Chapter 6
Linear Regression with Multiple Regressors

■ Solutions to Exercises

1. By equation (6.15) in the text, we know

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

Thus, that values of $\bar{R}^2$ are 0.175, 0.189, and 0.193 for columns (1)–(3).

2. (a) Workers with college degrees earn \$5.46/hour more, on average, than workers with only high school degrees.
   (b) Men earn \$2.64/hour more, on average, than women.

3. (a) On average, a worker earns \$0.29/hour more for each year he ages.
   (b) Sally's earnings prediction is $4.40 + 5.48 \times 1 - 2.62 \times 1 + 0.29 \times 29 = 15.67$ dollars per hour. Betsy's earnings prediction is $4.40 + 5.48 \times 1 - 2.62 \times 1 + 0.29 \times 34 = 17.12$ dollars per hour. The difference is 1.45

4. (a) Workers in the Northeast earn \$0.69 more per hour than workers in the West, on average, controlling for other variables in the regression. Workers in the Northeast earn \$0.60 more per hour than workers in the West, on average, controlling for other variables in the regression. Workers in the South earn \$0.27 less than workers in the West.
   (b) The regressor *West* is omitted to avoid perfect multicollinearity. If *West* is included, then the intercept can be written as a perfect linear function of the four regional regressors. Because of perfect multicollinearity, the OLS estimator cannot be computed.
   (c) The expected difference in earnings between Juanita and Jennifer is $-0.27 - 0.6 = -0.87$.

5. (a) \$23,400 (recall that *Price* is measured in \$1000s).
   (b) In this case $\Delta BDR = 1$ and $\Delta Hsize = 100$. The resulting expected change in price is $23.4 + 0.156 \times 100 = 39.0$ thousand dollars or \$39,000.
   (c) The loss is \$48,800.
   (d) From the text $\bar{R}^2 = 1 - \frac{n-1}{n-k-1}(1-R^2)$, so $R^2 = 1 - \frac{n-k-1}{n-1}(1-\bar{R}^2)$, thus, $R^2 = 0.727$.

6. (a) There are other important determinants of a country's crime rate, including demographic characteristics of the population.
   (b) Suppose that the crime rate is positively affected by the fraction of young males in the population, and that counties with high crime rates tend to hire more police. In this case, the size of the police force is likely to be positively correlated with the fraction of young males in the population leading to a positive value for the omitted variable bias so that $\hat{\beta}_1 > \beta_1$.

7. (a) The proposed research in assessing the presence of gender bias in setting wages is too limited. There might be some potentially important determinants of salaries: type of engineer, amount of work experience of the employee, and education level. The gender with the lower wages could reflect the type of engineer among the gender, the amount of work experience of the employee, or the education level of the employee. The research plan could be improved with the collection of additional data as indicated and an appropriate statistical technique for analyzing the data would be a multiple regression in which the dependent variable is wages and the independent variables would include a dummy variable for gender, dummy variables for type of engineer, work experience (time units), and education level (highest grade level completed). The potential importance of the suggested omitted variables makes a "difference in means" test inappropriate for assessing the presence of gender bias in setting wages.

   (b) The description suggests that the research goes a long way towards controlling for potential omitted variable bias. Yet, there still may be problems. Omitted from the analysis are characteristics associated with behavior that led to incarceration (excessive drug or alcohol use, gang activity, and so forth), that might be correlated with future earnings. Ideally, data on these variables should be included in the analysis as additional control variables.

8. Omitted from the analysis are reasons *why* the survey respondents slept more or less than average. People with certain chronic illnesses might sleep more than 8 hours per night. People with other illnesses might sleep less than 5 hours. This study says nothing about the *causal* effect of sleep on mortality.

9. For omitted variable bias to occur, two conditions must be true: X_1 (the included regressor) is correlated with the omitted variable, and the omitted variable is a determinant of the dependent variable. Since X_1 and X_2 are uncorrelated, the estimator of β_1 does not suffer from omitted variable bias.

10. (a)

$$\sigma^2_{\hat{\beta}_1} = \frac{1}{n}\left[\frac{1}{1-\rho^2_{X_1,X_2}}\right]\frac{\sigma^2_u}{\sigma^2_{X_1}}$$

Assume X_1 and X_2 are uncorrelated: $\rho^2_{X_1 X_2} = 0$

$$\sigma^2_{\hat{\beta}_1} = \frac{1}{400}\left[\frac{1}{1-0}\right]\frac{4}{6}$$

$$= \frac{1}{400}\cdot\frac{4}{6} = \frac{1}{600} = 0.00167$$

   (b) With $\rho_{X_1,X_2} = 0.5$

$$\sigma^2_{\hat{\beta}_1} = \frac{1}{400}\left[\frac{1}{1-0.5^2}\right]\frac{4}{6}$$

$$= \frac{1}{400}\left[\frac{1}{0.75}\right]\frac{4}{6} = .0022$$

(c) The statement correctly says that the larger is the correlation between X_1 and X_2 the larger is the variance of $\hat{\beta}_1$, however the recommendation "it is best to leave X_2 out of the regression" is incorrect. If X_2 is a determinant of Y, then leaving X_2 out of the regression will lead to omitted variable bias in $\hat{\beta}_1$.

