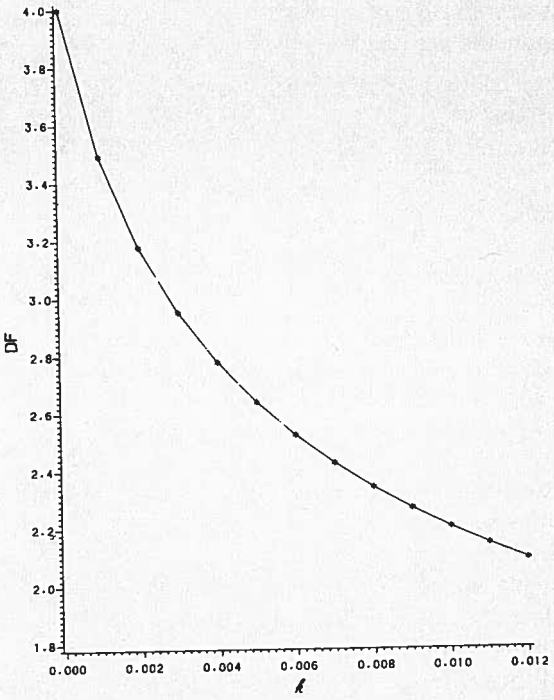
FIGURE 8.6 **df-trace on the tobacco data of Table 8.13**



PRINCIPAL COMPONENTS REGRESSION

Principal components regression represents another biased estimation technique for combating multicollinearity. With this method, we perform least squares estimation on a set of artificial variables called the *principal components* of the correlation matrix. Based on the nature of the analysis, we eliminate a certain number of the principal components to effect a substantial reduction in variance. The method varies somewhat in philosophy from ridge regression but, like ridge, gives biased estimates; when used successfully, this method results in estimation and prediction that is superior to OLS. Principal components are orthogonal to each other, so that it becomes quite easy to *attribute a specific amount of variance* to each.

Consider the matrix of normalized eigenvectors associated with the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_k$ of $\mathbf{X}^{*\prime}\mathbf{X}^*$ (correlation form). We know that $\mathbf{V}\mathbf{V}' = \mathbf{I}$ since $\mathbf{V}$ is an orthogonal matrix. Hence we can write the original regression model in the form

$$\mathbf{y} = \beta_0\mathbf{1} + \mathbf{X}^*\mathbf{V}\mathbf{V}'\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{8.16}$$

$$\mathbf{y} = \beta_0\mathbf{1} + \mathbf{Z}\boldsymbol{\alpha} + \boldsymbol{\varepsilon} \tag{8.17}$$

where $\mathbf{Z} = \mathbf{X}^*\mathbf{V}$ and $\boldsymbol{\alpha} = \mathbf{V}'\boldsymbol{\beta}$. $\mathbf{Z}$ is an $n \times k$ matrix and $\boldsymbol{\alpha}$ is a $k \times 1$ vector of new coefficients $\alpha_1, \alpha_2, \ldots, \alpha_k$. We can visualize the columns of $\mathbf{Z}$ (typical element z_{ij}) as representing readings on k *new variables*, the *principal components*. It is easy to see that the components are orthogonal to each other. We have

$$\begin{aligned}\mathbf{Z}'\mathbf{Z} &= (\mathbf{X}^*\mathbf{V})'(\mathbf{X}^*\mathbf{V}) \\ &= \mathbf{V}'\mathbf{X}^{*\prime}\mathbf{X}^*\mathbf{V} \\ &= \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_k)\end{aligned} \tag{8.18}$$

So, if regression is performed on the z's via the model in Eq. (8.17), the variances of the coefficients (the diagonal elements of $(\mathbf{Z}'\mathbf{Z})^{-1}$ apart from σ^2) are the reciprocals of eigenvalues. That is,

$$\frac{\text{Var}(\hat{\alpha}_j)}{\sigma^2} = \frac{1}{\lambda_j} \qquad (j = 1, 2, \ldots, k) \tag{8.19}$$

