Empirical Exercise: Chapter 6

Use the **Birthweight_Smoking** data set introduced in previous empirical exercise to answer the following questions.

a. Regress *Birthweight* on *Smoker*. What is the estimated effect of smoking on birth weight?
b. Regress *Birthweight* on *Smoker*, *Alcohol*, and *Nprevist*.
    i. Using the two conditions in **Key Concept 6.1**[1], explain why the exclusion of *Alcohol* and *Nprevist* could lead to omitted variable bias in the regression estimated in (a).
    ii. Is the estimated effect of smoking on birth weight substantially different from the regression that excludes *Alcohol* and *Nprevist*? Does the regression in (a) seem to suffer from omitted variable bias?
    iii. Jane smoked during her pregnancy, did not drink alcohol, and had 8 prenatal care visits. Use the regression to predict the birth weight of Jane's child.
    iv. Compute R^2 and . Why are they so similar?
    v. How should you interpret the coefficient on *Nprevist*? Does the coefficient measure a causal effect of prenatal visits on birth weight? If not, what does it measure?
c. Estimate the coefficient on *Smoking* for the multiple regression model in (b), using the three-step process in **Appendix 6.3**[2] (the Frisch–Waugh theorem). Verify that the three-step process yields the same estimated coefficient for *Smoking* as that obtained in (b).
d. An alternative way to control for prenatal visits is to use the binary variables *Tripre0* through *Tripre3*. Regress *Birthweight* on *Smoker*, *Alcohol*, *Tripre0*, *Tripre2*, and *Tripre3*.
    i. Why is *Tripre1* excluded from the regression? What would happen if you included it in the regression?
    ii. The estimated coefficient on *Tripre0* is large and negative. What does this coefficient measure? Interpret its value.
    iii. Interpret the value of the estimated coefficients on *Tripre2* and *Tripre3*.
    iv. Does the regression in (d) explain a larger fraction of the variance in birth weight than the regression in (b)?

1. **Key Concept 6.1**

   Omitted variable bias is the bias in the OLS estimator of the causal effect of *X* on *Y* that arises when the regressor, *X*, is correlated with an omitted variable. For omitted variable bias to occur, two conditions must be true:

   I. *X* is correlated with the omitted variable.
   II. The omitted variable is a determinant of the dependent variable, *Y*.

2. **Appendix 6.3**

   The Frisch–Waugh Theorem

Empirical Exercise: Chapter 6

The OLS estimator in multiple regression can be computed by a sequence of shorter regressions. Consider the multiple regression model in Equation (6.7).

(6.7)

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + u_i, i = 1, \ldots, n,$$

The OLS estimator of can be computed in three steps:

1. Regress X_1 on $X_2, X_3, \ldots, X_k$, and let $\tilde{X}_1$, denote the residuals from this regression;
2. Regress Y on $X_2, X_3, \ldots, X_k$, and let denote the residuals from this regression; and
3. Regress $\tilde{Y}$ on $\tilde{X}_1$,

where the regressions include a constant term (intercept). The Frisch–Waugh theorem states that the OLS coefficient in step 3 equals the OLS coefficient on in the multiple regression model [Equation (6.7)]

(6.7)

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + u_i, i = 1, \ldots, n,$$

This result provides a mathematical statement of how the multiple regression coefficient estimates the effect on Y of X_1, controlling for the other X's: Because the first two regressions (steps 1 and 2) remove from Y and X_1 their variation associated with the other X's, the third regression estimates the effect on Y of using what is left over after removing (controlling for) the effect of the other X's.