Chapter 7

Multiple Regression II

In this chapter, we take up some specialized topics that are unique to multiple regression. These include extra sums of squares, which are useful for conducting a variety of tests about the regression coefficients, the standardized version of the multiple regression model, and multicollinearity, a condition where the predictor variables are highly correlated.

7.1 Extra Sums of Squares

Basic Ideas

An extra sum of squares measures the marginal reduction in the error sum of squares when one or several predictor variables are added to the regression model, given that other predictor variables are already in the model. Equivalently, one can view an extra sum of squares as measuring the marginal increase in the regression sum of squares when one or several predictor variables are added to the regression model.

We first utilize an example to illustrate these ideas, and then we present definitions of extra sums of squares and discuss a variety of uses of extra sums of squares in tests about regression coefficients.

Example

Table 7.1 contains a portion of the data for a study of the relation of amount of body fat (Y) to several possible predictor variables, based on a sample of 20 healthy females 25–34 years old. The possible predictor variables are triceps skinfold thickness (X_1), thigh circumference (X_2), and midarm circumference (X_3). The amount of body fat in Table 7.1 for each of the 20 persons was obtained by a cumbersome and expensive procedure requiring the immersion of the person in water. It would therefore be very helpful if a regression model with some or all of these predictor variables could provide reliable estimates of the amount of body fat since the measurements needed for the predictor variables are easy to obtain.

Table 7.2 contains some of the main regression results when body fat (Y) is regressed (1) on triceps skinfold thickness (X_1) alone, (2) on thigh circumference (X_2) alone, (3) on X_1 and X_2 only, and (4) on all three predictor variables. To keep track of the regression model that is fitted, we shall modify our notation slightly. The regression sum of squares when X_1 only is in the model is, according to Table 7.2a, 352.27. This sum of squares will be denoted by $SSR(X_1)$. The error sum of squares for this model will be denoted by $SSE(X_1)$; according to Table 7.2a it is $SSE(X_1) = 143.12$.

Similarly, Table 7.2c indicates that when X_1 and X_2 are in the regression model, the regression sum of squares is $SSR(X_1, X_2) = 385.44$ and the error sum of squares is $SSE(X_1, X_2) = 109.95$.

Notice that the error sum of squares when X_1 and X_2 are in the model, $SSE(X_1, X_2) = 109.95$, is smaller than when the model contains only X_1, $SSE(X_1) = 143.12$. The difference is called an *extra sum of squares* and will be denoted by $SSR(X_2|X_1)$:

$$SSR(X_2|X_1) = SSE(X_1) - SSE(X_1, X_2)$$
$$= 143.12 - 109.95 = 33.17$$

TABLE 7.1 Basic Data—Body Fat Example.

Subject i	Triceps Skinfold Thickness X_{i1}	Thigh Circumference X_{i2}	Midarm Circumference X_{i3}	Body Fat Y_i
1	19.5	43.1	29.1	11.9
2	24.7	49.8	28.2	22.8
3	30.7	51.9	37.0	18.7
...	...	...	...	...
18	30.2	58.6	24.6	25.4
19	22.7	48.2	27.1	14.8
20	25.2	51.0	27.5	21.1

TABLE 7.2 Regression Results for Several Fitted Models—Body Fat Example.

(a) Regression of Y on X_1
$\hat{Y} = -1.496 + .8572X_1$

Source of Variation	SS	df	MS
Regression	352.27	1	352.27
Error	143.12	18	7.95
Total	495.39	19	

Variable	Estimated Regression Coefficient	Estimated Standard Deviation	t^*
X_1	$b_1 = .8572$	$s\{b_1\} = .1288$	6.66

(b) Regression of Y on X_2
$\hat{Y} = -23.634 + .8565X_2$

Source of Variation	SS	df	MS
Regression	381.97	1	381.97
Error	113.42	18	6.30
Total	495.39	19	

Variable	Estimated Regression Coefficient	Estimated Standard Deviation	t^*
X_2	$b_2 = .8565$	$s\{b_2\} = .1100$	7.79

(*continued*)

TABLE 7.2 (Continued).

(c) Regression of Y on X_1 and X_2

$\hat{Y} = -19.174 + .2224X_1 + .6594X_2$

Source of Variation	SS	df	MS
Regression	385.44	2	192.72
Error	109.95	17	6.47
Total	495.39	19	

Variable	Estimated Regression Coefficient	Estimated Standard Deviation	t^*
X_1	$b_1 = .2224$	$s\{b_1\} = .3034$	.73
X_2	$b_2 = .6594$	$s\{b_2\} = .2912$	2.26

(d) Regression of Y on X_1, X_2, and X_3

$\hat{Y} = 117.08 + 4.334X_1 - 2.857X_2 - 2.186X_3$

Source of Variation	SS	df	MS
Regression	396.98	3	132.33
Error	98.41	16	6.15
Total	495.39	19	

Variable	Estimated Regression Coefficient	Estimated Standard Deviation	t^*
X_1	$b_1 = 4.334$	$s\{b_1\} = 3.016$	1.44
X_2	$b_2 = -2.857$	$s\{b_2\} = 2.582$	-1.11
X_3	$b_3 = -2.186$	$s\{b_3\} = 1.596$	-1.37

This reduction in the error sum of squares is the result of adding X_2 to the regression model when X_1 is already included in the model. Thus, the extra sum of squares $SSR(X_2|X_1)$ measures the marginal effect of adding X_2 to the regression model when X_1 is already in the model. The notation $SSR(X_2|X_1)$ reflects this additional or extra reduction in the error sum of squares associated with X_2, given that X_1 is already included in the model.

The extra sum of squares $SSR(X_2|X_1)$ equivalently can be viewed as the marginal increase in the regression sum of squares:

$$
\begin{aligned}
SSR(X_2|X_1) &= SSR(X_1, X_2) - SSR(X_1) \\
&= 385.44 - 352.27 = 33.17
\end{aligned}
$$

The reason for the equivalence of the marginal reduction in the error sum of squares and the marginal increase in the regression sum of squares is the basic analysis of variance identity (2.50):

$$SSTO = SSR + SSE$$

Since $SSTO$ measures the variability of the Y_i observations and hence does not depend on the regression model fitted, any reduction in SSE implies an identical increase in SSR.

We can consider other extra sums of squares, such as the marginal effect of adding X_3 to the regression model when X_1 and X_2 are already in the model. We find from Tables 7.2c and 7.2d that:

$$SSR(X_3|X_1, X_2) = SSE(X_1, X_2) - SSE(X_1, X_2, X_3)$$
$$= 109.95 - 98.41 = 11.54$$

or, equivalently:

$$SSR(X_3|X_1, X_2) = SSR(X_1, X_2, X_3) - SSR(X_1, X_2)$$
$$= 396.98 - 385.44 = 11.54$$

We can even consider the marginal effect of adding several variables, such as adding both X_2 and X_3 to the regression model already containing X_1 (see Tables 7.2a and 7.2d):

$$SSR(X_2, X_3|X_1) = SSE(X_1) - SSE(X_1, X_2, X_3)$$
$$= 143.12 - 98.41 = 44.71$$

or, equivalently:

$$SSR(X_2, X_3|X_1) = SSR(X_1, X_2, X_3) - SSR(X_1)$$
$$= 396.98 - 352.27 = 44.71$$

Definitions

We assemble now our earlier definitions of extra sums of squares and provide some additional ones. As we noted earlier, an extra sum of squares always involves the difference between the error sum of squares for the regression model containing the X variable(s) already in the model and the error sum of squares for the regression model containing both the original X variable(s) and the new X variable(s). Equivalently, an extra sum of squares involves the difference between the two corresponding regression sums of squares.

Thus, we define:

$$SSR(X_1|X_2) = SSE(X_2) - SSE(X_1, X_2) \tag{7.1a}$$

or, equivalently:

$$SSR(X_1|X_2) = SSR(X_1, X_2) - SSR(X_2) \tag{7.1b}$$

If X_2 is the extra variable, we define:

$$SSR(X_2|X_1) = SSE(X_1) - SSE(X_1, X_2) \tag{7.2a}$$

or, equivalently:

$$SSR(X_2|X_1) = SSR(X_1, X_2) - SSR(X_1) \tag{7.2b}$$

Extensions for three or more variables are straightforward. For example, we define:

$$SSR(X_3|X_1, X_2) = SSE(X_1, X_2) - SSE(X_1, X_2, X_3) \tag{7.3a}$$

or:

$$SSR(X_3|X_1, X_2) = SSR(X_1, X_2, X_3) - SSR(X_1, X_2) \tag{7.3b}$$

and:

$$SSR(X_2, X_3|X_1) = SSE(X_1) - SSE(X_1, X_2, X_3) \tag{7.4a}$$

or:

$$SSR(X_2, X_3|X_1) = SSR(X_1, X_2, X_3) - SSR(X_1) \tag{7.4b}$$

Decomposition of *SSR* into Extra Sums of Squares

In multiple regression, unlike simple linear regression, we can obtain a variety of decompositions of the regression sum of squares SSR into extra sums of squares. Let us consider the case of two X variables. We begin with the identity (2.50) for variable X_1:

$$SSTO = SSR(X_1) + SSE(X_1) \tag{7.5}$$

where the notation now shows explicitly that X_1 is the X variable in the model. Replacing $SSE(X_1)$ by its equivalent in (7.2a), we obtain:

$$SSTO = SSR(X_1) + SSR(X_2|X_1) + SSE(X_1, X_2) \tag{7.6}$$

We now make use of the same identity for multiple regression with two X variables as in (7.5) for a single X variable, namely:

$$SSTO = SSR(X_1, X_2) + SSE(X_1, X_2) \tag{7.7}$$

Solving (7.7) for $SSE(X_1, X_2)$ and using this expression in (7.6) lead to:

$$SSR(X_1, X_2) = SSR(X_1) + SSR(X_2|X_1) \tag{7.8}$$

Thus, we have decomposed the regression sum of squares $SSR(X_1, X_2)$ into two marginal components: (1) $SSR(X_1)$, measuring the contribution by including X_1 alone in the model, and (2) $SSR(X_2|X_1)$, measuring the additional contribution when X_2 is included, given that X_1 is already in the model.