Note that the $\hat{\alpha}$'s are, indeed, least squares estimators. If all of the principal components are retained in the regression model, then all that has been accomplished by the transformation is essentially a *rotation of the regressor variables.* Even though the new variables are orthogonal, the same magnitude of variance (due to the ill-conditioning in $\mathbf{X}'\mathbf{X}$) is retained. In a sense, the total variance has merely been redistributed. If multicollinearity is severe, there will be at least one small eigenvalue. An elimination of one (at least one) principal component, that associated with the small eigenvalue, may *substantially* reduce the total variance in the model and thus produce an appreciably improved prediction equation.

What Are Principal Components?

The rotation of the regressor variables that produces the z's, the principal components, essentially allows for a new set of coefficients, the α's; these α's are defined so that we can directly attribute the variance of an estimator $\hat{\alpha}_j$ to a specific linear dependency. This is apparent from Eq. (8.19). Graphically, Figure 8.7 will allow the potential user of principal components regression to better understand what the z's actually are.

Clearly, the data indicate a strong association between x_1 and x_2. Now, this dependency will deposit its effect on the estimates of both β_1 and β_2. But consider the z-coordinate system. In fact, consider

$$\mathbf{Z} = \mathbf{X}^*\mathbf{V}$$

with a specific column of $\mathbf{Z}$ being given by

$$\mathbf{z}_j = \mathbf{X}^*\mathbf{v}_j \tag{8.20}$$

The elements in $\mathbf{z}_j$ are the data measured on the z_j axis, where for our illustration, $j = 1, 2$. The "variation" in the resulting z_j values is given by

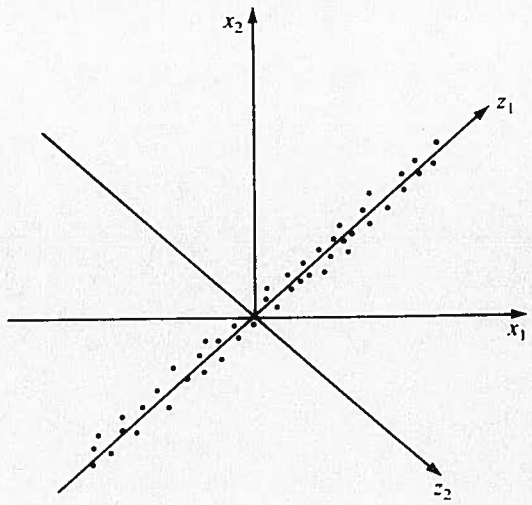
$$\mathbf{z}_j'\mathbf{z}_j = \mathbf{v}_j'(\mathbf{X}^{*\prime}\mathbf{X}^*)\mathbf{v}_j = \lambda_j \qquad (j = 1, 2)$$

Thus the regression on the principal components for this case would involve

$$\mathbf{Z}'\mathbf{Z} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \tag{8.21}$$

So, from Figure 8.7, λ_1 is the large eigenvalue, and λ_2 is small. The variance of coefficient $\hat{\alpha}_2$ is given by $\sigma^2(1/\lambda_2)$. The small variation in the $\mathbf{z}_2$ direction is responsible for the large variance in $\hat{\alpha}_2$. The variation in the data lies principally along the z_1 axis. Thus a large λ_1 allows $\text{Var}(\hat{\alpha}_1)$ to be relatively unaffected by the dependency between x_1 and x_2. From the foregoing illustration, the reader can understand that the transformation to the z's allows the linear dependencies, characterized by small eigenvalues, to be focused sharply on a small number of coefficients. This allows for ease in decision making regarding which z's are eliminated. Clearly in the case of Figure 8.7, the data have assumed the direction of z_1, and thus the principal component z_2 contributes essentially nothing to the regression. Thus z_2 would be eliminated, thereby reducing the variance contributed by $\hat{\alpha}_2$. So principal components regression is nothing more than variable screening in a regression on the principal components.

FIGURE 8.7 **Principal components for $k = 2$**



How Many Principal Components Are Eliminated?

