CHAPTER 7

# Multiple Regression Analysis

## 7.1 THE THREE-VARIABLE LINEAR MODEL

*Multiple regression analysis* is used for testing hypotheses about the relationship between a dependent variable $Y$ and two or more independent variables $X$ and for prediction. The three-variable linear regression model can be written as

$$Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + u_i \tag{7.1}$$

The additional assumption (to those of the simple regression model) is that there is no exact linear relationship between the $X$ values.

Ordinary least-squares (OLS) parameter estimates for Eq. (7.1) can be obtained by minimizing the sum of the squared residuals:

$$\sum e_i^2 = \sum (Y_i - \hat{Y}_i)^2 = \sum (Y_i - \hat{b}_0 - \hat{b}_1 X_{1i} - \hat{b}_2 X_{2i})^2$$

This gives the following three normal equations (see Prob. 7.2):

$$\sum Y_i = n\hat{b}_0 + \hat{b}_1 \sum X_{1i} + \hat{b}_2 \sum X_{2i} \tag{7.2}$$

$$\sum X_{1i} Y_i = \hat{b}_0 \sum X_{1i} + \hat{b}_1 \sum X_{1i}^2 + \hat{b}_2 \sum X_{1i} X_{2i} \tag{7.3}$$

$$\sum X_{2i} Y_i = \hat{b}_0 \sum X_{2i} + \hat{b}_1 \sum X_{1i} X_{2i} + \hat{b}_2 \sum X_{2i}^2 \tag{7.4}$$

which (when expressed in deviation form) can be solved simultaneously for $\hat{b}_1$ and $\hat{b}_2$, giving (see Prob. 7.3)

$$\hat{b}_1 = \frac{(\sum x_1 y)(\sum x_2^2) - (\sum x_2 y)(\sum x_1 x_2)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2} \tag{7.5}$$

$$\hat{b}_2 = \frac{(\sum x_2 y)(\sum x_1^2) - (\sum x_1 y)(\sum x_1 x_2)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2} \tag{7.6}$$

Then

$$\hat{b}_0 = \bar{Y} - \hat{b}_1 \bar{X}_1 - \hat{b}_2 \bar{X}_2 \tag{7.7}$$

Estimator $\hat{b}_1$ measures the change in $Y$ for a unit change in $X_1$ while holding $X_2$ constant. $\hat{b}_2$ is analogously defined. Estimators $\hat{b}_1$ and $\hat{b}_2$ are called *partial regression coefficients*. $\hat{b}_0$, $\hat{b}_1$, and $\hat{b}_2$ are BLUE (see Sec. 6.5).

**EXAMPLE 1.** Table 7.1 extends Table 6.1 and gives the bushels of corn per acre, $Y$, resulting from the use of various amounts of fertilizer $X_1$ and insecticides $X_2$, both in pounds per acre, from 1971 to 1980. Using Eqs. (*7.5*), (*7.6*), and (*7.7*), we get

$$\hat{b}_1 = \frac{(\sum x_1 y)(\sum x_2^2) - (\sum x_2 y)(\sum x_1 x_2)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2} = \frac{(956)(504) - (900)(524)}{(576)(504) - (524)^2} \cong 0.65$$

$$\hat{b}_2 = \frac{(\sum x_2 y)(\sum x_1^2) - (\sum x_1 y)(\sum x_1 x_2)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2} = \frac{(900)(576) - (956)(524)}{(576)(504) - (524)^2} \cong 1.11$$

$$\hat{b}_0 = \bar{Y} - \hat{b}_1 \bar{X}_1 - \hat{b}_2 \bar{X}_2 \cong 57 - (0.65)(18) - (1.11)(12) \cong 31.98$$

so that $\hat{Y}_i = 31.98 + 0.65X_{1i} + 1.11X_{2i}$. To estimate the regression parameters with three or more independent or explanatory variables, see Section 7.6.

## 7.2 TESTS OF SIGNIFICANCE OF PARAMETER ESTIMATES

In order to test for the statistical significance of the parameter estimates of the multiple regression, the variance of the estimates is required:

$$\operatorname{Var} \hat{b}_1 = \sigma_u^2 \frac{\sum x_2^2}{\sum x_1^2 \sum x_2^2 - (\sum x_1 x_2)^2} \tag{7.8}$$

$$\operatorname{Var} \hat{b}_2 = \sigma_u^2 \frac{\sum x_1^2}{\sum x_1^2 \sum x_2^2 - (\sum x_1 x_2)^2} \tag{7.9}$$

[$b_0$ is usually not of primary concern; see Prob. 7.7(*e*)]. Since $\sigma_u^2$ is unknown, the residual variance $s^2$ is used as an unbiased estimate of $\sigma_u^2$:

$$s^2 = \hat{\sigma}_u^2 = \frac{\sum e_i^2}{n-k} \tag{6.12}$$

where $k$ = number of parameter estimates.

