Regression Analysis (I)

Kutner's Applied Linear Statistical Models (5/E)

Chapter 7: Multiple Regression (II)

Thursday 09:10-12:00, 商館 260205

Han-Ming Wu

Department of Statistics, National Chengchi University

http://www.hmwu.idv.tw

Overview

1. Some specialized topics that are unique to multiple regression: (1) extra sums of squares, (2) the standardized version of the multiple regression model, and (3) multicollinearity.

7.1 Extra Sums of Squares

Basic Ideas

1. An extra sum of squares measures the _____ in the _____ _____ when one or several predictor variables are added to the regression model, given that other predictor variables are already in the model.

2. Equivalently, one can view an extra sum of squares as measuring the _____ _____ in the _____ sum of squares when one or several predictor variables are added to the regression model.

3. Example (Table 7.1) 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)(三頭肌皮下脂肪厚度), thigh circumference (X_2)(大腿圍), and midarm circumference (X_3) (中臂圍).

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

4. *Background and goal*: 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.

5. (Table 7.2) Conduct four regression results when body fat (Y) is regressed 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. The total sum of squares is _____.

   (a) (Table 7.2a) The regression sum of squares when X_1, only is in the model is, _____. The error sum of squares for this model is _____.

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

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

(b) (Table 7.2c) When X_1 and X_2 are in the regression model, the regression sum of squares is _____ and the error sum of squares is _____.

(c) Notice that the error sum of squares when X_1 and X_2 are in the model, _____, is smaller than when the model contains only X_1, _____.

(d) The difference is called an _____ and will be denoted by _____:

$$\begin{aligned} SSR(X_2|X_1) &= \_\_\_\_\_ \\ &= 385.44 - 352.27 = 33.17 \\ &= \_\_\_\_\_ \\ &= \_\_\_\_\_ \\ &= 143.12 - 109.95 = 33.17 \end{aligned}$$

This _____ in the error sum of squares is the result of _____ to the regression model when _____, is already included in the model.

(e) Thus, the extra sum of squares $SSR(X_2|X_1)$ measures the _____ (additional or extra reduction) of adding X_2 to the regression model when X_1, is already in the model.

(f) The reason for the equivalence of the _____ in the error sum of squares and the _____ in the regression sum of squares is the basic analysis of variance identity:

_____

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

6. (Tables 7.2c, 7.2d) 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.

_____ = _____ = 109.95−98.41 = 11.54

or, equivalently:

_____ = _____ = 396.98−385.44 = 11.54.

7. (table 7.2a, 7.2d) 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.

_____ = _____ = 143.12−98.41 = 44.71

or, equivalently:

_____ = _____ = 396.98−352.27 = 44.71

Definitions

1. An extra sum of squares always involves the _____ between the _____ _____ 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 _____ X variable(s) and the _____ X variable(s).

2. Equivalently, an extra sum of squares involves the difference between the two corresponding _____.

3. Thus, we define:

$$SSR(X_1|X_2) = \_\_\_\_\_ \qquad (7.1a)$$

or, equivalently:

$$SSR(X_1|X_2) = \_\_\_\_\_ \qquad (7.1b)$$

4. If X_2 is the extra variable, We define:

$$SSR(X_2|X_1) = \_\_\_\_\_ \qquad (7.2a)$$

or, equivalently:

$$SSR(X_2|X_1) = \_\_\_\_\_ \qquad (7.2b)$$

5. Extensions for three or more variables are straightforward:

$$SSR(X_3|X_1, X_2) = \_\_\_\_\_ \qquad (7.3a)$$

or:

$$SSR(X_3|X_1, X_2) = \_\_\_\_\_ \qquad (7.4b)$$

and

$$SSR(X_2, X_3|X_1) = \_\_\_\_\_ \qquad (7.4a)$$

or:

$$SSR(X_2, X_3|X_1) = \_\_\_\_\_ \qquad (7.4b)$$

Decomposition of SSR into Extra Sums of Squares

1. In multiple regression, we can obtain a _____ of decompositions of SSR into _____ sums of squares.

2. Consider the multiple regression model with two X variables:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i,\ i = 1, \cdot, n$$

3. Begin with the identity for X_1:

_____ (7.5)

when X_1 is the X variable in th model. Replacing $SSE(X_1)$ by its equivalent in (7.2a): _____, we obtain:

$$SSTO = \text{\_\_\_\_\_} \quad (7.6)$$

4. Use the same identity for multiple regression with two X variables as in (7.5) for a single X variable:

$$SSTO = \text{\_\_\_\_\_} \quad (7.7)$$

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

_____ (7.8)

5. We have decomposed $SSR(X_1, X_2)$ into two marginal components:

(a) _____: measuring the contribution by including X_1 alone in the model.

