CHAPTER 16

Ill-Conditioning in Regression Data

16.1. INTRODUCTION

Suppose we wish to fit the model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. The solution $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ would usually be sought. However, if $\mathbf{X}'\mathbf{X}$ is singular, we cannot perform the inversion and the normal equations do not have a unique solution. (An infinity of solutions exists instead.) When this happens, it stems from the fact that there is at least one linear combination of the columns of the $\mathbf{X}$ matrix that is zero. Or to put in another way, at least one column of $\mathbf{X}$ is linearly dependent on (i.e., is a linear combination of) the other columns. We would say that collinearity (or *multicollinearity*) exists among the columns of $\mathbf{X}$.

Confusingly, the words collinearity and multicollinearity are also often used in regression when there is a "near dependency" in the columns of $\mathbf{X}$. (What "near" means is also a problem.) This is why the situation when $\det(\mathbf{X}'\mathbf{X}) = 0$ is sometimes called "exact collinearity" or "exact multicollinearity." The adjective "exact" is needed only because of the weakening in the use of the nouns. In general, we assume that the word "multicollinearity" means a "near dependence" in the $\mathbf{X}$ columns. One can also say that the data are *ill-conditioned* in such a case. (Alternative phrases are *poorly conditioned* or *badly conditioned.*)

Ill-conditioning is undesirable in regression analysis. It usually leads to unreliable estimates of the regression coefficients, which then have large variances and covariances. Whether the data are well-conditioned or ill-conditioned depends on the way the data relate *to the specific model under consideration.* Data are often collected with a particular model in mind, in fact.

A Simple Example

Consider the data:

X	4	4	7	7	7.1	7.1
Y	19	20	37	39	36	38

From the plot in Figure 16.1 we see that, although the data are at three "distinct" locations, $X = 4$, 7, and 7.1, the locations 7 and 7.1 are relatively very close, compared to the gap between $X = 4$ and 7. With three X-sites, we usually would be able to fit a quadratic model $Y = \beta_0 + \beta_1 X + \beta_{11} X^2 + \epsilon$. If we try to do this here, we obtain normal equations $\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{Y}$ of the form

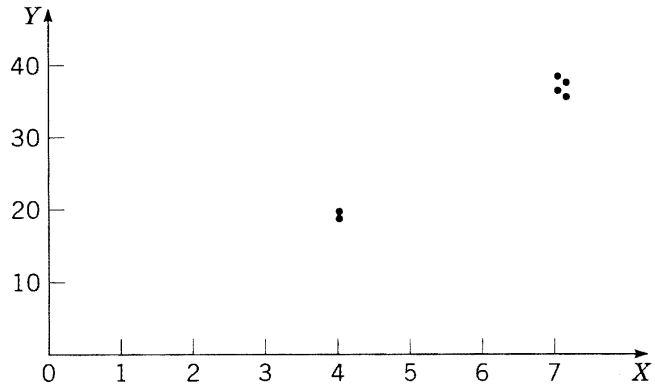

Figure 16.1. A simple example of data ill-conditioned for fitting a quadratic model $Y = \beta_0 + \beta_1 X + \beta_{11} X^2 + \epsilon$.

$$\begin{bmatrix} 6 & 36.2 & 230.82 \\ 36.2 & 230.82 & 1529.82 \\ 230.82 & 1529.82 & 10396.3362 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_{11} \end{bmatrix} = \begin{bmatrix} 189 \\ 1213.4 \\ 8078.34 \end{bmatrix}.$$

The $\mathbf{X}'\mathbf{X}$ matrix is close to singular. (One well-known regression program simply rejects entry of the X^2 term, in fact.) The fitted equation is

$$\hat{Y} = \underset{(113)}{-151} + \underset{(44)}{63.5X} \underset{(4.0)}{-5.2X^2}$$

with very large standard errors shown in the parentheses. Obviously the regression is meaningless, this arising from the strong near-dependence in the columns of $\mathbf{X}$. A look at Figure 16.1 makes it clear that a lot of different quadratic fits will pick up most of the variation here.

The data are well explained ($R^2 = 0.984$) by the least squares line $\hat{Y} = -4.03 + 5.89X$, with standard errors for the coefficients of 2.3 and 0.38, respectively. So the data are *not* inadequate for a straight line fit, but they are completely inadequate for a quadratic fit. Putting this another way, we can say that the data are well-conditioned for a straight line model fit, but very ill-conditioned for a quadratic model fit. Ill-conditioning in data is thus not just an attribute of the data alone. The level of conditioning of data depends on the model to which the data are supposed to relate. (One of the basic aspects of designing good experiments is, naturally enough, assuring that the data are adequate for actually estimating the model being considered for them. When the columns of $\mathbf{X}$ are orthogonal, the best possible conditioning is attained, and $\mathbf{X}'\mathbf{X}$ is a diagonal matrix.)

Demonstrating Dependence in X Via Regression

To show the dependence of the three columns in our example, we can regress any one column of $\mathbf{X}$ on all the others. If we use X^2 as the "response," for example, we get a fitted equation $\hat{X}^2 = -28.217(1) + 11.053(X)$, which accounts for over 99.99% of the variation about the mean of the X^2 values. This equation thus holds almost exactly and exhibits the near linear dependence of the three columns 1, X, and X^2.

We next discuss the concepts of *centering* and *scaling* data and then discuss how multicollinearity might be detected in a set of data.

16.2. CENTERING REGRESSION DATA

Suppose we have n observations on a response Y and r predictors $X_1, X_2, \ldots, X_r$. For the moment, let us assume that the X's are separately measured, unconnected variables (e.g., one X is not the square or logarithm of another, or a cross-product of two others). Suppose the model of interest is (for $u = 1, 2, \ldots, n$)

$$Y_u = \beta_0 + \beta_1 X_{1u} + \beta_2 X_{2u} + \cdots + \beta_r X_{ru} + \epsilon. \tag{16.2.1}$$

The $\mathbf{X}$ matrix for this fit consists of the columns:

$X_0 = 1$	X_1	X_2	$\cdots$	X_r
1	X_{11}	X_{21}	$\cdots$	X_{r1}
1	X_{12}	X_{22}	$\cdots$	X_{r2}
$\cdots$				
1	X_{1n}	X_{2n}	$\cdots$	X_{rn}
n	ΣX_{1u}	ΣX_{2u}		ΣX_{ru} (summed over $u = 1, 2, \ldots, n$)
1	$\overline{X}_1$	$\overline{X}_2$	$\cdots$	$\overline{X}_r$