Of course, the order of the X variables is arbitrary. Here, we can also obtain the decomposition:

$$SSR(X_1, X_2) = SSR(X_2) + SSR(X_1|X_2) \tag{7.9}$$

We show in Figure 7.1 schematic representations of the two decompositions of $SSR(X_1, X_2)$ for the body fat example. The total bar on the left represents $SSTO$ and presents decomposition (7.9). The unshaded component of this bar is $SSR(X_2)$, and the combined shaded area represents $SSE(X_2)$. The latter area in turn is the combination of the extra sum of squares $SSR(X_1|X_2)$ and the error sum of squares $SSE(X_1, X_2)$ when both X_1 and X_2 are included in the model. Similarly, the bar on the right in Figure 7.1 shows decomposition (7.8). Note in both cases how the extra sum of squares can be viewed either as a reduction in the error sum of squares or as an increase in the regression sum of squares when the second predictor variable is added to the regression model.

When the regression model contains three X variables, a variety of decompositions of $SSR(X_1, X_2, X_3)$ can be obtained. We illustrate three of these:

$$SSR(X_1, X_2, X_3) = SSR(X_1) + SSR(X_2|X_1) + SSR(X_3|X_1, X_2) \tag{7.10a}$$

$$SSR(X_1, X_2, X_3) = SSR(X_2) + SSR(X_3|X_2) + SSR(X_1|X_2, X_3) \tag{7.10b}$$

$$SSR(X_1, X_2, X_3) = SSR(X_1) + SSR(X_2, X_3|X_1) \tag{7.10c}$$

FIGURE 7.1 **Schematic Representation of Extra Sums of Squares—Body Fat Example.**

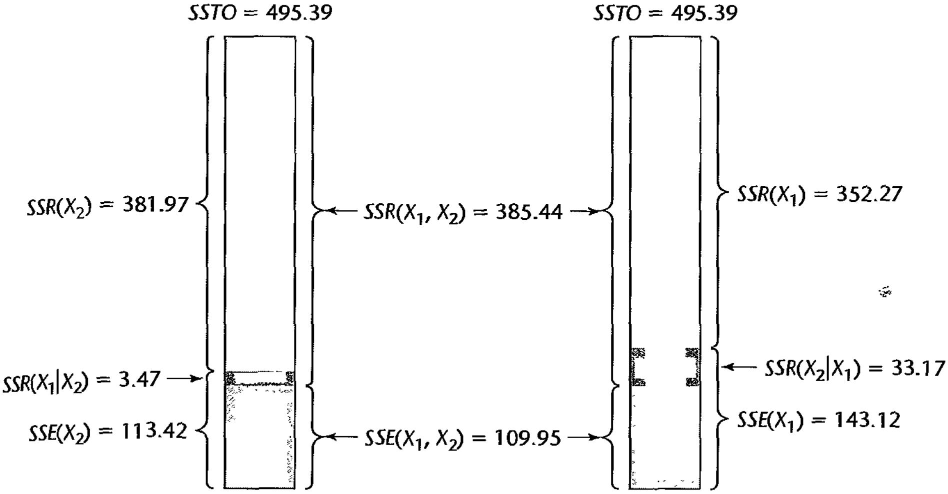


TABLE 7.3 **Example of ANOVA Table with Decomposition of *SSR* for Three *X* Variables.**

Source of Variation	*SS*	*df*	*MS*
Regression	$SSR(X_1, X_2, X_3)$	3	$MSR(X_1, X_2, X_3)$
X_1	$SSR(X_1)$	1	$MSR(X_1)$
$X_2 \mid X_1$	$SSR(X_2 \mid X_1)$	1	$MSR(X_2 \mid X_1)$
$X_3 \mid X_1, X_2$	$SSR(X_3 \mid X_1, X_2)$	1	$MSR(X_3 \mid X_1, X_2)$
Error	$SSE(X_1, X_2, X_3)$	$n-4$	$MSE(X_1, X_2, X_3)$
Total	$SSTO$	$n-1$	

It is obvious that the number of possible decompositions becomes vast as the number of X variables in the regression model increases.

ANOVA Table Containing Decomposition of *SSR*

ANOVA tables can be constructed containing decompositions of the regression sum of squares into extra sums of squares. Table 7.3 contains the ANOVA table decomposition for the case of three X variables often used in regression packages, and Table 7.4 contains this same decomposition for the body fat example. The decomposition involves single extra X variables.

Note that each extra sum of squares involving a single extra X variable has associated with it one degree of freedom. The resulting mean squares are constructed as usual. For example, $MSR(X_2|X_1)$ in Table 7.3 is obtained as follows:

$$MSR(X_2|X_1) = \frac{SSR(X_2|X_1)}{1}$$

Extra sums of squares involving two extra X variables, such as $SSR(X_2, X_3|X_1)$, have two degrees of freedom associated with them. This follows because we can express such an extra sum of squares as a sum of two extra sums of squares, each associated with one

TABLE 7.4 ANOVA Table with Decomposition of *SSR*—Body Fat Example with Three Predictor Variables.

Source of Variation	*SS*	*df*	*MS*
Regression	396.98	3	132.33
X_1	352.27	1	352.27
$X_2 \mid X_1$	33.17	1	33.17
$X_3 \mid X_1, X_2$	11.54	1	11.54
Error	98.41	16	6.15
Total	495.39	19	

degree of freedom. For example, by definition of the extra sums of squares, we have:

$$SSR(X_2, X_3|X_1) = SSR(X_2|X_1) + SSR(X_3|X_1, X_2) \tag{7.11}$$

The mean square $MSR(X_2, X_3|X_1)$ is therefore obtained as follows:

$$MSR(X_2, X_3|X_1) = \frac{SSR(X_2, X_3|X_1)}{2}$$

Many computer regression packages provide decompositions of SSR into single-degree-of-freedom extra sums of squares, usually in the order in which the X variables are entered into the model. Thus, if the X variables are entered in the order X_1, X_2, X_3, the extra sums of squares given in the output are:

$$SSR(X_1)$$
$$SSR(X_2|X_1)$$
$$SSR(X_3|X_1, X_2)$$

If an extra sum of squares involving several extra X variables is desired, it can be obtained by summing appropriate single-degree-of-freedom extra sums of squares. For instance, to obtain $SSR(X_2, X_3|X_1)$ in our earlier illustration, we would utilize (7.11) and simply add $SSR(X_2|X_1)$ and $SSR(X_3|X_1, X_2)$.

If the extra sum of squares $SSR(X_1, X_3|X_2)$ were desired with a computer package that provides single-degree-of-freedom extra sums of squares in the order in which the X variables are entered, the X variables would need to be entered in the order X_2, X_1, X_3 or X_2, X_3, X_1. The first ordering would give:

$$SSR(X_2)$$
$$SSR(X_1|X_2)$$
$$SSR(X_3|X_1, X_2)$$

The sum of the last two extra sums of squares will yield $SSR(X_1, X_3|X_2)$.

The reason why extra sums of squares are of interest is that they occur in a variety of tests about regression coefficients where the question of concern is whether certain X variables can be dropped from the regression model. We turn next to this use of extra sums of squares.

7.2 Uses of Extra Sums of Squares in Tests for Regression Coefficients

Test whether a Single $\beta_k = 0$

When we wish to test whether the term $\beta_k X_k$ can be dropped from a multiple regression model, we are interested in the alternatives:

$$H_0: \beta_k = 0$$
$$H_a: \beta_k \neq 0$$

We already know that test statistic (6.51b):

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

is appropriate for this test.