(b) _____: measuring the additional contribution when X_2 is included, given that X_1 is already in the model.

6. The order of the X variables is arbitrary:

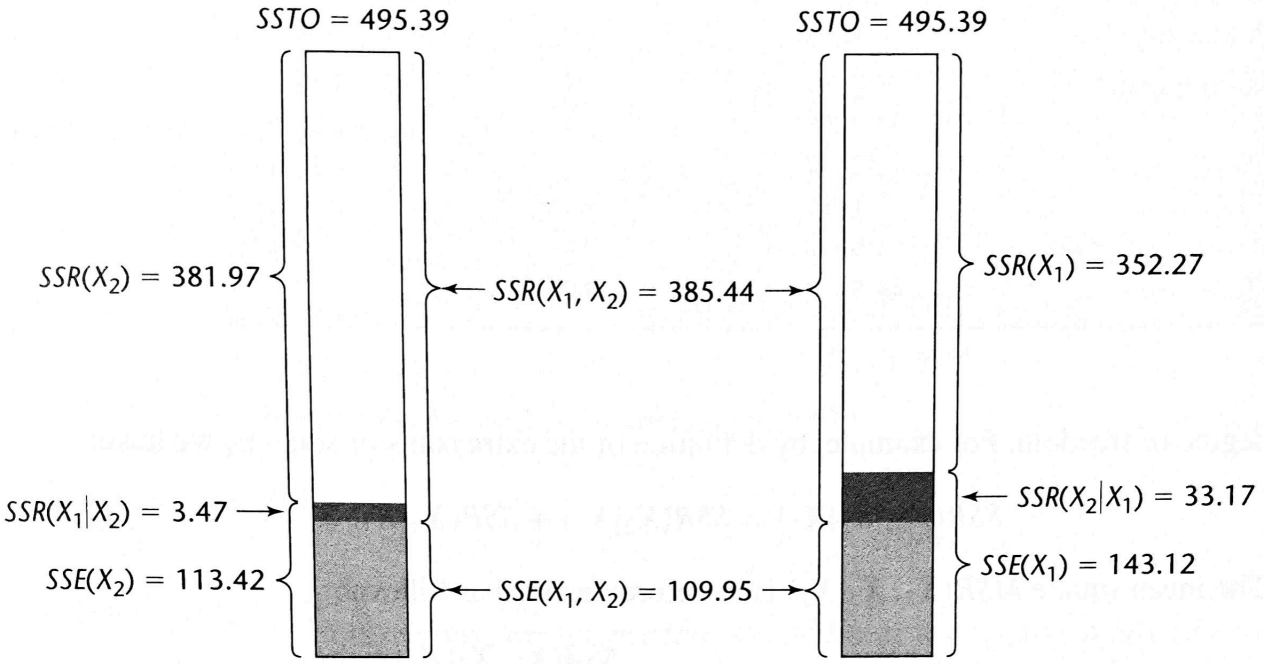
$$SSR(X_1, X_2) = \text{\_\_\_\_\_} \quad (7.9)$$

7. Example Body Fat Example

(a) A sample of $n = 20$ healthy females $25 - 34$ years old; Y: amount of body fat; X_1: triceps skinfold thickness; X_2: thigh circumference; X_3: midarm circumference.

(b) (Figure 7.1): The extra sum of squares can be viewed either as a _____ or as an _____ when the second predictor variable is added to the regression model.

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



8. 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:

$$SSR(X_1, X_2, X_3) = \_\_\_\_\_ \quad (7.10a)$$

$$SSR(X_1, X_2, X_3) = \_\_\_\_\_ \quad (7.10b)$$

$$SSR(X_1, X_2, X_3) = \_\_\_\_\_ \quad (7.10e)$$

9. The number of possible decompositions becomes _____ as the number of X variables in the regression model _____.

ANOVA Table Containing Decomposition of SSR

1. (Table 7.3, 7.4) ANOVA tables can be constructed containing decompositions of the regression sum of squares into extra sums of squares.

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	$SSTO$	$n-1$	

2. Note that each extra sum of squares involving a _____ has associated with it _____ degree of freedom.

3. 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: an extra sum of squares as a sum of two extra sums of squares, each associated with _____ degree of freedom.