(16.2.2)

with sums and averages

The $\mathbf{X}'\mathbf{X}$ matrix and $\mathbf{X}'\mathbf{Y}$ vector for this model are (all sums are over $u = 1, 2, \ldots, n$)

$$\begin{bmatrix} n & \Sigma X_{1u} & \Sigma X_{2u} & \cdots & \Sigma X_{ru} \\ \Sigma X_{1u} & \Sigma X_{1u}^2 & \Sigma X_{1u}X_{2u} & \cdots & \Sigma X_{1u}X_{ru} \\ \vdots & \vdots & \vdots & & \vdots \\ \Sigma X_{ru} & \Sigma X_{ru}X_{1u} & \Sigma X_{ru}X_{2u} & \cdots & \Sigma X_{ru}^2 \end{bmatrix} \quad \begin{bmatrix} \Sigma Y_u \\ \Sigma X_{1u}Y_u \\ \vdots \\ \Sigma X_{ru}Y_u \end{bmatrix} \tag{16.2.3}$$

The $\mathbf{X}'\mathbf{X}$ matrix in (16.2.3) will be singular if and only if the columns of (16.2.2) are linearly dependent, that is, if and only if there exist constants $c_0, c_1, \ldots, c_r$, say, not all zero, such that

$$c_0 + c_1 X_{1u} + c_2 X_{2u} + \cdots + c_r X_{ru} = 0 \tag{16.2.4}$$

for every $u = 1, 2, \ldots, n$. Or we can use vectors of (16.2.2) to write the same equation as

$$c_0\mathbf{1} + c_1\mathbf{X}_1 + c_2\mathbf{X}_2 + \cdots + c_r\mathbf{X}_r = 0. \tag{16.2.5}$$

Centering

We can rewrite the model (16.2.1) as

$$\begin{aligned} Y_u &= (\beta_0 + \beta_1\overline{X}_1 + \beta_2\overline{X}_2 + \cdots + \beta_r\overline{X}_r) \\ &\quad + \beta_1(X_{1u} - \overline{X}_1) + \cdots + \beta_r(X_{ru} - \overline{X}_r) + \epsilon, \qquad (16.2.6) \\ &= \beta_0' + \beta_1(X_{1u} - \overline{X}_1) + \cdots + \beta_r(X_{ru} - \overline{X}_r) + \epsilon, \qquad (16.2.7) \end{aligned}$$

where $n\overline{X}_i = \Sigma_u X_{iu}$ and where

$$\beta_0' = \beta_0 + \beta_1\overline{X}_1 + \cdots + \beta_r\overline{X}_r. \tag{16.2.8}$$

The **X** matrix for the model (16.2.7) consists of columns

$$\begin{matrix} 1 & (X_{11} - \overline{X}_1) & \cdots & (X_{r1} - \overline{X}_r) \\ 1 & (X_{12} - \overline{X}_1) & \cdots & (X_{r2} - \overline{X}_r) \\ \vdots & \vdots & & \vdots \\ 1 & (X_{1n} - \overline{X}_1) & \cdots & (X_{rn} - \overline{X}_r) \end{matrix} \tag{16.2.9}$$

Note that the first column is orthogonal to all the others now, and that the $\mathbf{X}'\mathbf{X}$ matrix and $\mathbf{X}'\mathbf{Y}$ vector for this fit are

$$\begin{bmatrix} n & 0 & 0 & \cdots & 0 \\ 0 & S_{11} & S_{12} & \cdots & S_{1r} \\ 0 & S_{21} & S_{22} & \cdots & S_{2r} \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & S_{r1} & S_{r2} & \cdots & S_{rr} \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} \Sigma Y_u \\ \Sigma(X_{1u} - \overline{X}_1)Y_u \\ \Sigma(X_{2u} - \overline{X}_2)Y_u \\ \vdots \\ \Sigma(X_{ru} - \overline{X}_r)Y_u \end{bmatrix}. \tag{16.2.10}$$

Here $S_{ij} = \Sigma(X_{iu} - \overline{X}_i)(X_{ju} - \overline{X}_j)$ and all summations are over $u = 1, 2, \ldots, n$ unless otherwise stated. The first normal equation thus immediately gives $b_0' = \overline{Y}$. Moreover, the remaining normal equations have the solution

$$\begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_r \end{bmatrix} = \begin{bmatrix} S_{11} & S_{12} & \cdots & S_{1r} \\ S_{21} & S_{22} & \cdots & S_{2r} \\ \vdots & \vdots & & \vdots \\ S_{r1} & S_{r2} & \cdots & S_{rr} \end{bmatrix}^{-1} \begin{bmatrix} \Sigma(X_{1u} - \overline{X}_1)(Y_u - \overline{Y}) \\ \Sigma(X_{2u} - \overline{X}_2)(Y_u - \overline{Y}) \\ \vdots \\ \Sigma(X_{ru} - \overline{X}_r)(Y_u - \overline{Y}) \end{bmatrix}. \tag{16.2.11}$$

Where do the extra $\overline{Y}$'s come from as we go from (16.2.10) to (16.2.11)? They come from the fact that

$$\Sigma(X_{iu} - \overline{X}_i)Y_u = \Sigma(X_{iu} - \overline{X}_i)(Y_u - \overline{Y}) \tag{16.2.12}$$

because the term we have subtracted is $\overline{Y}\Sigma(X_{iu} - \overline{X}_i) = 0$. All of this implies that, if we fit the model in the form

$$Y_u - \overline{Y} = \beta_1(X_{1u} - \overline{X}_1) + \cdots + \beta_r(X_{ru} - \overline{X}_r) + \epsilon' \tag{16.2.13}$$

to give estimates $b_1, b_2, \ldots, b_r$, and if [see (16.2.8) and the fact that $b_0' = \overline{Y}$] we recover b_0 from

$$b_0 = \overline{Y} - b_1\overline{X}_1 - b_2\overline{X}_2 - \cdots - b_r\overline{X}_r, \tag{16.2.14}$$

the estimates will be *exactly* the same as we would have obtained from fitting (16.2.1).

The data $(Y_u - \overline{Y}), (X_{1u} - \overline{X}_1), \ldots, (X_{ru} - \overline{X}_r)$ are said to be *centered around their average values*, or often just *centered*. (One can also "center" about values other than the mean, but we shall not use the term in that way.)