Unbiased estimates of the variance of $\hat{b}_0$ and $\hat{b}_1$ are then given by

$$s_{\hat{b}_1}^2 = \frac{\sum e_i^2}{n-k} \frac{\sum x_2^2}{\sum x_1^2 \sum x_2^2 - (\sum x_1 x_2)^2} \tag{7.10}$$

$$s_{\hat{b}_2}^2 = \frac{\sum e_i^2}{n-k} \frac{\sum x_1^2}{\sum x_1^2 \sum x_2^2 - (\sum x_1 x_2)^2} \tag{7.11}$$

so that $s_{\hat{b}_1}$ and $s_{\hat{b}_2}$ are the standard errors of the estimates. Tests of hypotheses about $b_1$ and $b_2$ are conducted as in Sec. 6.3.

**EXAMPLE 2.** Table 7.2 (an extension of Table 7.1) shows the additional calculations required to test the statistical significance of $\hat{b}_1$ and $\hat{b}_2$. The values for $\hat{Y}_i$ in Table 7.2 are obtained by substituting the values for $X_{1i}$ and $X_{2i}$ into the estimated OLS regression equation found in Example 1. (The values for $y_i^2$ are obtained by squaring $y_i$ from Table 7.1 and are to be used in Sec.7.3.) Using the values from Table 7.2 and 7.1, we get

$$s_{\hat{b}_1}^2 = \frac{\sum e_i^2}{n-k} \frac{\sum x_2^2}{\sum x_1^2 \sum x_2^2 - (\sum x_1 x_2)^2} = \frac{13.6704}{10-3} \frac{504}{(576)(504) - (524)^2} \cong 0.06 \qquad \text{and} \qquad s_{\hat{b}_1} \cong 0.24$$

$$s_{\hat{b}_2}^2 = \frac{\sum e_i^2}{n-k} \frac{\sum x_1^2}{\sum x_1^2 \sum x_2^2 - (\sum x_1 x_2)^2} = \frac{13.6704}{10-3} \frac{576}{(576)(504) - (524)^2} = 0.07 \qquad \text{and} \qquad s_{\hat{b}_2} \cong 0.27$$

**Table 7.1 Corn Produced with Fertilizer and Insecticide Used with Calculations for Parameter Estimation**

| Year | $Y$ | $X_1$ | $X_2$ | $y$ | $x_1$ | $x_2$ | $x_1y$ | $x_2y$ | $x_1x_2$ | $x_1^2$ | $x_2^2$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1971 | 40 | 6 | 4 | −17 | −12 | −8 | 204 | 136 | 96 | 144 | 64 |
| 1972 | 44 | 10 | 4 | −13 | −8 | −8 | 104 | 104 | 64 | 64 | 64 |
| 1973 | 46 | 12 | 5 | −11 | −6 | −7 | 66 | 77 | 42 | 36 | 49 |
| 1974 | 48 | 14 | 7 | −9 | −4 | −5 | 36 | 45 | 20 | 16 | 25 |
| 1975 | 52 | 16 | 9 | −5 | −2 | −3 | 10 | 15 | 6 | 4 | 9 |
| 1976 | 58 | 18 | 12 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1977 | 60 | 22 | 14 | 3 | 4 | 2 | 12 | 6 | 8 | 16 | 4 |
| 1978 | 68 | 24 | 20 | 11 | 6 | 8 | 66 | 88 | 48 | 36 | 64 |
| 1979 | 74 | 26 | 21 | 17 | 8 | 9 | 136 | 153 | 72 | 64 | 81 |
| 1980 | 80 | 32 | 24 | 23 | 14 | 12 | 322 | 276 | 168 | 196 | 144 |
| $n = 10$ | $\sum Y = 570$<br>$\bar{Y} = 57$ | $\sum X_1 = 180$<br>$\bar{X}_1 = 18$ | $\sum X_2 = 120$<br>$\bar{X}_2 = 12$ | $\sum y = 0$ | $\sum x_1 = 0$ | $\sum x_2 = 0$ | $\sum x_1y = 956$ | $\sum x_2y = 900$ | $\sum x_1x_2 = 524$ | $\sum x_1^2 = 576$ | $\sum x_2^2 = 504$ |

