TABLE 7.5 **Correlation Transformation and Fitted Standardized Regression Model—Dwaine Studios Example.**

(a) Original Data

Case i	Sales Y_i	Target Population X_{i1}	Per Capita Disposable Income X_{i2}
1	174.4	68.5	16.7
2	164.4	45.2	16.8
...	...	...	...
20	224.1	82.7	19.1
21	166.5	52.3	16.0
	$\bar{Y} = 181.90$	$\bar{X}_1 = 62.019$	$\bar{X}_2 = 17.143$
	$s_Y = 36.191$	$s_1 = 18.620$	$s_2 = .97035$

(b) Transformed Data

i	Y_i^*	X_{i1}^*	X_{i2}^*
1	−.04637	.07783	−.10205
2	−.10815	−.20198	−.07901
...	...	...	...
20	.26070	.24835	.45100
21	−.09518	−.11671	−.26336

(c) Fitted Standardized Model

$$\hat{Y}^* = .7484X_1^* + .2511X_2^*$$

When fitting the standardized regression model (7.45) to the transformed data, we obtain the fitted model in Table 7.5c:

$$\hat{Y}^* = .7484X_1^* + .2511X_2^*$$

The standardized regression coefficients $b_1^* = .7484$ and $b_2^* = .2511$ are shown in the SYSTAT regression output in Figure 6.5a on page 237, labeled STD COEF. We see from the standardized regression coefficients that an increase of one standard deviation of X_1 (target population) when X_2 (per capita disposable income) is fixed leads to a much larger increase in expected sales (in units of standard deviations of Y) than does an increase of one standard deviation of X_2 when X_1 is fixed.

To shift from the standardized regression coefficients b_1^* and b_2^* back to the regression coefficients for the model with the original variables, we employ (7.53). Using the data in Table 7.5, we obtain:

$$b_1 = \left(\frac{s_Y}{s_1}\right) b_1^* = \frac{36.191}{18.620}(.7484) = 1.4546$$

$$b_2 = \left(\frac{s_Y}{s_2}\right) b_2^* = \frac{36.191}{.97035}(.2511) = 9.3652$$

$$b_0 = \bar{Y} - b_1\bar{X}_1 - b_2\bar{X}_2 = 181.90 - 1.4546(62.019) - 9.3652(17.143) = -68.860$$

TABLE 7.5 **Correlation Transformation and Fitted Standardized Regression Model—Dwaine Studios Example.**

(a) Original Data

Case i	Sales Y_i	Target Population X_{i1}	Per Capita Disposable Income X_{i2}
1	174.4	68.5	16.7
2	164.4	45.2	16.8
...	...	...	...
20	224.1	82.7	19.1
21	166.5	52.3	16.0
	$\bar{Y} = 181.90$	$\bar{X}_1 = 62.019$	$\bar{X}_2 = 17.143$
	$s_Y = 36.191$	$s_1 = 18.620$	$s_2 = .97035$

(b) Transformed Data

i	Y_i^*	X_{i1}^*	X_{i2}^*
1	−.04637	.07783	−.10205
2	−.10815	−.20198	−.07901
...	...	...	...
20	.26070	.24835	.45100
21	−.09518	−.11671	−.26336

(c) Fitted Standardized Model

$$\hat{Y}^* = .7484X_1^* + .2511X_2^*$$

When fitting the standardized regression model (7.45) to the transformed data, we obtain the fitted model in Table 7.5c:

$$\hat{Y}^* = .7484X_1^* + .2511X_2^*$$

The standardized regression coefficients $b_1^* = .7484$ and $b_2^* = .2511$ are shown in the SYSTAT regression output in Figure 6.5a on page 237, labeled STD COEF. We see from the standardized regression coefficients that an increase of one standard deviation of X_1 (target population) when X_2 (per capita disposable income) is fixed leads to a much larger increase in expected sales (in units of standard deviations of Y) than does an increase of one standard deviation of X_2 when X_1 is fixed.

