CHAPTER 9

MULTICOLLINEARITY

9.1 INTRODUCTION

The use and interpretation of a multiple regression model often depends explicitly or implicitly on the estimates of the individual regression coefficients. Some examples of inferences that are frequently made include the following:

1. Identifying the relative effects of the regressor variables
2. Prediction and/or estimation
3. Selection of an appropriate set of variables for the model

If there is no linear relationship between the regressors, they are said to be **orthogonal**. When the regressors are orthogonal, inferences such as those illustrated above can be made relatively easily. Unfortunately, in most applications of regression, the regressors are not orthogonal. Sometimes the lack of orthogonality is not serious. However, in some situations the regressors are nearly perfectly linearly related, and in such cases the inferences based on the regression model can be misleading or erroneous. When there are **near-linear dependencies** among the regressors, the problem of **multicollinearity** is said to exist.

This chapter will extend the preliminary discussion of multicollinearity begun in Chapter 3 and discuss a variety of problems and techniques related to this problem. Specifically we will examine the causes of multicollinearity, some of its specific effects on inference, methods of detecting the presence of multicollinearity, and some techniques for dealing with the problem.

Introduction to Linear Regression Analysis, Fifth Edition. Douglas C. Montgomery, Elizabeth A. Peck, G. Geoffrey Vining.

© 2012 John Wiley & Sons, Inc. Published 2012 by John Wiley & Sons, Inc.

9.2 SOURCES OF MULTICOLLINEARITY

We write the multiple regression model as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where $\mathbf{y}$ is an $n \times 1$ vector of responses, $\mathbf{X}$ is an $n \times p$ matrix of the regressor variables, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown constants, and ε is an $n \times 1$ vector of random errors, with $\varepsilon_i \sim \text{NID}(0, \sigma^2)$. It will be convenient to assume that the regressor variables and the response have been centered and scaled to unit length, as in Section 3.9. Consequently, $\mathbf{X'X}$ is a $p \times p$ matrix of correlations[†] between the regressors and $\mathbf{X'y}$ is a $p \times 1$ vector of correlations between the regressors and the respouse.

Let the jth column of the $\mathbf{X}$ matrix be denoted $\mathbf{X}_j$, so that $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_p]$. Thus, $\mathbf{X}_j$ contains the n levels of the jth regressor variable. We may formally define multicollinearity in terms of the linear dependence of the columns of $\mathbf{X}$. The vectors $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_p$ are linearly dependent if there is a set of constants $t_1, t_2, \ldots, t_p$, not all zero, such that[‡]

$$\sum_{j=1}^{p} t_j \mathbf{X}_j = \mathbf{0} \tag{9.1}$$

If Eq. (9.1) holds exactly for a subset of the columns of $\mathbf{X}$, then the rank of the $\mathbf{X'X}$ matrix is less than p and $(\mathbf{X'X})^{-1}$ does not exist. However, suppose that Eq. (9.1) is approximately true for some subset of the columns of $\mathbf{X}$. Then there will be a near-linear dependency in $\mathbf{X'X}$ and the problem of multicollinearity is said to exist. Note that multicollinearity is a form of ill-conditioning in the $\mathbf{X'X}$ matrix. Furthermore, the problem is one of degree, that is, every data set will suffer from multicollinearity to some extent unless the columns of $\mathbf{X}$ are orthogonal ($\mathbf{X'X}$ is a diagonal matrix). Generally this will happen only in a designed experiment. As we shall see, the presence of multicollinearity can make the usual least-squares analysis of the regression model dramatically inadequate.

There are four primary **sources of multicollinearity**:

1. The data collection method employed
2. Constraints on the model or in the population
3. Model specification
4. An overdefined model

It is important to understand the differences among these sources of multicollinearity, as the recommendations for analysis of the data and interpretation of the resulting model depend to some extent on the cause of the problem (see Mason, Gunst, and Webster [1975] for further discussion of the source of multicollinearity).

[†] It is customary to refer to the off-diagonal elements of $\mathbf{X'X}$ as correlation coefficients, although the regressors are not necessarily random variables.

[‡] If the regressors are not centered, then 0 in Eq. (9.1) becomes a vector of constants m, not all necessarily equal to 0.

The **data collection method** can lead to multicollinearity problems when the analyst samples only a subspace of the region of the regressors defined (approximately) by Eq. (9.1). For example, consider the soft drink delivery time data discussed in Example 3.1. The space of the regressor variables "cases" and "distance," as well as the subspace of this region that has been sampled, is shown in the matrix of scatterplots, Figure 3.4. Note that the sample (cases, distance) pairs fall approximately along a straight line. In general, if there are more than two regressors, the data will lie approximately along a hyperplace defined by Eq. (9.1). In this example, observations with a small number of cases generally also have a short distance, while observations with a large number of cases usually also have a long distance. Thus, cases and distance are positively correlated, and if this positive correlation is strong enough, a multicollinearity problem will occur. Multicollinearity caused by the sampling technique is not inherent in the model or the population being sampled. For example, in the delivery time problem we could collect data with a small number of cases and a long distance. There is nothing in the physical structure of the problem to prevent this..

Constraints on the model or in the population being sampled can cause multicollinearity. For example, suppose that an electric utility is investigating the effect of family income (x_1) and house size (x_2) on residential electricity consumption. The levels of the two regressor variables obtained in the sample data are shown in Figure 9.1. Note that the data lie approximately along a straight line, indicating a potential multicollinearity problem. In this example a physical constraint in the population has caused this phenomenon, namely, families with higher incomes generally have larger homes than families with lower incomes. When physical constraints such as this are present, multicollinearity will exist **regardless** of the sampling method employed. Constraints often occur in problems involving production or chemical processes, where the regressors are the components of a product, and these components add to a constant.

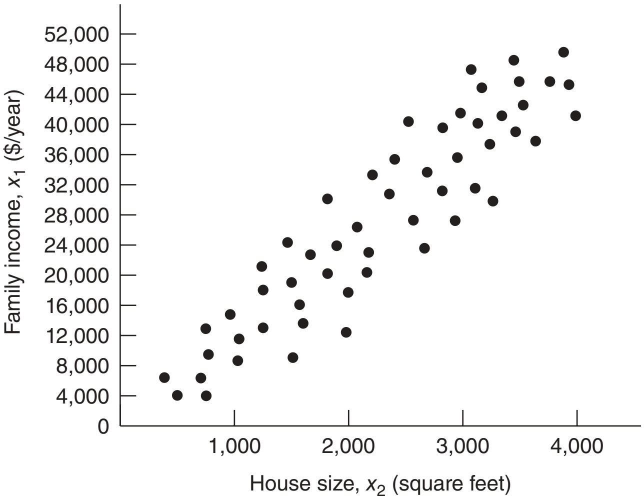

Figure 9.1 Levels of family income and house size for a study on residential electricity consumption.

Multicollinearity may also be induced by the **choice of model**. For example, we know from Chapter 7 that adding polynomial terms to a regression model causes ill-conditioning in $\mathbf{X'X}$. Furthermore, if the range of x is small, adding an x^2 term can result in significant multicollinearity. We often encounter situations such as these where two or more regressors are nearly linearly dependent, and retaining all these regressors may contribute to multicollinearity. In these cases some subset of the regressors is usually preferable from the standpoint of multicollinearity.

An **overdefined model** has more regressor variables than observations. These models are sometimes encountered in medical and behavioral research, where there may be only a small number of subjects (sample units) available, and information is collected for a large number of regressors on each subject. The usual approach to dealing with multicollinearity in this context is to eliminate some of the regressor variables from consideration. Mason, Gunst, and Webster [1975] give three specific recommendations: (1) redefine the model in terms of a smaller set of regressors, (2) perform preliminary studies using only subsets of the original regressors, and (3) use principal-component-type regression methods to decide which regressors to remove from the model. The first two methods ignore the interrelationships between the regressors and consequently can lead to unsatisfactory results. Principal-component regression will be discussed in Section 9.5.4, although not in the context of overdefined models.

9.3 EFFECTS OF MULTICOLLINEARITY

The presence of multicollinearity has a number of potentially serious effects on the least-squares estimates of the regression coefficients. Some of these effects may be easily demonstrated. Suppose that there are ouly two regressor variables, x_1 and x_2. The model, assuming that x_1, x_2, and y are scaled to unit length, is

$$y = \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

and the least-squares normal equations are

$$(\mathbf{X'X})\hat{\boldsymbol{\beta}} = \mathbf{X'y}$$

$$\begin{bmatrix} 1 & r_{12} \\ r_{12} & 1 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} r_{1y} \\ r_{2y} \end{bmatrix}$$

where r_{12} is the simple correlation between x_1 and x_2 and r_{jy} is the simple correlation between x_j and y, $j = 1, 2$. Now the inverse of $(\mathbf{X'X})$ is

$$\mathbf{C} = (\mathbf{X'X})^{-1} = \begin{bmatrix} \dfrac{1}{1-r_{12}^2} & \dfrac{-r_{12}}{1-r_{12}^2} \\ \dfrac{-r_{12}}{1-r_{12}^2} & \dfrac{1}{1-r_{12}^2} \end{bmatrix} \tag{9.2}$$

and the estimates of the regression coefficients are

$$\hat{\beta}_1 = \frac{r_{1y} - r_{12}r_{2y}}{1 - r_{12}^2}, \quad \hat{\beta}_2 = \frac{r_{2y} - r_{12}r_{1y}}{1 - r_{12}^2}$$

If there is strong multicollinearity between x_1 and x_2, then the correlation coefficient r_{12} will be large. From Eq. (9.2) we see that as $|r_{12}| \to 1$, $\text{Var}(\hat{\beta}_j) = C_{jj}\sigma^2 \to \infty$ and $\text{Cov}(\hat{\beta}_1, \hat{\beta}_2) = C_{12}\sigma^2 \to \pm\infty$ depending on whether $r_{12} \to +1$ or $r_{12} \to -1$. Therefore, strong multicollinearity between x_1 and x_2 results in **large variances and covariances** for the least-squares estimators of the regression coefficients.[†] This implies that different samples taken at the same x levels could lead to widely different estimates of the model parameters.

When there are more than two regressor variables, multicollinearity produces similar effects. It can be shown that the diagonal elements of the $\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1}$ matrix are

$$C_{jj} = \frac{1}{1 - R_j^2}, \quad j = 1, 2, \ldots, p \tag{9.3}$$

where R_j^2 is the coefficient of multiple determination from the regression of x_j on the remaining $p - 1$ regressor variables. If there is strong multicollinearity between x_j and any subset of the other $p - 1$, regressors, then the value of R_j^2 will be close to unity. Since the variance of $\hat{\beta}_j$ is $\text{Var}(\hat{\beta}_j) = C_{jj}\sigma^2 = (1 - R_j^2)^{-1}\sigma^2$, strong multicollinearity implies that the variance of the least-squares estimate of the regression coefficient β_j is very large. Generally, the covariance of $\hat{\beta}_i$ and $\hat{\beta}_j$ will also be large if the regressors x_i and x_j are involved in a multicollinear relationship.

Multicollinearity also tends to produce least-squares estimates $\hat{\beta}_j$ that are **too large** in absolute value. To see this, consider the squared distance from $\hat{\boldsymbol{\beta}}$ to the true parameter vector $\boldsymbol{\beta}$, for example,

$$L_1^2 = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$$

The expected squared distance, $E(L_1^2)$, is

$$\begin{aligned} E(L_1^2) &= E(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \sum_{j=1}^{p} E(\hat{\beta}_j - \beta_j)^2 \\ &= \sum_{j=1}^{p} \text{Var}(\hat{\beta}_j) = \sigma^2 \text{Tr}(\mathbf{X}'\mathbf{X})^{-1} \end{aligned} \tag{9.4}$$

where the trace of a matrix (abbreviated Tr) is just the sum of the main diagonal elements. When there is multicollinearity present, some of the eigenvalues of $\mathbf{X}'\mathbf{X}$ will be small. Since the trace of a matrix is also equal to the sum of its eigenvalues, Eq. (9.4) becomes

$$E(L_1^2) = \sigma^2 \sum_{j=1}^{p} \frac{1}{\lambda_j} \tag{9.5}$$

[†] Multlcollinearity is not the only cause of large variances and covariances of regression coefficients.

where $\lambda_j > 0, j = 1, 2, \ldots, p$, are the eigenvalues of $\mathbf{X'X}$. Thus, if the $\mathbf{X'X}$ matrix is ill-conditioned because of multicollinearity, at least one of the λ_j will be small, and Eq. (9.5) implies that the distance from the least-squares estimate $\hat{\boldsymbol{\beta}}$ to the true parameters $\boldsymbol{\beta}$ may be large. Equivalently we can show that

$$E\left(L_1^2\right) = E\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)'\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) = E\left(\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}} - 2\hat{\boldsymbol{\beta}}'\boldsymbol{\beta} + \boldsymbol{\beta}'\boldsymbol{\beta}\right)$$

or

$$E\left(\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}}\right) = \boldsymbol{\beta}'\boldsymbol{\beta} + \sigma^2 \mathrm{Tr}\left(\mathbf{X'X}\right)^{-1}$$

That is, the vector $\hat{\boldsymbol{\beta}}$ is generally longer than the vector $\boldsymbol{\beta}$. This implies that the method of least squares produces estimated regression coefficients that are too large in absolute value.