**Table 7.2. Corn-Fertilizer-Insecticide Calculations to Test Significance of Parameters**

| Year | $Y$ | $X_1$ | $X_2$ | $\hat{Y}$ | $e$ | $e^2$ | $y^2$ |
|---|---|---|---|---|---|---|---|
| 1971 | 40 | 6 | 4 | 40.32 | −0.32 | 0.1024 | 289 |
| 1972 | 44 | 10 | 4 | 42.92 | 1.08 | 1.1664 | 169 |
| 1973 | 46 | 12 | 5 | 45.33 | 0.67 | 0.4489 | 121 |
| 1974 | 48 | 14 | 7 | 48.85 | −0.85 | 0.7225 | 81 |
| 1975 | 52 | 16 | 9 | 52.37 | −0.37 | 0.1369 | 25 |
| 1976 | 58 | 18 | 12 | 57.00 | 1.00 | 1.0000 | 1 |
| 1977 | 60 | 22 | 14 | 61.82 | −1.82 | 3.3124 | 9 |
| 1978 | 68 | 24 | 20 | 69.78 | −1.78 | 3.1684 | 121 |
| 1979 | 74 | 26 | 21 | 72.19 | 1.81 | 3.2761 | 289 |
| 1980 | 80 | 32 | 24 | 79.42 | 0.58 | 0.3364 | 529 |
| $n = 10$ | | | | | $\sum e = 0$ | $\sum e^2 = 13.6704$ | $\sum y^2 = 1634$ |

Therefore, $t_1 = \hat{b}_1/s_{\hat{b}_1} \cong 0.65/0.24 \cong 2.70$, and $t_2 = \hat{b}_2/s_{\hat{b}_2} = 1.11/0.27 \cong 4.11$. Since both $t_1$ and $t_2$ exceed $t = 2.365$ with 7 df at the 5% level of significance (from App. 5), both $b_1$ and $b_2$ are statistically significant at the 5% level.

## 7.3 THE COEFFICIENT OF MULTIPLE DETERMINATION

The *coefficient of multiple determination* $R^2$ is defined as the proportion of the total variation in $Y$ "explained" by the multiple regression of $Y$ on $X_1$ and $X_2$, and (as shown in Sec. 6.4) it can be calculated by (see Prob. 7.14)

$$R^2 = \frac{\sum \hat{y}_i^2}{\sum y_i^2} = 1 - \frac{\sum e_i^2}{\sum y_i^2} = \frac{\hat{b}_1 \sum yx_1 + \hat{b}_2 \sum yx_2}{\sum y^2}$$

Since the inclusion of additional independent or explanatory variables is likely to increase the $\text{RSS} = \sum \hat{y}_i^2$ for the same $\text{TSS} = \sum y_i^2$ (see Sec. 6.4), $R^2$ increases. To factor in the reduction in the degrees of freedom as additional independent or explanatory variables are added, the *adjusted* $R^2$ *or* $\bar{R}^2$, is computed (see Prob. 7.16):

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-k} \tag{7.12}$$

where $n$ is the number of observations, and $k$ the number of parameters estimated.

**EXAMPLE 3.** $R^2$ for the corn-fertilizer-insecticide example can be found from Table 7.2:

$$R^2 = 1 - \frac{\sum e_i^2}{\sum y_i^2} = 1 - \frac{13.6704}{1634} \cong 1 - 0.0084 = 0.9916, \text{ or } 99.16\%$$

This compares with an $R^2$ of 97.10% in the simple regression, with fertilizer as the only independent or explanatory variable.

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-k} = 1 - (1 - 0.9916)\frac{10-1}{10-3} = 1 - 0.0084(1.2857) = 0.9892, \text{ or } 98.92\%$$

## 7.4 TEST OF THE OVERALL SIGNIFICANCE OF THE REGRESSION

The overall significance of the regression can be tested with the ratio of the explained to the unexplained *variance*. This follows an $F$ distribution (see Sec. 5.5) with $k-1$ and $n-k$ degrees of freedom, where $n$ is number of observations and $k$ is number of parameters estimated:

$$F_{k-1,n-k} = \frac{\sum \hat{y}_i^2/(k-1)}{\sum e_i^2/(n-k)} = \frac{R^2/(k-1)}{(1-R^2)/(n-k)} \tag{7.13}$$

If the calculated $F$ ratio exceeds the tabular value of $F$ at the specified level of significance and degrees of freedom (from App. 7), the hypothesis is accepted that the regression parameters are not all equal to zero and that $R^2$ is significantly different from zero.

In addition, the $F$ ratio can be used to test any linear restriction of regression parameters by using the form

$$F_{p,n-k} - \frac{\left(\dfrac{\sum e_{Ri}^2 - \sum e_i^2}{p}\right)}{\left(\dfrac{\sum e_i^2}{n-k}\right)}$$

where $p$ is the number of restriction being tested, $\sum e_{Ri}^2$ indicates the sum of squared residuals for the restricted regression where the restrictions are assumed to be true, and $\sum e_i^2$ indicates the sum of squared residuals for the unrestricted regression (i.e., the usual residuals). The null hypothesis is that the $p$ restrictions are true, in which case the residuals from the restricted and unrestricted models should be identical, and $F$ would take the value of zero. If the restrictions are not true, the unrestricted model will have lower errors, increasing the value of $F$. If $F$ exceeds the tabular value, the null hypothesis is rejected. This test will be used extensively in Sec. 11.6.