Equivalently, we can use the general linear test approach described in Section 2.8. We now show that this approach involves an extra sum of squares. Let us consider the first-order regression model with three predictor variables:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i \qquad \text{Full model} \tag{7.12}$$

To test the alternatives:

$$\begin{aligned} H_0&: \beta_3 = 0 \\ H_a&: \beta_3 \neq 0 \end{aligned} \tag{7.13}$$

we fit the full model and obtain the error sum of squares $SSE(F)$. We now explicitly show the variables in the full model, as follows:

$$SSE(F) = SSE(X_1, X_2, X_3)$$

The degrees of freedom associated with $SSE(F)$ are $df_F = n - 4$ since there are four parameters in the regression function for the full model (7.12).

The reduced model when H_0 in (7.13) holds is:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \qquad \text{Reduced model} \tag{7.14}$$

We next fit this reduced model and obtain:

$$SSE(R) = SSE(X_1, X_2)$$

There are $df_R = n - 3$ degrees of freedom associated with the reduced model.

The general linear test statistic (2.70):

$$F^* = \frac{SSE(R) - SSE(F)}{df_R - df_F} \div \frac{SSE(F)}{df_F}$$

here becomes:

$$F^* = \frac{SSE(X_1, X_2) - SSE(X_1, X_2, X_3)}{(n-3) - (n-4)} \div \frac{SSE(X_1, X_2, X_3)}{n-4}$$

Note that the difference between the two error sums of squares in the numerator term is the extra sum of squares (7.3a):

$$SSE(X_1, X_2) - SSE(X_1, X_2, X_3) = SSR(X_3|X_1, X_2)$$

Hence the general linear test statistic here is:

$$F^* = \frac{SSR(X_3|X_1, X_2)}{1} \div \frac{SSE(X_1, X_2, X_3)}{n-4} = \frac{MSR(X_3|X_1, X_2)}{MSE(X_1, X_2, X_3)} \tag{7.15}$$

We thus see that the test whether or not $\beta_3 = 0$ is a marginal test, given that X_1 and X_2 are already in the model. We also note that the extra sum of squares $SSR(X_3|X_1, X_2)$ has one degree of freedom associated with it, just as we noted earlier.

Test statistic (7.15) shows that we do not need to fit both the full model and the reduced model to use the general linear test approach here. A single computer run can provide a fit of the full model and the appropriate extra sum of squares.

Example

In the body fat example, we wish to test for the model with all three predictor variables whether midarm circumference (X_3) can be dropped from the model. The test alternatives are those of (7.13). Table 7.4 contains the ANOVA results from a computer fit of the full regression model (7.12), including the extra sums of squares when the predictor variables are entered in the order X_1, X_2, X_3. Hence, test statistic (7.15) here is:

$$F^* = \frac{SSR(X_3|X_1, X_2)}{1} \div \frac{SSE(X_1, X_2, X_3)}{n-4}$$
$$= \frac{11.54}{1} \div \frac{98.41}{16} = 1.88$$

For $\alpha = .01$, we require $F(.99; 1, 16) = 8.53$. Since $F^* = 1.88 \leq 8.53$, we conclude H_0, that X_3 can be dropped from the regression model that already contains X_1 and X_2.

Note from Table 7.2d that the t^* test statistic here is:

$$t^* = \frac{b_3}{s\{b_3\}} = \frac{-2.186}{1.596} = -1.37$$

Since $(t^*)^2 = (-1.37)^2 = 1.88 = F^*$, we see that the two test statistics are equivalent, just as for simple linear regression.

Comment

The F^* test statistic (7.15) to test whether or not $\beta_3 = 0$ is called a *partial F test* statistic to distinguish it from the F^* statistic in (6.39b) for testing whether *all* $\beta_k = 0$, i.e., whether or not there is a regression relation between Y and the set of X variables. The latter test is called the *overall F test*. ■

Test whether Several $\beta_k = 0$

In multiple regression we are frequently interested in whether several terms in the regression model can be dropped. For example, we may wish to know whether both $\beta_2 X_2$ and $\beta_3 X_3$ can be dropped from the full model (7.12). The alternatives here are:

$$\begin{aligned} &H_0\colon \beta_2 = \beta_3 = 0 \\ &H_a\colon \text{not both } \beta_2 \text{ and } \beta_3 \text{ equal zero} \end{aligned} \tag{7.16}$$

With the general linear test approach, the reduced model under H_0 is:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \varepsilon_i \qquad \text{Reduced model} \tag{7.17}$$

and the error sum of squares for the reduced model is:

$$SSE(R) = SSE(X_1)$$

This error sum of squares has $df_R = n - 2$ degrees of freedom associated with it.

The general linear test statistic (2.70) thus becomes here:

$$F^* = \frac{SSE(X_1) - SSE(X_1, X_2, X_3)}{(n-2) - (n-4)} \div \frac{SSE(X_1, X_2, X_3)}{n-4}$$

Again the difference between the two error sums of squares in the numerator term is an extra sum of squares, namely:

$$SSE(X_1) - SSE(X_1, X_2, X_3) = SSR(X_2, X_3|X_1)$$

Hence, the test statistic becomes:

$$F^* = \frac{SSR(X_2, X_3|X_1)}{2} \div \frac{SSE(X_1, X_2, X_3)}{n-4} = \frac{MSR(X_2, X_3|X_1)}{MSE(X_1, X_2, X_3)} \tag{7.18}$$

Note that $SSR(X_2, X_3|X_1)$ has two degrees of freedom associated with it, as we pointed out earlier.

Example

We wish to test in the body fat example for the model with all three predictor variables whether both thigh circumference (X_2) and midarm circumference (X_3) can be dropped from the full regression model (7.12). The alternatives are those in (7.16). The appropriate extra sum of squares can be obtained from Table 7.4, using (7.11):

$$\begin{aligned} SSR(X_2, X_3|X_1) &= SSR(X_2|X_1) + SSR(X_3|X_1, X_2) \\ &= 33.17 + 11.54 = 44.71 \end{aligned}$$

Test statistic (7.18) therefore is:

$$\begin{aligned} F^* &= \frac{SSR(X_2, X_3|X_1)}{2} \div MSE(X_1, X_2, X_3) \\ &= \frac{44.71}{2} \div 6.15 = 3.63 \end{aligned}$$

For $\alpha = .05$, we require $F(.95; 2, 16) = 3.63$. Since $F^* = 3.63$ is at the boundary of the decision rule (the P-value of the test statistic is .05), we may wish to make further analyses before deciding whether X_2 and X_3 should be dropped from the regression model that already contains X_1.

Comments

1. For testing whether a single β_k equals zero, two equivalent test statistics are available: the t^* test statistic and the F^* general linear test statistic. When testing whether several β_k equal zero, only the general linear test statistic F^* is available.

2. General linear test statistic (2.70) for testing whether several X variables can be dropped from the general linear regression model (6.7) can be expressed in terms of the coefficients of

multiple determination for the full and reduced models. Denoting these by R_F^2 and R_R^2, respectively, we have:

$$F^* = \frac{R_F^2 - R_R^2}{df_R - df_F} \div \frac{1 - R_F^2}{df_F} \tag{7.19}$$

Specifically for testing the alternatives in (7.16) for the body fat example, test statistic (7.19) becomes:

$$F^* = \frac{R_{Y|123}^2 - R_{Y|1}^2}{(n-2)-(n-4)} \div \frac{1 - R_{Y|123}^2}{n-4} \tag{7.20}$$

where $R_{Y|123}^2$ denotes the coefficient of multiple determination when Y is regressed on X_1, X_2, and X_3, and $R_{Y|1}^2$ denotes the coefficient when Y is regressed on X_1 alone.

We see from Table 7.4 that $R_{Y|123}^2 = 396.98/495.39 = .80135$ and $R_{Y|1}^2 = 352.27/495.39 = .71110$. Hence, we obtain by substituting in (7.20):

$$F^* = \frac{.80135 - .71110}{(20-2)-(20-4)} \div \frac{1 - .80135}{16} = 3.63$$

This is the same result as before. Note that $R_{Y|1}^2$ corresponds to the coefficient of simple determination R^2 between Y and X_1.

Test statistic (7.19) is not appropriate when the full and reduced regression models do not contain the intercept term β_0. In that case, the general linear test statistic in the form (2.70) must be used. ■

7.3 Summary of Tests Concerning Regression Coefficients

We have already discussed how to conduct several types of tests concerning regression coefficients in a multiple regression model. For completeness, we summarize here these tests as well as some additional types of tests.

Test whether All $\beta_k = 0$

This is the *overall F test* (6.39) of whether or not there is a regression relation between the response variable Y and the set of X variables. The alternatives are:

$$\begin{aligned} &H_0\colon \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0 \\ &H_a\colon \text{not all } \beta_k \ (k = 1, \ldots, p-1) \text{ equal zero} \end{aligned} \tag{7.21}$$

and the test statistic is:

$$\begin{aligned} F^* &= \frac{SSR(X_1, \ldots, X_{p-1})}{p-1} \div \frac{SSE(X_1, \ldots, X_{p-1})}{n-p} \\ &= \frac{MSR}{MSE} \end{aligned} \tag{7.22}$$

If H_0 holds, $F^* \sim F(p-1, n-p)$. Large values of F^* lead to conclusion H_a.

