CHAPTER 10

BIASED ESTIMATION OF REGRESSION COEFFICIENTS

10.1 INTRODUCTION

It was demonstrated in Chapter 9 that when multicollinearity is present in a set of predictor variables, the ordinary least squares estimates of the individual regression coefficients tend to be unstable and can lead to erroneous inferences. In this chapter, two alternative estimation methods that provide a more informative analysis of the data than the OLS method when multicollinearity is present are considered. The estimators discussed here are biased but tend to have more precision (as measured by mean square error) than the OLS estimators (see Draper and Smith (1998), McCallum (1970), and Hoerl and Kennard (1970)). These alternative methods do not reproduce the estimation data as well as the OLS method; the sum of squared residuals is not as small and, equivalently, the multiple correlation coefficient is not as large. However, the two alternatives have the potential to produce more precision in the estimated coefficients and smaller prediction errors when the predictions are generated using data other than those used for estimation.

Unfortunately, the criteria for deciding when these methods give better results than the OLS method depend on the true but unknown values of the model regression coefficients. That is, there is no completely objective way to decide when OLS should be replaced in favor of one of the alternatives. Nevertheless, when mul-

Regression Analysis by Example, Fourth Edition. By Samprit Chatterjee and Ali S. Hadi
Copyright © 2006 John Wiley & Sons, Inc.

ticollinearity is suspected, the alternative methods of analysis are recommended. The resulting estimated regression coefficients may suggest a new interpretation of the data that, in turn, can lead to a better understanding of the process under study.

The two specific alternatives to OLS that are considered are (1) principal components regression and (2) ridge regression. Principal components analysis was introduced in Chapter 9. It is assumed that the reader is familiar with that material. It will be demonstrated that the principal components estimation method can be interpreted in two ways; one interpretation relates to the nonorthogonality of the predictor variables, the other has to do with constraints on the regression coefficients. Ridge regression also involves constraints on the coefficients. The ridge method is introduced in this chapter and it is applied again in Chapter 11 to the problem of variable selection. Both methods, principal components and ridge regression, are examined using the French import data that were analyzed in Chapter 9.

10.2 PRINCIPAL COMPONENTS REGRESSION

The model under consideration is

$$\text{IMPORT} = \beta_0 + \beta_1 \cdot \text{DOPROD} + \beta_2 \cdot \text{STOCK} + \beta_3 \cdot \text{CONSUM} + \varepsilon. \quad (10.1)$$

The variables are defined in Section 9.3. Let $\bar{y}$ and $\bar{x}_j$ be the means of Y and X_j, respectively. Also, let s_y and s_j be the standard deviations of Y and X_j, respectively. The model of Equation (10.1) stated in terms of standardized variables (see Section 9.5) is

$$\tilde{Y} = \theta_1 \tilde{X}_1 + \theta_2 \tilde{X}_2 + \theta_3 \tilde{X}_3 + \varepsilon', \quad (10.2)$$

where $\tilde{Y} = (y_i - \bar{y})/s_y$ is the standardized version of the response variable and $\tilde{X}_j = (x_{ij} - \bar{x}_j)/s_j$ is the standardized version of the jth predictor variable. Many regression packages produce values for both the regular and standardized regression coefficients in (10.1) and (10.2), respectively. The estimated coefficients satisfy

$$\begin{aligned} \beta_j &= (s_y/s_j)\theta_j, \qquad j = 1, 2, 3, \\ \beta_0 &= \bar{y} - \beta_1 \bar{x}_1 - \beta_2 \bar{x}_2 - \beta_3 \bar{x}_3. \end{aligned} \quad (10.3)$$

The principal components of the standardized predictor variables are (see Equation (9.20))

$$\begin{aligned} C_1 &= 0.706\,\tilde{X}_1 + 0.044\,\tilde{X}_2 + 0.707\,\tilde{X}_3, \\ C_2 &= -0.036\,\tilde{X}_1 + 0.999\,\tilde{X}_2 - 0.026\,\tilde{X}_3, \\ C_3 &= -0.707\,\tilde{X}_1 - 0.007\,\tilde{X}_2 + 0.707\,\tilde{X}_3. \end{aligned} \quad (10.4)$$

These principal components were given in Table 9.13. The model in (10.2) may be written in terms of the principal components as

$$\tilde{Y} = \alpha_1 C_1 + \alpha_2 C_2 + \alpha_3 C_3 + \varepsilon'. \quad (10.5)$$

Table 10.1 Regression Results of Fitting Model (10.2) to the Import Data (1949–1959)

Variable	Coefficient	s.e.	t-test	p-value
$\tilde{X}_1$	−0.339	0.464	−0.73	0.4883
$\tilde{X}_2$	0.213	0.034	6.20	0.0004
$\tilde{X}_3$	1.303	0.464	2.81	0.0263
$n = 11$	$R^2 = 0.992$	$R_a^2 = 0.988$	$\hat{\sigma} = 0.034$	$d.f. = 7$

Table 10.2 Regression Results of Fitting Model (10.5) to the Import Data (1949–1959)

Variable	Coefficient	s.e.	t-test	p-value
C_1	0.690	0.024	28.70	< 0.0001
C_2	0.191	0.034	5.62	0.0008
C_3	1.160	0.656	1.77	0.1204
$n = 11$	$R^2 = 0.992$	$R_a^2 = 0.988$	$\hat{\sigma} = 0.034$	$d.f. = 7$

The equivalence of (10.2) and (10.5) follows since there is a unique relationship between the α's and θ's. In particular,

$$\begin{aligned} \alpha_1 &= 0.706\theta_1 + 0.044\theta_2 + 0.707\theta_3, \\ \alpha_2 &= -0.036\theta_1 + 0.999\theta_2 - 0.026\theta_3, \\ \alpha_3 &= -0.707\theta_1 - 0.007\theta_2 + 0.707\theta_3. \end{aligned} \tag{10.6}$$

Conversely,

$$\begin{aligned} \theta_1 &= 0.706\alpha_1 - 0.036\alpha_2 - 0.707\alpha_3, \\ \theta_2 &= 0.044\alpha_1 + 0.999\alpha_2 - 0.007\alpha_3, \\ \theta_3 &= 0.707\alpha_1 - 0.026\alpha_2 + 0.707\alpha_3. \end{aligned} \tag{10.7}$$

These same relationships hold for the least squares estimates, the $\hat{\alpha}$'s and $\hat{\theta}$'s of the α's and θ's, respectively. Therefore, the $\hat{\alpha}$'s and $\hat{\theta}$'s may be obtained by the regression of $\tilde{Y}$ against the principal components C_1, C_2, and C_3, or against the original standardized variables. The regression results of fitting models (10.2) and (10.5) to the import data are shown in Tables 10.1 and 10.2. From Table 10.1, the estimates of θ_1, θ_2, and θ_3 are −0.339, 0.213, and 1.303, respectively. Similarly, from Table 10.2, the estimates of α_1, α_2, and α_3 are 0.690, 0.191, and 1.160, respectively. The results in one of these tables can be obtained from the other table using (10.6) and (10.7).