**EXAMPLE 4.** To test the overall significance of the regression estimated in Example 1 at the 5% level, we can use $R^2 = 0.9916$ (from Example 3), so that

$$F_{2,7} = \frac{0.9916/2}{(1-0.9916)/7} \cong 413.17$$

Since the calculated value of $F$ exceeds the tabular value of $F = 4.74$ at the 5% level of significance and with df $= 2$ and 7 (from App. 7), the hypothesis is accepted that $b_1$ and $b_2$ are not both zero and that $R^2$ is significantly different from zero.

## 7.5 PARTIAL-CORRELATION COEFFICIENTS

The *partial-correlation coefficient* measures the net correlation between the dependent variable and one independent variable after excluding the common influence of (i.e., holding constant) the other independent variables in the model. For example, $r_{YX_1 \cdot X_2}$ is the partial correlation between $Y$ and $X_1$, after removing the influence of $X_2$ from both $Y$ and $X_1$ [see Prob. 7.23(*a*)]:

$$r_{YX_1 \cdot X_2} = \frac{r_{YX_1} - r_{YX_2} r_{X_1X_2}}{\sqrt{1 - r_{X_1X_2}^2}\sqrt{1 - r_{YX_2}^2}} \tag{7.14}$$

$$r_{YX_2 \cdot X_1} = \frac{r_{YX_2} - r_{YX_1} r_{X_1X_2}}{\sqrt{1 - r_{X_1X_2}^2}\sqrt{1 - r_{YX_1}^2}} \tag{7.15}$$

where $r_{YX_1}$ = simple-correlation coefficient between $Y$ and $X_1$, and $r_{YX_2}$ and $r_{X_1X_2}$ are analogously defined. Partial-correlation coefficients range in value from $-1$ to $+1$ (as do simple-correlation coefficients), have the sign of the corresponding estimated parameter, and are used to determine the relative importance of the different explanatory variables in a multiple regression.

**EXAMPLE 5.** Substituting the values from Tables 7.1 and 7.2 into Eq. (*6.18*) for the simple-correlation coefficient, we get

$$r_{YX_1} = \frac{\sum x_1 y}{\sqrt{\sum x_1^2}\sqrt{\sum y^2}} = \frac{956}{\sqrt{576}\sqrt{1634}} \cong 0.9854$$

$$r_{YX_2} = \frac{\sum x_2 y}{\sqrt{\sum x_2^2}\sqrt{\sum y^2}} = \frac{900}{\sqrt{504}\sqrt{1634}} \cong 0.9917$$

$$r_{X_1X_2} = \frac{\sum x_2 x_1}{\sqrt{\sum x_2^2}\sqrt{\sum x_1^2}} = \frac{524}{\sqrt{504}\sqrt{576}} \cong 0.9725$$

Thus
$$r_{YX_1 \cdot X_2} = \frac{r_{YX_1} - r_{YX_2} r_{X_1X_2}}{\sqrt{1 - r_{X_1X_2}^2}\sqrt{1 - r_{YX_2}^2}} = \frac{0.9854 - (0.9917)(0.9725)}{\sqrt{1 - 0.9725^2}\sqrt{1 - 0.9917^2}}$$
$$\cong 0.7023, \text{ or } 70.23\%$$

and
$$r_{YX_2 \cdot X_1} = \frac{r_{YX_2} - r_{YX_1} r_{X_1X_2}}{\sqrt{1 - r_{X_1X_2}^2}\sqrt{1 - r_{YX_1}^2}} = \frac{0.9917 - (0.9854)(0.9725)}{\sqrt{1 - 0.9725^2}\sqrt{1 - 0.9854^2}} \cong 0.8434, \text{ or } 84.34\%$$

Therefore, $X_2$ is more important than $X_1$ in explaining the variation $Y$.

**EXAMPLE 6.** The overall results of the corn-fertilizer-insecticide example can be summarized as

$$\hat{Y} = 31.98 + 0.65X_1 + 1.11X_2$$
$$t \text{ values } (2.70) \quad (4.11)$$
$$R^2 = 0.992 \qquad \bar{R}^2 = 0.989 \qquad F_{2,7} = 413.17$$
$$r_{YX_1 \cdot X_2} = 0.70 \qquad r_{YX_2 \cdot X_1} = 0.84$$

Even though results are usually obtained from the computer (see Chap. 12), it is crucial to work through a problem "by hand," as we have done, in order to clearly understand the procedure.