If the model contains terms other than simple, separate X's, for example, if it contains X_1^2 or X_1X_2 or $\ln X_1$, the same arguments as above apply except that the centering takes place around the column averages $\overline{X_1^2}$ (the average of the X_1^2 values) or $\overline{X_1X_2}$ (the average of the X_1X_2 values) or $\overline{\ln X_1}$ (the average of the $\ln X_1$ values).

Singularity and Centering

Suppose the centered X's are linearly dependent, that is, there exist constants $c_1, c_2, \ldots, c_r$ not all zero such that

$$c_1(X_{1u} - \overline{X}_1) + \cdots + c_r(X_{ru} - \overline{X}_r) = 0. \tag{16.2.15}$$

This implies that

$$-(c_1\overline{X}_1 + \cdots + c_r\overline{X}_r) + c_1X_{1u} + \cdots + c_rX_{ru} = 0, \tag{16.2.16}$$

which, if we denote the first term by c_0, is (16.2.4). Thus if the centered X's are dependent, the uncentered X's are also dependent. The reverse of this is not necessarily true, however. For example, suppose c_0 and c_1 are nonzero and

$$c_0 + c_1X_{1u} = 0 \tag{16.2.17}$$

is the only relationship among the $\mathbf{X}$-columns. This implies that all $X_{1u} = \overline{X}_1$ so that $X_{iu} - \overline{X} = 0$ for all u. Examining columns of centered $(X_{iu} - \overline{X}_i)$'s for linear relationships will thus be useless in detecting (16.2.17); recall that no other relationships exist by our assumption. More generally, two vectors $\mathbf{X}_1$ and $\mathbf{X}_2$, both close to the $\mathbf{1}$ vector, and thus to each other, could be such that the "residual" vectors $\mathbf{X}_1 - \overline{X}_1\mathbf{1}$ and $\mathbf{X}_2 - \overline{X}_2\mathbf{1}$ were orthogonal. Again, examination of the centered vectors would reveal no dependency relationship. For (a lot) more on this, see Chapter 6 (especially Section 6.3) of Belsley (1991).

The basic message from this section is that assessment of collinearity should be made on *un*centered X-data, including the $\mathbf{1}$ column, not centered data.

16.3. CENTERING AND SCALING REGRESSION DATA

Suppose we not only center the variables as in (16.2.9) but divide each column there, except the first, by the sum of squares of its members, namely, by $S_{11}^{1/2}, S_{22}^{1/2}, \ldots, S_{rr}^{1/2}$, respectively. Let us also do the same to the centered Y's (without worrying about the effect of this on the error structure of the new responses). If we write

$$x_{iu} = \frac{X_{iu} - \overline{X}_i}{S_{ii}^{1/2}}, \quad i = 1, 2, \ldots, r, \quad \text{and} \quad y_u = \frac{Y_u - \overline{Y}}{S_{YY}^{1/2}}, \tag{16.3.1}$$

where $S_{YY} = \Sigma(Y_u - \overline{Y})^2$, and substitute these in (16.2.13), we shall obtain

$$yS_{YY}^{1/2} = \beta_1S_{11}^{1/2}x_1 + \cdots + \beta_rS_{rr}^{1/2}x_r + \epsilon'. \tag{16.3.2}$$

Dividing through by $S_{YY}^{1/2}$, again without worrying about the effect of this on the error structure, gives

$$y = \alpha_1x_1 + \alpha_2x_2 + \cdots + \alpha_rx_r + \epsilon'', \tag{16.3.3}$$

where $\alpha_1 = \beta_1(S_{11}/S_{YY})^{1/2}, \ldots, \alpha_r = \beta_r(S_{rr}/S_{YY})^{1/2}$ are now coefficients to be estimated from the manipulated data $y, x_1, \ldots, x_r$, by least squares. The form of the normal equations for such a fit are

$$\begin{bmatrix} 1 & c_{12} & c_{13} & \cdots & c_{1r} \\ c_{21} & 1 & c_{23} & \cdots & c_{2r} \\ \vdots & \vdots & \vdots & & \vdots \\ c_{r1} & c_{r2} & c_{r3} & \cdots & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_r \end{bmatrix} = \begin{bmatrix} c_{1Y} \\ c_{2Y} \\ \vdots \\ c_{rY} \end{bmatrix}, \tag{16.3.4}$$

where

$$c_{ij} = \frac{S_{ij}}{(S_{ii}S_{jj})^{1/2}}, \qquad c_{jY} = \frac{S_{jY}}{(S_{jj}S_{YY})^{1/2}} \tag{16.3.5}$$

are correlations between (X_i, X_j) and (X_j, Y), respectively, with $S_{jY} = \Sigma_u(X_{ju} - \overline{X}_j)(Y_u - \overline{Y})$, and where $a_1, a_2, \ldots, a_r$ are estimates of $\alpha_1, \alpha_2, \ldots, \alpha_r$. Equations (16.3.4) are said to be *the normal equations in correlation form*. The $r \times r$ square matrix on the left of (16.3.4) is the correlation matrix of the predictors and will be referred to as **C** below.

It can be shown that, if we take the solutions $a_1, a_2, \ldots, a_r$ from (16.3.4), and calculate

$$\begin{aligned} b_1 &= a_1\left(\frac{S_{YY}}{S_{11}}\right)^{1/2}, \\ &\vdots \\ b_r &= a_r\left(\frac{S_{YY}}{S_{rr}}\right)^{1/2}, \\ b_0 &= \overline{Y} - b_1\overline{X}_1 - \cdots - b_r\overline{X}_r, \end{aligned} \tag{16.3.6}$$

we recover the appropriate solution to the original least squares problem. Why would we choose this route, however? Here are two reasons:

1. Workers in some fields argue that the coefficients $a_1, a_2, \ldots a_r$ are more meaningful to them in interpreting the regression. While our own experiences with physical science data do not confirm this, we would not argue against it. Readers can form their own opinions, depending on their field of interest. For some connected references, see Bring (1994).
2. Centering and scaling of the predictor variables is routine in a basic form of ridge regression discussed in Chapter 17. In that application, the response variable Y is usually not centered.

Centering and Scaling and Singularity

The inverse of the correlation matrix **C** on the left of (16.3.4) involves the determinant of this matrix. This determinant value lies between 0, when an exact dependence between the centered columns of the **X** matrix exists, as in (16.2.15) and (16.2.16), and 1, when all the correlations are zero. The size of the determinant is thus some guide to the extent of overall collinearity in the regression problem, although we are unsure how to calibrate it.