Although Equations (10.2) and (10.5) are equivalent, the C's in (10.5) are orthogonal. Observe, however, that the regression relationship given in terms of the principal components (Equation (10.5)) is not easily interpreted. The predictor variables of that model are linear combinations of the original predictor variables. The α's, unlike the θ's, do not have simple interpretations as marginal effects of the original predictor variables. Therefore, we use principal components regression only as a means for analyzing the multicollinearity problem. The final estimation results are always restated in terms of the θ's for interpretation.

10.3 REMOVING DEPENDENCE AMONG THE PREDICTORS

It has been mentioned that the principal components regression has two interpretations. We shall first use the principal components technique to reduce multicollinearity in the estimation data. The reduction is accomplished by using less than the full set of principal components to explain the variation in the response variable. Note that when all three principal components are used, the OLS solution is reproduced exactly by applying Equations (10.7).

The C's have sample variances $\lambda_1 = 1.999, \lambda_2 = 0.998$, and $\lambda_3 = 0.003$, respectively. Recall that the λ's are the eigenvalues of the correlation matrix of DOPROD, STOCK, and CONSUM. Since C_3 has variance equal to 0.003, the linear function defining C_3 is approximately equal to zero and is the source of multicollinearity in the data. We exclude C_3 and consider regressions of $\tilde{Y}$ against C_1 alone as well as against C_1 and C_2. We consider the two possible regression models

$$\tilde{Y} = \alpha_1 C_1 + \varepsilon \tag{10.8}$$

and

$$\tilde{Y} = \alpha_1 C_1 + \alpha_2 C_2 + \varepsilon. \tag{10.9}$$

Both models lead to estimates for all three of the original coefficients, θ_1, θ_2, and θ_3. The estimates are biased since some information (C_3 in Equation (10.9), C_2 and C_3 in Equation (10.8)) has been excluded in both cases.

The estimated values of α_1 or α_1 and α_2 may be obtained by regressing $\tilde{Y}$ in turn against C_1 and then against C_1 and C_2. However, a simpler computational method is available that exploits the orthogonality of C_1 C_2 and C_3.[1] For example, the same estimated value of α_1 will be obtained from regression using (10.5), (10.8), or (10.9). Similarly, the value of α_2 may be obtained from (10.5) or (10.9). It also follows that if we have the OLS estimates of the θ's, estimates of the α's may be obtained from Equations (10.6). Then principal components regression estimates of the θ's corresponding to (10.8) and (10.9) can be computed by referring back to Equations (10.7) and setting the appropriate α's to zero. The following example clarifies the process.

[1] In any regression equation where the full set of potential predictor variables under consideration are orthogonal, the estimated values of regression coefficients are not altered when subsets of these variables are either introduced or deleted.

Using $\alpha_1 = 0.690$ and $\alpha_2 = \alpha_3 = 0$ in Equations (10.7) yields estimated θ's corresponding to regression on only the first principal component, that is,

$$\begin{aligned} \hat{\theta}_1 &= 0.706 \times 0.690 = 0.487, \\ \hat{\theta}_2 &= 0.044 \times 0.690 = 0.030, \\ \hat{\theta}_3 &= 0.707 \times 0.690 = 0.487, \end{aligned} \tag{10.10}$$

which yields

$$\tilde{Y} = 0.487\tilde{X}_1 + 0.030\tilde{X}_2 + 0.487\tilde{X}_3.$$

The estimates using the first two principal components, as in (10.9), are obtained in a similar fashion using $\alpha_1 = 0.690$, $\alpha_2 = 0.191$, and $\alpha_3 = 0$ in (10.7). The estimated of the regression coefficients, β_0, β_1, β_2, and β_3, of the original variables in Equation (10.1), can be obtained by substituting θ_1, θ_2, and θ_3 in (10.3).

The estimates of the standardized and original regression coefficients using the three principal components models are shown in Table 10.3. It is evident that using different numbers of principal components gives substantially different results. It has already been argued that the OLS estimates are unsatisfactory. The negative coefficient of $\tilde{X}_1$ (DOPROD) is unexpected and cannot be sensibly interpreted. Furthermore, there is extensive multicollinearity which enters through the principal component, C_3. This variable has almost zero variance ($\lambda_3 = 0.003$) and is therefore approximately equal to zero. Of the two remaining principal components, it is fairly clear that the first one is associated with the combined effect of DOPROD and CONSUM. The second principal component is uniquely associated with STOCK. This conclusion is apparent in Table 10.3. The coefficients of DOPROD and CONSUM are completely determined from the regression of IMPORT on C_1 alone. These coefficients do not change when C_2 is used. The addition of C_2 causes the coefficient of STOCK to increase from 0.083 to 0.609. Also, R^2 increases from 0.952 to 0.988. Selecting the model based on the first two principal components, the resulting equation stated in original units is

$$\begin{aligned} \widehat{\text{IMPORT}} = \; & -9.106 + 0.073 \cdot \text{DOPROD} \\ & + 0.609 \cdot \text{STOCK} + 0.106 \cdot \text{CONSUM}. \end{aligned} \tag{10.11}$$

It provides a different and more plausible representation of the IMPORT relationship than was obtained from the OLS results. In addition, the analysis has led to an explicit quantification (in standardized variables) of the linear dependency in the predictor variables. We have $C_3 = 0$ or equivalently (from Equations (10.4))

$$-0.707\tilde{X}_1 - 0.007\tilde{X}_2 + 0.707\tilde{X}_3 \doteq 0.$$

The standardized values of DOPROD and CONSUM are essentially equal. This information can be useful qualitatively and quantitatively if Equation (10.11) is used for forecasting or for analyzing policy decisions.

Table 10.3 Estimated Regression Coefficients for the Standardized and Original Variables Using Different Numbers of Principal Components for IMPORT Data (1949–1959)

	First PC Equation (10.8)		First and Second PCs Equation (10.9)		All PCs Equation (10.5)	
Variable	Stand.	Original	Stand.	Original	Stand.	Original
Constant	0	−7.735	0	−9.106	0	−10.130
DOPROD	0.487	0.074	0.480	0.073	−0.339	−0.051
STOCK	0.030	0.083	0.221	0.609	0.213	0.587
CONSUM	0.487	0.107	0.483	0.106	1.303	0.287
$\hat{\sigma}$	0.232		0.121		0.108	
R^2	0.952		0.988		0.992	

10.4 CONSTRAINTS ON THE REGRESSION COEFFICIENTS

There is a second interpretation of the results of the principal components regression equation. The interpretation is linked to the notion of imposing constraints on the θ's which was introduced in Chapter 9. The estimates for Equation (10.9) were obtained by setting α_3 equal to zero in Equations (10.7). From (10.6), $\alpha_3 = 0$ implies that

$$-0.707\theta_1 - 0.007\theta_2 + 0.707\theta_3 = 0 \tag{10.12}$$

or $\theta_1 \doteq \theta_3$. In original units, Equation (10.12) becomes

$$-6.60\beta_1 + 4.54\beta_3 = 0 \tag{10.13}$$

or $\beta_1 = 0.69\beta_3$. Therefore, the estimates obtained by regression on C_1 and C_2 could have been obtained using OLS as in Chapter 9 with a linear constraint on the coefficients given by Equation (10.13).

