Econometrics – Problem Set #6

Junwoo Yang

May 14, 2021

The first four exercises refer to the table of estimated regressions on page 238, computed using data for 2015 from the Current Population Survey. The data set consists of information on 7178 full-time, full-year workers. The highest educational achievement for each worker was either a high school diploma or a bachelor's degree. The workers' ages ranged from 25 to 34 years. The data set also contains information on the region of the country where the person lived, marital status, and number of children. For the purposes of these exercises, let

$$
\begin{aligned}
AHE &= \text{average hourly earnings} \\
College &= \text{binary variable (1 if college, 0 if high school)} \\
Female &= \text{binary variable (1 if female, 0 if male)} \\
Age &= \text{age (in years)} \\
Northeast &= \text{binary variable (1 if Region = Northeast, 0 otherwise)} \\
Midwest &= \text{binary variable (1 if Region = Midwest, 0 otherwise)} \\
South &= \text{binary variable (1 if Region = South, 0 otherwise)} \\
West &= \text{binary variable (1 if Region = West, 0 otherwise)}
\end{aligned}
$$

Exercise 6.1. Compute $\overline{R}^2$ for each of the regressions.

Solution. solution 1

Exercise 6.2. Using the regression results in column (1):

(a) Do workers with college degrees earn more, on average, than workers with only high school diplomas? How much more?

(b) Do men earn more than women, on average? How much more?

Solution. (a)

(b)

Exercise 6.3. Using the regression results in column (2):

(a) Is age an important determinant of earnings? Explain.

(b) Sally is a 29-year-old female college graduate. Betsy is a 34-year-old female college graduate. Predict Sally's and Betsy's earnings.

Solution. (a)

(b)

Exercise 6.4. Using the regression results in column (3):

(a) Do there appear to be important regional differences?

(b) Why is the regressor *West* omitted from the regression? What would happen if it were included?

(c) Juanita is a 28-year-old female college graduate from the South. Jennifer is a 28-year-old female college graduate from the Midwest. Calculate the expected difference in earnings between Juanita and Jennifer.

Solution. (a)

(b)

(c)

Exercise 6.6. A researcher plans to study the causal effect of a strong legal system on the number of scandals in a country, using data from a random sample of countries in Asia. The researcher plans to regress the number of scandals on how strong a legal system is in the countries (an indicator variable taking the value 1 or 0, based on expert opinion).

(a) Do you think this regression suffers from omitted variable bias? Explain why. Which variables would you add to the regression?

(b) Using the expression for omitted variable bias given in Equation (6.1), assess whether the regression will likely over- or underestimate the effect of a strong legal system on the number of scandals in a country. That is, do you think that $\hat{\beta}_1 > \beta_1$ or $\hat{\beta}_1 < \beta_1$?

Solution. (a)

(b)

Exercise 6.8. A government study found that people who eat chocolate frequently weigh less than people who don't. Researchers questioned 1000 individuals from Cairo between the ages of 20 and 85 about their eating habits, and measured their weight and height. On average, participants ate chocolate twice a week and had a body mass index (BMI) of 28. There was an observed difference of five to seven pounds in weight between those who ate chocolate five times a week and those who did not eat any chocolate at all, with the chocolate eaters weighing less on average. Frequent chocolate eaters also consumed more calories, on average, than people who consumed less chocolate. Based on this summary, would you recommend that Egyptians who do not presently eat chocolate should consider eating chocolate up to five times a week if they want to lose weight? Why or why not? Explain.

Solution. solution 8

Exercise 6.11 (Optional). Consider the regression model $Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$ for $i = 1, \cdots, n$. Notice that there is no constant term in the regression. Following analysis like that used in Appendix 4.2:

(a) Specify the least squares function that is minimized by OLS.

(b) Compute the partial derivatives of the objective function with respect to b_1 and b_2.

(c) Suppose that $\sum_{i=1}^n X_{1i}X_{2i} = 0$. Show that $\hat{\beta}_1 = \sum_{i=1}^n X_{1i}Y_i / \sum_{i=1}^n X_{1i}^2$.

(d) Suppose that $\sum_{i=1}^n X_{1i}X_{2i} \neq 0$. Derive an expression for $\hat{\beta}_1$ as a function of the data (Y_i, X_{1i}, X_{2i}), $i = 1, \cdots, n$.

(e) Suppose that the model includes an intercept: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$. Show that the least squares estimators satisfy $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}_1 - \hat{\beta}_2 \overline{X}_2$.

(f) As in (e), suppose that the model contains an intercept. Also suppose that $\sum_{i=1}^n (X_{1i} - \overline{X}_1)(X_{2i} - \overline{X}_2) = 0$. Show that $\hat{\beta}_1 = \sum_{i=1}^n (X_{1i} - \overline{X}_1)(Y_i - \overline{Y}) / \sum_{i=1}^n (X_{1i} - \overline{X}_1)^2$. How does this compare to the OLS estimator of β_1 from the regression that omits X_2?

Solution. (a)

(b)

(c)

(d)

(e)

(f)

Empirical Exercise 6.1. Use the `Birthweight_Smoking` data set introduced in Empirical Exercise E5.3 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, 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 $\overline{R}^2$. 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 (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)?

Solution. (a)

(b) (i)

(ii)

(iii)

(iv)

(c)

(d) (i)

(ii)

(iii)

(iv)