The sizes of individual correlations provide no dependable guide to pairwise dependence of **X** columns in general, unfortunately, because of the effect of centering. As noted at the end of Section 6.2, two vectors that are nearly linearly dependent can be little correlated. Conversely, two highly correlated vectors can be orthogonal, as

the following extreme example, due to Belsley, shows: $\mathbf{X}_1 = (1, 1, 1, 3^{1/2})'$, $\mathbf{X}_2 = (-1, -1, -1, 3^{1/2})'$ are orthogonal vectors but perfectly correlated! So a high correlation in $\mathbf{C}$ is perhaps best regarded as an invitation to check in more detail. More generally, pairwise correlations typically provide little or no clue to more complicated dependencies between several columns of $\mathbf{X}$.

16.4. MEASURING MULTICOLLINEARITY

Several criteria have been used to check on multicollinearity. We refer the reader to Belsley (1991, particularly Sections 1.3, 1.4, 2.3, and 5.2), for an exceptionally detailed discussion, and concentrate here on the highlights adequate for an appreciation of what to do.

Suggestion 1. Check if some regression coefficients have the wrong sign, based on prior knowledge.

Suggestion 2. Check if predictors anticipated to be important based on prior knowledge have regression coefficients with small t-statistics.

Suggestion 3. Check if deletion of a row or column of the $\mathbf{X}$ matrix produces surprisingly large changes in the fitted model.

Suggestion 4. Check the correlations between all pairs of predictor variables to see if any are surprisingly high. These correlations can be obtained by viewing the correlation matrix $\mathbf{C}$ shown on the left of (16.3.4).

Suggestion 5. Examine the *variance inflation factors*, usually abbreviated to VIF. The name is due to D. W. Marquardt. When the model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_r X_r + \epsilon \tag{16.4.1}$$

is fitted by least squares, the variances of the estimates $b_1, b_2, \ldots, b_r$ are

$$V(b_i) = \mathrm{VIF}_i(\sigma^2/S_{ii}), \qquad i = 1, 2, \ldots, r \tag{16.4.2}$$

where $S_{ii} = \sum_{u=1}^{n} (X_{iu} - \overline{X}_i)^2$ is the usual corrected sum of squares of the $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots X_{in})'$ column. If an $\mathbf{X}_i$ column is orthogonal to all other columns of the $\mathbf{X}$ matrix, $\mathrm{VIF}_i = 1$. Thus VIF_i is a measure of how much σ^2/S_{ii} is inflated by the relationship of other columns of $\mathbf{X}$ to the $\mathbf{X}_i$ column. [As always, σ^2 would be replaced by s^2 to get an estimated $\hat{V}(b_i)$ and the square root of the result would be the appropriate standard error $\mathrm{se}(b_i)$.]

The VIFs are the diagonal elements of the inverse $\mathbf{C}^{-1}$ of the correlation matrix $\mathbf{C}$. The VIFs can be defined specifically in the following way. Suppose that R_i^2 is the multiple correlation coefficient obtained when the ith predictor variable column $\mathbf{X}_i$, $i = 1, 2, \ldots, r$, is regressed against all the remaining predictors $\mathbf{X}_j$ with $j \neq i$. (*Note*: The intercept column $\mathbf{X}_0 = \mathbf{1}$ is also in the regression but does not feature in R_i^2. Using centered X's also gives the same R_i^2 values, of course.) Then it can be shown that

$$\mathrm{VIF}_i = (1 - R_i^2)^{-1}. \tag{16.4.3}$$

Suggestion 6. Belsley's suggestion, described in Section 16.5.

Recommendations on Suggestions 1–6

1–3. These may or may not arise as a result of multicollinearity. Thus they are unreliable indicators of its presence or absence.

4. Each high pairwise correlation may indicate a (near) linear relationship between the pair of predictors involved or may not, as described at the end of Section 16.3. Low pairwise correlations can occur when multicollinearity involves several predictors, however, so low pairwise correlations are *not* a contrary indicator. Examination of correlations, while not decisive, is probably worthwhile, because it sometimes gives clues to the behavior of selection procedures.

5. When VIF_i is large, $1 - R_i^2$ is small and R_i^2 is close to 1. A large VIF_i thus indicates that there is a (near) dependence of the ith column of $\mathbf{X}$ on the other columns of $\mathbf{X}$, but excluding the $\mathbf{1}$ column, which, by the nature of the calculation made, is excluded. Thus even if all the usual VIFs were "not large" a dependency involving the $\mathbf{1}$ vector could exist, unseen by the VIFs. This could be overcome by calculating an uncentered R^2; see Belsley (1991, pp. 28–29). Obviously, how large a VIF value has to be to be "large enough" comes back to the question of when an R_i^2 is large enough and perhaps should be thought of in that manner. In some writings, specific numerical guidelines for VIF values are seen, but they are essentially arbitrary. Each person must decide for himself or herself.

6. See Section 16.5.

What Are the Relationships?

Even if suggestions 4 and/or 5 give indications of relationships, they do not in general tell how many different relationships might exist. It would then be necessary to discover this. The number of relationships can be found by finding the eigenvalues of the $\mathbf{X}'\mathbf{X}$ matrix. Zero eigenvalues indicate that an exact relationship exists between the columns of $\mathbf{X}$, and small eigenvalues (again the question "how small?" arises!) indicate approximate relationships.

It is also possible to use regression formulas to systematically transform the $\mathbf{X}$ matrix to a form in which all columns are orthogonal. A column transformed to all zeros would indicate that the column it replaces is exactly dependent on previous (former) columns, but not *which* previous columns, unless the regression equation were examined. A column transformed to "all small near-zero values" (again, how small?) would indicate an approximate dependency. See Appendix 16A for some additional details. To invoke a (near) dependency, it would be necessary to make some judgment on how small the length of a transformed vector should be.

Belsley (1991) has discussed the difficulties of all this extensively. His conclusions are that the method of the next section has merit as a way to discover and study multicollinearity. We think he is right.