Test whether a Single $\beta_k = 0$

This is a *partial F test* of whether a particular regression coefficient β_k equals zero. The alternatives are:

$$\begin{aligned} H_0&: \beta_k = 0 \\ H_a&: \beta_k \neq 0 \end{aligned} \tag{7.23}$$

and the test statistic is:

$$\begin{aligned} F^* &= \frac{SSR(X_k|X_1, \ldots, X_{k-1}, X_{k+1}, \ldots, X_{p-1})}{1} \div \frac{SSE(X_1, \ldots, X_{p-1})}{n-p} \\ &= \frac{MSR(X_k|X_1, \ldots, X_{k-1}, X_{k+1}, \ldots, X_{p-1})}{MSE} \end{aligned} \tag{7.24}$$

If H_0 holds, $F^* \sim F(1, n-p)$. Large values of F^* lead to conclusion H_a. Statistics packages that provide extra sums of squares permit use of this test without having to fit the reduced model.

An equivalent test statistic is (6.51b):

$$t^* = \frac{b_k}{s\{b_k\}} \tag{7.25}$$

If H_0 holds, $t^* \sim t(n-p)$. Large values of $|t^*|$ lead to conclusion H_a.

Since the two tests are equivalent, the choice is usually made in terms of available information provided by the regression package output.

Test whether Some $\beta_k = 0$

This is another *partial F test*. Here, the alternatives are:

$$\begin{aligned} H_0&: \beta_q = \beta_{q+1} = \cdots = \beta_{p-1} = 0 \\ H_a&: \text{not all of the } \beta_k \text{ in } H_0 \text{ equal zero} \end{aligned} \tag{7.26}$$

where for convenience, we arrange the model so that the last $p-q$ coefficients are the ones to be tested. The test statistic is:

$$\begin{aligned} F^* &= \frac{SSR(X_q, \ldots, X_{p-1}|X_1, \ldots, X_{q-1})}{p-q} \div \frac{SSE(X_1, \ldots, X_{p-1})}{n-p} \\ &= \frac{MSR(X_q, \ldots, X_{p-1}|X_1, \ldots, X_{q-1})}{MSE} \end{aligned} \tag{7.27}$$

If H_0 holds, $F^* \sim F(p-q, n-p)$. Large values of F^* lead to conclusion H_a.

Note that test statistic (7.27) actually encompasses the two earlier cases. If $q = 1$, the test is whether all regression coefficients equal zero. If $q = p-1$, the test is whether a single regression coefficient equals zero. Also note that test statistic (7.27) can be calculated without having to fit the reduced model if the regression package provides the needed extra sums of squares:

$$\begin{aligned} &SSR(X_q, \ldots, X_{p-1}|X_1, \ldots, X_{q-1}) \\ &\quad = SSR(X_q|X_1, \ldots, X_{q-1}) + \cdots + SSR(X_{p-1}|X_1, \ldots, X_{p-2}) \end{aligned} \tag{7.28}$$

Test statistic (7.27) can be stated equivalently in terms of the coefficients of multiple determination for the full and reduced models when these models contain the intercept term β_0, as follows:

$$F^* = \frac{R^2_{Y|1\cdots p-1} - R^2_{Y|1\cdots q-1}}{p-q} \div \frac{1 - R^2_{Y|1\cdots p-1}}{n-p} \tag{7.29}$$

where $R^2_{Y|1\cdots p-1}$ denotes the coefficient of multiple determination when Y is regressed on all X variables, and $R^2_{Y|1\cdots q-1}$ denotes the coefficient when Y is regressed on $X_1, \ldots, X_{q-1}$ only.

Other Tests

When tests about regression coefficients are desired that do not involve testing whether one or several β_k equal zero, extra sums of squares cannot be used and the general linear test approach requires separate fittings of the full and reduced models. For instance, for the full model containing three X variables:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i \qquad \text{Full model} \tag{7.30}$$

we might wish to test:

$$\begin{aligned} H_0&: \beta_1 = \beta_2 \\ H_a&: \beta_1 \neq \beta_2 \end{aligned} \tag{7.31}$$

The procedure would be to fit the full model (7.30), and then the reduced model:

$$Y_i = \beta_0 + \beta_c(X_{i1} + X_{i2}) + \beta_3 X_{i3} + \varepsilon_i \qquad \text{Reduced model} \tag{7.32}$$

where β_c denotes the common coefficient for β_1 and β_2 under H_0 and $X_{i1} + X_{i2}$ is the corresponding new X variable. We then use the general F^* test statistic (2.70) with 1 and $n - 4$ degrees of freedom.

Another example where extra sums of squares cannot be used is in the following test for regression model (7.30):

$$\begin{aligned} H_0&: \beta_1 = 3, \beta_3 = 5 \\ H_a&: \text{not both equalities in } H_0 \text{ hold} \end{aligned} \tag{7.33}$$

Here, the reduced model would be:

$$Y_i - 3X_{i1} - 5X_{i3} = \beta_0 + \beta_2 X_{i2} + \varepsilon_i \qquad \text{Reduced model} \tag{7.34}$$

Note the new response variable $Y - 3X_1 - 5X_3$ in the reduced model, since $\beta_1 X_1$ and $\beta_3 X_3$ are known constants under H_0. We then use the general linear test statistic F^* in (2.70) with 2 and $n - 4$ degrees of freedom.

7.4 Coefficients of Partial Determination

Extra sums of squares are not only useful for tests on the regression coefficients of a multiple regression model, but they are also encountered in descriptive measures of relationship called coefficients of partial determination. Recall that the coefficient of multiple determination, R^2, measures the proportionate reduction in the variation of Y achieved by the introduction of the entire set of X variables considered in the model. A *coefficient of partial determination*, in contrast, measures the marginal contribution of one X variable when all others are already included in the model.

Two Predictor Variables

We first consider a first-order multiple regression model with two predictor variables, as given in (6.1):

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

$SSE(X_2)$ measures the variation in Y when X_2 is included in the model. $SSE(X_1, X_2)$ measures the variation in Y when both X_1 and X_2 are included in the model. Hence, the relative marginal reduction in the variation in Y associated with X_1 when X_2 is already in the model is:

$$\frac{SSE(X_2) - SSE(X_1, X_2)}{SSE(X_2)} = \frac{SSR(X_1|X_2)}{SSE(X_2)}$$

This measure is the coefficient of partial determination between Y and X_1, given that X_2 is in the model. We denote this measure by $R^2_{Y1|2}$:

$$R^2_{Y1|2} = \frac{SSE(X_2) - SSE(X_1, X_2)}{SSE(X_2)} = \frac{SSR(X_1|X_2)}{SSE(X_2)} \tag{7.35}$$

Thus, $R^2_{Y1|2}$ measures the proportionate reduction in the variation in Y remaining after X_2 is included in the model that is gained by also including X_1 in the model.

The coefficient of partial determination between Y and X_2, given that X_1 is in the model, is defined correspondingly:

$$R^2_{Y2|1} = \frac{SSR(X_2|X_1)}{SSE(X_1)} \tag{7.36}$$

General Case

The generalization of coefficients of partial determination to three or more X variables in the model is immediate. For instance:

$$R^2_{Y1|23} = \frac{SSR(X_1|X_2, X_3)}{SSE(X_2, X_3)} \tag{7.37}$$

$$R^2_{Y2|13} = \frac{SSR(X_2|X_1, X_3)}{SSE(X_1, X_3)} \tag{7.38}$$

$$R^2_{Y3|12} = \frac{SSR(X_3|X_1, X_2)}{SSE(X_1, X_2)} \tag{7.39}$$

$$R^2_{Y4|123} = \frac{SSR(X_4|X_1, X_2, X_3)}{SSE(X_1, X_2, X_3)} \tag{7.40}$$

Note that in the subscripts to R^2, the entries to the left of the vertical bar show in turn the variable taken as the response and the X variable being added. The entries to the right of the vertical bar show the X variables already in the model.

Example

For the body fat example, we can obtain a variety of coefficients of partial determination. Here are three (Tables 7.2 and 7.4):

$$R^2_{Y2|1} = \frac{SSR(X_2|X_1)}{SSE(X_1)} = \frac{33.17}{143.12} = .232$$

$$R^2_{Y3|12} = \frac{SSR(X_3|X_1, X_2)}{SSE(X_1, X_2)} = \frac{11.54}{109.95} = .105$$

$$R^2_{Y1|2} = \frac{SSR(X_1|X_2)}{SSE(X_2)} = \frac{3.47}{113.42} = .031$$

We see that when X_2 is added to the regression model containing X_1 here, the error sum of squares $SSE(X_1)$ is reduced by 23.2 percent. The error sum of squares for the model containing both X_1 and X_2 is only reduced by another 10.5 percent when X_3 is added to the model. Finally, if the regression model already contains X_2, adding X_1 reduces $SSE(X_2)$ by only 3.1 percent.