## 7.6 MATRIX NOTATION

Calculations increase substantially as the number of independent variables increase. Matrix notation can aid in solving larger regressions algebraically. The following solution works with any number of independent variables, and is therefore extremely flexible. Students not familiar with linear algebra may skip this section with no loss of continuity.

The regression from Sec. 1 can be written with matrices as

$$Y = Xb + u$$

where
$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \qquad X = \begin{bmatrix} 1 & X_{11} & X_{21} \\ 1 & X_{12} & X_{22} \\ \vdots & \vdots & \vdots \\ 1 & X_{1n} & X_{2n} \end{bmatrix} \qquad b = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} \qquad u = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

$$\hat{b} = \begin{bmatrix} \hat{b}_0 \\ \hat{b}_1 \\ \hat{b}_2 \end{bmatrix} = (X'X)^{-1}X'Y$$

$$s_{\hat{b}}^2 = \begin{pmatrix} s_{\hat{b}_0}^2 & \text{cov}(b_0, b_1) & \text{cov}(b_0, b_2) \\ \text{cov}(b_0, b_1) & s_{\hat{b}_1}^2 & \text{cov}(b_1, b_2) \\ \text{cov}(b_0, b_2) & \text{cov}(b_1, b_2) & s_{\hat{b}_2}^2 \end{pmatrix} = \frac{e'e}{(n-k)}(X'X)^{-1}$$ (symmetrical, so lower and upper triangle are identical)

**EXAMPLE 7.** Recalculation of corn-fertilizer-insecticide example with matrices

$$\hat{b} = \left[ \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 6 & 10 & 12 & 14 & 16 & 18 & 22 & 24 & 26 & 32 \\ 4 & 4 & 5 & 7 & 9 & 12 & 14 & 20 & 21 & 24 \end{bmatrix} \begin{bmatrix} 1 & 6 & 4 \\ 1 & 10 & 4 \\ 1 & 12 & 5 \\ 1 & 14 & 7 \\ 1 & 16 & 9 \\ 1 & 18 & 12 \\ 1 & 22 & 14 \\ 1 & 24 & 20 \\ 1 & 26 & 21 \\ 1 & 37 & 24 \end{bmatrix} \right]^{-1}$$

$$\times \left[ \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 6 & 10 & 12 & 14 & 16 & 18 & 22 & 24 & 26 & 32 \\ 4 & 4 & 5 & 7 & 9 & 12 & 14 & 20 & 21 & 24 \end{bmatrix} \begin{bmatrix} 40 \\ 44 \\ 46 \\ 48 \\ 52 \\ 58 \\ 60 \\ 68 \\ 74 \\ 80 \end{bmatrix} \right]$$

$$\hat{b} = \begin{bmatrix} 1.36 & -0.18 & 0.16 \\ -0.18 & 0.03 & -0.03 \\ 0.16 & -0.03 & 0.04 \end{bmatrix} \begin{bmatrix} 570 \\ 11{,}216 \\ 7740 \end{bmatrix} = \begin{bmatrix} 31.98 \\ 0.65 \\ 1.11 \end{bmatrix}$$

therefore, $\hat{b}_0 = 31.98$, $\hat{b}_1 = 0.65$, and $\hat{b}_2 = 1.11$.

$$e = Y - X\hat{b} = \begin{bmatrix} 40 \\ 44 \\ 46 \\ 48 \\ 52 \\ 58 \\ 60 \\ 68 \\ 74 \\ 80 \end{bmatrix} - \begin{bmatrix} 1 & 6 & 4 \\ 1 & 10 & 4 \\ 1 & 12 & 5 \\ 1 & 14 & 7 \\ 1 & 16 & 9 \\ 1 & 18 & 12 \\ 1 & 22 & 14 \\ 1 & 24 & 20 \\ 1 & 26 & 21 \\ 1 & 32 & 24 \end{bmatrix} \begin{bmatrix} 31.98 \\ 0.65 \\ 1.11 \end{bmatrix} = \begin{bmatrix} -0.32 \\ 1.08 \\ 0.67 \\ -0.85 \\ -0.37 \\ 1.00 \\ -1.82 \\ -1.78 \\ 1.81 \\ 0.58 \end{bmatrix}$$

$$s^2_{\hat{b}} = \frac{13.6704}{(10-3)} \begin{bmatrix} 1.36 & -0.18 & 0.16 \\ -0.18 & 0.03 & -0.03 \\ 0.16 & -0.03 & 0.04 \end{bmatrix} = \begin{bmatrix} 2.66 & -0.35 & 0.31 \\ -0.34 & 0.06 & -0.07 \\ 0.31 & -0.07 & 0.07 \end{bmatrix}$$

therefore $s^2_{\hat{b}_0} = 2.66$, $s^2_{\hat{b}_1} = 0.06$, and $s^2_{\hat{b}_2} = 0.07$.