Recall that in Chapter 9 we conjectured that $\beta_1 = \beta_3$ as a prior constraint on the coefficients. It was argued that the constraint was the result of a qualitative judgment based on knowledge of the process under study. It was imposed without looking at the data. Now, using the data, we have found that principal components regression on C_1 and C_2 gives a result that is equivalent to imposing the constraint of Equation (10.13). The result suggests that the marginal effect of domestic production on imports is about 69% of the marginal effect of domestic consumption on imports.

To summarize, the method of principal components regression provides both alternative estimates of the regression coefficients as well as other useful information about the underlying process that is generating the data. The structure of linear dependence among the predictor variables is made explicit. Principal components with small variances (eigenvalues) exhibit the linear relationships among the original variables that are the source of multicollinearity. Also elimination of multicollinearity by dropping one or more principal components from the regression is equivalent to imposing constraints on the regression coefficients. It provides a constructive way of identifying those constraints that are consistent with the proposed model and the information contained in the data.

10.5 PRINCIPAL COMPONENTS REGRESSION: A CAUTION

We have seen in Chapter 9 that principal components analysis is an effective tool for the detection of multicollinearity. In this chapter we have used the principal components as an alternative to the least squares method to obtain estimates of the regression coefficients in the presence of multicollinearity. The method has worked to our advantage in the Import data, where the first two of the three principal components have succeeded in capturing most of the variability in the response variable (see Table 10.3). This analysis is not guaranteed to work for all data sets. In fact, the principal components regression can fail in accounting for the variability in the response variable. To illustrate this point Hadi and Ling (1998) use a data set known as the Hald's data and a constructed response variable U. The original data set can be found in Draper and Smith (1998), p. 348. It is given here in Table and can also be found in the book's Web site.[2] The data set has four predictor variables. The response variable U and the four PCs, $C_1, \ldots, C_4$, corresponding to the four predictor variables are given in Table 10.5. The variable U is already in a standardized form. The sample variances of the four PCs are $\lambda_1 = 2.2357$, $\lambda_2 = 1.5761$, $\lambda_3 = 0.1866$, and $\lambda_4 = 0.0016$. The condition number, $\kappa = \sqrt{\lambda_1/\lambda_4}$ $= \sqrt{2.236/0.002} = 37$, is large, indicating the presence of multicollinearity in the original data.

The regression results obtained from fitting the model

$$U = \alpha_1 C_1 + \alpha_2 C_2 + \alpha_3 C_3 + \alpha_4 C_4 + \varepsilon \qquad (10.14)$$

to the data are shown in Table 10.6. The coefficient of the last PC, C_4, is highly significant and all other three coefficients are not significant. Now if we drop C_4, the PC with the smallest variance, we obtain the results in Table 10.7. As it is clear from a comparison of Tables 10.6 and 10.7, all four PCs capture almost all the variability in U, while the first three account for none of the variability in U. Therefore, one should be careful before dropping any of the PCs.

Another problem with the principal component regression is that the results can be unduly influenced by the presence of high leverage point and outliers (see Chapter 4 for detailed discussion of outliers and influence). This is because the PCs are computed from the correlation matrix, which itself can be seriously affected by outliers in the data. A scatter plot of the response variable versus each of the PCs and the pairwise scatter plots of the PCs versus each other would point out outliers if they are present in the data. The scatter plot of U versus each of the PCs (Figure

[2] http://www.ilr.cornell.edu/˜hadi/RABE4

Table 10.4 Hald's Data

Y	X_1	X_2	X_3	X_4
78.5	7	26	6	60
74.3	1	29	15	52
104.3	11	56	8	20
87.6	11	31	8	47
95.9	7	52	6	33
109.2	11	55	9	22
102.7	3	71	17	6
72.5	1	31	22	44
93.1	2	54	18	22
115.9	21	47	4	26
83.8	1	40	23	34
113.3	11	66	9	12
109.4	10	68	8	12

Source: Draper and Smith (1998), p. 348

Table 10.5 A Response Variable U and a Set of Principal Components of Four Predictor Variables

U	C_1	C_2	C_3	C_4
0.955	1.467	1.903	−0.530	0.039
−0.746	2.136	0.238	−0.290	−0.030
−2.323	−1.130	0.184	−0.010	−0.094
−0.820	0.660	1.577	0.179	−0.033
0.471	−0.359	0.484	−0.740	0.019
−0.299	−0.967	0.170	0.086	−0.012
0.210	−0.931	−2.135	−0.173	0.008
0.558	2.232	−0.692	0.460	0.023
−1.119	0.352	−1.432	−0.032	−0.045
0.496	−1.663	1.828	0.851	0.020
0.781	1.641	−1.295	0.494	0.031
0.918	−1.693	−0.392	−0.020	0.037
0.918	−1.746	−0.438	−0.275	0.037

10.1) show that there are no outliers in the data and U is related only to C_4, which is consistent with the results in Tables 10.6 and 10.7. The pairwise scatter plots of the PCs versus each other (not shown) also show no outliers in the data. For other possible pitfalls of principal components regression see Hadi and Ling (1998).

Table 10.6 Regression Results Using All Four PCs of Hald's Data

Variable	Coefficient	s.e.	t-test	p-value
C_1	−0.002	0.001	−1.45	0.1842
C_2	−0.002	0.002	−1.77	0.1154
C_3	0.002	0.005	0.49	0.6409
C_4	24.761	0.049	502.00	< 0.0001
$n = 13$	$R^2 = 1.00$	$R_a^2 = 1.00$	$\hat{\sigma} = 0.0069$	$d.f. = 8$

Table 10.7 Regression Results Using the First Three PCs of Hald's Data

Variable	Coefficient	s.e.	t-test	p-value
C_1	−0.001	0.223	−0.01	0.9957
C_2	−0.000	0.266	−0.00	0.9996
C_3	0.002	0.772	0.00	0.9975
$n = 13$	$R^2 = 0.00$	$R_a^2 = -0.33$	$\hat{\sigma} = 1.155$	$d.f. = 9$