To shift from the standardized regression coefficients b_1^* and b_2^* back to the regression coefficients for the model with the original variables, we employ (7.53). Using the data in Table 7.5, we obtain:

$$b_1 = \left(\frac{s_Y}{s_1}\right) b_1^* = \frac{36.191}{18.620}(.7484) = 1.4546$$

$$b_2 = \left(\frac{s_Y}{s_2}\right) b_2^* = \frac{36.191}{.97035}(.2511) = 9.3652$$

$$b_0 = \bar{Y} - b_1\bar{X}_1 - b_2\bar{X}_2 = 181.90 - 1.4546(62.019) - 9.3652(17.143) = -68.860$$

The estimated regression function for the multiple regression model in the original variables therefore is:

$$\hat{Y} = -68.860 + 1.455X_1 + 9.365X_2$$

This is the same fitted regression function we obtained in Chapter 6, except for slight rounding effect differences. Here, b_1 and b_2 cannot be compared directly because X_1 is in units of thousands of persons and X_2 is in units of thousands of dollars.

Sometimes the standardized regression coefficients $b_1^* = .7484$ and $b_2^* = .2511$ are interpreted as showing that target population (X_1) has a much greater impact on sales than per capita disposable income (X_2) because b_1^* is much larger than b_2^*. However, as we will see in the next section, one must be cautious about interpreting any regression coefficient, whether standardized or not. The reason is that when the predictor variables are correlated among themselves, as here, the regression coefficients are affected by the other predictor variables in the model. For the Dwaine Studios data, the correlation between X_1 and X_2 is $r_{12} = .781$, as shown in the correlation matrix in Figure 6.4b on page 232.

The magnitudes of the standardized regression coefficients are affected not only by the presence of correlations among the predictor variables but also by the spacings of the observations on each of these variables. Sometimes these spacings may be quite arbitrary. Hence, it is ordinarily not wise to interpret the magnitudes of standardized regression coefficients as reflecting the comparative importance of the predictor variables.

Comments

1. Some computer packages present both the regression coefficients b_k for the model in the original variables as well as the standardized coefficients b_k^*, as in the SYSTAT output in Figure 6.5a. The standardized coefficients are sometimes labeled *beta coefficients* in printouts.

2. Some computer printouts show the magnitude of the determinant of the correlation matrix of the X variables. A near-zero value for this determinant implies both a high degree of linear association among the X variables and a high potential for roundoff errors. For two X variables, this determinant is seen from (7.54) to be $1 - r_{12}^2$, which approaches 0 as r_{12}^2 approaches 1.

3. It is possible to use the correlation transformation with a computer package that does not permit regression through the origin, because the intercept coefficient b_0^* will always be zero for data so transformed. The other regression coefficients will also be correct.

4. Use of the standardized variables (7.43) without the correlation transformation modification in (7.44) will lead to the same standardized regression coefficients as those in (7.52b) for the correlation-transformed variables. However, the elements of the $\mathbf{X}'\mathbf{X}$ matrix will not then be bounded between -1 and 1. ■

7.6 Multicollinearity and Its Effects

In multiple regression analysis, the nature and significance of the relations between the predictor or explanatory variables and the response variable are often of particular interest. Some questions frequently asked are:

1. What is the relative importance of the effects of the different predictor variables?
2. What is the magnitude of the effect of a given predictor variable on the response variable?
3. Can any predictor variable be dropped from the model because it has little or no effect on the response variable?

4. Should any predictor variables not yet included in the model be considered for possible inclusion?

If the predictor variables included in the model are (1) uncorrelated among themselves and (2) uncorrelated with any other predictor variables that are related to the response variable but are omitted from the model, relatively simple answers can be given to these questions. Unfortunately, in many nonexperimental situations in business, economics, and the social and biological sciences, the predictor or explanatory variables tend to be correlated among themselves and with other variables that are related to the response variable but are not included in the model. For example, in a regression of family food expenditures on the explanatory variables family income, family savings, and age of head of household, the explanatory variables will be correlated among themselves. Further, they will also be correlated with other socioeconomic variables not included in the model that do affect family food expenditures, such as family size.

When the predictor variables are correlated among themselves, *intercorrelation* or *multicollinearity* among them is said to exist. (Sometimes the latter term is reserved for those instances when the correlation among the predictor variables is very high.) We shall explore a variety of interrelated problems created by multicollinearity among the predictor variables. First, however, we examine the situation when the predictor variables are not correlated.