# Solved Problems

## THE THREE-VARIABLE LINEAR MODEL

**7.1** (*a*) Write the equation of the multiple regression linear model for the case of 2 and $k$ independent or explanatory variables. (*b*) State the assumptions of the multiple regression linear model.

(*a*) For the case of 2 independent or explanatory variables, we have

$$Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + u_i \tag{7.1}$$

For the case of $k$ independent or explanatory variables, we have

$$Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \cdots + b_k X_{ki} + u_i$$

where $X_{2i}$ represents, for example, the $i$th observation on independent variable $X_2$.

(*b*) The first five assumptions of the multiple regression linear model are exactly the same as those of the simple OLS regression model (see Prob. 6.4). That is, the first three assumptions can be summarized as $u_i \sim N(0, \sigma_u^2)$. The fourth assumption is $E(u_i u_j) = 0$ for $i \neq j$; and the fifth assumption is $E(X_i u_i) = 0$. The only additional assumption required for the multiple OLS regression linear model is that there is no exact linear relationship between the $X$s. If two or more explanatory variables are perfectly linearly correlated, it will be impossible to calculate OLS estimates of the parameters because the system of normal equations will contain two or more equations that are not independent. If two or more explanatory variables are highly but not perfectly linearly correlated, then OLS parameter estimates can be calculated, but the effect of each of the highly linearly correlated variables on the explanatory variable cannot be isolated (see Sec. 9.1).

**7.2** With the OLS procedure in the case of two independent or explanatory variables, derive (*a*) normal Eq. (*7.2*), (*b*) normal Eq. (*7.3*), and (*c*) normal Eq. (*7.4*). (The reader without knowledge of calculus can skip this problem.)

(*a*) Normal Eq. (*7.2*) is derived by minimizing $\sum e_i^2$ with respect to $\hat{b}_0$:

$$\begin{aligned} \frac{\partial e_i^2}{\partial \hat{b}_0} &= \frac{\partial \sum (Y_i - \hat{b}_0 - \hat{b}_1 X_{1i} - \hat{b}_2 X_{2i})^2}{\partial \hat{b}_0} = 0 \\ -2 \sum (Y_i - \hat{b}_0 - \hat{b}_1 X_{1i} - \hat{b}_2 X_{2i}) &= 0 \\ \sum Y_i = n\hat{b}_0 + \hat{b}_1 \sum X_{1i} + \hat{b}_2 \sum X_{2i} & \end{aligned} \tag{7.2}$$

(*b*) Normal Eq. (*7.3*) is derived by minimizing $\sum e_i^2$ with respect to $\hat{b}_1$:

$$\begin{aligned} \frac{\partial \sum e_i^2}{\partial \hat{b}_1} &= \frac{\partial \sum (Y_i - \hat{b}_0 - \hat{b}_1 X_{1i} - \hat{b}_2 X_{2i})^2}{\partial \hat{b}_1} = 0 \\ -2 \sum X_{1i} (Y_i - \hat{b}_0 - \hat{b}_1 X_{1i} - \hat{b}_2 X_{2i}) &= 0 \\ \sum X_{1i} Y_i = \hat{b}_0 \sum X_{1i} + \hat{b}_1 \sum x_{1i}^2 + \hat{b}_2 \sum X_{1i} X_{2i} & \end{aligned} \tag{7.3}$$

(*c*) Normal Eq. (*7.4*) is derived by minimizing $\sum e_i^2$ with respect to $\hat{b}_2$:

$$\begin{aligned} \frac{\partial \sum e_i^2}{\partial \hat{b}_2} &= \frac{\partial \sum (Y_i - \hat{b}_0 - \hat{b}_1 X_{1i} - \hat{b}_2 X_{2i})^2}{\partial \hat{b}_2} = 0 \\ -2 \sum X_{2i} (Y_i - \hat{b}_0 - \hat{b}_1 X_{1i} - \hat{b}_2 X_{2i}) &= 0 \\ \sum X_{2i} Y_i = \hat{b}_0 \sum X_{2i} + \hat{b}_1 \sum X_{1i} X_{2i} + \hat{b}_2 \sum X_{2i}^2 & \end{aligned} \tag{7.4}$$

**7.3** For the two independent or explanatory variable multiple linear regression model, (*a*) derive the normal equations in deviation form. (*Hint*: Start by deriving the expression for $\hat{y}_i$; the reader

without knowledge of calculus can skip this part of this problem.) (*b*) How are Eqs. (*7.5*), (*7.6*), and (*7.7*) derived for $\hat{b}_1$, $\hat{b}_2$, and $\hat{b}_0$?