While the method of least squares will generally produce poor estimates of the individual model parameters when strong multicollinearity is present, this does not necessarily imply that the fitted model is a poor predictor. If predictions are confined to regions of the x space where the multicollinearity holds approximately, the fitted model often produces satisfactory predictions. This can occur because the linear combination $\sum_{j=1}^{p} \beta_j x_{ij}$ may be estimated quite well, even though the individual parameters β_j are estimated poorly. That is, if the original data lie approximately along the hyperplane defined by Eq. (9.1), then future observations that also lie near this hyperplane can often be precisely predicted despite the inadequate estimates of the individual model parameters.

Example 9.1 The Acetylene Data

Table 9.1 presents data concerning the percentage of conversion of n-heptane to acetylene and three explanatory variables (Himmelblau [1970], Kunugi, Tamura, and Naito [1961], and Marquardt and Snee [1975]). These are typical chemical process data for which a full quadratic response surface in all three regressors is often considered to be an appropriate tentative model. A plot of contact time versus reactor temperature is shown in Figure 9.2. Since these two regressors are highly correlated, there are potential multicollinearity problems in these data.

The full quadratic model for the acetylene data is

$$\begin{aligned} P = {} & \gamma_0 + \gamma_1 T + \gamma_2 H + \gamma_3 C + \gamma_{12} TH + \gamma_{13} TC + \gamma_{23} HC \\ & + \gamma_{11} T^2 + \gamma_{22} H^2 + \gamma_{33} C^2 + \varepsilon \end{aligned}$$

where

$$P = \text{percentage of conversion}$$

$$T = \frac{\text{temperature} - 1212.50}{80.623}$$

$$H = \frac{\mathrm{H_2}\ (n\text{-heptane}) - 12.44}{5.662}$$

TABLE 9.1 Acetylene Data for Example 9.1

Observation	Conversion of *n*-Heptane to Acetylene (%)	Reactor Temperature (°C)	Ratio of H_2 to *n*-Heptane (mole ratio)	Contact Time (sec)
1	49.0	1300	7.5	0.0120
2	50.2	1300	9.0	0.0120
3	50.5	1300	11.0	0.0115
4	48.5	1300	13.5	0.0130
5	47.5	1300	17.0	0.0135
6	44.5	1300	23.0	0.0120
7	28.0	1200	5.3	0.0400
8	31.5	1200	7.5	0.0380
9	34.5	1200	11.0	0.0320
10	35.0	1200	13.5	0.0260
11	38.0	1200	17.0	0.0340
12	38.5	1200	23.0	0.0410
13	15.0	1100	5.3	0.0840
14	17.0	1100	7.5	0.0980
15	20.5	1100	11.0	0.0920
16	29.5	1100	17.0	0.0860

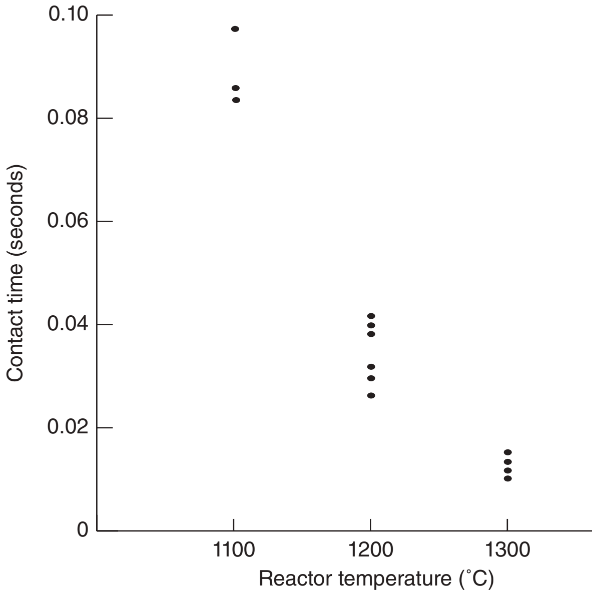

Figure 9.2 Contact time versus reactor temperature, acetylene data. (From Marquardt and Snee [1975], with permission of the publisher.)

and

$$C = \frac{\text{contact time} - 0.0403}{0.03164}$$

Each of the original regressors has been scaled using the unit normal scaling of Section 3.9 [subtracting the average (centering) and dividing by the standard deviation. The squared and cross-product terms are generated from the scaled linear terms. As we noted in Chapter 7, centering the linear terms is helpful in removing nonessential ill-conditioning when fitting polynomials. The least-squares fit is

$$\begin{aligned}\hat{P} = 35.897 + 4.019T + 2.781H - 8.031C - 6.457TH - 26.982TC \\ - 3.768HC - 12.54T^2 - 0.973H^2 - 11.594C^2\end{aligned}$$

The summary statistics for this model are displayed in Table 9.2. The regression coefficients are reported in terms of both the original centered regressors and standardized regressors.

The fitted values for the six points (A, B, E, F, I, and J) that define the boundary of the regressor variable hull of contact time and reactor temperature are shown in Figure 9.3 along with the corresponding observed values of percentage of conversion. The predicted and observed values agree very closely; consequently, the model seems adequate for interpolation within the range of the original data. Now consider using the model for extrapolation. Figure 9.3 (points C, D, G, and H) also shows predictions made at the corners of the region defined by the range of the original data. These points represent relatively mild extrapolation, since the original ranges of the regressors have not been exceeded. The predicted conversions at three of the four extrapolation points are negative, an obvious impossibility. It seems that the least-squares model fits the data reasonably well but extrapolates very poorly. A likely cause of this in view of the strong apparent correlation between contact time and reactor temperature is multicollinearity. In general, if a model is to extrapolate well, good estimates of the individual coefficients are required. When multicollinearity is suspected, the least-squares estimates of the regression coefficients may be very poor. This may seriously limit the usefulness of the regression model for inference and prediction. ■

9.4 MULTICOLLINEARITY DIAGNOSTICS

Several techniques have been proposed for **detecting multicollinearity**. We will now discuss and illustrate some of these diagnostic measures. Desirable characteristics of a diagnostic procedure are that it directly reflect the degree of the multicollinearity problem and provide information helpful in determining which regressors are involved.

9.4.1 Examination of the Correlation Matrix

A very simple measure of multicollinearity is inspection of the off-diagonal elements r_{ij} in $\mathbf{X'X}$. If regressors x_i and x_j are nearly linearly dependent, then $|r_{ij}|$ will

TABLE 9.2 Summary Statistics for the Least-Squares Acetylene Model

Term	Regression Coefficient	Standard Error	t_0	Standardized Regression Coefficient
Intercept	35.8971	1.0903	32.93	
T	4.0187	4.5012	0.89	0.3377
H	2.7811	0.3074	9.05	0.2337
C	–8.0311	6.0657	–1.32	–0.6749
TH	–6.4568	1.4660	–4.40	–0.4799
TC	–26.9818	21.0224	–1.28	–2.0344
HC	–3.7683	1.6554	–2.28	–0.2657
T^2	–12.5237	12.3239	–1.02	–0.8346
H^2	–0.9721	0.3746	–2.60	–0.0904
C^2	–11.5943	7.7070	–1.50	–1.0015

$MS_{\text{Res}} = 0.8126$, $R^2 = 0.998$, $F_0 = 289.72$.

When the response is standardized, $MS_{\text{Res}} = 0.00038$ for the least-squares model.

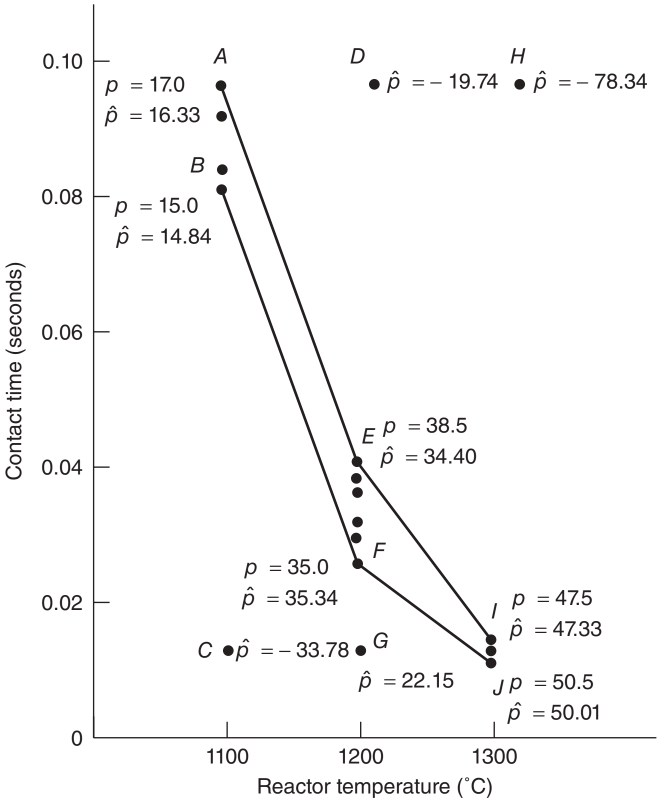

Figure 9.3 Predictions of percentage of conversion within the range of the data and extrapolation for the least-squares acetylene model. (Adapted from Marquardt and Snee [1975], with permission of the publisher.)

be near unity. To illustrate this procedure, consider the acetylene data from Example 9.1. Table 9.3 shows the nine regressor variables and the response in standardized form; that is, each of the variables has been centered by subtracting the mean for that variable and dividing by the square root of the corrected sum of squares for that variable. The **X′X** matrix in correlation form for the acetylene data is

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 1.000 & 0.224 & -0.958 & -0.132 & 0.443 & 0.205 & -0.271 & 0.031 & -0.577 \\ & 1.000 & -0.240 & 0.039 & 0.192 & -0.023 & -0.148 & 0.498 & -0.224 \\ & & 1.000 & 0.194 & -0.661 & -0.274 & 0.501 & -0.018 & 0.765 \\ & & & 1.000 & -0.265 & -0.975 & 0.246 & 0.398 & 0.274 \\ & & & & 1.000 & 0.323 & -0.972 & 0.126 & -0.972 \\ & & & & & 1.000 & -0.279 & -0.374 & 0.358 \\ & & & & & & 1.000 & -0.124 & 0.874 \\ & & & & & & & 1.000 & -0.158 \\ & & & & & & & & 1.000 \\ \text{Symmetric} & & & & & & & & \end{bmatrix}$$

The **X′X** matrix reveals the high correlation between reactor temperature (x_1) and contact time (x_3) suspected earlier from inspection of Figure 9.2, since $r_{13} = -0.958$. Furthermore, there are other large correlation coefficients between x_1x_2 and x_2x_3, x_1x_3 and x_1^2, and x_1^2 and x_3^2. This is not surprising as these variables are generated from the linear terms and they involve the highly correlated regressors x_1 and x_3. Thus, inspection of the correlation matrix indicates that there are several near-linear dependencies in the acetylene data.

Examining the simple correlations r_{ij} between the regressors is helpful in detecting near-linear dependence between **pairs of regressors** only. Unfortunately, when more than two regressors are involved in a near-linear dependence, there is no assurance that any of the pairwise correlations r_{ij} will be large. As an illustration, consider the data in Table 9.4. These data were artificially generated by Webster, Gunst, and Mason [1974]. They required that $\sum_{j=1}^{4} x_{ij} = 10$ for observations 2–12, while $\sum_{j=1}^{4} x_{1j} = 11$ for observation 1. Regressors 5 and 6 were obtained from a table of normal random numbers. The responses y_i were generated by the relationship

$$y_i = 10 + 2.0x_{i1} + 1.0x_{i2} + 0.2x_{i3} - 2.0x_{i4} + 3.0x_{i5} + 10.0x_{i6} + \varepsilon_i$$

where $\varepsilon_i \sim N(0, 1)$. The **X′X** matrix in correlation form for these data is

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 1.000 & 0.052 & -0.343 & -0.498 & 0.417 & -0.192 \\ & 1.000 & -0.432 & -0.371 & 0.485 & -0.317 \\ & & 1.000 & -0.355 & -0.505 & 0.494 \\ & & & 1.000 & -0.215 & -0.087 \\ & & & & 1.000 & -0.123 \\ & & & & & 1.000 \\ \text{Symmetric} & & & & & \end{bmatrix}$$

TABLE 9.3 **Standardized Acetylene Data**[a]

Observation, i	y	x_1	x_2	x_3	x_1x_2	x_1x_3	x_2x_3	x_1^2	x_2^2	x_3^2
1	.27979	.28022	–.22554	–.23106	–.33766	–.02085	.30952	.07829	–.04116	–.03452
2	.30583	.28022	–.15704	–.23106	–.25371	–.02085	.23659	.07829	–.13270	–.03452
3	.31234	.28022	–.06584	–.23514	–.14179	–.02579	.14058	.07829	–.20378	–.02735
4	.26894	.28022	.04817	–.22290	.00189	–.01098	.01960	.07829	–.21070	–.04847
5	.24724	.28022	.20777	–.21882	.19398	–.00605	–.14065	.07829	–.06745	–.05526
6	.18214	–.04003	.48139	–.23106	.52974	–.02085	–.44415	.07829	.59324	–.03452
7	.17590	–.04003	–.32577	–.00255	–.00413	.25895	.07300	–.29746	.15239	–.23548
8	–.09995	–.04003	–.22544	–.01887	–.02171	.26177	.08884	–.29746	–.04116	–.23418
9	–.03486	–.04003	–.06584	–.06784	–.04970	.27023	.08985	–.29746	–.20378	–.21822
10	–.02401	–.04003	.04817	–.11680	–.06968	.27869	.04328	–.29746	–.21070	–.18419
11	.04109	–.04003	.20777	–.05152	–.09766	.26741	.01996	–.29746	–.06745	–.22554
12	.05194	–.04003	.48139	.00561	–.14563	.25754	.08202	–.29746	.59329	–.23538
13	.45800	–.36029	–.32577	.35653	.45252	–.29615	–.46678	.32879	.15239	.24374
14	.41460	–.36029	–.22544	.47078	.29423	–.47384	–.42042	.32879	–.04116	.60000
15	–.33865	–.36029	–.06584	.42187	.04240	–.39769	–.05859	.32879	–.20378	.43527
16	–.14335	–.36029	.20777	.37285	–.38930	–.32153	–.42738	.32879	–.06745	.28861

[a] The standardized data were constructed from the centered and scaled form of the original data in Table 9.1.