The philosophy of principal component (pc) regression very much resembles the philosophy of least squares variable screening in general. Least squares estimation is conducted on the components, and if a component is eliminated, the resulting estimators of the coefficients of the original variables, the x's, are biased. Since

variance producing components are eliminated, variance is reduced, as in the case of least squares variable screening. The difficulty arises in the decision of how many (if any) components we should eliminate. The analyst has at his or her disposal the ordinary type of least squares criteria, s^2, PRESS, sum of absolute PRESS residuals, C_p, and the width of the confidence intervals on $E(y)$. These seem like natural criteria on which decisions should be based. Example 8.10 contains an illustration.

Transformation Back to Original Variables

Objections to principal components regression are quite often the result of the artificiality of the principal components themselves. Without a doubt, if principal components regression is used successfully, the analyst can expect the resulting model in the original variables to improve. Of course, computed statistics such as s^2, PRESS, etc. apply to the model that is transformed back to the original standardized variables. Suppose, for example, with k variables and hence k principal components, $r < k$ components are eliminated. From Eq. (8.17), with the retention of all components, we can write $\boldsymbol{\alpha} = \mathbf{V}'\boldsymbol{\beta}$, and hence

$$\boldsymbol{\beta} = \mathbf{V}\boldsymbol{\alpha} \tag{8.22}$$

Clearly then, if we eliminate the last r components, the least squares estimators of the regression coefficients for all k parameters (deleting principal components does not imply deletion of any of the original regressors) are given by

$$\mathbf{b}_{\text{pc}} = \begin{bmatrix} b_{1,\text{pc}} \\ b_{2,\text{pc}} \\ \vdots \\ b_{k,\text{pc}} \end{bmatrix} = [\mathbf{v}_1 \mathbf{v}_2 \cdots \mathbf{v}_{k-r}] \begin{bmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \\ \vdots \\ \hat{\alpha}_{k-r} \end{bmatrix} \tag{8.23}$$

Thus elimination of r principal components is tantamount to elimination of r eigenvectors and, of course, r of the α's. We are still assuming that the x's are centered and scaled so the constant term in the transformed model is $\bar{y}$.

Bias in Principal Components Coefficients

Suppose we consider the principal components procedure with r principal components eliminated and s components retained, where $s + r = k$. Also suppose we consider the matrix $\mathbf{V} = [\mathbf{v}_1 \mathbf{v}_2 \cdots \mathbf{v}_k]$ of normalized eigenvectors of $\mathbf{X}^{*\prime}\mathbf{X}^*$ partitioned into

$$\mathbf{V} = [\mathbf{V}_r \vdots \mathbf{V}_s]$$

and similarly consider the matrix $\boldsymbol{\Lambda}$ to be a diagonal matrix of eigenvalues of $\mathbf{X}^{*\prime}\mathbf{X}^*$. We partition $\boldsymbol{\Lambda}$ as

$$\boldsymbol{\Lambda} = \begin{bmatrix} \boldsymbol{\Lambda}_r & 0 \\ 0 & \boldsymbol{\Lambda}_s \end{bmatrix}$$

where Λ_r and Λ_s are diagonal matrices, with Λ_r containing the eigenvalues associated with the eliminated components. Since $\mathbf{V}'(\mathbf{X}^{*\prime}\mathbf{X}^*)\mathbf{V} = \mathbf{Z}'\mathbf{Z} = \Lambda$, the least squares estimates of the α's can be written

$$\hat{\boldsymbol{\alpha}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y} = \Lambda^{-1}\mathbf{V}'\mathbf{X}^{*\prime}\mathbf{y} \tag{8.24}$$

which implies that the estimator for the α's that are retained is given by

$$\hat{\boldsymbol{\alpha}}_s = \Lambda_s^{-1}\mathbf{V}_s'\mathbf{X}^{*\prime}\mathbf{y}$$