Comments

1. The coefficients of partial determination can take on values between 0 and 1, as the definitions readily indicate.

2. A coefficient of partial determination can be interpreted as a coefficient of simple determination. Consider a multiple regression model with two X variables. Suppose we regress Y on X_2 and obtain the residuals:

$$e_i(Y|X_2) = Y_i - \hat{Y}_i(X_2)$$

where $\hat{Y}_i(X_2)$ denotes the fitted values of Y when X_2 is in the model. Suppose we further regress X_1 on X_2 and obtain the residuals:

$$e_i(X_1|X_2) = X_{i1} - \hat{X}_{i1}(X_2)$$

where $\hat{X}_{i1}(X_2)$ denotes the fitted values of X_1 in the regression of X_1 on X_2. The coefficient of simple determination R^2 between these two sets of residuals equals the coefficient of partial determination $R^2_{Y1|2}$. Thus, this coefficient measures the relation between Y and X_1 when both of these variables have been adjusted for their linear relationships to X_2.

3. The plot of the residuals $e_i(Y|X_2)$ against $e_i(X_1|X_2)$ provides a graphical representation of the strength of the relationship between Y and X_1, adjusted for X_2. Such plots of residuals, called *added variable plots* or *partial regression plots*, are discussed in Section 10.1. ■

Coefficients of Partial Correlation

The square root of a coefficient of partial determination is called a *coefficient of partial correlation*. It is given the same sign as that of the corresponding regression coefficient in the fitted regression function. Coefficients of partial correlation are frequently used in practice, although they do not have as clear a meaning as coefficients of partial determination. One use of partial correlation coefficients is in computer routines for finding the best predictor variable to be selected next for inclusion in the regression model. We discuss this use in Chapter 9.

Example

For the body fat example, we have:

$$r_{Y2|1} = \sqrt{.232} = .482$$
$$r_{Y3|12} = -\sqrt{.105} = -.324$$
$$r_{Y1|2} = \sqrt{.031} = .176$$

Note that the coefficients $r_{Y2|1}$ and $r_{Y1|2}$ are positive because we see from Table 7.2c that $b_2 = .6594$ and $b_1 = .2224$ are positive. Similarly, $r_{Y3|12}$ is negative because we see from Table 7.2d that $b_3 = -2.186$ is negative.

Comment

Coefficients of partial determination can be expressed in terms of simple or other partial correlation coefficients. For example:

$$R^2_{Y2|1} = [r_{Y2|1}]^2 = \frac{(r_{Y2} - r_{12}r_{Y1})^2}{\left(1 - r^2_{12}\right)\left(1 - r^2_{Y1}\right)} \tag{7.41}$$

$$R^2_{Y2|13} = [r_{Y2|13}]^2 = \frac{(r_{Y2|3} - r_{12|3}r_{Y1|3})^2}{\left(1 - r^2_{12|3}\right)\left(1 - r^2_{Y1|3}\right)} \tag{7.42}$$

where r_{Y1} denotes the coefficient of simple correlation between Y and X_1, r_{12} denotes the coefficient of simple correlation between X_1 and X_2, and so on. Extensions are straightforward. ■

7.5 Standardized Multiple Regression Model

A standardized form of the general multiple regression model (6.7) is employed to control roundoff errors in normal equations calculations and to permit comparisons of the estimated regression coefficients in common units.

Roundoff Errors in Normal Equations Calculations

The results from normal equations calculations can be sensitive to rounding of data in intermediate stages of calculations. When the number of X variables is small—say, three or less—roundoff effects can be controlled by carrying a sufficient number of digits in intermediate calculations. Indeed, most computer regression programs use double-precision arithmetic in all computations to control roundoff effects. Still, with a large number of X variables, serious roundoff effects can arise despite the use of many digits in intermediate calculations.

Roundoff errors tend to enter normal equations calculations primarily when the inverse of $\mathbf{X'X}$ is taken. Of course, any errors in $(\mathbf{X'X})^{-1}$ may be magnified in calculating $\mathbf{b}$ and other subsequent statistics. The danger of serious roundoff errors in $(\mathbf{X'X})^{-1}$ is particularly great when (1) $\mathbf{X'X}$ has a determinant that is close to zero and/or (2) the elements of $\mathbf{X'X}$ differ substantially in order of magnitude. The first condition arises when some or all of the X variables are highly intercorrelated. We shall discuss this situation in Section 7.6.

The second condition arises when the X variables have substantially different magnitudes so that the entries in the $\mathbf{X'X}$ matrix cover a wide range, say, from 15 to 49,000,000. A solution for this condition is to transform the variables and thereby reparameterize the regression model into the standardized regression model.

The transformation to obtain the standardized regression model, called the *correlation transformation,* makes all entries in the $\mathbf{X}'\mathbf{X}$ matrix for the transformed variables fall between -1 and 1 inclusive, so that the calculation of the inverse matrix becomes much less subject to roundoff errors due to dissimilar orders of magnitudes than with the original variables.

Comment

In order to avoid the computational difficulties inherent in inverting the $\mathbf{X}'\mathbf{X}$ matrix, many statistical packages use an entirely different computational approach that involves decomposing the $\mathbf{X}$ matrix into a product of several matrices with special properties. The $\mathbf{X}$ matrix is often first modified by centering each of the variables (i.e., using the deviations around the mean) to further improve computational accuracy. Information on decomposition strategies may be found in texts on statistical computing, such as Reference 7.1. ■

Lack of Comparability in Regression Coefficients

A second difficulty with the nonstandardized multiple regression model (6.7) is that ordinarily regression coefficients cannot be compared because of differences in the units involved. We cite two examples.

1. When considering the fitted response function:

$$\hat{Y} = 200 + 20{,}000X_1 + .2X_2$$

one may be tempted to conclude that X_1 is the only important predictor variable, and that X_2 has little effect on the response variable Y. A little reflection should make one wary of this conclusion. The reason is that we do not know the units involved. Suppose the units are:

Y in dollars
X_1 in thousand dollars
X_2 in cents

In that event, the effect on the mean response of a \$1,000 increase in X_1 (i.e., a 1-unit increase) when X_2 is constant would be an increase of \$20,000. This is exactly the same as the effect of a \$1,000 increase in X_2 (i.e., a 100,000-unit increase) when X_1 is constant, despite the difference in the regression coefficients.

2. In the Dwaine Studios example of Figure 6.5, we cannot make any comparison between b_1 and b_2 because X_1 is in units of thousand persons aged 16 or younger, whereas X_2 is in units of thousand dollars of per capita disposable income.

Correlation Transformation

Use of the correlation transformation helps with controlling roundoff errors and, by expressing the regression coefficients in the same units, may be of help when these coefficients are compared. We shall first describe the correlation transformation and then the resulting standardized regression model.

The correlation transformation is a simple modification of the usual standardization of a variable. Standardizing a variable, as in (A.37), involves centering and scaling the variable. *Centering* involves taking the difference between each observation and the mean of all observations for the variable; *scaling* involves expressing the centered observations in units of the standard deviation of the observations for the variable. Thus, the usual standardizations of the response variable Y and the predictor variables $X_1, \ldots, X_{p-1}$ are as follows:

$$\frac{Y_i - \bar{Y}}{s_Y} \tag{7.43a}$$

$$\frac{X_{ik} - \bar{X}_k}{s_k} \qquad (k = 1, \ldots, p-1) \tag{7.43b}$$

where $\bar{Y}$ and $\bar{X}_k$ are the respective means of the Y and the X_k observations, and s_Y and s_k are the respective standard deviations defined as follows:

$$s_Y = \sqrt{\frac{\sum_i (Y_i - \bar{Y})^2}{n-1}} \tag{7.43c}$$

$$s_k = \sqrt{\frac{\sum_i (X_{ik} - \bar{X}_k)^2}{n-1}} \qquad (k = 1, \ldots, p-1) \tag{7.43d}$$

The correlation transformation is a simple function of the standardized variables in (7.43a, b):

$$Y_i^* = \frac{1}{\sqrt{n-1}}\left(\frac{Y_i - \bar{Y}}{s_Y}\right) \tag{7.44a}$$

$$X_{ik}^* = \frac{1}{\sqrt{n-1}}\left(\frac{X_{ik} - \bar{X}_k}{s_k}\right) \qquad (k = 1, \ldots, p-1) \tag{7.44b}$$

Standardized Regression Model

The regression model with the transformed variables Y^* and X_k^* as defined by the correlation transformation in (7.44) is called a *standardized regression model* and is as follows:

$$Y_i^* = \beta_1^* X_{i1}^* + \cdots + \beta_{p-1}^* X_{i,p-1}^* + \varepsilon_i^* \tag{7.45}$$

The reason why there is no intercept parameter in the standardized regression model (7.45) is that the least squares or maximum likelihood calculations always would lead to an estimated intercept term of zero if an intercept parameter were present in the model.

It is easy to show that the parameters $\beta_1^*, \ldots, \beta_{p-1}^*$ in the standardized regression model and the original parameters $\beta_0, \beta_1, \ldots, \beta_{p-1}$ in the ordinary multiple regression model (6.7) are related as follows:

$$\beta_k = \left(\frac{s_Y}{s_k}\right)\beta_k^* \qquad (k = 1, \ldots, p-1) \tag{7.46a}$$

$$\beta_0 = \bar{Y} - \beta_1 \bar{X}_1 - \cdots - \beta_{p-1}\bar{X}_{p-1} \tag{7.46b}$$

We see that the standardized regression coefficients β_k^* and the original regression coefficients β_k $(k = 1, \ldots, p-1)$ are related by simple scaling factors involving ratios of standard deviations.