16.5. BELSLEY'S SUGGESTION FOR DETECTING MULTICOLLINEARITY

1. Determine the n by p "$\mathbf{X}$ matrix" for the regression.
2. Column-equilibrate "$\mathbf{X}$." This means divide each column of "$\mathbf{X}$" by the square root of the sum of squares of its elements. In the column-equilibrated $\mathbf{X}$, then, each column will have sum of squares unity. (In what follows, we simply use the notation $\mathbf{X}$ for the column-equilibrated form, rather than introduce another letter.)
3. Find the *singular value decomposition* of $\mathbf{X}$, that is, find matrices $\mathbf{U}(n \text{ by } p)$, $\mathbf{D}(p \text{ by } p \text{ and diagonal})$, and $\mathbf{V}(p \text{ by } p)$ such that

$$\mathbf{X} = \mathbf{UDV}' \tag{16.5.1}$$

and where

$$\mathbf{U}'\mathbf{U} = \mathbf{V}'\mathbf{V} = \mathbf{V}\mathbf{V}' = \mathbf{I}_p. \tag{16.5.2}$$

The matrix $\mathbf{D}$ = diagonal $(\mu_1, \mu_2, \ldots, \mu_p)$. The μ_i are called the *singular values* of $\mathbf{X}$ and are non-negative, although some may be zero. The matrix $\mathbf{D}^2 = \mathbf{DD}$ = diagonal $(\mu_1^2, \mu_2^2, \ldots, \mu_p^2)$ = diagonal $(\lambda_1, \lambda_2, \ldots, \lambda_p) = \mathbf{\Lambda}$, say, is the matrix of the *eigenvalues* of $\mathbf{X}'\mathbf{X}$. Note the following algebra:

$$\mathbf{X}'\mathbf{X} = \mathbf{V}\,\mathbf{D}\mathbf{U}'\mathbf{U}\mathbf{D}\mathbf{V}' = \mathbf{V}\mathbf{D}^2\mathbf{V}' \tag{16.5.3}$$

so that

$$\mathbf{X}'\mathbf{X}\mathbf{V} = \mathbf{V}\mathbf{D}^2 = \mathbf{V}\mathbf{\Lambda}. \tag{16.5.4}$$

Thus the matrix $\mathbf{V}$ consists of the p eigenvectors of $\mathbf{X}'\mathbf{X}$. These could also be obtained in the more usual manner by solving the equation in λ of degree p

$$|\mathbf{X}'\mathbf{X} - \lambda\mathbf{I}| = 0 \tag{16.5.5}$$

for the roots $\lambda_1, \lambda_2, \ldots, \lambda_p$ and then solving

$$\mathbf{X}'\mathbf{X}\mathbf{v}_i = \lambda_i\mathbf{v}_i \tag{16.5.6}$$

for a normalized (sum of squares of elements = 1) vector $\mathbf{v}_i$, $i = 1, 2, \ldots, p$. Then $\mathbf{V} = (\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_p)$. The ordering of the columns of $\mathbf{V}$ is the same as the order selected for the corresponding $\lambda_1, \lambda_2, \ldots, \lambda_p$. Computer program outputs usually list the λ's in order of decreasing absolute magnitude. If $\mathbf{X}'\mathbf{X}$ is invertible,

$$(\mathbf{X}'\mathbf{X})^{-1} = (\mathbf{V}')^{-1}\mathbf{D}^{-2}\mathbf{V}^{-1} = \mathbf{V}\mathbf{D}^{-2}\mathbf{V}'. \tag{16.5.7}$$

4. Obtain the *p condition indices*. These are defined as

$$\eta_j = \mu_{\max}/\mu_j, \qquad j = 1, 2, \ldots, p. \tag{16.5.8}$$

Thus a (relatively) near-zero μ_j will be associated with a large condition index. A zero μ_j would imply the existence of an *exact* linear relationship among the columns of $\mathbf{X}$. (Zero $\lambda_j = \mu_j^2$ values sometimes appear as negative rounding errors, even though the eigenvalues λ_j cannot be negative for $\mathbf{X}'\mathbf{X}$.) (*Note*: The ratio $\mu_{\max}/\mu_{\min}$ is called the *condition number*.)

5. Decompose the variance structure of the parameter estimates. When $\mathbf{X}'\mathbf{X}$ can be inverted, we can write

$$\sigma^{-2}\mathbf{V}(\mathbf{b}) = (\mathbf{X}'\mathbf{X})^{-1} = \mathbf{V}\mathbf{D}^{-2}\mathbf{V}' \tag{16.5.9}$$

$$= \mathbf{V}\mathbf{\Lambda}^{-1}\mathbf{V}' \tag{16.5.10}$$

TABLE 16.1. Proportional Decompositions of $V(b_i)$

Condition Index	Proportions of $V(b_1)$	$V(b_2)$	$\cdots$	$V(b_p)$
η_1	q_{11}	q_{21}	$\cdots$	q_{p1}
η_2	q_{12}	q_{22}	$\cdots$	q_{p2}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
η_p	q_{1p}	q_{2p}	$\cdots$	q_{pp}
Column sums	1	1		1

from (16.5.7). Writing $\mathbf{V} = (v_{ij})$, we see that

$$\begin{aligned}\sigma^{-2}V(b_i) &= \frac{v_{i1}^2}{\mu_1^2} + \frac{v_{i2}^2}{\mu_2^2} + \cdots + \frac{v_{ip}^2}{\mu_p^2} \\ &= (q_{i1} + q_{i2} + \cdots + q_{ip}) \sum_{j=1}^{p} \frac{v_{ij}^2}{\mu_j^2},\end{aligned} \tag{16.5.11}$$

where the q's are simply the proportions (adding to 1) of the total amount of the expression represented by $\sigma^{-2}V(b_i)$. This enables the display of Table 16.1 to be formed.

6. Examine the table (of form 16.1) to detect possible dependencies. Large condition indices η_j indicate dependencies; large proportions q_{ij} within the corresponding rows indicate the **X** columns that are candidates for the dependencies. (Once again we enter the world of "How large is large?"). Low condition indices rows are examined also. High q_{ij} in these rows indicate the *non*-involvement in dependencies of the corresponding columns of **X**.

Computational Note: Use of (16.5.1) on **X**, as recommended by Belsley, gives more accurate answers, particularly in cases where exact or very nearly exact dependencies exist. Except for such cases, the more conveniently available (e.g., in MINITAB) eigenvalue programs can be used directly to get $\mathbf{\Lambda}$ and **V** from $\mathbf{X}'\mathbf{X}$. When $(\mathbf{X}'\mathbf{X})^{-1}$ can be computed, it is convenient to evaluate [see (16.5.9) and (16.5.10)]

$$(\mathbf{X}'\mathbf{X})^{-1} - \mathbf{V}\mathbf{\Lambda}^{-1}\mathbf{V}' \tag{16.5.12}$$

as a check; it should be **0**. If it is considerably off-zero, or if $(\mathbf{X}'\mathbf{X})^{-1}$ cannot be obtained, this is a signal that exact (or very close to exact) dependencies exist.