4. Many computer regression packages provide decompositions of SSR into _____-degree-of-freedom extra sums of squares, usually in the order in which the X variables are _____.

5. 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:

_____

6. 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 $SSR(X_2, X_3|X_1)$:

$$SSR(X_2, X_3|X_1) = \_\_\_\_\_.$$

7. The reason why extra sums of squares are of interest is that they occur in a variety of _____ about _____ where the question of concern is whether certain X variables can be dropped from the regression mode1.

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

Test whether a Single $\beta_k = 0$

1. Test whether the term $\beta_k X_k$ can be dropped from a multiple regression model,

   $$H_0: \_\_\_\_\_ \quad H_a: \_\_\_\_\_,$$

   the test statistic: _____ is appropriate for this test.

2. Use _____: consider the first-order regression model with three predictor variables:

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

   To test the alternatives:

   $$H_0: \beta_3 = 0 \quad H_a: \beta_3 \neq 0. \qquad (7.13)$$

3. The error sum of squares $SSE(F)$ for the full model:

   $$SSE(F) = \_\_\_\_\_, \qquad df_F = n - 4.$$

4. (*Reduced Model*) The reduced model when H_0 in (7.13) holds:

   _____ Reduced model (7.14)

   The error sum of squares $SSE(E)$ for the reduced model:

   $$SSE(R) = \_\_\_\_\_, \qquad df_R = n - 3.$$

5. The general linear test statistic:

   $$F^* = \_\_\_\_\_$$
   $$= \_\_\_\_\_$$
   $$= \_\_\_\_\_$$
   $$= \_\_\_\_\_ \qquad (7.15)$$

6. The test whether or not $\beta_3 = 0$ is a _____, given that X_1 and X_2 are already in the model.

7. 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 _____ can provide a fit of the full model and the appropriate extra sum of squares.

8. Example Body Fat Example

   (a) To test for the model with all three predictor variables whether midarm circumference (X_3) can be dropped from the model.

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	

   (b) (Table 7.4) ANOVA results 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) is:

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

$$= \text{\_\_\_\_\_}$$

   For $\alpha = 0.01$, we require _____. Since _____, we conclude _____, that X_3 can be dropped from the regression model that already contains X_1 and X_2.

   (c) (Table 7.2d) the t^* test statistic:

$$t^* = \text{\_\_\_\_\_}$$

   Since _____, we see that the two test statistics are _____, just as for simple linear regression.

9. The F^* test statistic (7.15) to test whether or not $\beta_3 = 0$ is called a _____ _____ 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 _____.

Test whether Several $\beta_k = 0$

1. 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:

$$H_0 : \_\_\_\_\_ \quad H_a : \text{ not both } \beta_2 \text{ and } \beta_3 \text{ equal zero} \qquad (7.16)$$

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

_____ Reduced model (7.17)

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

$$SSE(R) = \_\_\_\_\_ \qquad df_R = \_\_\_\_\_$$

3. The general linear test statistic:

$$F^* = \_\_\_\_\_$$

$$= \_\_\_\_\_$$

$$= \_\_\_\_\_$$

4. Example Body Fat Example

   (a) 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):

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

   (b) Test statistic (7.18) therefore:

   $F^* =$ _____ $=$ _____

(c) For $\alpha = 0.05$, we require _____. Since $F^* = 3.63$ is at the _____ of the decision rule (the P-value of the test statistic is _____), we may wish to make _____ before deciding whether X_2 and X_3 should be dropped from the regression model that already contains X_1.

7.3 Summary of Tests Concerning Regression Coefficients*

7.4 Coefficients of Partial Determination

1. Extra sums of squares are not only useful for _____ on the regression coefficients of a multiple regression model, but they are also encountered in descriptive measures of relationship called _____.

2. Recall: the coefficient of multiple determination, R^2, measures the _____ _____ in the variation of Y achieved by the introduction of the _____ of X variables considered in the model.

3. A coefficient of _____ determination measures the _____ of one X variable when all others are already included in the model.

Two Predictor Variables

1. Consider a first-order multiple regression model with two predictor variables:

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

(a) _____: measures the variation in Y when X_2 is included in the mode1.

(b) _____ measures the variation in Y when both X_1 and X_2 are included in the model.

2. (Recall) Coefficient of determination: _____.

3. The relative marginal reduction in the variation in Y associated with X_1 when X_2 is already in the model is:

$$R^2_{Y1|2} = \underline{\qquad\qquad} = \underline{\qquad\qquad}$$

This measure is the _____ between Y and X_1, given that X_2 is in the mode1. Denoted by _____.