Uncorrelated Predictor Variables

Table 7.6 contains data for a small-scale experiment on the effect of work crew size (X_1) and level of bonus pay (X_2) on crew productivity (Y). The predictor variables X_1 and X_2 are uncorrelated here, i.e., $r_{12}^2 = 0$, where r_{12}^2 denotes the coefficient of simple determination between X_1 and X_2. Table 7.7a contains the fitted regression function and the analysis of variance table when both X_1 and X_2 are included in the model. Table 7.7b contains the same information when only X_1 is included in the model, and Table 7.7c contains this information when only X_2 is in the model.

An important feature to note in Table 7.7 is that the regression coefficient for X_1, $b_1 = 5.375$, is the same whether only X_1 is included in the model or both predictor variables are included. The same holds for $b_2 = 9.250$. This is the result of the two predictor variables being uncorrelated.

TABLE 7.6 Uncorrelated Predictor Variables—Work Crew Productivity Example.

Case i	Crew Size X_{i1}	Bonus Pay (dollars) X_{i2}	Crew Productivity Y_i
1	4	2	42
2	4	2	39
3	4	3	48
4	4	3	51
5	6	2	49
6	6	2	53
7	6	3	61
8	6	3	60

TABLE 7.7 Regression Results when Predictor Variables Are Uncorrelated—Work Crew Productivity Example.

(a) Regression of Y on X_1 and X_2
$\hat{Y} = .375 + 5.375X_1 + 9.250X_2$

Source of Variation	*SS*	*df*	*MS*
Regression	402.250	2	201.125
Error	17.625	5	3.525
Total	419.875	7	

(b) Regression of Y on X_1
$\hat{Y} = 23.500 + 5.375X_1$

Source of Variation	*SS*	*df*	*MS*
Regression	231.125	1	231.125
Error	188.750	6	31.458
Total	419.875	7	

(c) Regression of Y on X_2
$\hat{Y} = 27.250 + 9.250X_2$

Source of Variation	*SS*	*df*	*MS*
Regression	171.125	1	171.125
Error	248.750	6	41.458
Total	419.875	7	

Thus, when the predictor variables are uncorrelated, the effects ascribed to them by a first-order regression model are the same no matter which other of these predictor variables are included in the model. This is a strong argument for controlled experiments whenever possible, since experimental control permits choosing the levels of the predictor variables so as to make these variables uncorrelated.

Another important feature of Table 7.7 is related to the error sums of squares. Note from Table 7.7 that the extra sum of squares $SSR(X_1|X_2)$ equals the regression sum of squares $SSR(X_1)$ when only X_1 is in the regression model:

$$\begin{aligned} SSR(X_1|X_2) &= SSE(X_2) - SSE(X_1, X_2) \\ &= 248.750 - 17.625 = 231.125 \\ SSR(X_1) &= 231.125 \end{aligned}$$

Similarly, the extra sum of squares $SSR(X_2|X_1)$ equals $SSR(X_2)$, the regression sum of squares when only X_2 is in the regression model:

$$\begin{aligned} SSR(X_2|X_1) &= SSE(X_1) - SSE(X_1, X_2) \\ &= 188.750 - 17.625 = 171.125 \\ SSR(X_2) &= 171.125 \end{aligned}$$

In general, when two or more predictor variables are uncorrelated, the marginal contribution of one predictor variable in reducing the error sum of squares when the other predictor variables are in the model is exactly the same as when this predictor variable is in the model alone.

Comment

To show that the regression coefficient of X_1 is unchanged when X_2 is added to the regression model in the case where X_1 and X_2 are uncorrelated, consider the following algebraic expression for b_1 in the first-order multiple regression model with two predictor variables:

$$b_1 = \frac{\dfrac{\sum(X_{i1} - \bar{X}_1)(Y_i - \bar{Y})}{\sum(X_{i1} - \bar{X}_1)^2} - \left[\dfrac{\sum(Y_i - \bar{Y})^2}{\sum(X_{i1} - \bar{X}_1)^2}\right]^{1/2} r_{Y2}r_{12}}{1 - r_{12}^2} \tag{7.56}$$

where, as before, r_{Y2} denotes the coefficient of simple correlation between Y and X_2, and r_{12} denotes the coefficient of simple correlation between X_1 and X_2.