11. (a)

$$\sum (Y_i - b_1 X_{1i} - b_2 X_{2i})^2$$

(b)

$$\frac{\partial \sum (Y_i - b_1 X_{1i} - b_2 X_{2i})^2}{\partial b_1} = -2\sum X_{1i}(Y_i - b_1 X_{1i} - b_2 X_{2i})$$

$$\frac{\partial \sum (Y_i - b_1 X_{1i} - b_2 X_{2i})^2}{\partial b_2} = -2\sum X_{2i}(Y_i - b_1 X_{1i} - b_2 X_{2i})$$

(c) From (b), $\hat{\beta}_1$ satisfies

$$\sum X_{1i}(Y_i - \hat{\beta}_1 X_{1i} - \hat{\beta}_1 X_{2i}) = 0$$

or

$$\hat{\beta}_1 = \frac{\sum X_{1i} Y_i - \hat{\beta}_2 \sum X_{1i} X_{2i}}{\sum X_{1i}^2}$$

and the result follows immediately.

(d) Following analysis as in (c)

$$\hat{\beta}_2 = \frac{\sum X_{2i} Y_i - \hat{\beta}_1 \sum X_{1i} X_{2i}}{\sum X_{2i}^2}$$

and substituting this into the expression for $\hat{\beta}_1$ in (c) yields

$$\hat{\beta}_1 = \frac{\sum X_{1i} Y \frac{\sum X_{2i} Y_i - \hat{\beta}_1 \sum X_{1i} X_{2i}}{\sum X_{2i}^2} \sum X_{1i} X_{2i}}{\sum X_{1i}^2}.$$

Solving for $\hat{\beta}_1$ yields:

$$\hat{\beta}_1 = \frac{\sum X_{2i}^2 \sum X_{1i} Y_i - \sum X_{1i} X_{2i} \sum X_{2i} Y_i}{\sum X_{1i}^2 \sum X_{2i}^2 - (\sum X_{1i} X_{2i})^2}$$

(e) The least squares objective function is $\sum (Y_i - b_0 - b_1 X_{1i} - b_2 X_{2i})^2$ and the partial derivative with respect to b_0 is

$$\frac{\partial \sum (Y_i - b_0 - b_1 X_{1i} - b_2 X_{2i})^2}{\partial b_0} = -2\sum (Y_i - b_0 - b_1 X_{1i} - b_2 X_{2i}).$$

Setting this to zero and solving for $\hat{\beta}_0$ yields: $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}_1 - \hat{\beta}_2 \overline{X}_2$.

(f)

$$\frac{\partial}{\partial \beta_1} = 0 \Rightarrow -\sum X_{1i}Y_i + \hat{\beta}_2 \sum X_{1i}Y_{2i} + \hat{\beta}_1 \sum X_{1i}^2 = 0$$

$$\hat{\beta}_1 \sum X_{1i}^2 = \sum X_{1i}Y_i - \hat{\beta}_2 \sum X_{1i}X_{2i}$$

$$\hat{\beta}_1 = \frac{\sum X_{1i}Y_i - \beta_2 \sum X_{1i}Y_{2i}}{\sum X_{1i}^2}$$

$$\frac{\partial 2}{\partial \beta_2} = 0 \Rightarrow -\sum X_{2i}Y_i + \hat{\beta}_1 \sum X_{1i}X_{2i} + \hat{\beta}_2 \sum X_{2i}^2 = 0$$

$$\hat{\beta}_2 \sum X_{2i}^2 = \sum X_{2i}Y_i - \hat{\beta}_1 \sum X_{1i}X_{2i}$$

$$\hat{\beta}_2 = \frac{\sum X_{2i}Y_i - \beta_1 \sum X_{1i}Y_{2i}}{\sum X_{2i}^2}$$

(g)

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u$$

$$u_i = Y_i - (\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i})$$

$$\sum_{i=1}^{n} u_i^2 = \sum [Y_i - (\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i})]^2$$

$$= \sum [(Y_i - \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i})(Y_i - \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i})]$$

$$= \sum [Y_i^2 - Y_i\beta_0 - Y_i\beta_1 X_{1i} - Y_i\beta_2 X_{2i} - \beta_0 Y_i + \beta_0^2 + \beta_0\beta_1 X_{1i} + \beta_0\beta_2 X_{2i}$$

$$- \beta_1 X_{1i}Y_i + \beta_1 X_{1i}\beta_0 + \beta_1^2 X_{1i}^2 + \beta_1 X_{1i}\beta_2 X_{2i} - \beta_2 X_{2i}Y_i + \beta_2 X_{2i}\beta_0 + \beta_2 X_{2i}\beta_1 X_{1i} + \beta_2^2 X_{2i}^2]$$

$$= \sum [Y_i^2 - 2\beta_0 Y_i - 2\beta_1 X_{1i}Y_i - 2\beta_2 X_{2i}Y_i + \beta_0^2 + 2\beta_0\beta_1 X_{1i} + 2\beta_0\beta_2 X_{2i}$$

$$+2\beta_1\beta_2 X_{1i}X_{2i} + \beta_1^2 X_{1i}^2 + \beta_2^2 X_{2i}^2]$$

$$= \frac{1}{n}\sum Y_i^2 - 2\beta_0\overline{Y} - 2\beta_1 \sum X_{1i}Y_i - 2\beta_2 \sum X_{2i}Y_i + \beta_0^2 + 2\beta_0\beta_1\overline{X}_1 + 2\beta_0\beta_2\overline{X}_2$$

$$+2\beta_1\beta_2 \sum X_{1i}X_{2i} + \beta_1^2 \sum X_{1i}^2 + \beta_2^2 \sum X_{2i}^2$$

$$\frac{\partial}{\partial \beta_0} = -2\overline{Y} + 2\hat{\beta}_0 + 2\hat{\beta}_1\overline{X}_1 + 2\hat{\beta}_2\overline{X}_2 = 0 \Rightarrow 2\hat{\beta}_0 = 2\overline{Y} - 2\hat{\beta}_1\overline{X}_1 - 2\hat{\beta}_2\overline{X}_2$$

$$\Rightarrow \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\overline{X}_1 - \hat{\beta}_2\overline{X}_2$$