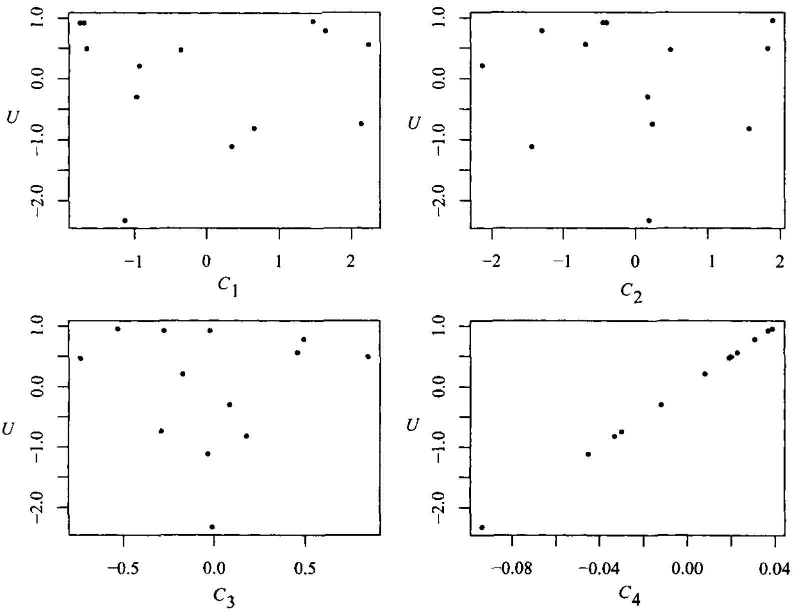

Figure 10.1 Scatter plots of U versus each of the PCs of the Hald's data.

10.6 RIDGE REGRESSION

Ridge regression[3] provides another alternative estimation method that may be used to advantage when the predictor variables are highly collinear. There are a number of alternative ways to define and compute ridge estimates (see the Appendix to this chapter). We have chosen to present the method associated with the *ridge trace*. It is a graphical approach and may be viewed as an exploratory technique. Ridge analysis using the ridge trace represents a unified approach to problems of detection and estimation when multicollinearity is suspected. The estimators produced are biased but tend to have a smaller mean squared error than OLS estimators (Hoerl and Kennard, 1970).

Ridge estimates of the regression coefficients may be obtained by solving a slightly altered form of the normal equations (introduced in Chapter 3). Assume that the standardized form of the regression model is given as:

$$\tilde{Y} = \theta_1 \tilde{X}_1 + \theta_2 \tilde{X}_2 + \cdots + \theta_p \tilde{X}_p + \varepsilon'. \tag{10.15}$$

The estimating equations for the ridge regression coefficients are

$$\begin{array}{ccccccccc}
(1+k)\theta_1 & + & r_{12}\ \theta_2 & + & \cdots & + & r_{1p}\ \theta_p & = & r_{1y}, \\
r_{21}\ \theta_1 & + & (1+k)\theta_2 & + & \cdots & + & r_{2p}\ \theta_p & = & r_{2y}, \\
\vdots & & \vdots & & & \vdots & & & \vdots \\
r_{p1}\ \theta_1 & + & r_{p2}\ \theta_2 & + & \cdots & + & (1+k)\theta_p & = & r_{py},
\end{array} \tag{10.16}$$

where r_{ij} is the correlation between the ith and jth predictor variables and r_{iy} is the correlation between the ith predictor variable and the response variable $\tilde{Y}$. The solution to (10.16), $\hat{\theta}_1, \ldots, \hat{\theta}_p$, is the set of estimated ridge regression coefficients. The ridge estimates may be viewed as resulting from a set of data that has been slightly altered. See the Appendix to this chapter for a formal treatment.

The essential parameter that distinguishes ridge regression from OLS is k. Note that when $k = 0$, the $\hat{\theta}$'s are the OLS estimates. The parameter k may be referred to as the bias parameter. As k increases from zero, bias of the estimates increases. On the other hand, the *total variance* (the sum of the variances of the estimated regression coefficients), is

$$\text{Total Variance}(k) = \sum_{j=1}^{p} Var(\hat{\theta}_j(k)) = \sigma^2 \sum_{j=1}^{p} \frac{\lambda_j}{(\lambda_j + k)^2}, \tag{10.17}$$

which is a decreasing function of k. The formula in (10.17) shows the effect of the ridge parameter on the total variance of the ridge estimates of the regression coefficients. Substituting $k = 0$ in (10.17), we obtain

$$\text{Total Variance}(0) = \sigma^2 \sum_{j=1}^{p} \frac{1}{\lambda_j}, \tag{10.18}$$

[3] Hoerl (1959) named the method *ridge regression* because of its similarity to ridge analysis used in his earlier work to study second-order response surfaces in many variables.

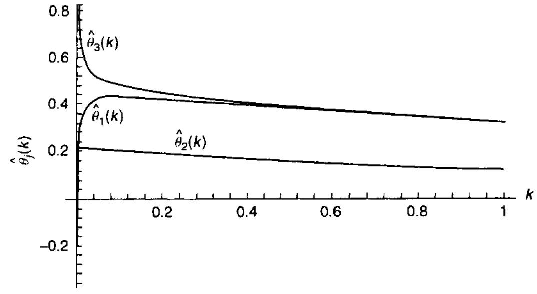

Figure 10.2 Ridge trace: IMPORT data (1949–1959).

which shows the effect of small eigenvalue on the total variance of the OLS estimates of the regression coefficients.

As k continues to increase without bound, the regression estimates all tend toward zero.[4] The idea of ridge regression is to pick a value of k for which the reduction in total variance is not exceeded by the increase in bias.

It has been shown that there is a positive value of k for which the ridge estimates will be stable with respect to small changes in the estimation data (Hoerl and Kennard, 1970). In practice, a value of k is chosen by computing $\hat{\theta}_1, \ldots, \hat{\theta}_p$ for a range of k values between 0 and 1 and plotting the results against k. The resulting graph is known as the *ridge trace* and is used to select an appropriate value for k. Guidelines for choosing k are given in the following example.

10.7 ESTIMATION BY THE RIDGE METHOD

A method for detecting multicollinearity that comes out of ridge analysis deals with the instability in the estimated coefficients resulting from slight changes in the estimation data. The instability may be observed in the *ridge trace*. The ridge trace is a simultaneous graph of the regression coefficients, $\hat{\theta}_1, \ldots, \hat{\theta}_p$, plotted against k for various values of k such as 0.001, 0.002, and so on. Figure 10.2 is the ridge trace for the IMPORT data. The graph is constructed from Table 10.8, which has the ridge estimated coefficients for 29 values of k ranging from 0 to 1. Typically, the values of k are chosen to be concentrated near the low end of the range. If the estimated coefficients show large fluctuations for small values of k, instability has been demonstrated and multicollinearity is probably at work.