As indicated earlier, we must view principal components regression as standard least squares model-building on the principal components. Since the principal components are orthogonal, we can use Eq. (4.7) to show that $\hat{\boldsymbol{\alpha}}_s$ is an unbiased estimator for $\boldsymbol{\alpha}_s$. Consider now Eq. (8.23) for $\mathbf{b}_{pc}$. We can write

$$\mathbf{b}_{pc} = \mathbf{V}_s\hat{\boldsymbol{\alpha}}_s \tag{8.25}$$

Thus

$$E(\mathbf{b}_{pc}) = \mathbf{V}_s\boldsymbol{\alpha}_s = \mathbf{V}_s\mathbf{V}_s'\boldsymbol{\beta}$$

Since $\mathbf{V}\mathbf{V}' = \mathbf{I} = \mathbf{V}_r\mathbf{V}_r' + \mathbf{V}_s\mathbf{V}_s'$,

$$\begin{aligned} E(\mathbf{b}_{pc}) &= [\mathbf{I} - \mathbf{V}_r\mathbf{V}_r']\boldsymbol{\beta} \\ &= \boldsymbol{\beta} - \mathbf{V}_r\mathbf{V}_r'\boldsymbol{\beta} \\ &= \boldsymbol{\beta} - \mathbf{V}_r\boldsymbol{\alpha}_r \end{aligned}$$

Thus the estimators of the p regression coefficients are biased by the quantity $\mathbf{V}_r\boldsymbol{\alpha}_r$, with $\boldsymbol{\alpha}_r$ being the vector of principal components that have been eliminated.

Variance in Principal Components Coefficients

As one would expect, the elimination of principal components results in a decrease in the variances of the regression coefficients in $\mathbf{b}_{pc}$. The magnitude of this decrease, as in the case of ridge regression, depends on the extent of the multicollinearity involved. It is relatively easy to determine, analytically, what the variance reduction is. If all components are retained, $\mathbf{b}_{pc}$ reduces to ordinary least squares, and hence

$$\begin{aligned} \frac{\text{Var } \mathbf{b}}{\sigma^2} &= (\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1} \\ &= \mathbf{V}\Lambda^{-1}\mathbf{V}' \\ &= \mathbf{V}_r\Lambda_r^{-1}\mathbf{V}_r' + \mathbf{V}_s\Lambda_s^{-1}\mathbf{V}_s' \end{aligned} \tag{8.26}$$

From Eq. (8.25), and using the fact that $\text{Var } \hat{\boldsymbol{\alpha}}_s = \Lambda_s^{-1}$, we have the variance-covariance matrix

$$\frac{\text{Var } \mathbf{b}_{pc}}{\sigma^2} = \mathbf{V}_s\Lambda_s^{-1}\mathbf{V}_s' \tag{8.27}$$

Thus the difference in the variance-covariance matrix for the OLS estimator and the principal components estimator is the quantity $\mathbf{V}_r\Lambda_r^{-1}\mathbf{V}_r'$. The diagonal

elements of this matrix are merely weighted sums of the *reciprocals* of the eigenvalues associated with the eliminated principal components. As a result, if the ignored principal components are associated with small eigenvalues, one may expect a substantial variance reduction.

Example 8.10 ***Principal Components Example***

Consider the following data set:

y	x_1	x_2	x_3	x_4
17.6	8.8	2589	83.1	158.2
10.9	8.5	1186	24.2	96.2
9.2	7.7	291	4.5	31.8
16.2	4.9	1276	9.1	95.0
10.1	9.6	6633	158.2	407.2
11.7	10.0	12125	132.2	404.6
17.9	11.5	36717	501.5	1180.6
21.1	11.6	43319	904.0	1807.5
14.7	11.2	10530	227.6	470.0
7.7	10.7	3931	66.6	151.4
8.4	10.0	1536	43.4	93.8
32.8	6.8	61400	1253.0	3293.4

A least squares regression gives the prediction equation

$$\hat{y} = 21.971979 - 1.277560x_1 + 0.00015026x_2 + 0.015533x_3 - 0.002854x_4$$

with $SS_{\text{Res}} = 63.17323$, $s^2 = 9.025$, PRESS = 670.303, and $R^2 = 0.8857$. The following are multicollinearity diagnostics (variance decomposition proportions and eigenvalues).