If X_1 and X_2 are uncorrelated, $r_{12} = 0$, and (7.56) reduces to:

$$b_1 = \frac{\sum(X_{i1} - \bar{X}_1)(Y_i - \bar{Y})}{\sum(X_{i1} - \bar{X}_1)^2} \qquad \text{when } r_{12} = 0 \tag{7.56a}$$

But (7.56a) is the estimator of the slope for the simple linear regression of Y on X_1, per (1.10a).

Hence, when X_1 and X_2 are uncorrelated, adding X_2 to the regression model does not change the regression coefficient for X_1; correspondingly, adding X_1 to the regression model does not change the regression coefficient for X_2. ■

Nature of Problem when Predictor Variables Are Perfectly Correlated

To see the essential nature of the problem of multicollinearity, we shall employ a simple example where the two predictor variables are perfectly correlated. The data in Table 7.8 refer to four sample observations on a response variable and two predictor variables. Mr. A was asked to fit the first-order multiple regression function:

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \tag{7.57}$$

TABLE 7.8
Example of Perfectly Correlated Predictor Variables.

Case i	X_{i1}	X_{i2}	Y_i	Fitted Values for Regression Function (7.58)	Fitted Values for Regression Function (7.59)
1	2	6	23	23	23
2	8	9	83	83	83
3	6	8	63	63	63
4	10	10	103	103	103

Response Functions:

$\hat{Y} = -87 + X_1 + 18X_2$ (7.58)

$\hat{Y} = -7 + 9X_1 + 2X_2$ (7.59)

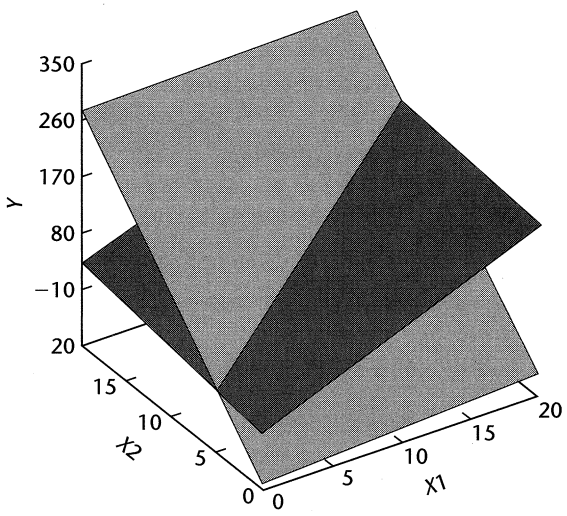


FIGURE 7.2
Two Response Planes That Intersect when $X_2 = 5 + .5X_1$.

He returned in a short time with the fitted response function:

$$\hat{Y} = -87 + X_1 + 18X_2 \qquad \textbf{(7.58)}$$

He was proud because the response function fits the data perfectly. The fitted values are shown in Table 7.8.

It so happened that Ms. B also was asked to fit the response function (7.57) to the same data, and she proudly obtained:

$$\hat{Y} = -7 + 9X_1 + 2X_2 \qquad \textbf{(7.59)}$$

Her response function also fits the data perfectly, as shown in Table 7.8.

Indeed, it can be shown that infinitely many response functions will fit the data in Table 7.8 perfectly. The reason is that the predictor variables X_1 and X_2 are perfectly related, according to the relation:

$$X_2 = 5 + .5X_1 \qquad \textbf{(7.60)}$$

Note that the fitted response functions (7.58) and (7.59) are entirely different response surfaces, as may be seen in Figure 7.2. The two response surfaces have the same fitted values only when they intersect. This occurs when X_1 and X_2 follow relation (7.60), i.e., when $X_2 = 5 + .5X_1$.

Thus, when X_1 and X_2 are perfectly related and, as in our example, the data do not contain any random error component, many different response functions will lead to the same perfectly fitted values for the observations and to the same fitted values for any other (X_1, X_2) combinations following the relation between X_1 and X_2. Yet these response functions are not the same and will lead to different fitted values for (X_1, X_2) combinations that do not follow the relation between X_1 and X_2.