What is evident from the trace or equivalently from Table 10.8 is that the estimated values of the coefficients θ_1 and θ_3 are quite unstable for small values of k. The estimate of θ_1 changes rapidly from an implausible negative value of -0.339 to a stable value of about 0.43. The estimate of θ_3 goes from 1.303 to stabilize at about

[4] Because the ridge method tends to shrink the estimates of the regression coefficients toward zero, ridge estimators are sometimes generically referred to as *shrinkage estimators*.

Table 10.8 Ridge Estimates $\hat{\theta}_j(k)$, as Functions of the Ridge Parameter k, for the IMPORT Data (1949–1959)

k	$\hat{\theta}_1(k)$	$\hat{\theta}_2(k)$	$\hat{\theta}_3(k)$
0.000	−0.339	0.213	1.303
0.001	−0.117	0.215	1.080
0.003	0.092	0.217	0.870
0.005	0.192	0.217	0.768
0.007	0.251	0.217	0.709
0.009	0.290	0.217	0.669
0.010	0.304	0.217	0.654
0.012	0.328	0.217	0.630
0.014	0.345	0.217	0.611
0.016	0.359	0.217	0.597
0.018	0.370	0.216	0.585
0.020	0.379	0.216	0.575
0.022	0.386	0.216	0.567
0.024	0.392	0.215	0.560
0.026	0.398	0.215	0.553
0.028	0.402	0.215	0.548
0.030	0.406	0.214	0.543
0.040	0.420	0.213	0.525
0.050	0.427	0.211	0.513
0.060	0.432	0.209	0.504
0.070	0.434	0.207	0.497
0.080	0.436	0.206	0.491
0.090	0.436	0.204	0.486
0.100	0.436	0.202	0.481
0.200	0.426	0.186	0.450
0.300	0.411	0.173	0.427
0.400	0.396	0.161	0.408
0.500	0.381	0.151	0.391
0.600	0.367	0.142	0.376
0.700	0.354	0.135	0.361
0.800	0.342	0.128	0.348
0.900	0.330	0.121	0.336
1.000	0.319	0.115	0.325

0.50. The coefficient of $\tilde{X}_2$ (STOCK), θ_2 is unaffected by the multicollinearity and remains stable throughout at about 0.21.

The next step in the ridge analysis is to select a value of k and to obtain the corresponding estimates of the regression coefficients. If multicollinearity is a serious problem, the ridge estimators will vary dramatically as k is slowly increased from zero. As k increases, the coefficients will eventually stabilize. Since k is a bias parameter, it is desirable to select the smallest value of k for which stability occurs since the size of k is directly related to the amount of bias introduced. Several methods have been suggested for the choice of k. These methods include:

1. *Fixed Point.* Hoerl, Kennard, and Baldwin (1975) suggest estimating k by

$$k = \frac{p\hat{\sigma}^2(0)}{\sum_{j=1}^{p}[\hat{\theta}_j(0)]^2}, \tag{10.19}$$

where $\hat{\theta}_1(0), \ldots, \hat{\theta}_p(0)$ are the least squares estimates of $\theta_1, \ldots, \theta_p$ when the model in (10.15) is fitted to the data (i.e., when $k = 0$), and $\hat{\sigma}^2(0)$ is the corresponding residual mean square.

2. *Iterative Method.* Hoerl and Kennard (1976) propose the following iterative procedure for selecting k: Start with the initial estimate of k in (10.19). Denote this value by k_0. Then, calculate

$$k_1 = \frac{p\hat{\sigma}^2(0)}{\sum_{j=1}^{p}[\hat{\theta}_j(k_0)]^2}. \tag{10.20}$$

Then use k_1 to calculate k_2 as

$$k_2 = \frac{p\hat{\sigma}^2(0)}{\sum_{j=1}^{p}[\hat{\theta}_j(k_1)]^2}. \tag{10.21}$$

Repeat this process until the difference between two successive estimates of k is negligible.

3. *Ridge Trace.* The behavior of $\hat{\theta}_j(k)$ as a function of k is easily observed from the ridge trace. The value of k selected is the smallest value for which all the coefficients $\hat{\theta}_j(k)$ are stable. In addition, at the selected value of k, the residual sum of squares should remain close to its minimum value. The variance inflation factors,[5] $\text{VIF}_j(k)$, should also get down to less than 10. (Recall that a value of 1 is a characteristic of an orthogonal system and a value less than 10 would indicate a non-collinear or stable system.)

[5] The formula for $\text{VIF}_j(k)$ is given in the Appendix to this chapter.

4. *Other Methods.* Many other methods for estimating k have been suggested in the literature. See, for example, Marquardt (1970), Mallows (1973), Goldstein and Smith (1974), McDonald and Galarneau (1975), Dempster et al. (1977), and Wahba, Golub, and Health (1979). The appeal of the ridge trace, however, lies in its graphical representation of the effects that multicollinearity has on the estimated coefficients.

For the IMPORT data, the fixed point formula in (10.19) gives

$$k = \frac{3 \times 0.0101}{(-0.339)^2 + (0.213)^2 + (1.303)^2} = 0.0164. \qquad (10.22)$$

The iterative method gives the following sequence: $k_0 = 0.0164$, $k_1 = 0.0161$, and $k_2 = 0.0161$. So, it converges after two iterations to $k = 0.0161$. The ridge trace in Figure 10.2 (see also Table 10.8) appears to stabilize for k around 0.04. We therefore have three estimates of k (0.0164, 0.0161, and 0.04).

From Table 10.8, we see that at any of these values the improper negative sign on the estimate of θ_1 has disappeared and the coefficient has stabilized (at 0.359 for k = 0.016 and at 0.42 for k = 0.04). From Table 10.9, we see that the sum of squared residuals (SSE(k)) has only increased from 0.081 at $k = 0$ to 0.108 at $k = 0.016$, and to 0.117 at $k = 0.04$. Also, the variance inflation factors, $\text{VIF}_1(k)$ and $\text{VIF}_3(k)$, decreased from about 185 to values between 1 and 4. It is clear that values of k in the interval (0.016 to 0.04) appear to be satisfactory.

The estimated coefficients from the model stated in standardized and original variables units are summarized in Table 10.10. The original coefficient $\hat{\beta}_j$ is obtained from the standardized coefficient $\hat{\theta}_j$ using (10.3). For example, $\hat{\beta}_1$ is calculated by

$$\hat{\beta}_{1j} = (s_y/s_1)\hat{\theta}_1 = (4.5437/29.9995)(0.4196) = 0.0635.$$

Thus, the resulting model in terms for the original variables fitted by ridge method using $k = 0.04$ is

$$\begin{aligned} \text{IMPORT} \;=\; & -8.5537 + 0.0635 \cdot \text{DOPROD} \\ & +\; 0.5859 \cdot \text{STOCK} + 0.1156 \cdot \text{CONSUM}. \end{aligned}$$

The equation gives a plausible representation of the relationship. Note that the final equation for these data is not particularly different from the result obtained by using the first two principal components (see Table 10.3), although the two computational methods appear to be very different.