X′X Matrix for Transformed Variables

In order to be able to study the special nature of the **X′X** matrix and the least squares normal equations when the variables have been transformed by the correlation transformation, we need to decompose the correlation matrix in (6.67) containing all pairwise correlation coefficients among the response and predictor variables $Y, X_1, X_2, \ldots, X_{p-1}$ into two matrices.

1. The first matrix, denoted by $\mathbf{r}_{XX}$, is called the *correlation matrix of the X variables*. It has as its elements the coefficients of simple correlation between all pairs of the X variables. This matrix is defined as follows:

$$\underset{(p-1)\times(p-1)}{\mathbf{r}_{XX}} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1,p-1} \\ r_{21} & 1 & \cdots & r_{2,p-1} \\ \vdots & \vdots & & \vdots \\ r_{p-1,1} & r_{p-1,2} & \cdots & 1 \end{bmatrix} \tag{7.47}$$

Here, r_{12} again denotes the coefficient of simple correlation between X_1 and X_2, and so on. Note that the main diagonal consists of 1s because the coefficient of simple correlation between a variable and itself is 1. The correlation matrix $\mathbf{r}_{XX}$ is symmetric; remember that $r_{kk'} = r_{k'k}$. Because of the symmetry of this matrix, computer printouts frequently omit the lower or upper triangular block of elements.

2. The second matrix, denoted by $\mathbf{r}_{YX}$, is a vector containing the coefficients of simple correlation between the response variable Y and each of the X variables, denoted again by r_{Y1}, r_{Y2}, etc.:

$$\underset{(p-1)\times 1}{\mathbf{r}_{YX}} = \begin{bmatrix} r_{Y1} \\ r_{Y2} \\ \vdots \\ r_{Y,p-1} \end{bmatrix} \tag{7.48}$$

Now we are ready to consider the **X′X** matrix for the transformed variables in the standardized regression model (7.45). The **X** matrix here is:

$$\underset{n\times(p-1)}{\mathbf{X}} = \begin{bmatrix} X_{11}^* & \cdots & X_{1,p-1}^* \\ X_{21}^* & \cdots & X_{2,p-1}^* \\ \vdots & & \vdots \\ X_{n1}^* & \cdots & X_{n,p-1}^* \end{bmatrix} \tag{7.49}$$

Remember that the standardized regression model (7.45) does not contain an intercept term; hence, there is no column of 1s in the **X** matrix. It can be shown that the **X′X** matrix for the transformed variables is simply the correlation matrix of the X variables defined in (7.47):

$$\underset{(p-1)\times(p-1)}{\mathbf{X}'\mathbf{X}} = \mathbf{r}_{XX} \tag{7.50}$$

Since the **X′X** matrix for the transformed variables consists of coefficients of correlation between the X variables, all of its elements are between -1 and 1 and thus are of the same order of magnitude. As we pointed out earlier, this can be of great help in controlling roundoff errors when inverting the **X′X** matrix.

Comment

We illustrate that the $\mathbf{X'X}$ matrix for the transformed variables is the correlation matrix of the X variables by considering two entries in the matrix:

1. In the upper left corner of $\mathbf{X'X}$ we have:

$$\sum (X_{i1}^*)^2 = \sum \left(\frac{X_{i1} - \bar{X}_1}{\sqrt{n-1}\, s_1} \right)^2 = \frac{\sum (X_{i1} - \bar{X}_1)^2}{n-1} \div s_1^2 = 1$$

2. In the first row, second column of $\mathbf{X'X}$, we have:

$$\begin{aligned} \sum X_{i1}^* X_{i2}^* &= \sum \left(\frac{X_{i1} - \bar{X}_1}{\sqrt{n-1}\, s_1} \right)\left(\frac{X_{i2} - \bar{X}_2}{\sqrt{n-1}\, s_2} \right) \\ &= \frac{1}{n-1} \frac{\sum (X_{i1} - \bar{X}_1)(X_{i2} - \bar{X}_2)}{s_1 s_2} \\ &= \frac{\sum (X_{i1} - \bar{X}_1)(X_{i2} - \bar{X}_2)}{\left[\sum (X_{i1} - \bar{X}_1)^2 \sum (X_{i2} - \bar{X}_2)^2\right]^{1/2}} \end{aligned}$$

But this equals r_{12}, the coefficient of correlation between X_1 and X_2, by (2.84). ■

Estimated Standardized Regression Coefficients

The least squares normal equations (6.24) for the ordinary multiple regression model:

$$\mathbf{X'Xb} = \mathbf{X'Y}$$

and the least squares estimators (6.25):

$$\mathbf{b} = (\mathbf{X'X})^{-1}\mathbf{X'Y}$$

can be expressed simply for the transformed variables. It can be shown that for the transformed variables, $\mathbf{X'Y}$ becomes:

$$\underset{(p-1)\times 1}{\mathbf{X'Y}} = \mathbf{r}_{YX} \tag{7.51}$$

where $\mathbf{r}_{YX}$ is defined in (7.48) as the vector of the coefficients of simple correlation between Y and each X variable. It now follows from (7.50) and (7.51) that the least squares normal equations and estimators of the regression coefficients of the standardized regression model (7.45) are as follows:

$$\mathbf{r}_{XX}\mathbf{b} = \mathbf{r}_{YX} \tag{7.52a}$$

$$\mathbf{b} = \mathbf{r}_{XX}^{-1}\mathbf{r}_{YX} \tag{7.52b}$$

where:

$$\underset{(p-1)\times 1}{\mathbf{b}} = \begin{bmatrix} b_1^* \\ b_2^* \\ \vdots \\ b_{p-1}^* \end{bmatrix} \tag{7.52c}$$

The regression coefficients $b_1^*, \ldots, b_{p-1}^*$ are often called *standardized regression coefficients*.

The return to the estimated regression coefficients for regression model (6.7) in the original variables is accomplished by employing the relations:

$$b_k = \left(\frac{s_Y}{s_k}\right) b_k^* \qquad (k = 1, \ldots, p-1) \tag{7.53a}$$

$$b_0 = \bar{Y} - b_1\bar{X}_1 - \cdots - b_{p-1}\bar{X}_{p-1} \tag{7.53b}$$

Comment

When there are two X variables in the regression model, i.e., when $p-1=2$, we can readily see the algebraic form of the standardized regression coefficients. We have:

$$\mathbf{r}_{XX} = \begin{bmatrix} 1 & r_{12} \\ r_{12} & 1 \end{bmatrix} \tag{7.54a}$$

$$\mathbf{r}_{YX} = \begin{bmatrix} r_{Y1} \\ r_{Y2} \end{bmatrix} \tag{7.54b}$$

$$\mathbf{r}_{XX}^{-1} = \frac{1}{1-r_{12}^2}\begin{bmatrix} 1 & -r_{12} \\ -r_{12} & 1 \end{bmatrix} \tag{7.54c}$$

Hence, by (7.52b) we obtain:

$$\mathbf{b} = \frac{1}{1-r_{12}^2}\begin{bmatrix} 1 & -r_{12} \\ -r_{12} & 1 \end{bmatrix}\begin{bmatrix} r_{Y1} \\ r_{Y2} \end{bmatrix} = \frac{1}{1-r_{12}^2}\begin{bmatrix} r_{Y1} - r_{12}r_{Y2} \\ r_{Y2} - r_{12}r_{Y1} \end{bmatrix} \tag{7.55}$$

Thus:

$$b_1^* = \frac{r_{Y1} - r_{12}r_{Y2}}{1-r_{12}^2} \tag{7.55a}$$

$$b_2^* = \frac{r_{Y2} - r_{12}r_{Y1}}{1-r_{12}^2} \tag{7.55b}$$

■

Example

Table 7.5a repeats a portion of the original data for the Dwaine Studios example in Figure 6.5b, and Table 7.5b contains the data transformed according to the correlation transformation (7.44). We illustrate the calculation of the transformed data for the first case, using the means and standard deviations in Table 7.5a (differences in the last digit of the transformed data are due to rounding effects):

$$Y_1^* = \frac{1}{\sqrt{n-1}}\left(\frac{Y_1 - \bar{Y}}{s_Y}\right) = \frac{1}{\sqrt{21-1}}\left(\frac{174.4 - 181.90}{36.191}\right) = -.04634$$

$$X_{11}^* = \frac{1}{\sqrt{n-1}}\left(\frac{X_{11} - \bar{X}_1}{s_1}\right) = \frac{1}{\sqrt{21-1}}\left(\frac{68.5 - 62.019}{18.620}\right) = .07783$$

$$X_{12}^* = \frac{1}{\sqrt{n-1}}\left(\frac{X_{12} - \bar{X}_2}{s_2}\right) = \frac{1}{\sqrt{21-1}}\left(\frac{16.7 - 17.143}{.97035}\right) = -.10208$$

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 $\mathbf{X'X}$ 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.

