mathbf{X})^-$ merely leads to one of the infinity of solutions provided by (20A.7). (This point is true in the general case also.)

2. Only the two most obvious solutions arise, those assuming that $b_0 = \overline{Y}$ or $b_0 = 0$. To get other solutions [which all still satisfy (20A.7)], other choices of $(\mathbf{X}'\mathbf{X})^-$ are needed. It is, however, pointless to follow this up, as other choices essentially ask us to make other assumptions about the b's. Obviously we could apply any assumption we chose *directly* to the general solution (20A.7) and not use a generalized inverse at all.

What Should One Do?

Our overall recommendation is that using a generalized inverse for a practical regression problem is usually a waste of time. Four alternative choices are:

1. Keep the original data but modify the model to make the new $\mathbf{X}'\mathbf{X}$ nonsingular.
2. Keep the original model but get more data to make the new $\mathbf{X}'\mathbf{X}$ nonsingular.
3. Keep both data and model and decide to implement sensible linear restrictions on the parameters to make the new $\mathbf{X}'\mathbf{X}$ nonsingular. [Computer programs typically set equal to zero all parameters associated with the later (in sequence, as given to the computer) dependent columns of the original $\mathbf{X}$ matrix.]
4. Add nonlinear restrictions and solve the least squares problem subject to them. Ridge regression is an example of this.

Choice 3 will often be the most practical.

EXERCISES FOR CHAPTER 20

A. Fit the model $Y = \beta_0 + \beta_1 X + \epsilon$ by least squares to the four data points $(X, Y) = (1, 1.4)$, $(2, 2.2)$, $(3, 2.3)$, $(4, 3.1)$.

1. Write down the SS function $S(\beta_0, \beta_1)$.
2. Find the least squares estimate $\mathbf{b}$.
3. Find the vector $\hat{\mathbf{Y}}$ of fitted values and $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}$.
4. Find the projection matrix $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.
5. For the sample space (E_4 space), make a plot, as best you can, showing what you have done, and identifying all relevant details.
6. Write down an analysis of variance table "appropriate for checking $H_0: \beta_0 = \beta_1 = 0$ versus H_1: "not so," for this regression situation. Specify the test statistic and get its observed value. Identify what this means in your figure also.
7. Repeat 6 for $H_0: \beta_0 = 1, \beta_1 = 0.5$.
8. The two vectors of $\mathbf{X}$ and the four vectors of the least squares projection matrix $\mathbf{P}$ span the same space. What specific linear combinations of the two vectors of $\mathbf{X}$ will give $\mathbf{P}$?
9. What specific linear combinations of the four vectors of $\mathbf{P}$ will give $\mathbf{X}$?
10. The vectors of $\mathbf{X}$ span the estimation space. Write $\mathbf{X} = (\mathbf{X}_0, \mathbf{X}_1)$, where $\mathbf{X}_0 = (1, 1, 1, 1)'$. Find $\mathbf{X}_{1.0}$, a vector orthogonal to $\mathbf{X}_0$, such that $(\mathbf{X}_0, \mathbf{X}_{1.0})$ also spans the estimation space.
11. Hence (see 10) or otherwise, find a suitable sum of squares and F-value for testing $H_0: \beta_1 = 0$, irrespective of the value of β_0.

B. Consider (very carefully) the least squares regression problem with model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where

$$\mathbf{X} = \begin{matrix} \\ 1 \\ 1 \\ 1 \\ 1 \end{matrix} \begin{matrix} \mathbf{1} & \mathbf{X}_1 & \mathbf{X}_2 \\ \end{matrix} \begin{bmatrix} 1 & -3 & 1 \\ 1 & -1 & 7 \\ 1 & 1 & 13 \\ 1 & 3 & 19 \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}, \quad \mathbf{Y} = \begin{bmatrix} 4 \\ 7 \\ 2 \\ 3 \end{bmatrix}.$$

1. Write down the normal equations.
2. Find a general solution to the normal equations.
3. Determine $\hat{\mathbf{Y}}$.
4. Make a few appropriate comments.