TABLE 9.4 Unstandardized Regressor and Response Variables from Webster, Gunst, and Mason [1974]

Observation, i	y_i	x_{i1}	x_{i2}	x_{i3}	x_{i4}	x_{i5}	x_{i6}
1	10.006	8.000	1.000	1.000	1.000	0.541	–0.099
2	9.737	8.000	1.000	1.000	0.000	0.130	0.070
3	15.087	8.000	1.000	1.000	0.000	2.116	0.115
4	8.422	0.000	0.000	9.000	1.000	–2.397	0.252
5	8.625	0.000	0.000	9.000	1.000	–0.046	0.017
6	16.289	0.000	0.000	9.000	1.000	0.365	1.504
7	5.958	2.000	7.000	0.000	1.000	1.996	–0.865
8	9.313	2.000	7.000	0.000	1.000	0.228	–0.055
9	12.960	2.000	7.000	0.000	1.000	1.380	0.502
10	5.541	0.000	0.000	0.000	10.000	–0.798	–0.399
11	8.756	0.000	0.000	0.000	10.000	0.257	0.101
12	10.937	0.000	0.000	0.000	10.000	0.440	0.432

None of the pairwise correlations r_{ij} are suspiciously large, and consequently we have no indication of the near-linear dependence among the regressors. Generally, inspection of the r_{ij} is not sufficient for detecting anything more complex than pairwise multicollinearity.

9.4.2 Variance Inflation Factors

We observed in Chapter 3 that the diagonal elements of the $\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1}$ matrix are very useful in detecting multicollinearity. Recall from Eq. (9.3) that C_{jj}, the jth diagonal element of $\mathbf{C}$, can be written as $C_{jj} = (1 - R_j^2)^{-1}$, where R_j^2 is the coefficient of determination obtained when x_j is regressed on the remaining $p - 1$ regressors. If x_j is nearly orthogonal to the remaining regressors, R_j^2 is small and C_{jj} is close to unity, while if x_j is nearly linearly dependent on some subset of the remaining regressors, R_j^2 is near unity and C_{jj} is large. Since the variance of the jth regression coefficients is $C_{jj}\sigma^2$, we can view C_{jj} as the factor by which the variance of $\hat{\beta}_j$ is increased due to near-linear dependences among the regressors. In Chapter 3 we called

$$\text{VIF}_j = C_{jj} = (1 - R_j^2)^{-1}$$

the **variance inflation factor**. This terminology is due to Marquardt [1970]. The VIF for each term in the model measures the combined effect of the dependences among the regressors on the variance of that term. One or more large VIFs indicate multicollinearity. Practical experience indicates that if any of the VIFs exceeds 5 or 10, it is an indication that the associated regression coefficients are poorly estimated because of multicollinearity.

The VIFs have another interesting interpretation. The length of the normal theory confidence interval on the jth regression coefficient may be written as

$$L_j = 2(C_{jj}\sigma^2)^{1/2} t_{\alpha/2,n-p-1}$$

and the length of the corresponding interval based on an **orthogonal reference design** with the same sample size and root-mean-square (rms) values [i.e., rms $= \sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2/n$ is a measure of the spread of the regressor x_j] as the original design is

$$L^* = 2\sigma t_{\alpha/2,n-p-1}$$

The ratio of these two confidence intervals is $L_j/L^* = C_{jj}^{1/2}$. Thus, the square root of the jth VIF indicates how much longer the confidence interval for the jth regression coefficient is because of multicollinearity.

The VIFs for the acetylene data are shown in panel A of Table 9.5. These VIFs are the main diagonal elements of $(\mathbf{X}'\mathbf{X})^{-1}$, assuming that the linear terms in the model are centered and the second-order terms are generated directly from the linear terms. The maximum VIF is 6565.91, so we conclude that a multicollinearity problem exists. Furthermore, the VIFs for several of the other cross-product and squared variables involving x_1 and x_3 are large. Thus, the VIFs can help identify which regressors are involved in the multicollinearity. Note that the VIFs in polynomial models are affected by centering the linear terms. Panel B of Table 9.5 shows the VIFs for the acetylene data, assuming that the linear terms are not centered. These VIFs are much larger than those for the centered data. Thus centering the linear terms in a polynomial model removes some of the nonessential ill-conditioning caused by the choice of origin for the regressors.

The VIFs for the Webster, Gunst, and Mason data are shown in panel C of Table 9.5. Since the maximum VIF is 297.14, multicollinearity is clearly indicated. Once again, note that the VIFs corresponding to the regressors involved in the multicollinearity are much larger than those for x_5 and x_6.

9.4.3 Eigensystem Analysis of X′X

The characteristic roots or **eigenvalues** of **X′X**, say $\lambda_1, \lambda_2, \ldots, \lambda_p$, can be used to measure the extent of multicollinearity in the data.[†] If there are one or more

TABLE 9.5 VIFs for Acetylene Data and Webster, Gunst, and Mason Data

Data, (A) Acetylene Centered Term VIF	Data, (B) Acetylene Uncentered Term VIF	Data, (C) Webster, Gunst, and Mason Term VIF
$x_1 = 374$	$x_1 = 2{,}856{,}749$	$x_1 = 181.83$
$x_2 = 1.74$	$x_2 = 10{,}956.1$	$x_2 = 161.40$
$x_3 = 679.11$	$x_3 = 2{,}017{,}163$	$x_3 = 265.49$
$x_1x_2 = 31.03$	$x_1x_2 = 2{,}501{,}945$	$x_4 = 297.14$
$x_1x_3 = 6565.91$	$x_1x_3 = 65.73$	$x_5 = 1.74$
$x_2x_3 = 35.60$	$x_2x_3 = 12{,}667.1$	$x_6 = 1.44$
$x_1^2 = 1762.58$	$x_1^2 = 9802.9$	
$x_2^2 = 3.17$	$x_2^2 = 1{,}428{,}092$	
$x_3^2 = 1158.13$	$x_3^2 = 240.36$	
Maximum VIF = 6565.91	Maximum VIF = 2,856,749	Maximum VIF = 297.14

[†] Recall that the eigenvalues of a $p \times p$ matrix **A** are the p roots of the equation $|\mathbf{A} - \lambda\mathbf{I}| = 0$. Eigenvalues are almost always calculated by computer routines. Methods for computing eigenvalues and eigenvectors are discussed in Smith et al. [1974], Stewart [1973], and Wilkinson [1965].

near-linear dependences in the data, then one or more of the characteristic roots will be small. One or more small eigenvalues imply that there are near-linear dependences among the columns of **X**. Some analysts prefer to examine the **condition number** of **X′X**, defined as

$$\kappa = \frac{\lambda_{\max}}{\lambda_{\min}} \tag{9.6}$$

This is just a measure of the spread in the eigenvalue spectrum of **X′X**. Generally, if the condition number is less than 100, there is no serious problem with multicollinearity. Condition numbers between 100 and 1000 imply moderate to strong multicollinearity, and if κ exceeds 1000, severe multicollinearity is indicated.

The **condition indices** of the **X′X** matrix are

$$\kappa_j = \frac{\lambda_{\max}}{\lambda_j}, \quad j = 1, 2, \ldots, p$$

Clearly the largest condition index is the condition number defined in Eq. (9.6). The number of condition indices that are large (say ≥ 1000) is a useful measure of the number of near-linear dependences in **X′X**.

The eigenvalues of **X′X** for the acetylene data are $\lambda_1 = 4.2048$, $\lambda_2 = 2.1626$, $\lambda_3 = 1.1384$, $\lambda_4 = 1.0413$, $\lambda_5 = 0.3845$, $\lambda_6 = 0.0495$, $\lambda_7 = 0.0136$, $\lambda_8 = 0.0051$, and $\lambda_9 = 0.0001$. There are four very small eigenvalues, a symptom of seriously ill-conditioned data. The condition number is

$$\kappa = \frac{\lambda_{\max}}{\lambda_{\min}} = \frac{4.2048}{0.0001} = 42{,}048$$

which indicates severe multicollinearity. The condition indices are

$$\kappa_1 = \frac{4.2048}{4.2048} = 1, \qquad \kappa_2 = \frac{4.2048}{2.1626} = 1.94, \qquad \kappa_3 = \frac{4.2048}{1.1384} = 3.69$$

$$\kappa_4 = \frac{4.2048}{1.0413} = 4.04, \qquad \kappa_5 = \frac{4.2048}{0.3845} = 10.94, \qquad \kappa_6 = \frac{4.2048}{0.0495} = 84$$

$$\kappa_7 = \frac{4.2048}{0.0136} = 309.18, \quad \kappa_8 = \frac{4.2048}{0.0051} = 824.47, \quad \kappa_9 = \frac{4.2048}{0.0001} = 42{,}048$$

Since one of the condition indices exceeds 1000 (and two others exceed 100), we conclude that there is at least one strong near-linear dependence in the acetylene data. Considering that x_1 is highly correlated with x_3 and the model contains both quadratic and cross-product terms in x_1 and x_3, this is, of course, not surprising.

The eigenvalues for the Webster, Gunst, and Mason data are $\lambda_1 = 2.4288$, $\lambda_2 = 1.5462$, $\lambda_3 = 0.9221$, $\lambda_4 = 0.7940$, $\lambda_5 = 0.3079$, and $\lambda_6 = 0.0011$. The small eigenvalue indicates the near-linear dependence in the data. The condition number is

$$\kappa = \frac{\lambda_{\max}}{\lambda_{\min}} = \frac{2.4288}{0.0011} = 2188.11$$

which also indicates strong multicollinearity. Only one condition index exceeds 1000, so we conclude that there is only one near-linear dependence in the data.

Eigensystem analysis can also be used to identify the nature of the near-linear dependences in data. The **X′X** matrix may be decomposed as

$$\mathbf{X}'\mathbf{X} = \mathbf{T}\boldsymbol{\Lambda}\mathbf{T}'$$

where $\boldsymbol{\Lambda}$ is a $p \times p$ diagonal matrix whose main diagonal elements are the **eigenvalues** λ_j $(j = 1, 2, \ldots, p)$ of **X′X** and **T** is a $p \times p$ orthogonal matrix whose columns are the eigenvectors of **X′X**. Let the columns of **T** be denoted by $\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_p$. If the eigenvalue λ_j is close to zero, indicating a near-linear dependence in the data, the elements of the associated eigenvector $\mathbf{t}_j$ describe the nature of this linear dependence. Specifically the elements of the vector $\mathbf{t}_j$ are the coefficients $t_1, t_2, \ldots, t_p$ in Eq. (9.1).

Table 9.6 displays the eigenvectors for the Webster, Gunst, and Mason data. The smallest eigenvalue is $\lambda_6 = 0.0011$, so the elements of the eigenvector $\mathbf{t}_6$ are the coefficients of the regressors in Eq. (9.1). This implies that

$$-0.44768x_1 - 0.42114x_2 - 0.54169x_3 - 0.57337x_4 - 0.00605x_5 - 0.00217x_6 = 0$$

Assuming that –0.00605 and –0.00217 are approximately zero and rearranging terms gives

$$x_1 \simeq -0.941x_2 - 1.120x_3 - 1.281x_4$$

That is, the first four regressors add approximately to a constant. Thus, the elements of $\mathbf{t}_6$ directly reflect the relationship used to generate x_1, x_2, x_3, and x_4.

Belsley, Kuh, and Welsch [1980] propose a similar approach for diagnosing multicollinearity. The $n \times p$ **X** matrix may be decomposed as

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{T}'$$

where **U** is $n \times p$, **T** is $p \times p$, $\mathbf{U}'\mathbf{U} = \mathbf{I}$, $\mathbf{T}'\mathbf{T} = \mathbf{I}$, and **D** is a $p \times p$ diagonal matrix with nonnegative diagonal elements $\mu, j = 1, 2, \ldots, p$. The μ_j are called the **singular values** of **X** and **X** = **UDT′** is called the **singular-value decomposition** of **X**. The singular-value decomposition is closely related to the concepts of eigenvalues and eigenvectors, since $\mathbf{X}'\mathbf{X} = (\mathbf{U}\mathbf{D}\mathbf{T}')'\mathbf{U}\mathbf{D}\mathbf{T}' = \mathbf{T}\mathbf{D}^2\mathbf{T}' = \mathbf{T}\boldsymbol{\Lambda}\mathbf{T}'$, so that the squares of the singular values of **X** are the eigenvalues of **X′X**. Here **T** is the matrix of eigenvectors of **X′X**

TABLE 9.6 Eigenvectors for the Webster, Gunst, and Mason Data

t_1	t_2	t_3	t_4	t_5	t_6
–.39072	–.33968	.67980	.07990	–.25104	–.44768
–.45560	–.05392	–.70013	.05769	–.34447	–.42114
.48264	–.45333	–.16078	.19103	.45364	–.54169
.18766	.73547	.13587	–.27645	.01521	–.57337
–.49773	–.09714	–.03185	–.56356	.65128	–.00605
.35195	–.35476	–.04864	–.74818	–.43375	–.00217

defined earlier, and **U** is a matrix whose columns are the eigenvectors associated with the p nonzero eigenvalues of $\mathbf{XX}'$.