4. $R^2_{Y1|2}$ measures the _____ in the variation in Y remaining after X_2 is included in the model that is _____ by also including X_1 in the model.

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

_____

General Case

1. The generalization of coefficients of partial determination to three or more X variables in the model:

$$R^2_{Y1|23} = \underline{\qquad\qquad} \tag{7.37}$$

$$R^2_{Y2|13} = \underline{\qquad\qquad} \tag{7.38}$$

$$\underline{\qquad\qquad} = \frac{SSR(X_3|X_1, X_2)}{SSE(X_1, X_2)} \tag{7.39}$$

$$\underline{\qquad\qquad} = \frac{SSR(X_4|X_1, X_2, X_3)}{SSE(X_1, X_2, X_3)} \tag{7.40}$$

2. Example Body Fat Example

   (a) Example: we can obtain a variety of coefficients of partial determination. (Tables 7.2 and 7.4):

$$R^2_{Y2|1} = \underline{\qquad\qquad}$$

$$R^2_{Y3|12} = \underline{\qquad\qquad\qquad\qquad\qquad\qquad}$$

$$R^2_{Y1|2} = \underline{\qquad\qquad\qquad\qquad\qquad\qquad}$$

(b) When X_2 is added to the regression model containing X_1, the _____ sum of squares _____ is reduced by _____.

(c) SSE for the model containing both X_1 and X_2 is only reduced by another _____ percent when $X_3$is added to the model.

(d) If the regression model already contains X_2, adding X_1 reduces _____ by only _____.

Coefficients of Partial Correlation

1. The _____ of a coefficient of partial determination is called a _____ _____.

2. One use of partial correlation coefficients is in computer routines for finding the _____ to be selected next for inclusion in the regression model.

3. For the body fat example, we have:

$$\begin{aligned} r_{Y2|1} &= \sqrt{0.232} = 0.482 \\ r_{Y3|12} &= \sqrt{0.105} = -0.324 \\ r_{Y1|2} &= \sqrt{0.031} = 0.176 \end{aligned}$$

4. The coefficients $r_{Y2|1}$ and $r_{Y1|2}$ are positive because we see from Table 7.2c that $b_2 = 0.6594$ and $b_1 = 0.2224$ are _____. Similarly, $r_{Y3|12}$ is negative because we see from Table 7.2d that $b_3 = -2.186$ is _____.

7.5 Standardized Multiple Regression Model*

7.6 Multicollinearity and Its Effects

1. In multiple regression analysis, some questions frequently asked:

(a) What is the _____ of the effects of the different predictor variables?

(b) What is the _____ of the effect of a given predictor variable on the response variable?

(c) Can any predictor variable be _____ from the model because it has little or no effect on the response variable?

(d) Should any predictor variables not yet included in the model be considered for _____?

2. In many nonexperimental situations in business, economics, and the social and biological sciences, the _____ tend to be _____ among themselves and _____ that are related to the response variable but are not included in the model.

3. 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 _____ 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.

4. When the predictor variables are correlated among themselves, _____ or _____ among them is said to exist.

Uncorrelated Predictor Variables

1. (Table 7.6) The 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 (_____).

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

2. (Table 7.7a (7.7b) (7.7c)) The fitted regression function and the analysis of variance table when both X_1 and X_2 are (only (X_1) (X_2) is) included in the model.

3. (Table 7.7) The regression coefficient for X_1, _____, is the _____ whether only X_1 is included in the model or both predictor variables are included. The same holds for _____.

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	

4. When the predictor variables are _____, the effects ascribed to them by a first-order regression model are the _____ no matter which other of these predictor variables are included in the model.

5. 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:

$$\begin{aligned} SSR(X_1|X_2) &= \_\_\_\_\_ \\ &= \_\_\_\_\_ \\ SSR(X_1) &= \_\_\_\_\_ \end{aligned}$$

6. 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:

$$\begin{aligned} SSR(X_2|X_1) &= \underline{\qquad\qquad\qquad\qquad} \\ &= \underline{\qquad\qquad\qquad\qquad} \\ SSR(X_2) &= \underline{\qquad\qquad} \end{aligned}$$

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

8. See **Comment** on page 281 for the proof: 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

1. (Table 7.8) Example The data refer to four sample observations on a response variable and two predictor variables. The first-order multiple regression function fit:

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2.$$

TABLE 7.8 Example of Perfectly Correlated Predictor Variables.