(*a*)
$$\hat{Y}_i = \hat{b}_0 + \hat{b}_1 X_{1i} + \hat{b}_2 X_{2i}$$
$$\bar{Y} = \hat{b}_0 + \hat{b}_1 \bar{X}_1 + \hat{b}_2 \bar{X}_2$$

Subtracting, we get

$$\hat{y}_i = \hat{Y}_i - \bar{Y} = \hat{b}_1 x_{1i} + \hat{b}_2 x_{2i}$$

Therefore, $e_i = y_i - \hat{y}_i = y_i - \hat{b}_1 x_{1i} - \hat{b}_2 x_{2i}$

$$\sum e_i^2 = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - \hat{b}_1 x_{1i} - \hat{b}_2 x_{2i})^2$$
$$\frac{\partial \sum e_i^2}{\partial \hat{b}_1} = \frac{\partial \sum (y_i - \hat{b}_1 x_{1i} - \hat{b}_2 x_{2i})^2}{\partial \hat{b}_1} = 0$$
$$-2 \sum x_{1i}(y_i - \hat{b}_1 x_{1i} - \hat{b}_2 x_{2i}) = 0$$
$$\sum x_{1i} y_i = \hat{b}_1 \sum x_{1i}^2 + \hat{b}_2 \sum x_{1i} x_{2i} \tag{7.16}$$
$$\frac{\partial \sum e_i^2}{\partial \hat{b}_2} = \frac{\partial \sum (y_i - \hat{b}_1 x_{1i} - \hat{b}_2 x_{2i})^2}{\partial \hat{b}_2} = 0$$
$$-2 \sum x_{2i}(y_i - \hat{b}_1 x_{1i} - \hat{b}_2 x_{2i}) = 0$$
$$\sum x_{2i} y_i = \hat{b}_1 \sum x_{1i} x_{2i} + \hat{b}_2 \sum x_{2i}^2 \tag{7.17}$$

(*b*) Equations (*7.5*) and (*7.6*) to calculate $\hat{b}_1$ and $\hat{b}_2$, respectively, are obtained by solving Eqs. (*7.16*) and (*7.17*) simultaneously. It is always possible to calculate $\hat{b}_1$ and $\hat{b}_2$, except if there is an exact linear relationship between $X_1$ and $X_2$ or if the number of observations on each variable of the model is 3 or fewer. Parameter $\hat{b}_0$ can then be calculated by substituting into Eq. (*7.7*) the values of $\hat{b}_1$ and $\hat{b}_2$ [calculated with Eqs. (*7.5*) and (*7.6*)] and $\bar{Y}$, $\bar{X}_1$, and $\bar{X}_2$ (calculated from the given values of the problem).

**7.4** With reference to multiple regression analysis with two independent or explanatory variables, indicate the meaning of (*a*) $\hat{b}_0$, (*b*) $\hat{b}_1$, (*c*) $\hat{b}_2$. (*d*) Are $\hat{b}_0$, $\hat{b}_1$, and $\hat{b}_2$ BLUE?

(*a*) Parameter $b_0$ is the constant term or intercept of the regression and gives the estimated value of $Y_i$, when $X_{1i} = X_{2i} = 0$.

(*b*) Parameter $b_1$ measures the change in $Y$ for each one-unit change in $X_1$ while holding $X_2$ constant. Slope parameter $b_1$ is a partial regression coefficient because it corresponds to the partial derivative of $Y$ with respect to $X_1$, or $\partial Y/\partial X_1$.

(*c*) Parameter $b_2$ measures the change in $Y$ for each one-unit change in $X_2$ while holding $X_1$ constant. Slope parameter $b_2$ is the second partial regression coefficient because it corresponds to the partial derivative of $Y$ with respect to $X_2$, or $\partial Y/\partial X_2$.

(*d*) Since $\hat{b}_0$, $\hat{b}_1$, and $\hat{b}_2$ are obtained by the OLS method, they are also best linear unbiased estimators (BLUE; see Sec. 6.5). That is, $E(\hat{b}_0) = b_0$, $E(\hat{b}_1) = b_1$, and $E(\hat{b}_2) = b_2$, and $s_{\hat{b}_0}$, $s_{\hat{b}_1}$, and $s_{\hat{b}_2}$ are lower than for any other unbiased linear estimator. Proof of these properties is very cumbersome without the use of matrix algebra, so they are not provided here.

**7.5** Table 7.3 gives the real per capita income in thousands of U.S. dollars $Y$ with the percentage of the labor force in agriculture $X_1$ and the average years of schooling of the population over 25 years of age $X_2$ for 15 developed countries in 1981. (*a*) Find the least-squares regression equation of $Y$ on $X_1$ and $X_2$. (*b*) Interpret the results of part *a*.