Ill-conditioning in **X** is reflected in the size of the singular values. There will be one small singular value for each near-linear dependence. The extent of ill-conditioning depends on how small the singular value is relative to the maximum singular value $\mu_{\max}$. SAS follows Belsley, Kuh, and Welsch [1980] and defines the **condition indices** of the **X** matrix as

$$\eta_j = \frac{\mu_{\max}}{\mu_j}, \quad j = 1, 2, \ldots, p$$

The largest value for η_j is the condition number of **X**. Note that this approach deals directly with the data matrix **X**, with which we are principally concerned, not the matrix of sums of squares and cross products $\mathbf{X}'\mathbf{X}$. A further advantage of this approach is that algorithms for generating the singular-value decomposition are more stable numerically than those for eigensystem analysis, although in practice this is not likely to be a severe handicap if one prefers the eigensystem approach.

The covariance matrix of $\hat{\boldsymbol{\beta}}$ is

$$\operatorname{Var}\left(\hat{\boldsymbol{\beta}}\right) = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1} = \sigma^2 \mathbf{T}\boldsymbol{\Lambda}^{-1}\mathbf{T}'$$

and the variance of the jth regression coefficient is the jth diagonal element of this matrix, or

$$\operatorname{Var}\left(\hat{\beta}_j\right) = \sigma^2 \sum_{i=1}^{p} \frac{t_{ji}^2}{\mu_i^2} = \sigma^2 \sum_{i=1}^{p} \frac{t_{ji}^2}{\lambda_i}$$

Note also that apart from σ^2, the jth diagonal element of $\mathbf{T}\boldsymbol{\Lambda}^{-1}\mathbf{T}'$ is the jth VIF, so

$$\text{VIF}_j = \sum_{i=1}^{p} \frac{t_{ji}^2}{\mu_i^2} = \sum_{i=1}^{p} \frac{t_{ji}^2}{\lambda_i}$$

Clearly, one or more small singular values (or small eigenvalues) can dramatically inflate the variance of $\hat{\beta}_j$. Belsley, Kuh, and Welsch suggest using **variance decomposition proportions**, defined as

$$\pi_{ij} = \frac{t_{ji}^2/\mu_i^2}{\text{VIF}_j}, \quad j = 1, 2, \ldots, p$$

as measures of multicollinearity. If we array the π_{ij} in a $p \times p$ matrix π, then the elements of each column of π are just the proportions of the variance of each $\hat{\beta}_j$ (or each VIF) contributed by the ith singular value (or eigenvalue). If a high proportion of the variance for two or more regression coefficients is associated with one small singular value, multicollinearity is indicated. For example, if π_{32} aud π_{34} are large, the third singular value is associated with a multicollinearity that is inflating the variances of $\hat{\beta}_2$ and $\hat{\beta}_4$ Condition indices greater than 30 and variance decomposition proportions greater than 0.5 are recommended guidelines.

TABLE 9.7 Variance Decomposition Proportions for the Webster, Gunst, and Mason [1974] Data

		Condition		Variance Decomposition Proportions					
Number	Eigenvalue	Indices		X_1	X_2	X_3	X_4	X_5	X_6
				A. Regressors Centered					
1	2.42879	1.00000		0.0003	0.0005	0.0004	0.0000	0.0531	0.0350
2	1.54615	1.25334		0.0004	0.0000	0.0005	0.0012	0.0032	0.0559
3	0.92208	1.62297		0.0028	0.0033	0.0001	0.0001	0.0006	0.0018
4	0.79398	1.74900		0.0000	0.0000	0.0002	0.0003	0.2083	*004845*
5	0.30789	2.80864		0.0011	0.0024	0.0025	0.0000	0.7175	004199
6	0.00111	46.86052		0.9953	0.9937	0.9964	0.9984	0.0172	0.0029
				B. Regressors Not Centered					
1	2.63287	1.00000	0.0001	0.0003	0.0003	0.0001	0.0001	0.0217	0.0043
2	1.82065	1.20255	0.0000	0.0001	0.0002	0.0005	0.0000	0.0523	0.0949
3	1.03335	159622	0.0000	0.0002	0.0000	0.0002	0.0013	0.0356	0.1010
4	0.65826	1.99994	0.0000	0.0005	0.0000	0.0005	0.0003	0.1906	0.3958
5	0.60573	2.08485	0.0000	0.0025	0.0035	0.0001	0.0001	0.0011	0.0002
6	0.24884	3.25280	0.0000	0.0012	0.0023	0.0028	0.0000	0.6909	0.4003
7	0.00031	92.25341	0.9999	0.9953	0.9936	0.9959	0.9983	0.0178	0.0034

Table 9.7 displays the condition indices of **X** (η_j) and the variance-decomposition proportions (the π_{ij}) for the Webster, Gunst, and Mason data. In panel A of this table we have centered the regressors so that these variables are $(x_{ij} - \bar{x}_j)$, $j = 1, 2, \ldots, 6$. In Section 9.4.2 we observed that the VIFs in a polynomial model are affected by centering the linear terms in the model before generating the higher order polynomial terms. Centering will also affect the variance decomposition proportions (and also the eigenvalues and eigenvectors). Essentially, centering removes any nonessential ill-conditioning resulting from the intercept.

Notice that there is only one large condition index ($\eta_6 = 46.86 > 30$), so there is one dependence in the columns of **X**. Furthermore, the variance decomposition proportions π_{61}, π_{62}, π_{63}, and π_{64} all exceed 0.5, indicating that the first four regressors are involved in a multicollinear relationship. This is essentially the same information derived previously from examining the eigenvalues.

Belsley, Kuh, and Welsch [1980] suggest that the regressors should be scaled to unit length but not centered when computing the variance decomposition proportions so that the role of the intercept in near-linear dependences can be diagnosed. This option is displayed in panel B of Table 9.7. Note that the effect of this is to increase the spread in the eigenvalues and make the condition indices larger.

There is some controversy about whether regression data should be centered when diagnosing multicollinearity using either the eigensystem analysis or the variance decomposition proportion approach. Centering makes the intercept orthogonal to the other regressors, so we can view centering as an operation that removes ill-conditioning that is due to the model's constant term. If the intercept has no physical interpretation (as is the case in many applications of regression in engineering and the physical sciences), then ill-conditioning caused by the constant term is truly "nonessential," and thus centering the regressors is entirely appropriate.

However, if the intercept has interpretative value, then centering is not the best approach. Clearly the answer to this question is problem specific. For excellent discussions of this point, see Brown [1977] and Myers [1990].

9.4.4 Other Diagnostics

There are several other techniques that are occasionally useful in diagnosing multicollinearity. The **determinant** of $\mathbf{X}'\mathbf{X}$ can be used as an index of multicollinearity. Since the $\mathbf{X}'\mathbf{X}$ matrix is in correlation form, the possible range of values of the determinant is $0 \le |\mathbf{X}'\mathbf{X}| \le 1$. If $|\mathbf{X}'\mathbf{X}| = 1$, the regressors are orthogonal, while if $|\mathbf{X}'\mathbf{X}| = 0$, there is an exact linear dependence among the regressors. The degree of multicollinearity becomes more severe as $|\mathbf{X}'\mathbf{X}|$ approaches zero. While this measure of multicollinearity is easy to apply, it does not provide any information on the source of the multicollinearity.

Willan and Watts [1978] suggest another interpretation of this diagnostic. The joint $100(1 - \alpha)$ percent confidence region for $\boldsymbol{\beta}$ based on the observed data is

$$\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)' \mathbf{X}'\mathbf{X}\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right) \le p\hat{\sigma}^2 F_{\alpha,p,n-p-1}$$

while the corresponding confidence region for $\hat{\boldsymbol{\beta}}$ based on the orthogonal reference design described earlier is

$$\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)' \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right) \le p\hat{\sigma}^2 F_{\alpha,p,n-p-1}$$

The orthogonal reference design produces the smallest joint confidence region for fixed sample size and rms values and a given α. The ratio of the volumes of the two confidence regions is $|\mathbf{X}'\mathbf{X}|^{1/2}$, so that $|\mathbf{X}'\mathbf{X}|^{1/2}$ measures the loss of estimation power due to multicollinearity. Put another way, $100(|\mathbf{X}'\mathbf{X}|^{1/2} - 1)$ reflects the percentage increase in the volume of the joint confidence region for $\boldsymbol{\beta}$ because of the near-linear dependences in $\mathbf{X}$. For example, if $|\mathbf{X}'\mathbf{X}| = 0.25$, then the volume of the joint confidence region is $100[(0.25)^{-1/2} - 1] = 100\%$ larger than it would be if an orthogonal design had been used.

The F statistic for significance of regression and the individual t (or partial F) statistics can sometimes indicate the presence of multicollinearity. Specifically, if the overall F statistic is significant but the individual t statistics are all nonsignificant, multicollinearity is present. Unfortunately, many data sets that have significant multicollinearity will not exhibit this behavior, and so the usefulness of this measure of multicollinearity is questionable.

The **signs** and **magnitudes** of the regression coefficients will sometimes provide an indication that multicollinearity is present. In particular, if adding or removing a regressor produces large changes in the estimates of the regression coefficients, multicollinearity is indicated. If the deletion of one or more data points results in large changes in the regression coefficients, there may be multicollinearity present. Finally, if the signs or magnitudes of the regression coefficients in the regression model are contrary to prior expectation, we should be alert to possible multicollinearity. For example, the least-squares model for the acetylene data has large standardized regression coefficients for the x_1x_3 interaction and for the squared terms x_1^2 and x_3^2. It is somewhat unusual for quadratic models to display large regression coefficients for the higher order terms, and so this may be an indication of multicollinearity. However, one should be cautious in using the signs and magnitudes of the regression coefficients as indications of multicollinearity, as many seriously ill-conditioned data sets do not exhibit behavior that is obviously unusual in this respect.

We believe that the VIFs and the procedures based on the eigenvalues of $\mathbf{X}'\mathbf{X}$ are the best currently available multicollinearity diagnostics. They are easy to compute, straightforward to interpret, and useful in investigating the specific nature of the multicollinearity. For additional information on these and other methods of detecting multicollinearity, see Belsley, Kuh, and Welsch [1980], Farrar and Glauber [1997], and Willan and Watts [1978].

9.4.5 SAS and R Code for Generating Multicollinearity Diagnostics

The appropriate SAS code for generating the multicollinearity diagnostics for the acetylene data is

```
proc reg;
model conv = t h c t2 h2·c2 th tc hc / corrb vif collin;
```

The corrb option prints the variance–covariance matrix of the estimated coefficients in correlation form. The vif option prints the VIFs. The collin option prints the singular-value analysis including the condition numbers and the variance decomposition proportions. SAS uses the singular values to compute the condition numbers. Some other software packages use the eigenvalues, which are the squares of the singular values. The collin option includes the effect of the intercept on the diagnostics. The option collinoint performs the singular-value analysis excluding the intercept.

The collinearity diagnostics in R require the packages “perturb” and “car”. The R code to generate the collinearity diagnostics for the delivery data is:

```
deliver.model <- lm(time~cases+dist, data=deliver)
print(vif(deliver.model))
print(colldiag(deliver.model))
```

9.5 METHODS FOR DEALING WITH MULTICOLLINEARITY

Several techniques have been proposed for dealing with the problems caused by multicollinearity. The general approaches include collecting additional data, model respecification, and the use of estimation methods other than least squares that are specifically designed to combat the problems induced by multicollinearity.

9.5.1 Collecting Additional Data

Collecting additional data has been suggested as the best method of combating multicollinearity (e.g., see Farrar and Glauber [1967] and Silvey [1969]). The additional data should be collected in a manner designed to break up the multicollinearity in the existing data. For example, consider the delivery time data first introduced

Example 3.1. A plot of the regressor cases (x_1) versus distance (x_2) is shown in the matrix of scatterplots, Figure 3.4. We have remarked previously that most of these data lie along a line from low values of cases and distance to high values of cases and distance, and consequently there may be some problem with multicollinearity. This could be avoided by collecting some additional data at points designed to break up any potential multicollinearity, that is, at points where cases are small and distance is large and points where cases are large and distance is small.

Unfortunately, collecting additional data is not always possible because of **economic constraints** or because the process being studied is **no longer available** for sampling. Even when additional data are available it may be inappropriate to use if the new data extend the range of the regressor variables far beyond the analyst′s region of interest. Furthermore, if the new data points are unusual or atypical of the process being studied, their presence in the sample could be highly influential on the fitted model. Finally, note that collecting additional data is not a viable solution to the multicollinearity problem when the multicollinearity is due to constraints on the model or in the population. For example, consider the factors family income (x_1) and house size (x_2) plotted in Figure 9.1. Collection of additional data would be of little value here, since the relationship between family income and house size is a structural characteristic of the population. Virtually all the data in the population will exhibit this behavior.