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

Response Functions:
$\hat{Y} = -87 + X_1 + 18X_2$ (7.58)
$\hat{Y} = -7 + 9X_1 + 2X_2$ (7.59)

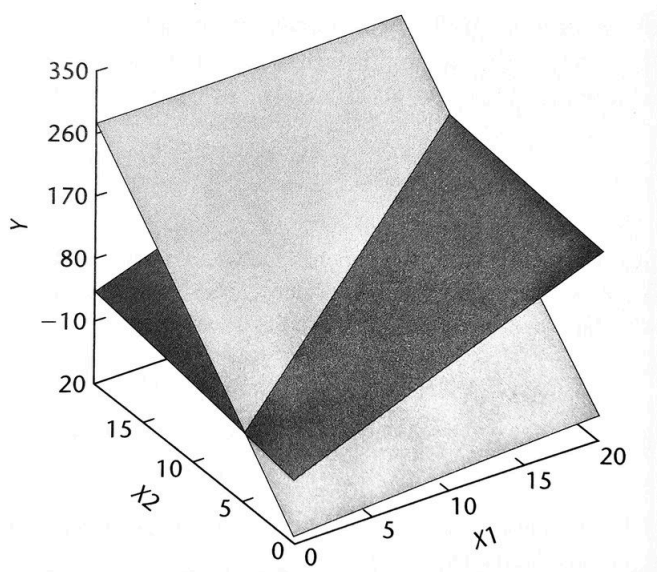
$$\text{Mr. A} : \hat{Y} = -87 + X_1 + 18X_2 \quad \text{(perfect fit)} \quad (7.58)$$
$$\text{Mr. B} : \hat{Y} = -7 + 9X_1 + 2X_2 \quad \text{(perfect fit)} \quad (7.59)$$

2. It can be shown that _____ will fit the data in Table 7.8 perfectly. The reason is that the predictor variables X_1, and X_2 are perfectly related:

$$X_2 = 5 + 0.5X_1 \qquad (7.60)$$

3. (Figure 7.2) The fitted response functions (7.58) and (7.59) are entirely different response surfaces. The two response surfaces have _____ only when they _____.

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



4. Two key implications of this example are:

   (a) The perfect relation between X_1, and X_2 did not inhibit our ability to obtain a _____ to the data.

   (b) Since many different response functions provide the same good fit, we cannot _____ anyone set of _____ as reflecting the effects of the different predictor variables.

Effects of Multicollinearity

1. The fact that some or all predictor variables are correlated among themselves (a) does not, in general, inhibit our ability to obtain a _____ (b) nor does it tend

to affect _____ or _____,
provided these inferences are made within the region of observations.

2. The estimated _____ tend to have _____ 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.

3. Many of the estimated regression coefficients individually may be _____ _____ even though a definite statistical relation exists between the response variable and the set of predictor variables.

4. The common _____ 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 _____ when multicollinearity exists.

5. Example The Body Fat Example

   (a) (Table 7.1): A sample of 20 healthy females $25-34$ years old, Y: amount of body fat, X_1: triceps skinfold thickness, X_2: thigh circumference, X_3: midarm circumference. (Table 7.2): The regression results for different fitted models.

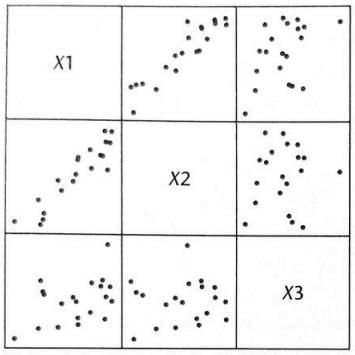
   (b) (Figure 7.3) The scatter plot matrix and the _____ matrix of the predictor variables: predictor variables X_1 and X_2 are highly correlated _____.

   (c) $r_{13} = 0.458$ and $r_{23} = 0.085$.

   (d) The _____ when X_3 is regressed on X_1 and X_2 is 0.998: X_3 is highly correlated with X_1 and X_2 together.

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

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

6. **Effects on Regression Coefficients**.

   (a) The regression coefficient for X_1, triceps skinfold thickness, _____ 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

   (b) The story is the same for the regression coefficient for X_2. The regression coefficient b_2 even _____ when X_3 is added to the model that includes X_1 and X_2.

   (c) *Important conclusion*: When predictor variables are correlated, the regression coefficient of anyone variable _____ 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 _____ or _____ effect, given whatever other correlated predictor variables are included in the model.

7. **Effects on Extra Sums of Squares**.