Two key implications of this example are:

1. The perfect relation between X_1 and X_2 did not inhibit our ability to obtain a good fit to the data.

2. Since many different response functions provide the same good fit, we cannot interpret any one set of regression coefficients as reflecting the effects of the different predictor variables. Thus, in response function (7.58), $b_1 = 1$ and $b_2 = 18$ do not imply that X_2 is the key predictor variable and X_1 plays little role, because response function (7.59) provides an equally good fit and its regression coefficients have opposite comparative magnitudes.

Effects of Multicollinearity

In practice, we seldom find predictor variables that are perfectly related or data that do not contain some random error component. Nevertheless, the implications just noted for our idealized example still have relevance.

1. The fact that some or all predictor variables are correlated among themselves does not, in general, inhibit our ability to obtain a good fit nor does it tend to affect inferences about mean responses or predictions of new observations, provided these inferences are made within the region of observations. (Figure 6.3 on p. 231 illustrates the concept of the region of observations for the case of two predictor variables.)

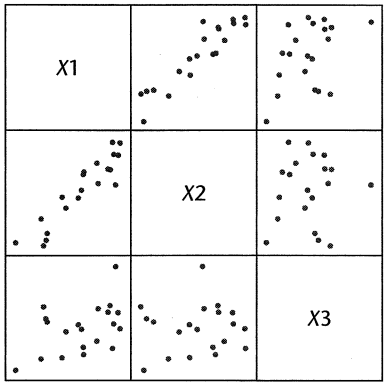
2. The counterpart in real life to the many different regression functions providing equally good fits to the data in our idealized example is that the estimated regression coefficients tend to have large sampling variability when the predictor variables are highly correlated. Thus, the estimated regression coefficients tend to vary widely from one sample to the next when the predictor variables are highly correlated. As a result, only imprecise information may be available about the individual true regression coefficients. Indeed, many of the estimated regression coefficients individually may be statistically not significant even though a definite statistical relation exists between the response variable and the set of predictor variables.

3. The common interpretation of a regression coefficient as measuring the change in the expected value of the response variable when the given predictor variable is increased by one unit while all other predictor variables are held constant is not fully applicable when multicollinearity exists. It may be conceptually feasible to think of varying one predictor variable and holding the others constant, but it may not be possible in practice to do so for predictor variables that are highly correlated. For example, in a regression model for predicting crop yield from amount of rainfall and hours of sunshine, the relation between the two predictor variables makes it unrealistic to consider varying one while holding the other constant. Therefore, the simple interpretation of the regression coefficients as measuring marginal effects is often unwarranted with highly correlated predictor variables.

We illustrate these effects of multicollinearity by returning to the body fat example. A portion of the basic data was given in Table 7.1, and regression results for different fitted models were presented in Table 7.2. Figure 7.3 contains the scatter plot matrix and the correlation matrix of the predictor variables. It is evident from the scatter plot matrix that predictor variables X_1 and X_2 are highly correlated; the correlation matrix of the X variables shows that the coefficient of simple correlation is $r_{12} = .924$. On the other hand, X_3 is not so highly related to X_1 and X_2 individually; the correlation matrix shows that the correlation coefficients are $r_{13} = .458$ and $r_{23} = .085$. (But X_3 is highly correlated with X_1 and X_2 together; the coefficient of multiple determination when X_3 is regressed on X_1 and X_2 is .998.)

FIGURE 7.3
Scatter Plot Matrix and Correlation Matrix of the Predictor Variables—Body Fat Example.

(a) Scatter Plot Matrix of *X* Variables



(b) Correlation Matrix of *X* Variables

$$\mathbf{r}_{XX} = \begin{bmatrix} 1.0 & .924 & .458 \\ .924 & 1.0 & .085 \\ .458 & .085 & 1.0 \end{bmatrix}$$

Effects on Regression Coefficients. Note from Table 7.2 that the regression coefficient for X_1, triceps skinfold thickness, varies markedly depending on which other variables are included in the model:

Variables in Model	b_1	b_2
X_1	.8572	—
X_2	—	.8565
X_1, X_2	.2224	.6594
X_1, X_2, X_3	4.334	−2.857

The story is the same for the regression coefficient for X_2. Indeed, the regression coefficient b_2 even changes sign when X_3 is added to the model that includes X_1 and X_2.