9.5.2 Model Respecification

Multicollinearity is often caused by the choice of model, such as when two highly correlated regressors are used in the regression equation. In these situations some **respecification** of the regression equation may lessen the impact of multicollinearity. One approach to model respecification is to redefine the regressors. For example, if x_1, x_2, and x_3 are nearly linearly dependent, it may be possible to find some function such as $x = (x_1 + x_2)/x_3$ or $x = x_1x_2x_3$ that preserves the information content in the original regressors but reduces the ill-conditioning.

Another widely used approach to model respecification is **variable elimination**. That is, if x_1, x_2 and x_3 are nearly linearly dependent, eliminating one regressor (say x_3) may be helpful in combating multicollinearity. Variable elimination is often a highly effective technique. However, it may not provide a satisfactory solution if the regressors dropped from the model have significant explanatory power relative to the response y. That is, eliminating regressors to reduce multicollinearity may damage the predictive power of the model. Care must be exercised in variable selection because many of the selection procedures are seriously distorted by multicollinearity, and there is no assurance that the final model will exhibit any lesser degree of multicollinearity than was present in the original data. We discuss appropriate variable elimination techniques in Chapter 10.

9.5.3 Ridge Regression

When the method of least squares is applied to nonorthogonal data, very poor estimates of the regression coefficients can be obtained. We saw in Section 9.3 that the variance of the least-squares estimates of the regression coefficients may be considerably inflated, and the length of the vector of least-squares parameter

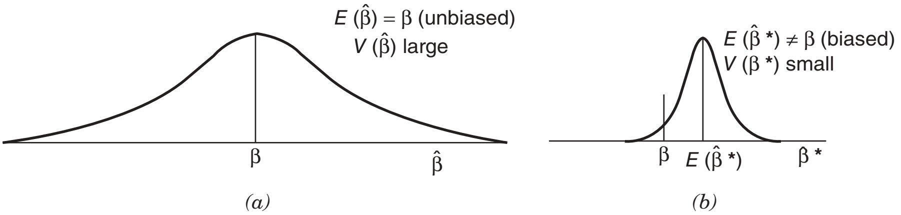

Figure 9.4 Sampling distribution of (*a*) nnbiased and (*b*) biased estimators of β. (Adapted from Marquardt and Snee [1975], with permission of the publisher.)

estimates is too long on the average. This implies that the absolute value of the least-squares estimates are too large and that they are very unstable, that is, their magnitudes and signs may change considerably given a different sample.

The problem with the method of least squares is the requirement that $\hat{\boldsymbol{\beta}}$ be an **unbiased estimator** of $\boldsymbol{\beta}$. The Gauss-Markov property referred to in Section 3.2.3 assures us that the least-squares estimator has minimum variance in the class of unbiased linear estimators, but there is no guarantee that this variance will be small. The situation is illustrated in Figure 9.4*a*, where the sampling distribution of $\hat{\boldsymbol{\beta}}$, the unbiased estimator of $\boldsymbol{\beta}$, is Shown. The variance of $\hat{\boldsymbol{\beta}}$ is large, implying that confidence intervals on $\boldsymbol{\beta}$ would be wide and the point estimate $\hat{\boldsymbol{\beta}}$ is very unstable.

One way to alleviate this problem is to drop the requirement that the estimator of $\boldsymbol{\beta}$ be unbiased. Suppose that we can find a **biased estimator** of $\boldsymbol{\beta}$, say $\hat{\boldsymbol{\beta}}^*$, that has a smaller variance than the unbiased estimator $\hat{\boldsymbol{\beta}}$. The mean square error of the estimator $\hat{\boldsymbol{\beta}}^*$ is defined as

$$\text{MSE}\left(\hat{\boldsymbol{\beta}}^*\right) = E\left(\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}\right)^2 = \text{Var}\left(\hat{\boldsymbol{\beta}}^*\right) + \left[E\left(\hat{\boldsymbol{\beta}}^*\right) - \boldsymbol{\beta}\right]^2$$

or

$$\text{MSE}\left(\hat{\boldsymbol{\beta}}^*\right) = \text{Var}\left(\hat{\boldsymbol{\beta}}^*\right) + \left(\text{bias in } \hat{\boldsymbol{\beta}}^*\right)^2$$

Note that the MSE is just the expected squared distance from $\hat{\boldsymbol{\beta}}^*$ to $\boldsymbol{\beta}$ [see Eq. (9.4)]. By allowing a small amount of bias in $\hat{\boldsymbol{\beta}}^*$, the variance of $\hat{\boldsymbol{\beta}}^*$ can be made small such that the MSE of $\hat{\boldsymbol{\beta}}^*$ is less than the variance of the unbiased estimator $\boldsymbol{\beta}$. Figure 9.4*b* illustrates a situation where the variance of the biased estimator is considerably smaller than the variance of the unbiased estimator (Figure 9.4*a*). Consequently, confidence intervals on $\boldsymbol{\beta}$ would be much narrower using the biased estimator. The small variance for the biased estimator also implies that $\hat{\boldsymbol{\beta}}^*$ is a more stable estimator of $\boldsymbol{\beta}$ than is the unbiased estimator $\hat{\boldsymbol{\beta}}$.

A number of procedures have been developed for obtaining biased estimators of regression coefficients. One of these procedures is **ridge regression**, originally proposed by Hoerl and Kennard [1970a, b]. The ridge estimator is found by solving a slightly modified version of the normal equations. Specifically we define the ridge estimator $\hat{\boldsymbol{\beta}}_{\text{R}}$ as the solution to

$$(\mathbf{X}'\mathbf{X} + k\mathbf{I})\hat{\boldsymbol{\beta}}_{\text{R}} = \mathbf{X}'\mathbf{y}$$

or

$$\hat{\boldsymbol{\beta}}_{\mathrm{R}} = (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}$$

where $k \geq 0$ is a constant selected by the analyst. The procedure is called ridge regression because the underlying mathematics are similar to the method of ridge analysis used earlier by Hoerl [1959] for describing the behavior of second-order response surfaces. Note that when $k = 0$, the ridge estimator is the least-squares estimator.

The ridge estimator is a linear transformation of the least-squares estimator since

$$\hat{\boldsymbol{\beta}}_{\mathrm{R}} = (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{y} = (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{Z}_k\hat{\boldsymbol{\beta}}$$

Therefore, since $E\left(\hat{\boldsymbol{\beta}}_{\mathrm{R}}\right) = E\left(\mathbf{Z}_k\hat{\boldsymbol{\beta}}\right) = \mathbf{Z}_k\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}_{\mathrm{R}}$ is a biased estimator of $\boldsymbol{\beta}$. We usually refer to the constant k as the **biasing parameter**. The covariance matrix of $\hat{\boldsymbol{\beta}}_{\mathrm{R}}$ is

$$\operatorname{Var}\left(\hat{\boldsymbol{\beta}}_{\mathrm{R}}\right) = \sigma^2(\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}$$

The mean square error of the ridge estimator is

$$\begin{aligned} \operatorname{MSE}\left(\hat{\boldsymbol{\beta}}_{\mathrm{R}}\right) &= \operatorname{Var}\left(\hat{\boldsymbol{\beta}}_{\mathrm{R}}\right) + \left(\text{bias in } \hat{\boldsymbol{\beta}}_{\mathrm{R}}\right)^2 \\ &= \sigma^2 \operatorname{Tr}\left[(\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\right] + k^2\boldsymbol{\beta}'(\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-2}\boldsymbol{\beta} \\ &= \sigma^2 \sum_{j=1}^{p} \frac{\lambda_j}{(\lambda_j + k)^2} + k^2\boldsymbol{\beta}'(\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-2}\boldsymbol{\beta} \end{aligned}$$

where $\lambda_1, \lambda_2, \ldots, \lambda_p$ are the eigenvalues of $\mathbf{X}'\mathbf{X}$. The first term on the right-hand side of this equation is the sum of variances of the parameters in $\hat{\boldsymbol{\beta}}_{\mathrm{R}}$ and the second term is the square of the bias. If $k > 0$, note that the bias in $\hat{\boldsymbol{\beta}}_{\mathrm{R}}$ increases with k. However, the variance decreases as k increases.

In using ridge regression we would like to choose a value of k such that the reduction in the variance term is greater than the increase in the squared bias. If this can be done, the mean square error of the ridge estimator $\hat{\boldsymbol{\beta}}_{\mathrm{R}}$ will be less than the variance of the least-squares estimator $\hat{\boldsymbol{\beta}}$. Hoerl and Kennard proved that there exists a nonzero value of k for which the MSE of $\hat{\boldsymbol{\beta}}_{\mathrm{R}}$ is less than the variance of the least-squares estimator $\hat{\boldsymbol{\beta}}$, provided that $\boldsymbol{\beta}'\boldsymbol{\beta}$ is bounded. The residual sum of squares is

$$\begin{aligned} SS_{\mathrm{Res}} &= \left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\mathrm{R}}\right)'\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\mathrm{R}}\right) \\ &= \left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)'\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right) + \left(\hat{\boldsymbol{\beta}}_{\mathrm{R}} - \hat{\boldsymbol{\beta}}\right)'\mathbf{X}'\mathbf{X}\left(\hat{\boldsymbol{\beta}}_{\mathrm{R}} - \hat{\boldsymbol{\beta}}\right) \end{aligned} \qquad (9.7)$$

Since the first term on the right-hand side of Eq. (9.7) is the residual sum of squares for the least-squares estimates $\hat{\boldsymbol{\beta}}$, we see that as k increases, the residual sum of squares increases. Consequently, because the total sum of squares is fixed, R^2 decreases as k increases. Therefore, the ridge estimate will not necessarily provide the best "fit" to the data, but this should not overly concern us, since we are more interested in obtaining a stable set of parameter estimates. The ridge estimates may result in an equation that does a better job of predicting future observations than would least squares (although there is no conclusive **proof** that this will happen).

Hoed and Kennard have suggested that an appropriate value of k may be determined by inspection of the **ridge trace**. The ridge trace is a plot of the elements of $\hat{\boldsymbol{\beta}}_{\mathrm{R}}$ versus k for values of k usually in the interval 0–1. Marquardt and Snee [1975] suggest using up to about 25 values of k, spaced approximately logarithmically over the interval [0, 1]. If multicollinearity is severe, the instability in the regression coefficients will be obvious from the ridge trace. As k is increased, some of the ridge estimates will vary dramatically. At some value of k, the ridge estimates $\hat{\boldsymbol{\beta}}_{\mathrm{R}}$ will stabilize. The objective is to select a reasonably small value of k at which the ridge estimates $\hat{\boldsymbol{\beta}}_{\mathrm{R}}$ are stable. Hopefully this will produce a set of estimates with smaller MSE than the least-squares estimates.

Example 9.2 The Acetylene Data

To obtain the ridge solution for the acetylene data, we must solve the equations $(\mathbf{X}'\mathbf{X}+k\mathbf{I})\hat{\boldsymbol{\beta}}_{\mathrm{R}}=\mathbf{X}'\mathbf{y}$ for several values $0 \le k \le 1$, with $\mathbf{X}'\mathbf{X}$ and $\mathbf{X}'\mathbf{y}$ in correlation form. The ridge trace is shown in Figure 9.5, and the ridge coefficients for several values of k are listed in Table 9.8. This table also shows the residual mean square and R^2 for each ridge model. Notice that as k increases, MS_{Res} increases and R^2 decreases. The ridge trace illustrates the instability of the least-squares solution, as there are large changes in the regression coefficients for small values of k. However, the coefficients stabilize rapidly as k increases.

Judgment is required to interpret the ridge trace and select an appropriate value of k. We want to choose k large enough to provide stable coefficients, but not unnecessarily large ones, as this introduces additional bias and increases the residual mean square. From Figure 9.5 we see that reasonable coefficient stability is achieved in the region $0.008 < k < 0.064$ without a severe increase in the residual mean square (or loss in R^2). If we choose $k = 0.032$, the ridge regression model is

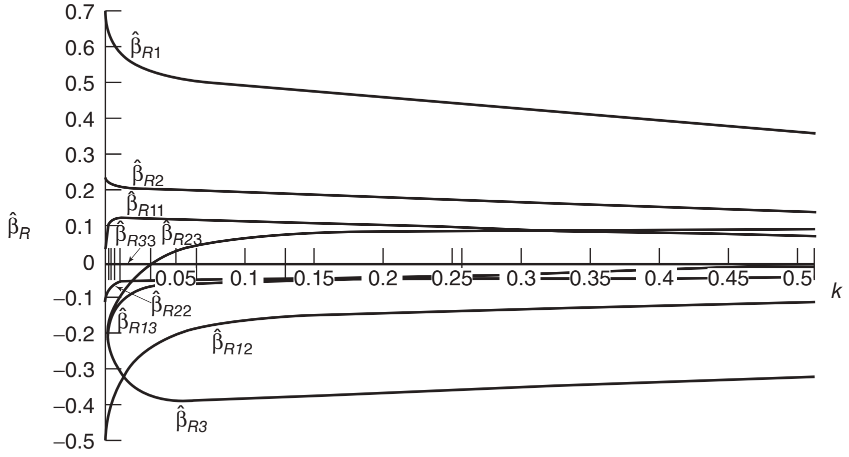

Figure 9.5 Ridge trace for acetylene data using nine regressors.