10.8 RIDGE REGRESSION: SOME REMARKS

Ridge regression provides a tool for judging the stability of a given body of data for analysis by least squares. In highly collinear situations, as has been pointed out, small changes (perturbations) in the data cause very large changes in the

Table 10.9 Residual Sum of Squares, SSE(k), and Variance Inflation Factors, VIF$_j(k)$, as Functions of the Ridge Parameter k, for the IMPORT Data (1949–1959)

k	SSE(k)	VIF$_1(k)$	VIF$_2(k)$	VIF$_3(k)$
0.000	0.0810	186.11	1.02	186.00
0.001	0.0837	99.04	1.01	98.98
0.003	0.0911	41.80	1.00	41.78
0.005	0.0964	23.00	0.99	22.99
0.007	0.1001	14.58	0.99	14.57
0.009	0.1027	10.09	0.98	10.09
0.010	0.1038	8.60	0.98	8.60
0.012	0.1056	6.48	0.98	6.48
0.014	0.1070	5.08	0.97	5.08
0.016	0.1082	4.10	0.97	4.10
0.018	0.1093	3.39	0.97	3.39
0.020	0.1102	2.86	0.96	2.86
0.022	0.1111	2.45	0.96	2.45
0.024	0.1118	2.13	0.95	2.13
0.026	0.1126	1.88	0.95	1.88
0.028	0.1132	1.67	0.95	1.67
0.030	0.1139	1.50	0.94	1.50
0.040	0.1170	0.98	0.93	0.98
0.050	0.1201	0.72	0.91	0.72
0.060	0.1234	0.58	0.89	0.58
0.070	0.1271	0.49	0.87	0.49
0.080	0.1310	0.43	0.86	0.43
0.090	0.1353	0.39	0.84	0.39
0.100	0.1400	0.35	0.83	0.35
0.200	0.2052	0.24	0.69	0.24
0.300	0.2981	0.20	0.59	0.20
0.400	0.4112	0.18	0.51	0.18
0.500	0.5385	0.17	0.44	0.17
0.600	0.6756	0.15	0.39	0.15
0.700	0.8191	0.14	0.35	0.14
0.800	0.9667	0.13	0.31	0.13
0.900	1.1163	0.12	0.28	0.12
1.000	1.2666	0.11	0.25	0.11

Table 10.10 OLS and Ridge Estimates of the Regression Coefficients for IMPORT Data (1949–1959)

	OLS ($k = 0$)		Ridge ($k = 0.04$)	
Variable	Standardized Coefficients	Original Coefficients	Standardized Coefficients	Original Coefficients
Constant	0	−10.1300	0	−8.5537
DOPROD	−0.3393	−0.0514	0.4196	0.0635
STOCK	0.2130	0.5869	0.2127	0.5859
CONSUM	1.3027	0.2868	0.5249	0.1156
	$R^2 = 0.992$		$R^2 = 0.988$	

estimated regression coefficients. Ridge regression will reveal this condition. Least squares regression should be used with caution in these situations. Ridge regression provides estimates that are more robust than least squares estimates for small perturbations in the data. The method will indicate the sensitivity (or the stability) of the least squares coefficients to small changes in the data.

The ridge estimators are stable in the sense that they are not affected by slight variations in the estimation data. Because of the smaller mean square error property, values of the ridge estimated coefficients are expected to be closer than the OLS estimates to the true values of the regression coefficients. Also, forecasts of the response variable corresponding to values of the predictor variables not included in the estimation set tend to be more accurate.

The estimation of the bias parameter k is rather subjective. There are many methods for estimating k but there is no consensus as to which method is preferable. Regardless of the method of choice for estimating the ridge parameter k, the estimated parameter can be affected by the presence of outliers in the data. Therefore a careful checking for outliers should accompany any method for estimating k to ensure that the obtained estimate is not unduly influenced by outliers in the data.

As with the principal components method, the criteria for deciding when the ridge estimators are superior to the OLS estimators depend on the values of the true regression coefficients in the model. Although these values cannot be known, we still suggest that ridge analysis is useful in cases where extreme multicollinearity is suspected. The ridge coefficients can suggest an alternative interpretation of the data that may lead to a better understanding of the process under study.

Another practical problem with ridge regression is that it has not been implemented in some statistical packages. If a statistical package does not have a routine for ridge regression, ridge regression estimates can be obtained from the standard least squares package by using a slightly altered data set. Specifically, the ridge estimates of the regression coefficients can be obtained from the regression of Y^* on $X_1^*, \ldots, X_p^*$. The new response variable Y^* is obtained by augmenting $\tilde{Y}$ by p new fictitious observations, each of which is equal to zero. Similarly, the new predictor variable X_j^* is obtained by augmenting $\tilde{X}_j$ by p new fictitious observations, each of which is equal to zero except the one in the jth position which is equal to $\sqrt{k}$, where k is the chosen value of the ridge parameter. It can be shown that the ridge estimates $\hat{\theta}_1(k)$, ..., $\hat{\theta}_p(k)$ are obtained by the least squares regression of Y^* on $X_1^*, \ldots, X_p^*$ without having a constant term in the model.

10.9 SUMMARY

Both alternative estimation methods, ridge regression and principal components regression, provide additional information about the data being analyzed. We have seen that the eigenvalues of the correlation matrix of predictor variables play an important role in detecting multicollinearity and in analyzing its effects. The regression estimates produced by these methods are biased but may be more accurate than OLS estimates in terms of mean square error. It is impossible to evaluate the gain in accuracy for a specific problem since a comparison of the two methods to OLS requires knowledge of the true values of the coefficients. Nevertheless, when severe multicollinearity is suspected, we recommend that at least one set of estimates in addition to the OLS estimates be calculated. The estimates may suggest an interpretation of the data that were not previously considered.

There is no strong theoretical justification for using principal components or ridge regression methods. We recommend that the methods be used in the presence of severe multicollinearity as a visual diagnostic tool for judging the suitability of the data for least squares analysis. When principal components or ridge regression analysis reveal the instability of a particular data set, the analyst should first consider using least squares regression on a reduced set of variables (as indicated in Chapter 9). If least squares regression is still unsatisfactory (high VIFs, coefficients with wrong signs, large condition number), only then should principal components or ridge regression be used.

EXERCISES

10.1 Longley's (1967) data set is a classic example of multicollinear data. The data (Table 10.11) consist of a response variable S and six predictor variables $X_1, \ldots, X_6$. The data can be found in the book's Web site. The initial model

$$S = \beta_0 + \beta_1 X_1 + \ldots + \beta_6 X_6 + \varepsilon, \qquad (10.23)$$

in terms of the original variables, can be written in terms of the standardized variables as

$$\tilde{S} = \theta_1 \tilde{X}_1 + \ldots + \theta_6 \tilde{X}_6 + \varepsilon'. \qquad (10.24)$$

(a) Fit the model (10.24) to the data using least squares. What conclusion can you draw from the data?