The important conclusion we must draw is: When predictor variables are correlated, the regression coefficient of any one variable depends on which other predictor variables are included in the model and which ones are left out. Thus, a regression coefficient does not reflect any inherent effect of the particular predictor variable on the response variable but only a marginal or partial effect, given whatever other correlated predictor variables are included in the model.

Comment

Another illustration of how intercorrelated predictor variables that are omitted from the regression model can influence the regression coefficients in the regression model is provided by an analyst who was perplexed about the sign of a regression coefficient in the fitted regression model. The analyst had found in a regression of territory company sales on territory population size, per capita income, and some other predictor variables that the regression coefficient for population size was negative, and this conclusion was supported by a confidence interval for the regression coefficient. A consultant noted that the analyst did not include the major competitor's market penetration as a predictor variable in the model. The competitor was most active and effective in territories with large populations, thereby

keeping company sales down in these territories. The result of the omission of this predictor variable from the model was a negative coefficient for the population size variable. ■

Effects on Extra Sums of Squares. When predictor variables are correlated, the marginal contribution of any one predictor variable in reducing the error sum of squares varies, depending on which other variables are already in the regression model, just as for regression coefficients. For example, Table 7.2 provides the following extra sums of squares for X_1:

$$SSR(X_1) = 352.27$$
$$SSR(X_1|X_2) = 3.47$$

The reason why $SSR(X_1|X_2)$ is so small compared with $SSR(X_1)$ is that X_1 and X_2 are highly correlated with each other and with the response variable. Thus, when X_2 is already in the regression model, the marginal contribution of X_1 in reducing the error sum of squares is comparatively small because X_2 contains much of the same information as X_1.

The same story is found in Table 7.2 for X_2. Here $SSR(X_2|X_1) = 33.17$, which is much smaller than $SSR(X_2) = 381.97$. The important conclusion is this: When predictor variables are correlated, there is no unique sum of squares that can be ascribed to any one predictor variable as reflecting its effect in reducing the total variation in Y. The reduction in the total variation ascribed to a predictor variable must be viewed in the context of the other correlated predictor variables already included in the model.

Comments

1. Multicollinearity also affects the coefficients of partial determination through its effects on the extra sums of squares. Note from Table 7.2 for the body fat example, for instance, that X_1 is highly correlated with Y:

$$R^2_{Y1} = \frac{SSR(X_1)}{SSTO} = \frac{352.27}{495.39} = .71$$

However, the coefficient of partial determination between Y and X_1, when X_2 is already in the regression model, is much smaller:

$$R^2_{Y1|2} = \frac{SSR(X_1|X_2)}{SSE(X_2)} = \frac{3.47}{113.42} = .03$$

The reason for the small coefficient of partial determination here is, as we have seen, that X_1 and X_2 are highly correlated with each other and with the response variable. Hence, X_1 provides only relatively limited additional information beyond that furnished by X_2.

2. The extra sum of squares for a predictor variable after other correlated predictor variables are in the model need not necessarily be smaller than before these other variables are in the model, as we found in the body fat example. In special cases, it can be larger. Consider the following special data set and its correlation matrix:

Y	X_1	X_2
20	5	25
20	10	30
0	5	5
1	10	10

$$\begin{array}{c} \\ Y \\ X_1 \\ X_2 \end{array} \begin{array}{c} \begin{array}{ccc} Y & X_1 & X_2 \end{array} \\ \begin{bmatrix} 1.0 & .026 & .976 \\ & 1.0 & .243 \\ & & 1.0 \end{bmatrix} \end{array}$$

Here, Y and X_2 are highly positively correlated, but Y and X_1 are practically uncorrelated. In addition, X_1 and X_2 are moderately positively correlated. The extra sum of squares for X_1 when it is the only variable in the model for this data set is $SSR(X_1) = .25$, but when X_2 already is in the model the extra sum of squares is $SSR(X_1|X_2) = 18.01$. Similarly, we have for these data:

$$SSR(X_2) = 362.49 \qquad SSR(X_2|X_1) = 380.25$$

The increase in the extra sums of squares with the addition of the other predictor variable in the model is related to the special situation here that X_1 is practically uncorrelated with Y but moderately correlated with X_2, which in turn is highly correlated with Y. The general point even here still holds—the extra sum of squares is affected by the other correlated predictor variables already in the model.