TABLE 9.8 **Coefficients at Various Values of *k***

k	.000	.001	.002	.004	.008	.016	.032	.064	.128	.256	.512
$\beta_{R.1}$	.3377	.6770	.6653	.6362	.6003	.5672	.5392	.5122	.4806	.4379	.3784
$\beta_{R.2}$	.2337	.2242	.2222	.2199	.2173	.2148	.2117	.2066	.1971	.1807	.1554
$\beta_{R.3}$	−.6749	−.2129	−.2284	−.2671	−.3134	−.3515	−.3735	−.3800	−.3724	−.3500	−.3108
$\beta_{R.12}$	−.4799	−.4479	−.4258	−.3913	−.3437	−.2879	−.2329	−.1862	−.1508	−.1249	−.1044
$\beta_{R.13}$	−2.0344	−.2774	−.1887	−.1350	−.1017	−.0809	−.0675	−.0570	−.0454	−.0299	−.0092
$\beta_{R.23}$	−.2675	−.2173	−.1920	−.1535	−.1019	−.0433	.0123	.0562	.0849	.0985	.0991
$\beta_{R.11}$	−.8346	.0643	.1035	.1214	.1262	.1254	.1249	.1258	.1230	.1097	.0827
$\beta_{R.22}$	−.0904	−.0732	−.0682	−.0621	−.0558	−.0509	−.0481	−.0464	−.0444	−.0406	−.0341
$\beta_{R.33}$	−1.0015	−.2451	−.1853	−.1313	−.0825	−.0455	−.0267	−.0251	−.0339	−.0464	−.0586
MS_{Res}	.00038	.00047	.00049	.00054	.00062	.00074	.00094	.00127	.00206	.00425	.01002
R^2	.998	.997	.997	.997	.996	.996	.994	.992	.988	.975	.940

$$\begin{aligned}\hat{y} = 0.5392x_1 + 0.2117x_2 - 0.3735x_3 - 0.2329x_1x_2 - 0.0675x_1x_3 \\ + 0.0123x_2x_3 + 0.1249x_1^2 - 0.0481x_2^2 - 0.0267x_3^2\end{aligned}$$

Note that in this model the estimates of β_{13}, β_{11}, and β_{23} are considerably smaller than the least-squares estimates and the original negative estimates of β_{23} and β_{11} are now positive. The ridge model expressed in terms of the original regressors is

$$\begin{aligned}\hat{P} = 0.7598 + 0.1392T + 0.0547H - 0.0965C - 0.0680TH - 0.0194TC \\ + 0.0039CH + 0.0407T^2 - 0.0112H^2 - 0.0067C^2\end{aligned}$$

Figure 9.6 shows the performance of the ridge model in prediction for both interpolation (points A, B, E, F, I, and J) and extrapolation (points C, D, G, and H). Comparing Figures 9.6 and 9.3, we note that the ridge model predicts as well as the nine-term least-squares model at the boundary of the region covered by the data. However, the ridge model gives much more realistic predictions when extrapolating than does least squares. We conclude that ridge regression has produced a model that is superior to the original least squares fit.

The ridge regression estimates may be computed by using an ordinary least-squares computer program and augmenting the standardized data as follows:

$$\mathbf{X}_\mathrm{A} = \begin{bmatrix} \mathbf{X} \\ \sqrt{k}\mathbf{I}_p \end{bmatrix}, \quad \mathbf{y}_\mathrm{A} = \begin{bmatrix} \mathbf{y} \\ \mathbf{0}_p \end{bmatrix}$$

where $\sqrt{k}\mathbf{I}_p$ is a $p \times p$ diagonal matrix with diagonal elements equal to the square root of the biasing parameter and $\mathbf{0}_p$ is a $p \times 1$ vector of zeros. The ridge estimates are then computed from

$$\hat{\boldsymbol{\beta}}_\mathrm{R} = (\mathbf{X}_\mathrm{A}'\mathbf{X}_\mathrm{A})^{-1}\mathbf{X}_\mathrm{A}'\mathbf{y}_\mathrm{A} = (\mathbf{X}'\mathbf{X} + k\mathbf{I}_p)^{-1}\mathbf{X}'\mathbf{y}$$

Table 9.9 shows the augmented matrix $\mathbf{X}_\mathbf{A}$ and vector $\mathbf{y}_\mathbf{A}$ required to produce the ridge solution for the acetylene data with $k = 0.032$. ■

Some Other Properties of Ridge Regression Figure 9.7 illustrates the geometry of ridge regression for a two-regressor problem. The point $\hat{\boldsymbol{\beta}}$ at the center of the ellipses corresponds to the least-squares solution, where the residual sum of squares takes on its minimum value. The small ellipse represents the locus of points in the β_1, β_2 plane where the residual sum of squares is constant at some value greater than the minimum. The ridge estimate $\hat{\boldsymbol{\beta}}_\mathrm{R}$ is the shortest vector from the origin that produces a residual sum of squares equal to the value represented by the small ellipse. That is, the ridge estimate $\hat{\boldsymbol{\beta}}_\mathrm{R}$ produces the vector of regression coefficients with the smallest norm consistent with a specified increase in the residual sum of squares. We note that the ridge estimator shrinks the least-squares

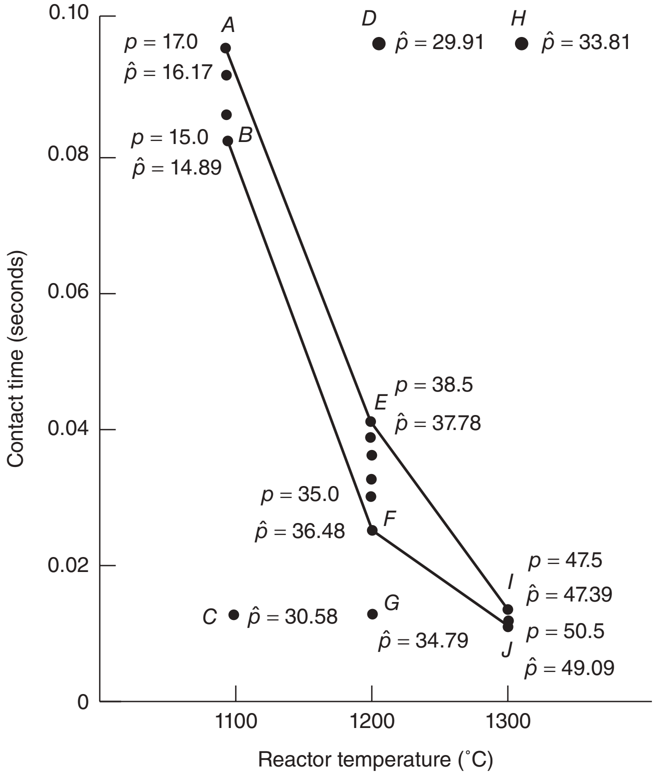

Figure 9.6 Performance of the ridge model with $k = 0.032$ in prediction and extrapolation for the acetylene data. (Adapted from Marquardt and Snee [1975], with permission of the publisher.)

estimator toward the origin. Consequently, ridge estimators (and other biased estimators generally) are sometimes called **shrinkage** estimators. Hocking [1976] has observed that the ridge estimator shrinks the least-squares estimator with respect to the contours of $\mathbf{X'X}$. That is, $\hat{\boldsymbol{\beta}}_R$ is the solution to

$$\underset{\beta}{\text{Minimize}}\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)'\mathbf{X}'\mathbf{X}\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)$$

$$\text{subject to } \boldsymbol{\beta}'\boldsymbol{\beta} \le d^2$$

where the radius d depends on k.

Many of the properties of the ridge estimator assume that the value of k is fixed. In practice, since k is estimated from the data by inspection of the ridge trace, k is **stochastic**. It is of interest to ask if the optimality properties cited by Hoerl and Kennard hold if k is stochastic. Several authors have shown through simulations that ridge regression generally offers improvement in mean square error over least squares when k is estimated from the data. Theobald [1974] has generalized the

TABLE 9.9 Augmented Matrix $\mathbf{X}_A$ and Vector $\mathbf{y}_A$ for Generating the Ridge Solution for the Acetylene Data with $k = 0.032$

$$\mathbf{x}_A = \begin{bmatrix}
.280224 & -.22544 & -.23106 & -.33766 & -.02085 & 309525 & .078278 & -.04116 & -.03452 \\
.280224 & -.15704 & -.23106 & -.25371 & -.02085 & .236588 & .078278 & -.1327 & -.03452 \\
.280224 & -.06584 & -.23514 & -.14179 & -.02579 & .140577 & .078278 & -.20378 & -.02735 \\
.280224 & .048167 & -.2229 & -.00189 & -.01098 & .0196 & .078278 & -.2107 & -.04847 \\
.280224 & .207774 & -.21882 & .193976 & -.00605 & -.14065 & .078278 & -.06745 & -.05526 \\
.280224 & .481385 & -.23106 & .529744 & -.02085 & -.44415 & .078278 & .593235 & -.03452 \\
-.04003 & -.32577 & -.00255 & -.00413 & .258949 & .073001 & -.29746 & .152387 & -.23548 \\
-.04003 & -.22544 & -.01887 & -.02171 & .261769 & .088842 & -.29746 & -.04116 & -.23418 \\
-.04003 & -.06584 & -.06784 & -.0497 & -.270231 & .089856 & -.29746 & -.20378 & -.21822 \\
-.04003 & .048167 & -.1168 & -.06968 & .278693 & .043276 & -.29746 & -.2107 & -.18419 \\
-.04003 & .207774 & -.05152 & -.09766 & .267411 & .019961 & -.29746 & -.06745 & -.22554 \\
-.04003 & .481385 & .005609 & -.14563 & .257539 & .0832021 & -.29746 & .593235 & -.23538 \\
-.36029 & -.32577 & .356528 & .452517 & -.29615 & -.46678 & .328768 & .152387 & .243742 \\
-.36029 & -.22544 & .470781 & .294227 & -.47384 & -.42042 & .328768 & -.04116 & .599999 \\
-.36029 & -.06584 & .421815 & .042401 & -.39769 & -.05859 & .328768 & -.20378 & .435271 \\
-.36029 & .207774 & .37285 & -.3893 & -.32153 & .427375 & .328768 & -.06745 & .288613 \\
.17888 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & .17888 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & .17888 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & .17888 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & .17888 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & .17888 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & .17888 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & .17888 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & .17888
\end{bmatrix}
\qquad
\mathbf{y}_A = \begin{bmatrix}
.27979 \\ .305829 \\ .312339 \\ .26894 \\ .24724 \\ .182141 \\ -.1759 \\ -.09995 \\ -.03486 \\ -.02401 \\ .041094 \\ .051944 \\ -.0458 \\ -.04146 \\ -.33865 \\ -.14335 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0
\end{bmatrix}$$

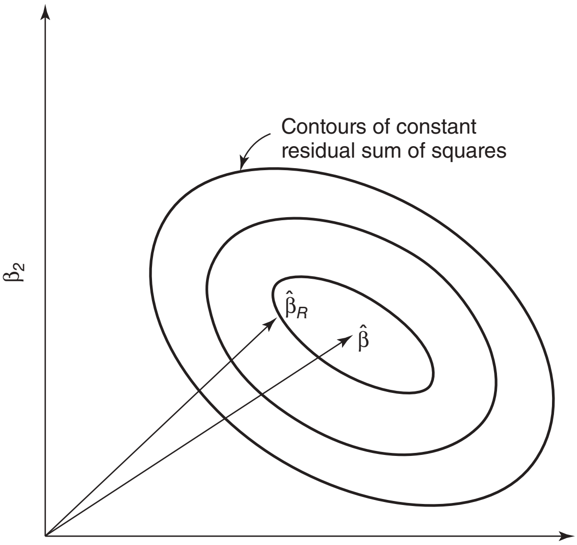

Figure 9.7 A geometrical interpretation of ridge regression.

conditions under which ridge regression leads to smaller MSE than least squares. The expected improvement depends on the orientation of the $\boldsymbol{\beta}$ vector relative to the eigenvectors of $\mathbf{X'X}$. The expected improvement is greatest when $\boldsymbol{\beta}$ coincides with the eigenvector associated with the largest eigenvalue of $\mathbf{X'X}$. Other interesting results appear in Lowerre [1974] and Mayer and Willke [1973].

Obenchain [1977] has shown that nonstochastically shrunken ridge estimators yield the same t and F statistics for testing hypotheses as does least squares. Thus, although ridge regression leads to biased point estimates, it does not generally require a new distribution theory. However, distributional properties are still unknown for stochastic choices of k. One would assume that when k is small, the usual normal-theory inference would be approximately applicable.

Relationship to Other Estimators Ridge regression is closely related to **Bayesian Estimation**. Generally, if prior information about $\boldsymbol{\beta}$ can be described by a p-variate normal distribution with mean vector $\boldsymbol{\beta}_0$ and covariance matrix $\mathbf{V}_0$, then the Bayes estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}}_{\mathrm{B}} = \left(\frac{1}{\sigma^2}\mathbf{X'X} + \mathbf{V}_0^{-1}\right)^{-1}\left(\frac{1}{\sigma^2}\mathbf{X'y} + \mathbf{V}_0^{-1}\boldsymbol{\beta}_0\right)$$

The use of Bayesian methods in regression is discussed in Leamer [1973, 1978] and Zellner [1971]. Two major drawbacks of this approach are that the data analyst must make an explicit statement about the form of the prior distribution and the statistical theory is not widely understood. However, if we choose the prior mean $\boldsymbol{\beta}_0 = \mathbf{0}$ and $\mathbf{V}_0 = \sigma_0^2\mathbf{I}$, then we obtain

$$\hat{\boldsymbol{\beta}}_{\mathrm{B}} + (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{y} \equiv \hat{\boldsymbol{\beta}}_{\mathrm{R}}, \quad 2k = \frac{\sigma^2}{\sigma_0^2}$$

the usual ridge estimator. In effect, the method of least squares can be viewed as a Bayes estimator using an unbounded uniform prior distribution for $\boldsymbol{\beta}$. The ridge estimator results from a prior distribution that places weak boundedness conditions on $\boldsymbol{\beta}$. Also see Lindley and Smith [1972].