(b) From the results you obtained from the model in (10.24), obtain the least squares estimated regression coefficients in model (10.23).

Table 10.11 Longley (1967) Data

Y	X_1	X_2	X_3	X_4	X_5	X_6
60323	830	234289	2356	1590	107608	1947
61122	885	259426	2325	1456	108632	1948
60171	882	258054	3682	1616	109773	1949
61187	895	284599	3351	1650	110929	1950
63221	962	328975	2099	3099	112075	1951
63639	981	346999	1932	3594	113270	1952
64989	990	365385	1870	3547	115094	1953
63761	1000	363112	3578	3350	116219	1954
66019	1012	397469	2904	3048	117388	1955
67857	1046	419180	2822	2857	118734	1956
68169	1084	442769	2936	2798	120445	1957
66513	1108	444546	4681	2637	121950	1958
68655	1126	482704	3813	2552	123366	1959
69564	1142	502601	3931	2514	125368	1960
69331	1157	518173	4806	2572	127852	1961
70551	1169	554894	4007	2827	130081	1962

(c) Now fit the model in (10.23) to the data using least squares and verify that the obtained results are consistent with those obtained above.

(d) Compute the correlation matrix of the six predictor variables and the corresponding scatter plot matrix. Do you see any evidence of collinearity?

(e) Compute the corresponding PCs, their sample variances, and the condition number. How many different sets of multicollinearity exist in the data? What are the variables involved in each set?

(f) Based on the number of PCs you choose to retain, obtain the PC estimates of the coefficients in (10.23) and (10.24).

(g) Using the ridge method, construct the ridge trace. What value of k do you recommend to be used in the estimation of the parameters in (10.23) and (10.24)? Use the chosen value of k and compute the ridge estimates of the regression coefficients in (10.23) and (10.24).

(h) Compare the estimates you obtained by the three methods. Which one would you recommend? Explain.

10.2 Repeat Exercise 10.1 using the Hald's data discussed in Section 10.5 but using the original response variable Y and the four predictors $X_1, \ldots, X_4$. The data appear in Table 10.4.

10.3 From your analysis of the Longley and Hald data sets, do you observe the sort of problems pointed out in Section 10.5?

Appendix: Ridge Regression

In this appendix we present ridge regression method in matrix notation.

A. The Model

The regression model can be expressed as

$$\mathbf{Y} = \mathbf{Z}\boldsymbol{\theta} + \boldsymbol{\varepsilon}, \tag{A.1}$$

where $\mathbf{Y}$ is an $n \times 1$ vector of observations on the response variable, $\mathbf{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_p)$ is an $n \times p$ matrix of n observations on p predictor variables, $\boldsymbol{\theta}$ is a $p \times 1$ vector of regression coefficients, and $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of random errors. It is assumed that $E(\boldsymbol{\varepsilon}) = \mathbf{0}$, $E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T) = \sigma^2\mathbf{I}$, where $\mathbf{I}$ is the identity matrix of order n. It is also assumed, without loss of generality, that $\mathbf{Y}$ and $\mathbf{Z}$ have been centered and scaled so that $\mathbf{Z}^T\mathbf{Z}$ and $\mathbf{Z}^T\mathbf{Y}$ are matrices of correlation coefficients.[6]

The least squares estimator for $\boldsymbol{\theta}$ is $\hat{\boldsymbol{\theta}} = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{Y}$. It can be shown that

$$E[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})] = \sigma^2 \sum_{j=1}^{p} \lambda_j^{-1}, \tag{A.2}$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ are the eigenvalues of $\mathbf{Z}^T\mathbf{Z}$. The left-hand side of (A.2) is called the *total mean square error*. It serves as a composite measure of the squared distance of the estimated regression coefficients from their true values.

B. Effect of Multicollinearity

It was argued in Chapter 9 and in the Appendix to Chapter 9 that multi-collinearity is synonymous with small eigenvalues. It follows from Equation (A.2) that when one or more of the λ's are small, the total mean square error of $\hat{\boldsymbol{\theta}}$ is large, suggesting imprecision in the least squares estimation method. The ridge regression approach is an attempt to construct an alternative estimator that has a smaller total mean square error value.

C. Ridge Regression Estimators

Hoerl and Kennard (1970) suggest a class of estimators indexed by a parameter $k > 0$. The estimator is (for a given value of k)

$$\hat{\boldsymbol{\theta}}(k) = (\mathbf{Z}^T\mathbf{Z} + k\mathbf{I})^{-1}\mathbf{Z}^T\mathbf{Y} = (\mathbf{Z}^T\mathbf{Z} + k\mathbf{I})^{-1}\mathbf{Z}^T\mathbf{Z}\hat{\boldsymbol{\theta}}. \tag{A.3}$$

The expected value of $\hat{\boldsymbol{\theta}}(k)$ is

$$E[\hat{\boldsymbol{\theta}}(k)] = (\mathbf{Z}^T\mathbf{Z} + k\mathbf{I})^{-1}\mathbf{Z}^T\mathbf{Z}\boldsymbol{\theta} \tag{A.4}$$

[6] Note that Z_j is obtained by transforming the original predictor variable X_j by $z_{ij} = (x_{ij} - \bar{x}_j)/\sqrt{\sum(x_{ij} - \bar{x}_j)^2}$. Thus, Z_j is centered and scaled to have unit length, that is, $\sum z_{ij}^2 = 1$.

and the variance-covariance matrix is

$$Var[\hat{\boldsymbol{\theta}}(k)] = (\mathbf{Z}^T\mathbf{Z} + k\mathbf{I})^{-1}\mathbf{Z}^T\mathbf{Z}(\mathbf{Z}^T\mathbf{Z} + k\mathbf{I})^{-1}\sigma^2. \tag{A.5}$$

The variance inflation factor, $\text{VIF}_j(k)$, as a function of k is the jth diagonal element of the matrix $(\mathbf{Z}^T\mathbf{Z} + k\mathbf{I})^{-1}\mathbf{Z}^T\mathbf{Z}(\mathbf{Z}^T\mathbf{Z} + k\mathbf{I})^{-1}$.