When $SSR(X_1|X_2) > SSR(X_1)$, as in the example just cited, the variable X_2 is sometimes called a *suppressor variable*. Since $SSR(X_2|X_1) > SSR(X_2)$ in the example, the variable X_1 would also be called a suppressor variable. ■

Effects on $s\{b_k\}$. Note from Table 7.2 for the body fat example how much more imprecise the estimated regression coefficients b_1 and b_2 become as more predictor variables are added to the regression model:

Variables in Model	$s\{b_1\}$	$s\{b_2\}$
X_1	.1288	—
X_2	—	.1100
X_1, X_2	.3034	.2912
X_1, X_2, X_3	3.016	2.582

Again, the high degree of multicollinearity among the predictor variables is responsible for the inflated variability of the estimated regression coefficients.

Effects on Fitted Values and Predictions. Notice in Table 7.2 for the body fat example that the high multicollinearity among the predictor variables does not prevent the mean square error, measuring the variability of the error terms, from being steadily reduced as additional variables are added to the regression model:

Variables in Model	*MSE*
X_1	7.95
X_1, X_2	6.47
X_1, X_2, X_3	6.15

Furthermore, the precision of fitted values within the range of the observations on the predictor variables is not eroded with the addition of correlated predictor variables into the regression model. Consider the estimation of mean body fat when the only predictor variable in the model is triceps skinfold thickness (X_1) for $X_{h1} = 25.0$. The fitted value and its estimated standard deviation are (calculations not shown):

$$\hat{Y}_h = 19.93 \qquad s\{\hat{Y}_h\} = .632$$

When the highly correlated predictor variable thigh circumference (X_2) is also included in the model, the estimated mean body fat and its estimated standard deviation are as follows

for $X_{h1} = 25.0$ and $X_{h2} = 50.0$:

$$\hat{Y}_h = 19.36 \qquad s\{\hat{Y}_h\} = .624$$

Thus, the precision of the estimated mean response is equally good as before, despite the addition of the second predictor variable that is highly correlated with the first one. This stability in the precision of the estimated mean response occurred despite the fact that the estimated standard deviation of b_1 became substantially larger when X_2 was added to the model (Table 7.2). The essential reason for the stability is that the covariance between b_1 and b_2 is negative, which plays a strong counteracting influence to the increase in $s^2\{b_1\}$ in determining the value of $s^2\{\hat{Y}_h\}$ as given in (6.79).

When all three predictor variables are included in the model, the estimated mean body fat and its estimated standard deviation are as follows for $X_{h1} = 25.0$, $X_{h2} = 50.0$, and $X_{h3} = 29.0$:

$$\hat{Y}_h = 19.19 \qquad s\{\hat{Y}_h\} = .621$$

Thus, the addition of the third predictor variable, which is highly correlated with the first two predictor variables together, also does not materially affect the precision of the estimated mean response.

Effects on Simultaneous Tests of β_k. A not infrequent abuse in the analysis of multiple regression models is to examine the t^* statistic in (6.51b):

$$t^* = \frac{b_k}{s\{b_k\}}$$

for each regression coefficient in turn to decide whether $\beta_k = 0$ for $k = 1, \ldots, p-1$. Even if a simultaneous inference procedure is used, and often it is not, problems still exist when the predictor variables are highly correlated.

Suppose we wish to test whether $\beta_1 = 0$ and $\beta_2 = 0$ in the body fat example regression model with two predictor variables of Table 7.2c. Controlling the family level of significance at .05, we require with the Bonferroni method that each of the two t tests be conducted with level of significance .025. Hence, we need $t(.9875; 17) = 2.46$. Since both t^* statistics in Table 7.2c have absolute values that do not exceed 2.46, we would conclude from the two separate tests that $\beta_1 = 0$ and that $\beta_2 = 0$. Yet the proper F test for H_0: $\beta_1 = \beta_2 = 0$ would lead to the conclusion H_a, that not both coefficients equal zero. This can be seen from Table 7.2c, where we find $F^* = MSR/MSE = 192.72/6.47 = 29.8$, which far exceeds $F(.95; 2, 17) = 3.59$.