Methods for Choosing k Much of the controversy concerning ridge regression centers around the choice of the biasing parameter k. Choosing k by inspection of the ridge trace is a subjective procedure requiring judgment on the part of the analyst. Several authors have proposed procedures for choosing k that are more analytical. Hoerl, Kennard, and Baldwin [1975] have suggested that an appropriate choice for k is

$$k = \frac{p\hat{\sigma}^2}{\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}}} \tag{9.8}$$

where $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are found from the least-squares solution. They showed via simulation that the resulting ridge estimator had significant improvement in MSE over least squares. In a subsequent paper, Hoerl and Kennard [1976] proposed an iterative estimation procedure based on Eq. (11.8). McDonald and Galarneau [1975] suggest choosing k so that

$$\hat{\boldsymbol{\beta}}'_R\hat{\boldsymbol{\beta}}_R = \hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}} - \sigma^2 \sum_{j=1}^{p} \left(\frac{1}{\lambda_j}\right)$$

A drawback to this procedure is that k may be negative, Mallows [1973] suggested a graphical procedure for selecting k based on a modification of his C_p statistic. Another approach chooses k to minimize a modification of the PRESS statistic. Wahba, Golub, and Health [1979] suggest choosing k to minimize a cross-validation statistic.

There are many other possibilities for choosing k. For example, Marquardt [1970] has proposed using a value of k such that the maximum VIP is between 1 and 10, preferably closer to 1. Other methods of choosing k have been suggested by Dempster, Schatzoff, and Wermuth [1971], Goldstein and Smith [1974], Lawless and Wang [1976], Lindley and Smith [1972], and Obenchain [1975]. Hoerl and Kennard [1970a] proposed an extension of standard ridge regression that allows separate k's for each regression. This is called **generalized ridge regression**. There is no guarantee that these methods are superior to straightforward inspection of the ridge trace.

9.5.4 Principal-Component Regression

Biased estimators of regression coefficients can also be obtained by using a procedure known as **principal-component regression**. Consider the canonical form of the model,

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}$$

where

$$\mathbf{Z} = \mathbf{XT}, \quad \boldsymbol{\alpha} = \mathbf{T}'\boldsymbol{\beta}, \quad \mathbf{T}'\mathbf{X}'\mathbf{XT} = \mathbf{Z}'\mathbf{Z} = \boldsymbol{\Lambda}$$

Recall that $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p)$ is a $p \times p$ diagonal matrix of the eigenvalues of $\mathbf{X}'\mathbf{X}$ and $\mathbf{T}$ is a $p \times p$ orthogonal matrix whose columns are the eigenvectors associated with $\lambda_1, \lambda_2, \ldots, \lambda_p$. The columns of $\mathbf{Z}$, which define a new set of orthogonal regressors, such as

$$\mathbf{Z} = [\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_p]$$

are referred to as **principal components.**

The least-squares estimator of $\hat{\boldsymbol{\alpha}}$ is

$$\hat{\boldsymbol{\alpha}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y} = \boldsymbol{\Lambda}^{-1}\mathbf{Z}'\mathbf{y}$$

and the covariance matrix of $\hat{\boldsymbol{\alpha}}$ is

$$\text{Var}(\hat{\boldsymbol{\alpha}}) = \sigma^2(\mathbf{Z}'\mathbf{Z})^{-1} = \sigma^2\boldsymbol{\Lambda}^{-1}$$

Thus, a small eigenvalue of $\mathbf{X}'\mathbf{X}$ means that the variance of the corresponding orthogonal regression coefficient will be large. Since

$$\mathbf{Z}'\mathbf{Z} = \sum_{i=1}^{p}\sum_{j=1}^{p}\mathbf{Z}_i\mathbf{Z}_j' = \boldsymbol{\Lambda}$$

we often refer to the eigenvalue λ_j as the variance of the jth principal component. If all the λ_j are equal to unity, the **original** regressors are orthogonal, while if a λ_j is exactly equal to zero, this implies a perfect linear relationship between the **original** regressors. One or more of the λ_j near zero implies that multicollinearity is present. Note also that the covariance matrix of the standardized regression coefficients $\hat{\boldsymbol{\beta}}$ is

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \text{Var}(\mathbf{T}\hat{\boldsymbol{\alpha}}) = \mathbf{T}\boldsymbol{\Lambda}^{-1}\mathbf{T}'\sigma^2$$

This implies that the variance of $\hat{\beta}_j$ is $\hat{\sigma}^2\left(\sum_{j=1}^{p} t_{ji}^2/\lambda_i\right)$. Therefore, the variance of $\hat{\beta}_j$ is a linear combination of the reciprocals of the eigenvalues. This demonstrates how one or more small eigenvalues can destroy the precision of the least-squares estimate $\hat{\beta}_j$.

We have observed previously how the eigenvalues and eigenvectors of $\mathbf{X}'\mathbf{X}$ provide specific information on the nature of the multicollinearity. Since $\mathbf{Z} = \mathbf{XT}$, we have

$$\mathbf{Z}_i = \sum_{j=1}^{p} t_{ji}\mathbf{X}_j \tag{9.9}$$

where $\mathbf{X}_j$ is the jth column of the $\mathbf{X}$ matrix and t_{ji} are the elements of the ith column of $\mathbf{T}$ (the ith eigenvector of $\mathbf{X}'\mathbf{X}$). If the variance of the ith principal component (λ_j) is small, this implies that $\mathbf{Z}_i$ is nearly constant, and Eq. (9.9) indicates that there is a linear combination of the **original regressors** that is nearly constant. This is the definition of multicollinearity, that is, the t_{ji} are the constants in Eq. (9.1). Therefore, Eg. (9.9) explains why the elements of the eigenvector associated with a smaIl eigenvalue of $\mathbf{X'X}$ identify the regressors involved in the multicollinearity.

The principal-component regression approach combats multicollinearity by using less than the full set of principal components in the model. To obtain the principal-component estimator, assume that the regressors are arranged in order of decreasing eigenvalues, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$. Suppose that the last s of these eigenvalues are approximately equal to zero. In principal-component regression the principal components corresponding to near-zero eigenvalues are removed from the analysis and least squares applied to the remaining components. That is,

$$\hat{\boldsymbol{\alpha}}_{\mathrm{PC}} = \mathbf{B}\hat{\boldsymbol{\alpha}}$$

where $b_1 = b_2 = \cdots = b_{p-s} = 1$ and $b_{p-s+1} = b_{p-s+2} = \cdots = b_p = 0$. Thus, the principal-component estimator is

$$\hat{\boldsymbol{\alpha}}_{\mathrm{PC}} = \left[\begin{array}{c} \hat{\alpha}_1 \\ \hat{\alpha}_2 \\ \vdots \\ \hat{\alpha}_{p-s} \\ \hdashline 0 \\ 0 \\ \vdots \\ 0 \end{array}\right] \begin{array}{l} \\ \\ \\ p-s \text{ components} \\ s \text{ components} \\ \\ \\ \\ \end{array}$$

or in terms of the standardized regressors

$$\hat{\boldsymbol{\beta}}_{\mathrm{PC}} = \mathbf{T}\hat{\boldsymbol{\alpha}}_{\mathrm{PC}} = \sum_{j=1}^{p-s} \lambda_j^{-1}\mathbf{t}_j'\mathbf{X}'\mathbf{y}\mathbf{t}_j \tag{9.10}$$

A simulation study by Gunst and Mason [1977] showed that principal-component regression offers considerable improvement over least squares when the data are ill-conditioned. They also point out that another advantage of principal components is that exact distribution theory and variable selection procedures are available (see Mansfield, Webster, and Gunst [1977]). Some computer packages will perform principal-component regression.

Example 9.3 Principal-Component Regression for the Acetylene Data

We illustrate the use of principal-component regression for the acetylene data. We begin with the linear transformation $\mathbf{Z} = \mathbf{XT}$ that transforms the original standardized regressors into an orthogonal set of variables (the principal components). The

TABLE 9.10 Matrix *T* of Eigenvectors and Eigenvalnes λ_j for the Acetylene Data

Eigenvectors									Eigenvalues λ_j
.3387	.1057	.6495	.0073	.1428	−.2488	−.2077	−.5436	.1768	4.20480
.1324	.3391	−.0068	−.7243	−5843	.0205	−.0102	−.0295	−.0035	2.16261
−.4137	−.0978	−.4696	−.0718	−.0182	.0160	−.1468	−.7172	.2390	1.13839
−.2191	.5403	.0897	.3612	−.1661	.3733	−.5885	.0909	.0003	1.04130
.4493	.0860	−.2863	.1912	−.0943	.0333	.0575	.1543	.7969	0.38453
.2524	—.5172	−.0570	−.3447	.2007	.3232	−.6209	.1280	.0061	0.04951
−.4056	−.0742	.4404	−.2230	.1443	.5393	.3233	.0565	.4087	0.01363
.0258	.5316	−.2240	−.3417	.7342	−.0705	−.0057	.0761	.0050	0.00513
−.4667	−.0969	.1421	−.1337	−.0350	−.6299	−.3089	.3631	.3309	0.00010

eigenvalues λ_j and the **T** matrix for the acetylene data are shown in Table 9.10. This matrix indicates that the relationship between z_1 (for example) and the standardized regressors is

$$\begin{aligned} z_1 &= 0.3387x_1 + 0.1324x_2 - 0.4137x_3 - 0.2191x_1x_2 + 0.4493x_1x_3 \\ &\quad + 0.2524x_2x_3 - 0.4056x_1^2 + 0.0258x_2^2 - 0.4667x_3^2 \end{aligned}$$

The relationships between the remaining principal components $z_2, z_3, \ldots, z_9$ and the standardized regressors are determined similarly. Table 9.11 shows the elements of the **Z** matrix (sometimes called the principal-component scores).

The principal-component estimator reduces the effects of multicollinearity by using a subset of the principal components in the model. Since there are four small eigenvalues for the acetylene data, this implies that there are four principal components that should be deleted. We will exclude z_6, z_7, z_8, and z_9 and consider regressions involving only the first five principal components.

Suppose we consider a regression model involving only the first principal component, as in

$$y = \alpha_1 z_1 + \varepsilon$$

The fitted model is

$$\hat{y} = -0.35225z_1$$

or $\hat{\boldsymbol{\alpha}}'_{\text{PC}} = [-0.35225, 0, 0, 0, 0, 0, 0, 0, 0]$. The coefficients in terms of the standardized regressors are found from $\hat{\boldsymbol{\beta}}_{\text{PC}} = \mathbf{T}\hat{\boldsymbol{\alpha}}_{\text{PC}}$. Panel A of Table 9.11 shows the resulting standardized regression coefficients as well as the regression coefficients in terms of the original centered regressors. Note that even though only one principal component is included, the model produces estimates for all nine standardized regression coefficients.