Example 1. The original "**X** matrix" is

$$\begin{bmatrix} -74 & 80 & 18 & -56 & -112 \\ 14 & -69 & 21 & 52 & 104 \\ 66 & -72 & -5 & 764 & 1528 \\ -12 & 66 & -30 & 4096 & 8192 \\ 3 & 8 & -7 & -13276 & -26552 \\ 4 & -12 & 4 & 8421 & 16842 \end{bmatrix}. \tag{16.5.13}$$

This matrix is called "the modified Bauer matrix" by Belsley who took four columns given by F. L. Bauer (in his article on pp. 119–133 of *Handbook for Automatic*

TABLE 16.2. Proportional Decompositions of $V(b_i)$ for the "Modified Bauer Data"

Condition[a] Index	Proportions of $V(b_1)$	$V(b_2)$	$V(b_3)$	$V(b_4)$	$V(b_5)$
1	.000	.000	.000	.000	.000
1	.005	.005	.000	.000	.000
1	.001	.001	.047	.000	.000
16	.994	.994	.953	.000	.000
5799	.000	.000	.000	1.000	1.000
Column sums	1.000	1.000	1.000	1.000	1.000

[a] Rounded to the nearest integer.

Computation, Volume II: Linear Algebra, edited by J. H. Wilkinson and C. Reisch, published by Springer-Verlag, 1971) and added a fifth column equal to double the fourth. So "the $\mathbf{X}'\mathbf{X}$" is obviously singular. Dividing each entry by the appropriate root sum of squares gives 10^{-3} times the following column-equilibrated figures:

$$10^3\mathbf{X} = \begin{bmatrix} -733 & 553 & 430 & -3 & -3 \\ 139 & -477 & 501 & 3 & 3 \\ 654 & -498 & -119 & 47 & 47 \\ -119 & 456 & -716 & 252 & 252 \\ 30 & 55 & -167 & -816 & -816 \\ 40 & -83 & 95 & 518 & 518 \end{bmatrix}. \tag{16.5.14}$$

The figures are rounded for display. The equivalence of the last two columns is obvious. One could be dropped immediately, once this is noticed, but we continue with both in. The resulting display of the type shown in Table 16.1 is as in Table 16.2.

Comparison with Belsley's (1991, Exhibit 5.8, p. 148) table shows a large difference in the numerical value of the fifth condition index, undoubtedly because his computations are more accurate than ours, but no difference in any other figure nor in the message transmitted. The condition index 5799 (Belsley gets 8×10^{19}; in practical terms, both numbers are infinity!) indicates a relationship between X_4 and X_5. The next largest index 16 indicates a possible relationship between X_1, X_2, and X_3. We now pick one variable from each set, X_5 from the first, X_1 from the second, say, and regress these as "responses" against the other three variables X_2, X_3, and X_4 to show up the nature of the column dependencies or near dependencies. We obtain the exact relationship

$$X_5 = 0.0X_2 + 0.0X_3 + 2X_5$$

and the approximate relationship

$$X_1 = -0.701X_2 - 1.269X_3 - 0.0X_4.$$

Of the total (not corrected for the mean) SS variation in X_5, 100% is accounted for by the first equation; and for X_1, 98.20% is accounted for by the second equation.

Example 2. The Hald Data (see Appendix 15A). The original "**X** matrix" consists of five columns of $[1, X_1, X_2, X_3, X_4]$. After dividing each entry by the corresponding column root sum of squares we get 10^{-3} times the following figures:

$$10^3\mathbf{X} = \begin{bmatrix} 277 & 207 & 143 & 125 & 489 \\ 277 & 30 & 160 & 313 & 424 \\ 277 & 326 & 308 & 167 & 163 \\ 277 & 326 & 171 & 167 & 383 \\ 277 & 207 & 286 & 125 & 269 \\ 277 & 326 & 303 & 188 & 179 \\ 277 & 89 & 391 & 355 & 49 \\ 277 & 30 & 171 & 459 & 359 \\ 277 & 59 & 297 & 376 & 179 \\ 277 & 622 & 259 & 84 & 212 \\ 277 & 30 & 220 & 480 & 277 \\ 277 & 326 & 363 & 188 & 98 \\ 277 & 296 & 374 & 167 & 98 \end{bmatrix}. \tag{16.5.15}$$

The variances of the b's decompose into the proportions shown in Table 16.3. As always, each column sums to 1, apart from rounding error in the $V(b_3)$ column.

Obviously only one linear relationship is indicated and it links all the columns. Via regression we get the almost exact relationship $X_0 = 0.0103\,X_1 + 0.0103\,X_2 + 0.0105\,X_3 + 0.0101\,X_4$, accounting for 99.99% of the uncorrected SS of 13. Clearly this is showing up the original near-mixture relationship that leads to the fact that (nearly) $X_1 + X_2 + X_3 + X_4 = 1 \equiv 0.277X_0$ for the equilibrated columns of (16.5.15).

Example 3. The Steam Data (see Appendix 1A). We proceed directly to Table 16.4. Again each column sums to 1 (i.e., 1000 times 10^{-3}) apart from rounding error. We have dropped the decimal points by using a factor of 10^{-3} and replaced 000 by a dash to simplify the look at the table and to show an alternative display format. We recall that $X_1 = Y$ is the response, so that the factors are numbered 0 for the b_0 term, and 2–10 for the nine (other) X's. Working up from the bottom of the table, line by line, we need to check possible relations between

TABLE 16.3. Proportional Decompositions of $V(b_i)$ for the Hald Data of Appendix 15A

Condition[a] Index	Proportions of $V(b_0)$	$V(b_1)$	$V(b_2)$	$V(b_3)$	$V(b_4)$
1	0.000	0.000	0.000	0.000	0.000
3	0.000	0.010	0.000	0.003	0.000
4	0.000	0.001	0.000	0.002	0.002
10	0.000	0.057	0.003	0.046	0.001
250	1.000	0.932	0.997	0.950	0.997

[a] Rounded to the nearest integer.