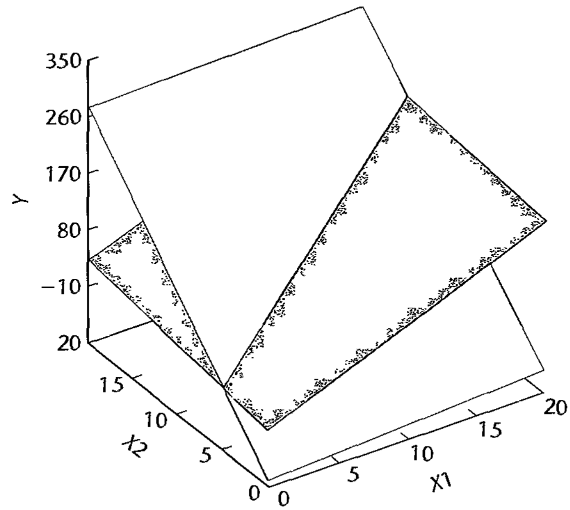
Case i	X_{i1}	X_{i2}	Y_i	Fitted Values for Regression Function (7.58)	(7.59)
1	2	6	23	23	23
2	8	9	83	83	83
3	6	8	63	63	63
4	10	10	103	103	103

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:

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

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:

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

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:

$$X_2 = 5 + .5X_1 \tag{7.60}$$

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.)

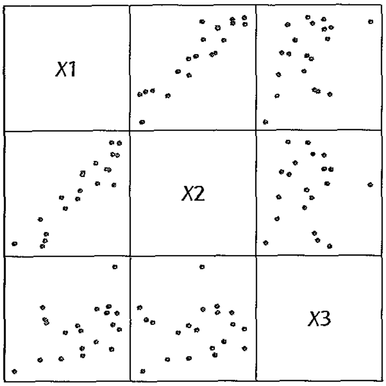
2. The counterpart in real life to the many different regression functions providing equally good fits to the data in our idealized example is that the estimated regression coefficients tend to have large sampling variability when the predictor variables are highly correlated. Thus, the estimated regression 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

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

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. ■

Need for More Powerful Diagnostics for Multicollinearity

As we have seen, multicollinearity among the predictor variables can have important consequences for interpreting and using a fitted regression model. The diagnostic tool considered here for identifying multicollinearity—namely, the pairwise coefficients of simple correlation between the predictor variables—is frequently helpful. Often, however, serious multicollinearity exists without being disclosed by the pairwise correlation coefficients. In Chapter 10, we present a more powerful tool for identifying the existence of serious multicollinearity. Some remedial measures for lessening the effects of multicollinearity will be considered in Chapter 11.

Cited Reference

7.1. Kennedy, W. J., Jr., and J. E. Gentle. *Statistical Computing*. New York: Marcel Dekker, 1980.

Problems

7.1. State the number of degrees of freedom that are associated with each of the following extra sums of squares: (1) $SSR(X_1|X_2)$; (2) $SSR(X_2|X_1, X_3)$; (3) $SSR(X_1, X_2|X_3, X_4)$; (4) $SSR(X_1, X_2, X_3|X_4, X_5)$.

7.2. Explain in what sense the regression sum of squares $SSR(X_1)$ is an extra sum of squares.

7.3. Refer to **Brand preference** Problem 6.5.

a. Obtain the analysis of variance table that decomposes the regression sum of squares into extra sums of squares associated with X_1 and with X_2, given X_1.

b. Test whether X_2 can be dropped from the regression model given that X_1 is retained. Use the F^* test statistic and level of significance .01. State the alternatives, decision rule, and conclusion. What is the P-value of the test?

*7.4. Refer to **Grocery retailer** Problem 6.9.

a. Obtain the analysis of variance table that decomposes the regression sum of squares into extra sums of squares associated with X_1; with X_3, given X_1; and with X_2, given X_1 and X_3.

b. Test whether X_2 can be dropped from the regression model given that X_1 and X_3 are retained. Use the F^* test statistic and $\alpha = .05$. State the alternatives, decision rule, and conclusion. What is the P-value of the test?

c. Does $SSR(X_1) + SSR(X_2|X_1)$ equal $SSR(X_2) + SSR(X_1|X_2)$ here? Must this always be the case?

*7.5. Refer to **Patient satisfaction** Problem 6.15.

a. Obtain the analysis of variance table that decomposes the regression sum of squares into extra sums of squares associated with X_2; with X_1, given X_2; and with X_3, given X_2 and X_1.

b. Test whether X_3 can be dropped from the regression model given that X_1 and X_2 are retained. Use the F^* test statistic and level of significance .025. State the alternatives, decision rule, and conclusion. What is the P-value of the test?

*7.6. Refer to **Patient satisfaction** Problem 6.15. Test whether both X_2 and X_3 can be dropped from the regression model given that X_1 is retained. Use $\alpha = .025$. State the alternatives, decision rule, and conclusion. What is the P-value of the test?

7.7. Refer to **Commercial properties** Problem 6.18.

a. Obtain the analysis of variance table that decomposes the regression sum of squares into extra sums of squares associated with X_4; with X_1, given X_4; with X_2, given X_1 and X_4; and with X_3, given X_1, X_2 and X_4.

b. Test whether X_3 can be dropped from the regression model given that X_1, X_2 and X_4 are retained. Use the F^* test statistic and level of significance .01. State the alternatives, decision rule, and conclusion. What is the P-value of the test?

7.8. Refer to **Commercial properties** Problems 6.18 and 7.7. Test whether both X_2 and X_3 can be dropped from the regression model given that X_1 and X_4 are retained; use $\alpha = .01$. State the alternatives, decision rule, and conclusion. What is the P-value of the test?

*7.9. Refer to **Patient satisfaction** Problem 6.15. Test whether $\beta_1 = -1.0$ and $\beta_2 = 0$; use $\alpha = .025$. State the alternatives, full and reduced models, decision rule, and conclusion.

7.10. Refer to **Commercial properties** Problem 6.18. Test whether $\beta_1 = -.1$ and $\beta_2 = .4$; use $\alpha = .01$. State the alternatives, full and reduced models, decision rule, and conclusion.

7.11. Refer to the work crew productivity example in Table 7.6.

a. Calculate R^2_{Y1}, R^2_{Y2}, R^2_{12}, $R^2_{Y1|2}$, $R^2_{Y2|1}$, and R^2. Explain what each coefficient measures and interpret your results.

b. Are any of the results obtained in part (a) special because the two predictor variables are uncorrelated?

7.12. Refer to **Brand preference** Problem 6.5. Calculate R^2_{Y1}, R^2_{Y2}, R^2_{12}, $R^2_{Y1|2}$, $R^2_{Y2|1}$, and R^2. Explain what each coefficient measures and interpret your results.

*7.13. Refer to **Grocery retailer** Problem 6.9. Calculate R^2_{Y1}, R^2_{Y2}, R^2_{12}, $R^2_{Y1|2}$, $R^2_{Y2|1}$, $R^2_{Y2|13}$, and R^2. Explain what each coefficient measures and interpret your results.

*7.14. Refer to **Patient satisfaction** Problem 6.15.

a. Calculate R^2_{Y1}, $R^2_{Y1|2}$, and $R^2_{Y1|23}$. How is the degree of marginal linear association between Y and X_1 affected, when adjusted for X_2? When adjusted for both X_2 and X_3?

b. Make a similar analysis to that in part (a) for the degree of marginal linear association between Y and X_2. Are your findings similar to those in part (a) for Y and X_1?

7.15. Refer to **Commercial properties** Problems 6.18 and 7.7. Calculate R^2_{Y4}, R^2_{Y1}, $R^2_{Y1|4}$, R^2_{14}, $R^2_{Y2|14}$, $R^2_{Y3|124}$, and R^2. Explain what each coefficient measures and interpret your results. How is the degree of marginal linear association between Y and X_1 affected, when adjusted for X_4?

7.16. Refer to **Brand preference** Problem 6.5.

a. Transform the variables by means of the correlation transformation (7.44) and fit the standardized regression model (7.45).

b. Interpret the standardized regression coefficient b_1^*.

c. Transform the estimated standardized regression coefficients by means of (7.53) back to the ones for the fitted regression model in the original variables. Verify that they are the same as the ones obtained in Problem 6.5b.

*7.17. Refer to **Grocery retailer** Problem 6.9.

a. Transform the variables by means of the correlation transformation (7.44) and fit the standardized regression model (7.45).

b. Calculate the coefficients of determination between all pairs of predictor variables. Is it meaningful here to consider the standardized regression coefficients to reflect the effect of one predictor variable when the others are held constant?

c. Transform the estimated standardized regression coefficients by means of (7.53) back to the ones for the fitted regression model in the original variables. Verify that they are the same as the ones obtained in Problem 6.10a.