Eigenvalue	Portion Intercept	Portion b_1	Portion b_2	Portion b_3	Portion b_4
4.022	0.0009	0.0009	0.0010	0.0004	0.0004
0.932897	0.0077	0.0073	0.0019	0.0010	0.0010
0.030779	0.2296	0.1932	0.1141	0.0001	0.0476
0.010341	0.1778	0.1989	0.8239	0.2644	0.0061
0.003661	0.5839	0.5998	0.0592	0.7340	0.9448

Condition Number = 1,098.6069

These eigenvalues of $\mathbf{X}'\mathbf{X}$ are for a scaled $\mathbf{X}$ matrix (not centered). The variance inflation factors are 2.080, 34.723, 79.155, and 82.967 for the coefficients $b_1, b_2, b_3,$

and b_4, respectively. The correlation matrix is given by

$$\mathbf{X}^{*\prime}\mathbf{X}^{*} = \begin{bmatrix} 1 & 0.13141 & 0.08008 & -0.01470 \\ 0.13141 & 1 & 0.98161 & 0.97364 \\ 0.08008 & 0.98161 & 1 & 0.98871 \\ -0.01470 & 0.97364 & 0.98871 & 1 \end{bmatrix}$$

The eigenvalues of the correlation matrix[1] are given by $\lambda_1 = 2.9692$, $\lambda_2 = 1.00464$, $\lambda_3 = 0.019438$, and $\lambda_4 = 0.0067049$ and the resulting matrix of eigenvectors is as follows:

$$\mathbf{V} = \begin{bmatrix} -0.05768 & -0.99267 & -0.07732 & 0.07280 \\ -0.57623 & -0.03408 & 0.81466 & -0.05591 \\ -0.57817 & 0.01934 & -0.45462 & -0.67725 \\ -0.57476 & 0.11432 & -0.35167 & 0.73000 \end{bmatrix}$$

Three regression coefficients, b_2, b_3, and b_4, appear to be affected by collinearity. The simple correlations indicate that x_2, x_3, and x_4 are involved in pairwise correlations. The variance decomposition proportions (noncentered diagnostics) indicate one or perhaps two collinearities that are causing difficulties. One collinearity (eigenvalue = 0.003661) is responsible for a substantial portion of the variance of b_3 and b_4, while a second dependency (eigenvalue = 0.010341) is responsible for a majority (82.39%) of the variance of b_2 and a modest portion of the variance of b_3. If principal components regression is successful, it will involve the elimination of one or perhaps both of the principal components associated with the two dependencies.

The $\mathbf{Z}$ matrix of principal components is found by $\mathbf{Z} = \mathbf{X}^{*}\mathbf{V}$, where $\mathbf{X}^{*}$ is centered and scaled. $\mathbf{X}^{*}$ is a 12×4 matrix without the column of ones, i.e., with the model shown in Eq. (8.16). The principal components are columns of

$$\mathbf{Z} = \begin{array}{c} \begin{matrix} z_1 & z_2 & z_3 & z_4 \end{matrix} \\ \begin{bmatrix} 0.29068 & 0.05460 & -0.02032 & -0.01064 \\ 0.34169 & 0.09626 & -0.00699 & 0.00326 \\ 0.37607 & 0.21148 & 0.00506 & -0.00908 \\ 0.37836 & 0.62388 & 0.04049 & -0.02813 \\ 0.17280 & -0.05490 & -0.03332 & 0.01247 \\ 0.13435 & -0.11679 & 0.03738 & 0.02485 \\ -0.38528 & -0.31647 & 0.10644 & 0.00762 \\ -0.72778 & -0.30655 & -0.02028 & -0.06018 \\ 0.08479 & -0.28827 & -0.03513 & 0.00534 \\ 0.27142 & -0.22518 & -0.01950 & 0.01554 \\ 0.31803 & -0.12368 & -0.02620 & 0.00882 \\ -1.25514 & 0.44561 & -0.02761 & 0.03011 \end{bmatrix} \end{array}$$

[1]Note how the eigenvalues for the case of centering and scaling differ from those that occur when only scaling is used.