   (a) When predictor variables are correlated, the marginal contribution of anyone predictor variable in reducing the error sum of squares _____, depending on which other variables are already in the regression model, just as for regression coefficients.

   (b) (Table 7.2) Consider the following extra sums of squares for X_1:

$$SSR(X_1) = 352.27 \qquad SSR(X_1|X_2) = 3.47.$$

   The reason why $SSR(X_1|X_2)$ is so small compared with $SSR(X_1)$ is that X_1 and X_2 are _____ with each other and with the response variable.

   (c) When X_2 is already in the regression model, the marginal contribution of X_1 in reducing the error sum of squares is _____ because X_2 contains much of the _____ as X_1.

(d) The same story is found in Table 7.2 for X_2. Here $SSR(X_2|X_1) =$ _____, which is much smaller than $SSR(X_2) =$ _____.

(e) *Important conclusion*: When predictor variables are correlated, there is _____ _____ sum of squares that can be ascribed to anyone 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 _____ variables already included in the model.

8. **Effects on** $s(b_k)$.

(a) (Table 7.2 for the body fat example) how much more _____ 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

(b) The _____ of multicollinearity among the predictor variables is responsible for the _____ of the estimated regression coefficients.

9. **Effects on Fitted Values and Predictions**.

(a) (Table 7.2 for the body fat example) the high multicollinearity among the predictor variables _____, measuring the variability of the error terms, from being _____ 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

(b) The _____ within the range of the observations on the predictor variables is _____ with the addition of correlated predictor variables into the regression model.

(c) Example 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):

$$\hat{Y}_h = 19.93, \quad s(\hat{Y}_h) = 0.632$$

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 $X_{h2} = 50.0$:

$$\hat{Y}_h = 19.36 \quad s(\hat{Y}_h) = 0.624$$

Thus, the _____ is equally good as before, despite the addition of the second predictor variable that is highly correlated with the first one.

(d) The essential reason for the _____ is that the _____ _____ is negative, which plays a strong _____ influence to the increase in $s^2(b_1)$, in determining the value of $s^2(\hat{Y}_h)$ as given in (6.79).

$$s^2\{\hat{Y}_h\} = s^2\{b_0\} + X_{h1}^2 s^2\{b_1\} + X_{h2}^2 s^2\{b_2\} + 2X_{h1} s\{b_0, b_1\} + 2X_{h2} s\{b_0, b_2\} + 2X_{h1} X_{h2} s\{b_1, b_2\} \quad (6.79)$$

10. **Effects on Simultaneous Tests of β_k.** Paradox of t-test and F-test:

(a) (The Body Fat Example) test whether _____ and _____. Controlling the family level of significance at 0.05, we require with the _____ that each of the two t tests be conducted with level of significance _____.

(b) Hence, we need _____. Since both t^* statistics in Table 7.2c have absolute values that do not exceed 2.46, we would conclude from the two _____ tests that $\beta_1 = 0$ and that $\beta_2 = 0$.

(c) (Table 7.2c) Yet the proper F test for _____ would lead to the _____ that not both coefficients equal zero. We find $F^* = MSR/MSE = 192.72/6.47 = 29.8$, which far exceeds $F_{(0.95;2,17)} = 3.59$.

(d) The reason for this apparently paradoxical result is that each _____ is a _____, as we have seen in (7.15) from the perspective of the general linear test approach.

(e) Thus, a _____ 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 $\beta_1 = 0$.

(f) Similarly, we are led to conclude $\beta_2 = 0$ here because _____ is small, indicating that X_2 does not provide much more additional information when X_1 is already in the model.

(g) But the two tests of the marginal effects of _____ are not equivalent to testing whether there is a regression relation between Y and the two predictor variables.

(h) The reason is that the reduced model for each of the separate tests contains the _____, whereas the reduced model for testing whether _____ $\beta_1 = 0$ and $\beta_2 = 0$ would contain _____ predictor variable. The proper F test shows that there is a definite regression relation here between Y and X_1 and X_2.

Need for More Powerful Diagnostics for Multicollinearity

1. The diagnostic tool for identifying multicollinearity: the pairwise _____ _____ between the predictor variables is frequently helpful.
2. (Chapter 10) more powerful tool for identifying the existence of serious multicollinearity.
3. (Chapter 11) Some remedial measures for lessening the effects of multicollinearity.

😍 TA Class

- **Problems**: 7.2, 7.3, 7.6, 7.11, 7.24.
- **Exercises**: 7.31