The residual sum of squares can be written as

$$\begin{aligned} \text{SSE}(k) &= (\mathbf{Y} - \mathbf{Z}\hat{\boldsymbol{\theta}}(k))^T(\mathbf{Y} - \mathbf{Z}\hat{\boldsymbol{\theta}}(k)) \\ &= (\mathbf{Y} - \mathbf{Z}\hat{\boldsymbol{\theta}})^T(\mathbf{Y} - \mathbf{Z}\hat{\boldsymbol{\theta}}) + (\hat{\boldsymbol{\theta}}(k) - \hat{\boldsymbol{\theta}})^T\mathbf{Z}^T\mathbf{Z}(\hat{\boldsymbol{\theta}}(k) - \hat{\boldsymbol{\theta}}). \end{aligned} \tag{A.6}$$

The total mean square error is

$$\begin{aligned} \text{TMSE}(k) &= E[(\hat{\boldsymbol{\theta}}(k) - \boldsymbol{\theta})^T(\hat{\boldsymbol{\theta}}(k) - \boldsymbol{\theta})] \\ &= \sigma^2 \, \text{trace}[(\mathbf{Z}^T\mathbf{Z} + k\mathbf{I})^{-1}\mathbf{Z}^T\mathbf{Z}(\mathbf{Z}^T\mathbf{Z} + k\mathbf{I})^{-1}] \\ &\quad + k^2\boldsymbol{\theta}^T(\mathbf{Z}^T\mathbf{Z} + k\mathbf{I})^{-2}\boldsymbol{\theta} \\ &= \sigma^2\sum_{j=1}^{p}\lambda_j(\lambda_j + k)^{-2} + k^2\boldsymbol{\theta}^T(\mathbf{Z}^T\mathbf{Z} + k\mathbf{I})^{-2}\boldsymbol{\theta}. \end{aligned} \tag{A.7}$$

Note that the first term on the right-hand side of Equation (A.7) is the sum of the variances of the components of $\hat{\boldsymbol{\theta}}(k)$ (total variance) and the second term is the square of the bias. Hoerl and Kennard (1970) prove that there exists a value of $k > 0$ such that

$$E[(\hat{\boldsymbol{\theta}}(k) - \boldsymbol{\theta})^T(\hat{\boldsymbol{\theta}}(k) - \boldsymbol{\theta})] < E[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})],$$

that is, the mean square error of the ridge estimator, $\hat{\boldsymbol{\theta}}(k)$, is less than the mean square error of the OLS estimator, $\hat{\boldsymbol{\theta}}$. Hoerl and Kennard (1970) suggest that an appropriate value of k may be selected by observing the ridge trace and some complementary summary statistics for $\hat{\boldsymbol{\theta}}(k)$ such as $\text{SSE}(k)$ and $\text{VIF}_j(k)$. The value of k selected is the smallest value for which $\hat{\boldsymbol{\theta}}(k)$ is stable. In addition, at the selected value of k, the residual sum of squares should remain close to its minimum value, and the variance inflation factors are less than 10, as discussed in Chapter 9.

Ridge estimators have been generalized in several ways. They are sometimes generically referred to as *shrinkage estimators*, because these procedures tend to shrink the estimates of the regression coefficients toward zero. To see one possible generalization, consider the regression model restated in terms of the principal components, $\mathbf{C} = (\mathbf{C}_1, \ldots, \mathbf{C}_p)$, discussed in the Appendix to Chapter 9. The general model takes the form

$$\mathbf{Y} = \mathbf{C}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}, \tag{A.8}$$

where

$$\mathbf{C} = \mathbf{Z}\mathbf{V}, \qquad \boldsymbol{\alpha} = \mathbf{V}^T\boldsymbol{\theta}, \tag{A.9}$$

$$\mathbf{V}^T\mathbf{Z}^T\mathbf{Z}\mathbf{V} = \boldsymbol{\Lambda}, \qquad \mathbf{V}^T\mathbf{V} = \mathbf{V}\mathbf{V}^T = \mathbf{I},$$

and

$$\mathbf{\Lambda} = \begin{pmatrix} \lambda_1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & \lambda_2 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_{p-1} & 0 \\ 0 & 0 & 0 & \cdots & 0 & \lambda_p \end{pmatrix}, \quad \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p,$$

is a diagonal matrix consisting of the ordered eigenvalues of $\mathbf{Z}^T\mathbf{Z}$. The total mean square error in (A.7) becomes

$$\begin{aligned} \text{TMSE}(k) &= E[(\hat{\boldsymbol{\theta}}(k) - \boldsymbol{\theta})^T(\hat{\boldsymbol{\theta}}(k) - \boldsymbol{\theta})] \\ &= \sigma^2 \sum_{j=1}^{p} \frac{\lambda_j}{(\lambda_j + k)^2} + \sum_{j=1}^{p} \frac{k^2\alpha_j^2}{(\lambda_j + k)^2}, \end{aligned} \tag{A.10}$$

where $\alpha^T = (\alpha_1, \alpha_2, \ldots, \alpha_p)$. Instead of taking a single value for k, we can consider several different values k, say $k_1, k_2, \ldots, k_p$. We consider separate ridge parameters (i.e., shrinkage factors) for each of the regression coefficients. The quantity k, instead of being a scalar, is now a vector and denoted by $\mathbf{k}$. The total mean square error given in (A.10) now becomes

$$\begin{aligned} \text{TMSE}(\mathbf{k}) &= E[(\hat{\boldsymbol{\theta}}(\mathbf{k}) - \boldsymbol{\theta})^T(\hat{\boldsymbol{\theta}}(\mathbf{k}) - \boldsymbol{\theta})] \\ &= \sigma^2 \sum_{j=1}^{p} \frac{\lambda_j}{(\lambda_j + k_j)^2} + \sum_{j=1}^{p} \frac{k_j^2\alpha_j^2}{(\lambda_j + k_j)^2}. \end{aligned} \tag{A.11}$$

The total mean square error given in (A.11) is minimized by taking $k_j = \sigma^2/\alpha_j^2$. An iterative estimation procedure is suggested. At Step 1, k_j is computed by using ordinary least squares estimates for σ^2 and α_j. Then a new value of $\hat{\alpha}(\mathbf{k})$ is computed,

$$\hat{\alpha}(\mathbf{k}) = (\mathbf{C}^T\mathbf{C} + \mathbf{K})^{-1}\mathbf{C}^T\mathbf{Y},$$

where $\mathbf{K}$ is a diagonal matrix with diagonal elements $k_1, \ldots, k_p$ from Step 1. The process is repeated until successive changes in the components of $\hat{\alpha}(k)$ are negligible. Then, using Equation (A.9), the estimate of $\boldsymbol{\theta}$ is

$$\hat{\boldsymbol{\theta}}(\mathbf{k}) = \mathbf{V}\hat{\alpha}(\mathbf{k}). \tag{A.12}$$

The two ridge-type estimators (one value of k, several values of k) defined previously, as well as other related alternatives to ordinary least-squares estimation, are discussed by Dempster et al. (1977). The different estimators are compared and evaluated by Monte Carlo techniques. In general, the choice of the best estimation method for a particular problem depends on the specific model and data. Dempster et al. (1977) hint at an analysis that could be used to identify the best estimation method for a given set of data. At the present time, our preference is for the simplest version of the ridge method, a single ridge parameter k, chosen after an examination of the ridge trace.