The results of adding the other principal components z_2, z_3, z_4, and z_5 to the model one at a time are displayed in panels B, C, D, and E, respectively, of Table 9.12. We see that using different numbers of principal components in the model produces

TABLE 9.11 Matrix Z = XT for the Acetylene Data

Observation	Z_1	Z_2	Z_3	$Z_4(=Z_{x_1x_2})$	$Z_5(=Z_{x_1x_3})$	$Z_6(=Z_{x_2x_3})$	$Z_7\left(=Z_{x_1^2}\right)$	$Z_8\left(=Z_{x_2^2}\right)$	$Z_9\left(=Z_{x_3^2}\right)$
1	.5415	−1.0347	1.0487	−.1880	1.7389	−.6593	.6492	.7822	.2402
2	.4846	−.8830	1.1638	−.0468	.8909	−.3874	.5067	.2045	−.1939
3	.4046	−.6129	.2914	.0676	−.0025	−.1631	.2187	−.0898	−1.6609
4	.3388	−.1513	3.3176	.1315	−.7526	.3579	.1269	−1.2150	.9250
5	.2353	.6905	1.2785	−.0089	−1.0842	.6884	−.4181	−1.2768	1.6754
6	.0310	2.7455	.9535	−.7783	.2235	.2093	−1.1200	1.3128	−1.1453
7	.5940	−.0165	−1.0885	1.1554	1.5790	.1926	−1.3363	−.4626	.5964
8	.6385	−.2399	−.9170	1.0916	.3634	.4238	−1.2453	−.7138	−.3611
9	.7139	−.3558	−.7151	.8354	−.9374	.3207	−.6525	.5144	−.7716
10	.7436	−.2228	−.6170	.5668	−1.4297	−.4038	.5657	2.5203	1.4085
11	.7668	.1034	−.8626	−.0706	−1.3472	−.3706	1.5958	−.8815	−1.3485
12	.8726	1.1054	−1.5272	−1.8442	.8129	−.9285	.8411	−.8981	.7053
13	−1.7109	.8164	−.3702	1.2052	.8885	1.9123	2.0708	.2251	−.1036
14	−2.1618	.1860	−.1026	.5619	−.1290	−2.5588	−.3380	−.1080	.8652
15	−1.6050	−.6784	−.2117	−.3325	−.7456	−.0658	−.8259	−.4662	−1.0012
16	−.8875	−1.4521	−.6417	−2.3461	−.0690	1.4324	−.6387	.5524	.1699

TABLE 9.12 Principal Components Regression for the Acetylene Data

	Principal Components in Model									
	A		B		C		D		E	
Parameter	z_1		z_1, z_2		z_1, z_2, z_3		z_1, z_2, z_3, z_4		z_1, z_2, z_3, z_4, z_5	
	Standarized Estimate	Original Estimate	Standardized Estimate	Original Estimate	Standardized Estimate	Original Estimate	Standardized Estimate	Original Estimate	Standardized Estimate	Original Estimate
β_0	.0000	42.1943	.0000	42.2219	.0000	36.6275	.0000	34.6688	.0000	34.7517
β_1	.1193	1.4194	.1188	1.4141	.5087	6.0508	.5070	6.0324	.5056	6.0139
β_2	.0466	.5530	.0450	.5346	.0409	.4885	.2139	2.5438	.2195	2.6129
β_3	–.1457	–1.7327	–.1453	–1.7281	–.4272	–5.0830	–.4100	–4.8803	–.4099	4.8757
β_{12}	–.0772	–1.0369	–.0798	–1.0738	–.0260	–.3502	–.1123	–1.5115	–.1107	–1.4885
β_{13}	.1583	2.0968	.1578	2.0922	–.0143	–.1843	–.0597	–.7926	–.0588	–.7788
β_{23}	.0889	1.2627	.0914	1.2950	.0572	.8111	.1396	1.9816	.1377	1.9493
β_{11}	–.1429	–2.1429	–.1425	–2.1383	.1219	1.8295	.1751	2.6268	.1738	2.6083
β_{22}	.0091	.0968	.0065	.0691	–.1280	–1.3779	–.0460	–.4977	–.0533	–.5760
β_{33}	–.1644	–1.9033	–.1639	–1.8986	–.0786	–.9125	–.0467	–.5392	–.0463	–.5346
R^2	.5217		.5218		.9320		.9914		.9915	
MS_{Res}	.079713		.079705		.011333		.001427		.00142	

substantially different estimates of the regression coefficients. Furthermore, the principal-component estimates differ considerably from the least-squares estimates (e.g., see Table 9.8). However, the principal-component procedure with either four or five components included results in coefficient estimates that do not differ dramatically from those produced by the other biased estimation metbods (refer to the ordinary ridge regression estimates in Table 9.9. Principal-component analysis shrinks the large least-squares estimates of β_{13} and β_{33} and changes the sign of the original negative least-squares estimate of β_{11}. The five-component model does not substantially degrade the fit to the original data as there has been little loss in R^2 from the least-squares model. Thus, we conclude that the relationship based on the first five principal components provides a more plausible model for the acetylene data than was obtained via ordinary least squares.

Marquardt [1970] suggested a generalization of principal-component regression. He felt that the assumption of an integral rank for the **X** matrix is too restrictive and proposed a "fractional rank" estimator that allows the rank to be a piecewise continuous function.

Hawkins [1973] and Webster et al. [1974] developed latent root procedures following the same philosophy as principal components. Gunst, Webster, and Mason [1976] and Gunst and Masou [1977] indicate that latent root regression may provide considerable improvement in mean square error over least squares. Gunst [1979] points out that latent root regression can produce regression coefficients that are very sinillar to those found by principal components, particularly when there are only one or two strong multicollinearities in **X**. A number of large-sample properties of latent root regression are in White and Gunst [1979]. ■

9.5.5 Comparison and Evaluation of Biased Estimators

A number of Monte Carlo simulation studies have been conducted to examine the effectiveness of biased estimators and to attempt to determine which procedures perform best. For example, see McDonald and Galarneau [1975], Hoerl and Kennard [1976], Hoerl, Kennard, and Baldwin [1975] (who compare least squares and ridge), Gunst et al. [1976] (latent root versus least squares), Lawless [1978], Hemmerle and Brantle [1978] (ridge, generalized ridge, and least squares), Lawless and Wang [1976] (least squares, ridge, and principal components), Wichern and Churchill [1978], Gibbons [1979] (various forms of ridge), Gunst and Mason [1977] (ridge, principal components, latent root, and others), and Dempster et al. [1977]. The Dempster et al. [1977] study compared 57 different estimators for 160 different model configurations. While no single procedure emerges from these studies as best overall, there is considerable evidence indicating the superiority of biased estimation to least squares if multicollinearity is present. Our own preference in practice is for ordinary ridge regression with k selected by inspection of the ridge trace. The procedure is straightforward and easy to implement on a standard least-squares computer program, and the analyst can learn to interpret the ridge trace very quickly. It is also occasionally useful to find the "optimum" value of k suggested by Hoerl, Kennard, and Baldwin [1975] and the iteratively estimated "optimum" k of Hoed and Kennard [1976] and compare the resulting models with the one obtained via the ridge trace.

The use of biased estimators in regression is not without controversy. Several authors have been critical of ridge regression and other related biased estimation techniques. Conniffe and Stone [1973, 1975] have criticized the use of the ridge trace to select the biasing parameter, since $\hat{\boldsymbol{\beta}}_{\mathrm{R}}$ will change slowly and eventually stabilize as k increases even for orthogonal regressors. They also claim that if the data are not adequate to support a least-squares analysis, then it is unlikely that ridge regression will be of any substantive help, since the parameter estimates will be nonsensical. Marquardt and Snee [1975] and Smith and Goldstein [1975] do not accept these conclusions and feel that biased estimators are a valuable tool for the data analyst confronted by ill-conditioned data. Several authors have noted that while we can prove that there exists a k such that the mean square error of the ridge estimator is always less than the mean square error of the least-squares estimator, there is no assurance that the ridge trace (or any other method that selects the biasing parameter stochastically by analysis of the data) produces the optimal k.

Draper and Van Nostrand [1977a, b, 1979] are also critical of biased estimators. They find fault with a number of the technical details of the simulation studies used as the basis of claims of improvement in MSE for biased estimation, suggesting that the simulations have been designed to favor the biased estimators. They note that ridge regression is really only appropriate in situations where external information is added to a least-squares problem. This may take the form of either the Bayesian formulation and interpretation of the procedure or a constrained least-squares problem in which the constraints on $\boldsymbol{\beta}$ are chosen to reflect the analyst's knowledge of the regression coefficients to "improve the conditioning" of the data.

Smith and Campbell [1980] suggest using explicit Bayesian analysis or mixed estimation to resolve multicollinearity· problems. They reject ridge methods as weak and imprecise because they only loosely incorporate prior beliefs and information into the analysis. When explicit prior information is known, then Bayesian or mixed estimation should certainly be used. However, often the prior information is not easily reduced to a specific prior distribution, and ridge regression methods offer a method to incorporate, at least approximately, this knowledge.

There has also been some controversy surrounding whether the regressors and the response should be centered and scaled so that $\mathbf{X}'\mathbf{X}$ and $\mathbf{X}'\mathbf{y}$ are in correlation form. This results in an artificial removal of the intercept from the model. Effectively the intercept in the ridge model is estimated by $\bar{y}$. Hoerl and Kennard [1970a, b] use this approach, as do Marquardt and Snee [1975], who note that centering tends to minimize any nonessential ill-conditioning when fitting polynomials. On the other hand, Brown [1977] feels that the variables should not be centered, as centering affects only the intercept estimate and not the slopes. Belsley, Kuh, and Welsch [1980] suggest not centering the regressors so that the role of the intercept in any near-linear dependences may be diagnosed. Centering and scaling allow the analyst to think of the parameter estimates as standardized regression coefficients, which is often intuitively appealing. Furthermore, centering the regressors can remove nonessential ill-conditioning, thereby reducing variance inflation in the parameter estimates. Consequently, we recommend both centering and scaling the data.

Despite the objections noted, we believe that biased estimation methods are useful techniques that the analyst should consider when dealing with multicollinearity. Biased estimation methods certainly compare very favorably to other methods for handling multicollinearity, such as variable elimination. As Marquardt and Snee [1975] note, it is often better to use some of the information in all of the regressors, as ridge regression does, than to use all of the information in some regressors and none of the information in others, as variable elimination does. Furthermore, variable elimination can be thought of as a form of biased estimation because subset regression models often produce biased estimates of the regression coefficients. In effect, variable elimination often shrinks the vector of parameter estimates, as does ridge regression. We do not recommend the mechanical or automatic use of ridge regression without thoughtful study of the data and careful analysis of the adequacy of the final model. Properly used, biased estimation methods are a valuable tool in the data analyst's kit.

9.6 USING SAS TO PERFORM RIDGE AND PRINCIPAL-COMPONENT REGRESSION

Table 9.14 gives the SAS code to perform ridge regression for the acetylene data. The lines immediately prior to the cards statement center and scale the linear terms. The other statements create the interaction and pure quadratic terms. The option

```
ridge = 0.006 to 0.04 by .002
```

on the first proc reg statement creates the series of *k*'s to be used for the ridge trace. Typically, we would start the range of values for *k* at 0, which would yield the ordinary least-squares (OLS) estimates. Unfortunately, for the acetylene data the OLS estimates greatly distort the ridge trace plot to the point that it is very difficult to select a good choice for *k*. The statement

```
plot / ridgeplot nomodel;
```

creates the actual ridge trace. The option

```
ridge = .032
```

on the second proc reg statement fixes the value of *k* to 0.032.

Table 9.15 gives the additional SAS code to perform principal-component regression. The statement

```
proc princomp data=acetylene out=pc_acetylene std,
```

sets up the principal-component analysis and creates an output data data set called

```
pc_acetylene.
```

The std option standardizes the principal-component scores to unit variance. The statement

TABLE 9.14 SAS Code to Perform Ridge Regression for Acetylene Data

```
data acetylene;
input conv t h c;
t =(t - 1212.5) / 80.623;
h =(h - 12.44) / 5.662;
c =(c - 0.0403) / 0.03164;
th = t*h;
tc = t*c;
hc = h*c;
t2 = t*t;
h2 = h*h;
c2 = c*c;
cards;
49.0 1300  7.5 0.0120
50.2 1300  9.0 0.0120
50.5 1300 11.0 0.0115
48.5 1300 13.5 0.0130
47.5 1300 17.0 0.0135
44.5 1300 23.0 0.0120
28.0 1200  5.3 0.0400
31.5 1200  7.5 0.0380
34.5 1200 11.0 0.0320
35.0 1200 13.5 0.0260
38.0 1200 17.0 0.0340
38.5 1200 23.0 0.0410
15.0 1100  5.3 0.0840
17.0 1100  7.5 0.0980
20.5 1100 11.0 0.0920
29.5 1100 17.0 0.0860
proc reg outest = b ridge = 0.006 to 0.04 by .002;
model conv = t h c t2 h2 c2 th tc hc / noprint;
plot / ridgeplot nomodel;
run;
proc reg outest = b2 data = acetylene ridge =.032;
model conv = t h c t2 h2 c2 th tc hc; run;proc print data = b2i
run;
```

```
var t h c th tc hc t2 h2 c2;
```

specifies the specific variables from which to create the principal components. In this case, the variables are all of the regressors. The statement

```
ods select eigenvectors eigenvalues;
```

creates the eigenvectors and eigenvalues. The other two ods statements set up the output. This procedure creates the principal component, names them prinl, prin2, and so on, and stores them in the output data set, which in this example is

TABLE 9.15 SAS Code to Perform Principal-Component Regression for Acetylene Data

```
proc princomp data = acetylene out = pc_acetylene std;
var t h c th tc he t2 h2 c2;
ods select eigenvectors eigenvalues;
ods trace on;
ods show;
run;
proc reg data = pc_acetylene;
model conv = prinl prin2 prin3 prin4 prin5 prin6 prin7 prin8 prin9 / vif;
run;
proc reg data = pc_acetylene;
model conv = prinl;
run;
proc reg data = pc_acetylene;
model conv = prinl prin2;
run;
```

pc_acetylene

The remainder of the code illustrates how to use proc reg with the principal components as the regressors. SAS does not automatically convert the resulting regression equation in the principal components back to the original variables. The analyst must perform this calculation using the appropriate eigenvectors.

PROBLEMS

9.1 Consider the soft drink delivery time data in Example 3.1.

a. Find the simple correlation between cases (x_1) an distance (x_2).

b. Find the variance inflation factors.

c. Find the condition number of $\mathbf{X'X}$. Is there evidence of multicollinearity in these data?

9.2 Consider the Hald cement data in Table B.21.

a. From the matrix of correlations between the regressors, would you suspect that multicollinearity is present?

b. Calculate the variance inflation factors.

c. Find the eigenvalues of $\mathbf{X'X}$.

d. Find the condition number of $\mathbf{X'X}$.

9.3 Using the Hald cement data (Example 10.1), find the eigenvector associated with the smallest eigenvalue of $\mathbf{X'X}$. Interpret the elements of this vector. What can you say about the source of multicollinearity in these data?

9.4 Find the condition indices and the variance decomposition proportions for the Hald cement data (Table B.21), assuming centered regressors. What can you say about multicollinearity in these data?