C. To be estimable, a linear combination $\mathbf{c}'\boldsymbol{\beta}$ (say) of the elements of $\boldsymbol{\beta}$ must have a $\mathbf{c}'$ vector that is a linear combination of the rows of $\mathbf{X}$. If $\mathbf{c}'$ cannot be expressed that way, $\mathbf{c}'\boldsymbol{\beta}$ is not estimable. If the regression is of full rank, all $\mathbf{c}'\boldsymbol{\beta}$ are estimable because $\boldsymbol{\beta}$ itself can be uniquely estimated. If the regression is not full rank, then some $\mathbf{c}'\boldsymbol{\beta}$ are estimable and some are not. With this in mind, consider the following:

Four observations of Y are taken at each of $X = -1,0,1$ with the intention of fitting the cubic model $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$ via least squares, under the usual error assumptions.

1. What is the projection matrix $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'$.
2. Is $\mathbf{P}$ unique? Is $\mathbf{PY}$ unique?
3. Which of the following are estimable?

$$\beta_0 + \beta_2, \quad \beta_1, \quad \beta_0 - \beta_1, \quad \beta_1 + \beta_3, \quad \beta_0 + \beta_1 + \beta_2 + \beta_3 + \beta_4.$$

D. Consider the least squares fit $Y = \beta_1 X_1 + \beta_2 X_2 + \epsilon$ (no intercept) to the data $(X_1, X_2, Y) = (1, 2, 19)$, $(2, 1, 13)$, and $(0, 0, 16)$. Using axes (U_1, U_2, U_3), say, for three-dimensional space, do the following:

1. Draw a diagram showing the basic least squares fit. On this diagram, name the various spaces and explain how they are defined, and label any points with actual coordinates.
2. Show what the numbers in the ANOVA table mean in your figure.
3. The regression is not a significant one. What feature(s) of the data make(s) it so?
4. Is there anything "special" about the estimation space?
5. Evaluate $\mathbf{X}_{2\cdot 1}$ and draw a new diagram explaining what the ANOVA numbers mean in this diagram. $\mathbf{X}_{2\cdot 1}$ is the part of $\mathbf{X}_2$ orthogonal to $\mathbf{X}_1$.
6. Suppose that, instead of the number 16, we substitute 2. Geometrically, what does that do?
7. Provide a new ANOVA table and a new F-value for the situation when the number 16 is replaced by 2.

E. 1. Find the (unique) projection matrix $\mathbf{P}$ onto a space Ω spanned by the vectors of $\mathbf{A}$, where

$$\mathbf{A} = \begin{bmatrix} 1 & -3 & 1 \\ 1 & -1 & -1 \\ 1 & 1 & -1 \\ 1 & 3 & 1 \end{bmatrix}.$$

2. Find a basis for the space orthogonal to Ω, that is, find vectors that span that orthogonal space.
3. Find the projections into Ω of all the vectors you gave in (2), and specify the space spanned by the $\mathbf{P}$ you gave in (1).

F. Which of the matrices below are generalized inverses of the matrix $\mathbf{11}'$, namely,

$$\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}?$$

1. $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$.

2. $\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$.

3. $\begin{bmatrix} 0.25 & 0.25 \\ 0.25 & 0.25 \end{bmatrix}$.

4. $\begin{bmatrix} 0.1 & 0.2 \\ 0.3 & 0.4 \end{bmatrix}$.

G. A straight line $Y = \beta_0 + \beta_1 X + \epsilon$ is to be fitted to the data below. Show a "tree diagram" of the allocation of the 10 degrees of freedom and find 10 orthogonal vectors that span the whole space, saying which correspond to your degrees of freedom split-up, and so divide the whole space up into three distinct orthogonal spaces.

Y	1	X
Y_1	1	−2
Y_2	1	−2
Y_3	1	−1
Y_4	1	−1
Y_5	1	0
Y_6	1	0
Y_7	1	1
Y_8	1	1
Y_9	1	2
Y_{10}	1	2

H. If $\mathbf{P}$ is a projection matrix for a regression with a β_0 in the model, its row and column sums should all be 1. Explain geometrically why this must *obviously* be true?