The reason for this apparently paradoxical result is that each t^* test is a marginal test, as we have seen in (7.15) from the perspective of the general linear test approach. Thus, a small $SSR(X_1|X_2)$ here indicates that X_1 does not provide much additional information beyond X_2, which already is in the model; hence, we are led to the conclusion that $\beta_1 = 0$. Similarly, we are led to conclude $\beta_2 = 0$ here because $SSR(X_2|X_1)$ is small, indicating that X_2 does not provide much more additional information when X_1 is already in the model. But the two tests of the marginal effects of X_1 and X_2 together are not equivalent to testing whether there is a regression relation between Y and the two predictor variables. The reason is that the reduced model for each of the separate tests contains the other predictor variable, whereas the reduced model for testing whether both $\beta_1 = 0$ and $\beta_2 = 0$ would contain

neither predictor variable. The proper F test shows that there is a definite regression relation here between Y and X_1 and X_2.

The same paradox would be encountered in Table 7.2d for the regression model with three predictor variables if three simultaneous tests on the regression coefficients were conducted at family level of significance .05.

Comments

1. It was noted in Section 7.5 that a near-zero determinant of $\mathbf{X}'\mathbf{X}$ is a potential source of serious roundoff errors in normal equations calculations. Severe multicollinearity has the effect of making this determinant come close to zero. Thus, under severe multicollinearity, the regression coefficients may be subject to large roundoff errors as well as large sampling variances. Hence, it is particularly advisable to employ the correlation transformation (7.44) in normal equations calculations when multicollinearity is present.

2. Just as high intercorrelations among the predictor variables tend to make the estimated regression coefficients imprecise (i.e., erratic from sample to sample), so do the coefficients of partial correlation between the response variable and each predictor variable tend to become erratic from sample to sample when the predictor variables are highly correlated.

3. The effect of intercorrelations among the predictor variables on the standard deviations of the estimated regression coefficients can be seen readily when the variables in the model are transformed by means of the correlation transformation (7.44). Consider the first-order model with two predictor variables:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \tag{7.61}$$

This model in the variables transformed by (7.44) becomes:

$$Y_i^* = \beta_1^* X_{i1}^* + \beta_2^* X_{i2}^* + \varepsilon_i^* \tag{7.62}$$

The $(\mathbf{X}'\mathbf{X})^{-1}$ matrix for this standardized model is given by (7.50) and (7.54c):

$$(\mathbf{X}'\mathbf{X})^{-1} = \mathbf{r}_{XX}^{-1} = \frac{1}{1 - r_{12}^2} \begin{bmatrix} 1 & -r_{12} \\ -r_{12} & 1 \end{bmatrix} \tag{7.63}$$

Hence, the variance-covariance matrix of the estimated regression coefficients is by (6.46) and (7.63):

$$\boldsymbol{\sigma}^2\{\mathbf{b}\} = (\sigma^*)^2 \mathbf{r}_{XX}^{-1} = (\sigma^*)^2 \frac{1}{1 - r_{12}^2} \begin{bmatrix} 1 & -r_{12} \\ -r_{12} & 1 \end{bmatrix} \tag{7.64}$$

where $(\sigma^*)^2$ is the error term variance for the standardized model (7.62). We see that the estimated regression coefficients b_1^* and b_2^* have the same variance here:

$$\sigma^2\{b_1^*\} = \sigma^2\{b_2^*\} = \frac{(\sigma^*)^2}{1 - r_{12}^2} \tag{7.65}$$

and that each of these variances become larger as the correlation between X_1 and X_2 increases. Indeed, as X_1 and X_2 approach perfect correlation (i.e., as r_{12}^2 approaches 1), the variances of b_1^* and b_2^* become larger without limit.

4. We noted in our discussion of simultaneous tests of the regression coefficients that it is possible that a set of predictor variables is related to the response variable, yet all of the individual tests on the regression coefficients will lead to the conclusion that they equal zero because of the multicollinearity among the predictor variables. This apparently paradoxical result is also possible under special circumstances when there is no multicollinearity among the predictor variables. The special circumstances are not likely to be found in practice, however. ■