TABLE 16.4. Proportional Decompositions of $V(b_i)$ for the Steam Data of Appendix 1A

Condition[a] Index	10^{-3} Times Proportions of $V(b_i)$ for i =									
	0	2	3	4	5	6	7	8	9	10
1	—	—	—	—	—	—	1	—	—	—
4	—	—	—	—	—	—	113	4	—	—
7	—	—	—	—	—	—	135	—	7	3
13	—	2	4	—	—	4	114	118	—	40
25	—	7	16	—	—	18	7	10	—	721
31	5	4	5	3	10	66	345	404	5	6
43	1	50	27	2	1	574	230	129	10	8
96	23	612	595	13	2	160	20	106	14	18
148	—	13	134	789	134	138	1	74	809	139
253	970	311	220	193	853	40	34	156	155	66

[a] Rounded to nearest integer.

$$\begin{aligned} &X_0 \text{ and } X_5 \quad (\text{Condition index } 253) \\ &X_4 \text{ and } X_9 \quad (148) \\ &X_2 \text{ and } X_3 \quad (96) \end{aligned} \tag{16.5.16}$$

The condition indices just quoted (253, 148, and 96) are all of the same order of magnitude, as will be described more fully after this example. This indicates possible *coexisting near dependencies*. For example, variables X_0 and X_5, which are linked by the strongest indicated near dependency (condition index 253), might also be involved (but masking, i.e., concealing, each other) in the other near dependencies; and similarly for the other pairs of variables. Thus, while the most simple pairwise involvements of variables (X_0, X_5), (X_4, X_9), and (X_2, X_3) *might* be all that is indicated by Table 16.3 (and we note, in passing, that $r_{49} = 0.990$ and $r_{23} = 0.944$), it is incumbent on us to investigate more complicated possibilities. We select one variable from each set (here, each pair), say, X_5, X_9, and X_2, and then regress these against all of the remaining seven X's. This leads to the fitted equations

$$X_5 = 33.011X_0 + 2.717X_3 - 0.158X_4 - 0.033X_6 - 0.025X_7 - 0.025X_8 - 0.280X_{10} \quad (99.95\%),$$

$$X_9 = -51.278X_0 + 14.962X_3 + 14.202X_4 - 0.379X_6 + 0.017X_7 + 0.134X_8 - 1.290X_{10} \quad (99.78\%),$$

and

$$X_2 = 1.947X_0 + 6.480X_3 - 0.002X_4 - 0.040X_6 - 0.008X_7 - 0.008X_8 + 0.073X_{10} \quad (99.81\%).$$

The percentages of the uncorrected SS accounted for are shown in parentheses after the equations. An examination of the usual t-statistics for the regression coefficients confirms the major entanglement within the set (X_0, X_5) and suggests that the second set should be expanded from (X_4, X_9) to (X_0, X_4, X_9), and that the third set should be expanded from (X_2, X_3) to (X_0, X_2, X_3).

Comments

In the end, the choice of which X's would be retained for regression purposes would usually be made on the basis of practicalities such as ease of measurement, value in

terms of assessing subsequent regression equations, and a desire to keep an intercept term $\beta_0 X_0$ always in the final fitted equation. We noted in Section 15.6 that a useful equation for the steam data could be obtained by using the set of predictors (X_0, X_2, X_8) or (X_0, X_2, X_6, X_8). Both equations include X_0 because it was forced to be in all the selection procedures, and neither equation includes X_5 (entangled with X_0), nor the entangled with X_0 pair (X_4, X_9) nor X_3 (entangled with X_0 and X_2). So the steam data results in Chapter 15 and in this chapter present complementary and consistent pictures.

How Large Is a "Large" Condition Index?

The assessment of when a condition index is "large" is difficult. Let us talk about the sequences in our examples:

Example 1: 1, 1, 1, 16, 5799.

Example 2: 1, 3, 4, 10, 250.

Example 3: 1, 4, 7, 13, 25, 31, 43, 96, 148, 253.

Belsley notes (1991, p. 139) that condition indices increase in order of magnitude along a scale of 1, 3, 10, 30, 100, 300, 1000, and so on. Moreover, his experiments show that the number 30 provides a reasonable threshold for indicating the presence of collinearity; that is, condition indices of 30 or more can be considered large. With this in mind, Belsley suggests (pp. 139–141) assessing a table of condition indices by beginning with the largest ones. If the largest condition index is 5 or 10, "collinearity is not really a major problem," while if it is in the range of 30–100, "then there are collinearity problems." If the largest condition index is in the range of 1000–3000, there are severe collinearity problems, and "even ones like 30 are not necessarily of major concern." Belsley also suggests that gaps "in the 10/30 progression" are of interest. Interpreting all this in our examples makes us focus on 5799 and 16 in Example 1; on 250 and 10 in Example 2; and in Example 3 (a more difficult case) to back up along 253, 148, 96, until we see that the "big" proportions in the body of the table are dying out. Thus we lose interest in the index 43. In spite of the fact that 43 might otherwise appear big, it is simply dwarfed by the larger numbers below it. Going back up the sequence too far would not be a mistake but would involve one in looking at linear combinations of X's that are not close to zero and then discarding them as possibilities.

For a more detailed discussion the reader should consult *Conditioning Diagnostics, Collinearity and Weak Data in Regression*, by D. A. Belsley, Wiley, 1991. The methods described by Belsley, and discussed in this section, can be implemented by using a package available in the SAS computing system.

APPENDIX 16A. TRANSFORMING X MATRICES TO OBTAIN ORTHOGONAL COLUMNS

The **X** matrix in a regression problem must be such that none of the columns can be expressed as a linear combination of the other columns. This effectively implies also that there must be at least as many rows not dependent on other rows, as there are parameters to estimate, or else a dependence will appear in the columns also. As an example, suppose observations Y are recorded at only three levels of X, namely, $X = a$, b, and c, but that the model $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$ is postulated. The **X** matrix takes the form

$$\begin{bmatrix} 1 & a & a^2 & a^3 \\ 1 & b & b^2 & b^3 \\ 1 & c & c^2 & c^3 \end{bmatrix}$$

and the columns are dependent since (column 4) $-$ $(a + b + c)$ (column 3) $+$ $(ab + bc + ca)$ (column 2) $-$ abc(column 1) $= 0$. To spot such a dependence in a general regression problem is often very difficult. When it exists the $\mathbf{X'X}$ matrix will always be singular and thus cannot be inverted. When the columns of the $\mathbf{X}$ matrix are almost dependent, the $\mathbf{X'X}$ matrix will be almost singular and difficulties in inversion, including large round-off errors, are likely.