(*a*) Table 7.4 shows the calculations required to estimate the parameters of the OLS regression equation of $Y$ on $X_1$ and $X_2$.

**Table 7.3 Per Capita Income, Labor Force in Agriculture, and Years of Schooling**

| $n$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $Y$ | 6 | 8 | 8 | 7 | 7 | 12 | 9 | 8 | 9 | 10 | 10 | 11 | 9 | 10 | 11 |
| $X_1$ | 9 | 10 | 8 | 7 | 10 | 4 | 5 | 5 | 6 | 8 | 7 | 4 | 9 | 5 | 8 |
| $X_2$ | 8 | 13 | 11 | 10 | 12 | 16 | 10 | 10 | 12 | 14 | 12 | 16 | 14 | 10 | 12 |

$$\hat{b}_1 = \frac{(\sum x_1 y)(\sum x_2^2) - (\sum x_2 y)(\sum x_1 x_2)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2} = \frac{(-28)(74) - (38)(-12)}{(60)(74) - (-12)^2}$$
$$= \frac{-2072 + 456}{4440 - 144} \cong -0.38$$
$$\hat{b}_2 = \frac{(\sum x_2 y)(\sum x_1^2) - (\sum x_1 y)(\sum x_1 x_2)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2} = \frac{(38)(60) - (-28)(-12)}{(60)(74) - (-12)^2}$$
$$= \frac{2280 - 336}{4440 - 144} \cong 0.45$$
$$\hat{b}_0 = \bar{Y} - \hat{b}_1 \bar{X}_1 - \hat{b}_2 \bar{X}_2 \cong 9 - (-0.38)(7) - (0.45)(12) = 9 + 2.66 - 5.40 \cong 6.26$$

Thus the estimated OLS regression equation of $Y$ on $X_1$ and $X_2$ is

$$\hat{Y}_i = 6.26 - 0.38X_{1i} + 0.45X_{2i}$$

(*b*) The estimated OLS regression equation indicates that the level of real per capita income $Y$ is inversely related to the percentage of the labor force in agriculture $X_1$ but directly related to the years of schooling of the population over 25 years (as might have been anticipated). Specifically, $\hat{b}_1$ indicates that a 1 percentage point decline in the labor force in agriculture is associated with an increase in per capita income of 380 U.S. dollars while holding $X_2$ constant. However, an increase of 1 year of schooling for the population over 25 years of age is associated with an increase in per capita income of 450 U.S. dollars, while holding $X_1$ constant. When $X_{1i} = X_{2i} = 0$, $\hat{Y}_i = \hat{b}_0 = 6.26$.

**7.6** Table 7.5 extends Table 6.11 and gives the per capita GDP (gross domestic product) to the nearest \$100 ($Y$) and the percentage of the economy represented by agriculture ($X_1$), and the male literacy rate ($X_2$) reported by the World Bank World Development Indicators for 1999 for 15 Latin American countries. (*a*) Find the least-squares regression equation of $Y$ on $X_1$ and $X_2$. (*b*) Interpret the results of part *a* and compare them with those of Prob. 6.30.

(*a*) Table 7.6 shows the calculations required to estimate the parameters of the OLS regression equation of $Y$ on $X_1$ and $X_2$.

$$\hat{b}_1 = \frac{(\sum x_1 y)(\sum x_2^2) - (\sum x_2 y)(\sum x_1 x_2)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2} = \frac{(-1149)(1093.7335) - (1637.7335)(-543)}{(442)(1093.7335) - (-543)^2} \cong -1.95$$
$$\hat{b}_2 = \frac{(\sum x_2 y)(\sum x_1^2) - (\sum x_1 y)(\sum x_1 x_2)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2} = \frac{(1637.7335)(442) - (-1149)(-543)}{(442)(1093.7335) - (-543)^2} \cong 0.53$$
$$\hat{b}_0 = \bar{Y} - \hat{b}_1 \bar{X}_1 - \hat{b}_2 \bar{X}_2 = 30.53 - (-1.95)(11) - (0.53)(88.53) = 5.06$$

Thus the estimated OLS regression equation of $Y$ on $X_1$ and $X_2$ is

$$\hat{Y} = 5.06 - 1.95X_1 + 0.53X_2$$

(*b*) The estimated OLS equation indicates that the level of per capita income $Y$ is inversely related to the percentage of the economy represented by agriculture $X_1$ but directly related to the literacy rate of the male population (as might have been anticipated). Specifically, $\hat{b}_1$ indicates that a 1 point decline in the percentage of the economy represented by agriculture is associated with an increase in per capita