*7.18. Refer to **Patient satisfaction** Problem 6.15.

a. Transform the variables by means of the correlation transformation (7.44) and fit the standardized regression model (7.45).

b. Calculate the coefficients of determination between all pairs of predictor variables. Do these indicate that it is meaningful here to consider the standardized regression coefficients as indicating the effect of one predictor variable when the others are held constant?

c. Transform the estimated standardized regression coefficients by means of (7.53) back to the ones for the fitted regression model in the original variables. Verify that they are the same as the ones obtained in Problem 6.15c.

7.19. Refer to **Commercial properties** Problem 6.18.

a. Transform the variables by means of the correlation transformation (7.44) and fit the standardized regression model (7.45).

b. Interpret the standardized regression coefficient b_2^*.

c. Transform the estimated standardized regression coefficients by means of (7.53) back to the ones for the fitted regression model in the original variables. Verify that they are the same as the ones obtained in Problem 6.18c.

7.20. A speaker stated in a workshop on applied regression analysis: "In business and the social sciences, some degree of multicollinearity in survey data is practically inevitable." Does this statement apply equally to experimental data?

7.21. Refer to the example of perfectly correlated predictor variables in Table 7.8.

a. Develop another response function, like response functions (7.58) and (7.59), that fits the data perfectly.

b. What is the intersection of the infinitely many response surfaces that fit the data perfectly?

7.22. The progress report of a research analyst to the supervisor stated: "All the estimated regression coefficients in our model with three predictor variables to predict sales are statistically significant. Our new preliminary model with seven predictor variables, which includes the three variables of our smaller model, is less satisfactory because only two of the seven regression coefficients are statistically significant. Yet in some initial trials the expanded model is giving more precise sales predictions than the smaller model. The reasons for this anomaly are now being investigated." Comment.

7.23. Two authors wrote as follows: "Our research utilized a multiple regression model. Two of the predictor variables important in our theory turned out to be highly correlated in our data set. This made it difficult to assess the individual effects of each of these variables separately. We retained both variables in our model, however, because the high coefficient of multiple determination makes this difficulty unimportant." Comment.

7.24. Refer to **Brand preference** Problem 6.5.

a. Fit first-order simple linear regression model (2.1) for relating brand liking (Y) to moisture content (X_1). State the fitted regression function.

b. Compare the estimated regression coefficient for moisture content obtained in part (a) with the corresponding coefficient obtained in Problem 6.5b. What do you find?

c. Does $SSR(X_1)$ equal $SSR(X_1|X_2)$ here? If not, is the difference substantial?

d. Refer to the correlation matrix obtained in Problem 6.5a. What bearing does this have on your findings in parts (b) and (c)?

*7.25. Refer to **Grocery retailer** Problem 6.9.

a. Fit first-order simple linear regression model (2.1) for relating total hours required to handle shipment (Y) to total number of cases shipped (X_1). State the fitted regression function.

b. Compare the estimated regression coefficient for total cases shipped obtained in part (a) with the corresponding coefficient obtained in Problem 6.10a. What do you find?

c. Does $SSR(X_1)$ equal $SSR(X_1|X_2)$ here? If not, is the difference substantial?

d. Refer to the correlation matrix obtained in Problem 6.9c. What bearing does this have on your findings in parts (b) and (c)?

*7.26. Refer to **Patient satisfaction** Problem 6.15.

a. Fit first-order linear regression model (6.1) for relating patient satisfaction (Y) to patient's age (X_1) and severity of illness (X_2). State the fitted regression function.

b. Compare the estimated regression coefficients for patient's age and severity of illness obtained in part (a) with the corresponding coefficients obtained in Problem 6.15c. What do you find?

c. Does $SSR(X_1)$ equal $SSR(X_1|X_3)$ here? Does $SSR(X_2)$ equal $SSR(X_2|X_3)$?

d. Refer to the correlation matrix obtained in Problem 6.15b. What bearing does it have on your findings in parts (b) and (c)?

7.27. Refer to **Commercial properties** Problem 6.18.

a. Fit first-order linear regression model (6.1) for relating rental rates (Y) to property age (X_1) and size (X_4). State the fitted regression function.

b. Compare the estimated regression coefficients for property age and size with the corresponding coefficients obtained in Problem 6.18c. What do you find?

c. Does $SSR(X_4)$ equal $SSR(X_4|X_3)$ here? Does $SSR(X_1)$ equal $SSR(X_1|X_3)$?

d. Refer to the correlation matrix obtained in Problem 6.18b. What bearing does this have on your findings in parts (b) and (c)?

Exercises

7.28. a. Define each of the following extra sums of squares: (1) $SSR(X_5|X_1)$; (2) $SSR(X_3, X_4|X_1)$; (3) $SSR(X_4|X_1, X_2, X_3)$.

b. For a multiple regression model with five X variables, what is the relevant extra sum of squares for testing whether or not $\beta_5 = 0$? whether or not $\beta_2 = \beta_4 = 0$?

7.29. Show that:

a. $SSR(X_1, X_2, X_3, X_4) = SSR(X_1) + SSR(X_2, X_3|X_1) + SSR(X_4|X_1, X_2, X_3)$.

b. $SSR(X_1, X_2, X_3, X_4) = SSR(X_2, X_3) + SSR(X_1|X_2, X_3) + SSR(X_4|X_1, X_2, X_3)$.

7.30. Refer to **Brand preference** Problem 6.5.

a. Regress Y on X_2 using simple linear regression model (2.1) and obtain the residuals.

b. Regress X_1 on X_2 using simple linear regression model (2.1) and obtain the residuals.

c. Calculate the coefficient of simple correlation between the two sets of residuals and show that it equals $r_{Y1|2}$.

7.31. The following regression model is being considered in a water resources study:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \beta_4 \sqrt{X_{i3}} + \varepsilon_i$$

State the reduced models for testing whether or not: (1) $\beta_3 = \beta_4 = 0$, (2) $\beta_3 = 0$, (3) $\beta_1 = \beta_2 = 5$, (4) $\beta_4 = 7$.

7.32. The following regression model is being considered in a market research study:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1}^2 + \varepsilon_i$$

State the reduced models for testing whether or not: (1) $\beta_1 = \beta_3 = 0$, (2) $\beta_0 = 0$, (3) $\beta_3 = 5$, (4) $\beta_0 = 10$, (5) $\beta_1 = \beta_2$.

7.33. Show the equivalence of the expressions in (7.36) and (7.41) for $R^2_{Y2|1}$.

7.34. Refer to the work crew productivity example in Table 7.6.

a. For the variables transformed according to (7.44), obtain: (1) $\mathbf{X'X}$, (2) $\mathbf{X'Y}$, (3) $\mathbf{b}$, (4) $\mathbf{s}^2\{\mathbf{b}\}$.

b. Show that the standardized regression coefficients obtained in part (a3) are related to the regression coefficients for the regression model in the original variables according to (7.53).

7.35. Derive the relations between the β_k and β_k^* in (7.46a) for $p - 1 = 2$.

7.36. Derive the expression for $\mathbf{X'Y}$ in (7.51) for standardized regression model (7.30.) for $p-1 = 2$.

Projects

7.37. Refer to the **CDI** data set in Appendix C.2. For predicting the number of active physicians (Y) in a county, it has been decided to include total population (X_1) and total personal income (X_2) as predictor variables. The question now is whether an additional predictor variable would be helpful in the model and, if so, which variable would be most helpful. Assume that a first-order multiple regression model is appropriate.

a. For each of the following variables, calculate the coefficient of partial determination given that X_1 and X_2 are included in the model: land area (X_3), percent of population 65 or older (X_4), number of hospital beds (X_5), and total serious crimes (X_6).

b. On the basis of the results in part (a), which of the four additional predictor variables is best? Is the extra sum of squares associated with this variable larger than those for the other three variables?

c. Using the F^* test statistic, test whether or not the variable determined to be best in part (*b*) is helpful in the regression model when X_1 and X_2 are included in the model; use $\alpha = .01$. State the alternatives, decision rule, and conclusion. Would the F^* test statistics for the other three potential predictor variables be as large as the one here? Discuss.

7.38. Refer to the **SENIC** data set in Appendix C.1. For predicting the average length of stay of patients in a hospital (Y), it has been decided to include age (X_1) and infection risk (X_2) as predictor variables. The question now is whether an additional predictor variable would be helpful in the model and, if so, which variable would be most helpful. Assume that a first-order multiple regression model is appropriate.

a. For each of the following variables, calculate the coefficient of partial determination given that X_1 and X_2 are included in the model: routine culturing ratio (X_3), average daily census (X_4), number of nurses (X_5), and available facilities and services (X_6).

b. On the basis of the results in part (a), which of the four additional predictor variables is best? Is the extra sum of squares associated with this variable larger than those for the other three variables?

c. Using the F^* test statistic, test whether or not the variable determined to be best in part (*b*) is helpful in the regression model when X_1 and X_2 are included in the model; use $\alpha = .05$. State the alternatives, decision rule, and conclusion. Would the F^* test statistics for the other three potential predictor variables be as large as the one here? Discuss.