One procedure that can be programmed and used as a routine check on $\mathbf{X}$ matrices consists of successively transforming the columns so that each new column is orthogonal to all previously transformed columns. If a column dependence exists we shall eventually obtain a new column that consists entirely of zeros. If the columns are nearly dependent, a new column will contain all very small numbers perhaps with some zeros. The column transformation takes the following form:

$$\begin{aligned} \mathbf{Z}_{iT} &= (\mathbf{I} - \mathbf{Z}(\mathbf{Z'Z})^{-1}\mathbf{Z'})\mathbf{Z}_i \\ &= \mathbf{Z}_i - \mathbf{Z}(\mathbf{Z'Z})^{-1}\mathbf{Z'Z}_i \end{aligned} \tag{16A.1}$$

where $\mathbf{Z}$ = the matrix of column vectors already transformed,

$\mathbf{Z}_i$ = the next column vector of $\mathbf{X}$ to be transformed, and

$\mathbf{Z}_{iT}$ = the transformed vector orthogonal to vectors already in $\mathbf{Z}$.

Note that $\mathbf{Z}_{iT}$ is actually the residual vector of $\mathbf{Z}_i$ after $\mathbf{Z}_i$ has been regressed on the columns of $\mathbf{Z}$.

The column transformation continues until a dependence or near dependence occurs, which causes $\mathbf{Z'Z}$ to be singular or so ill-conditioned that $(\mathbf{Z'Z})^{-1}$ cannot be obtained. Removal of the current last column of $\mathbf{Z}$ would allow the process to continue.

Example

We illustrate using a special case that will lead us to the orthogonal polynomials (see Chapter 22) for $n = 5$. Suppose values of Y are recorded at $X = 1, 2, 3, 4$, and 5 and the model

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon \tag{16A.2}$$

is postulated. The original $\mathbf{X}$ matrix is

$$\mathbf{X} = \begin{array}{c} \begin{matrix} 1 & X & X^2 & X^3 \end{matrix} \\ \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 4 & 8 \\ 1 & 3 & 9 & 27 \\ 1 & 4 & 16 & 64 \\ 1 & 5 & 25 & 125 \end{bmatrix} \end{array}. \tag{16A.3}$$

We set

$$\mathbf{Z}_{1T} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = \mathbf{Z} \tag{16A.4}$$

at this stage. (One column vector must be chosen to begin the process.) Choose

$$\mathbf{Z}_2 = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{bmatrix}. \tag{16A.5}$$

Then by Eq. (16A.1),

$$\begin{aligned} \mathbf{Z}_{2T} &= \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{bmatrix} - \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} (5)^{-1}(15) \\ &= \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{bmatrix} - \begin{bmatrix} 3 \\ 3 \\ 3 \\ 3 \\ 3 \end{bmatrix} = \begin{bmatrix} -2 \\ -1 \\ 0 \\ 1 \\ 2 \end{bmatrix}. \end{aligned} \tag{16A.6}$$

At this stage

$$\mathbf{Z} = \begin{bmatrix} 1 & -2 \\ 1 & -1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix} = [\mathbf{Z}_{1T}, \mathbf{Z}_{2T}] \tag{16A.7}$$

and the third column of $\mathbf{X}$ is used as $\mathbf{Z}_3$. We find that

$$\begin{aligned} (\mathbf{Z}'\mathbf{Z})^{-1} &= \begin{bmatrix} 5 & 0 \\ 0 & 10 \end{bmatrix}^{-1} = \begin{bmatrix} \frac{1}{5} & 0 \\ 0 & \frac{1}{10} \end{bmatrix}, \\ \mathbf{Z}'\mathbf{Z}_i &= \begin{bmatrix} 55 \\ 60 \end{bmatrix}, \qquad \text{so} \quad (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Z}_i = \begin{bmatrix} 11 \\ 6 \end{bmatrix}. \end{aligned} \tag{16A.8}$$

Hence

$$\mathbf{Z}_{3T} = \begin{bmatrix} 1 \\ 4 \\ 9 \\ 16 \\ 25 \end{bmatrix} - \begin{bmatrix} 11 - 12 \\ 11 - 6 \\ 11 + 0 \\ 11 + 6 \\ 11 + 12 \end{bmatrix} = \begin{bmatrix} 2 \\ -1 \\ -2 \\ -1 \\ 2 \end{bmatrix}. \tag{16A.9}$$

We leave the evaluation of $\mathbf{Z}_{4T}$ as an exercise. The final, orthogonal matrix is

$$\begin{bmatrix} 1 & -2 & 2 & -1.2 \\ 1 & -1 & -1 & 2.4 \\ 1 & 0 & -2 & 0 \\ 1 & 1 & -1 & -2.4 \\ 1 & 2 & 2 & 1.2 \end{bmatrix}. \tag{16A.10}$$

Note that the first three columns are ψ_0, ψ_1, and ψ_2—the orthogonal polynomials of zero, first, and second order for $n = 5$. The fourth column is 1.2 times ψ_3, the orthogonal polynomial of third order for $n = 5$. See Chapter 22.

The above procedure is usually known as a Gram–Schmidt orthogonalization of the columns. Another way of tackling this problem is to find the eigenvalues (or characteristic values, or latent roots—all terms mean the same thing) of $\mathbf{X}'\mathbf{X}$. See Section 16.5. Zero eigenvalues arise if linear dependencies exist among the $\mathbf{Z}$'s and small eigenvalues (small in relation to the range 0 to 1) are indicative of possible near dependencies. See, for example, p. 78 of Snee (1973). For related algorithms see Clayton (1971) (Fortran) and Farebrother (1974) (Algol 60).

EXERCISES FOR CHAPTER 16

A. Can we use the data below to get a unique fit to the model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$? If not, what model(s) can be fitted?

X_1	X_2	X_3	Y
1	−2	4	81
2	−7	11	88
4	3	5	94
7	1	13	95
8	−1	17	123

B. Can we use the data below to get a unique fit to the model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$? If not, what model(s) can be fitted?

X_1	X_2	X_3	Y
−4	1	3	7.4
3	2	−5	14.7
1	3	−4	13.9
4	4	−8	18.2
−3	5	−2	12.1
−1	6	−5	14.8

C. Can we use the data below to fit, uniquely, the model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{11} X_1^2 + \beta_{22} X_2^2 + \beta_{12} X_1 X_2 + \epsilon$? If not, what model(s) can be fitted?

X_1	X_2	Y
−1	−1	38
1	−1	45
−1	1	41
1	1	40
0	0	47
0	0	42
0	0	48

D. Consider the data below for the model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$. Center and scale the X_1, X_2, and Y columns, and write down the normal equations for the centered and scaled model. Solve these. What is the value of the determinant of the correlation matrix?

	X_1	X_2	Y
	−2	−4	12
	−1	−1	9
	0	0	9
	1	1	14
	2	4	16
Sum	0	0	60
SS	10	34	758