The fourth column of **Z** is the principal component associated with the smallest eigenvalue. Notice the small variation in the z_{4i}.

The application of principal components regression involves the removal, initially, of z_4 with the response y being regressed against the remaining components. This regression gives the following coefficients of the **z**'s:

$$\hat{\alpha}_1 = -12.0121$$
$$\hat{\alpha}_2 = 7.57925$$
$$\hat{\alpha}_3 = 2.79077$$

Ordinary least squares procedures are applied to this regression with the computation of the residual sum of squares and PRESS statistic as discussed in Chapters 3 and 4. The results are

$$SS_{\text{Res}} = 66.4176$$
$$s^2 = 8.302$$
$$\text{PRESS} = 116.007$$

The regression with the reduced number of components has resulted in a slight increase in the residual SS but a substantial reduction in PRESS from 670.303 to 116.007 and a reduction in s^2 from 9.025 to 8.302. The improvement in the model is illustrated by the following PRESS residuals.

	Before Deletion	After Deletion
Observation	PRESS Residual	PRESS Residual
1	6.6604	6.8011
2	−0.5697	−0.6516
3	−3.6473	−3.3503
4	1.5293	2.6342
5	−2.3068	−2.5677
6	−0.2959	−0.9254
7	3.8907	2.7974
8	−17.3040	−0.1931
9	4.2692	4.0924
10	−2.2844	−2.6002
11	−1.7473	−1.9560
12	16.2328	−2.9183

Equation (8.23) can then be used to determine the estimates of the coefficients in terms of the centered and scaled regressors; the constant term for the regression is $\bar{y}$. This is followed by a transformation to the coefficients of the natural variables as

described in Section 8.3. The results are given by

$$b_{1,\text{pc}} = -1.04099$$
$$b_{2,\text{pc}} = 0.000132085$$
$$b_{3,\text{pc}} = 0.00436507$$
$$b_{4,\text{pc}} = 0.00209076$$

The constant term obtained through this transformation is $b_{0,\text{pc}} = 19.849$.

The deletion of a single principal component seems to have produced a regression that is somewhat superior to the OLS regression; the deletion of another principal component is tempting. The principal component that was eliminated was associated with a dependency, which was damaging to coefficients of regressors x_3 and x_4. The other infected coefficient is b_2. From the variance decomposition proportions, it would seem that the deletion of another principal component might be at least marginally effective. Such an analysis does slightly reduce PRESS (to a value of 100.001 with $SS_{\text{Res}} = 66.569$ and $s^2 = 7.396$). We shall not show the details here. Obviously, continuing the deletion of principal components will result in eventual deterioration of the regression.

What Is the Appropriate Order of Elimination of Principal Components?

Through our work with Example 8.10, we have implied that the strategy of elimination of principal components should be to begin by discarding the principal component associated with the smallest eigenvalue. Often this will be the appropriate plan of action, the rationale being that this is the least informative component and thus deposits the largest amount of variance. However, a more acceptable strategy is to treat the principal component reduction as if it were a standard variable screening problem (which it actually *is*). Now, of course, the principal components, as regressors, are orthogonal, and thus a reasonable set of statistics to dictate the order of reduction are the t-statistics given by

$$t = \frac{\hat{\alpha}_j}{s_{\hat{\alpha}j}} = \frac{\hat{\alpha}_j \sqrt{\lambda_j}}{s}$$

In other words, the t-values should be rank ordered and components be considered for elimination beginning with the *smallest t-value*, in magnitude. Obviously many times the rank order of t-values will be that which is the descending order of the eigenvalues, but this is not necessarily the case.

EXERCISES FOR CHAPTER 8

8.1 Thirty firms were chosen for a study of the effect of several factors on firm return on assets, y, for 1982. The data are as follows: