Problem Set 6

Junwoo Yang

May 24, 2021

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

Answer. By the definition of $\overline{R}^2$,

$$\overline{R}^2 = 1 - \frac{n-1}{n-k-1}\frac{SSR}{TSS} = 1 - \frac{n-1}{n-k-1}(1-R^2),$$

Those are 0.1647672, 0.1816579, 0.1843181 in order. □

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?

Answer. (a) Workers with college degrees earn \$10.47/hour more, on average, than workers with only high school diplomas, holding sex constant.

(b) Men earn \$4.69/hour more, on average, than women, holding educational attainment constant.

□

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.

Answer. (a) Workers earn \$0.61 hourly per year on average.

(b) The prediction of Sally's earnings is $0.11 + 10.44 - 4.56 + 0.61 \times 29 = 23.68$ dollars per hour. The prediction of Betsy's earnings is $0.11 + 10.44 - 4.56 + 0.61 \times 34 = 26.73$ dollars per hour. The difference is same with $5\hat{\beta}_3 = 5 \times 0.61 = 3.05$.

□

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.

Answer. (a) Compared to workers in West, workers in Northeast earn \$0.74 more per hour, workers in Midwest \$1.54 less, workers in South \$0.44 less, on average, controlling for other variables in the regression.

(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 $22.49 - 21.39 = 1.1$.

□

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 $\hat{\beta}_1 < \beta_1$?

Answer. (a) Yes. There are many factors that determine the number of scandals taking place in a country, including the presence of an effective judiciary. Other missing variables include the level of education in the country (which will dictate how many individuals practice law) and the level of technological development.

(b) Suppose that the number of scandals in an Asian country is negatively affected by the share of lawyers in the population, such that with an increase in the rate of occurrence of scandals, the lawyers are more likely to create a stronger legal system. In this case, a strong legal system is likely to be positively correlated with the fraction of lawyers in the population, leading to a negative value for the omitted variable bias so that $\hat{\beta}_1 < \beta_1$. If the effect of a strong legal system, β_1, is negative, the effect will likely overstated by omitting the fraction of lawyers in the population.

□

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.

Answer. Omitted from the analysis are reasons why the survey respondents ate more or less chocolate than average. People who spend a lot of time working out may have higher metabolism, which allows them to eat more chocolate without putting on weight. On the other hand, people who are overweight may eat less chocolate in order to avoid putting on more weight. This study says nothing about the causal effect of chocolate consumption on weight. □

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?

Answer. (a) $\sum_{i=1}^n (Y_i - b_1 X_{1i} - b_2 X_{2i})^2$

(b) The partial derivatives are

$$\frac{\partial}{\partial b_1}\sum_{i=1}^{n}(Y_i - b_1X_{1i} - b_2X_{2i})^2 = -2\sum_{i=1}^{n}(Y_i - b_1X_{1i} - b_2X_{2i})X_{1i}$$
$$\frac{\partial}{\partial b_2}\sum_{i=1}^{n}(Y_i - b_1X_{1i} - b_2X_{2i})^2 = -2\sum_{i=1}^{n}(Y_i - b_1X_{1i} - b_2X_{2i})X_{1i}$$

(c) From (b), $\hat{\beta}_1$ satisfies $\sum X_{1i}(Y_i - b_1X_{1i} - b_2X_{2i}) = 0$, and so

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

(d) Following analysis as in (c)

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

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 $\sum(Y_i - b_0 - b_1X_{1i} - b_2X_{2i})^2$ and the partial derivative with respect to b_0 is

$$\frac{\partial\sum(Y_i - b_0 - b_1X_{1i} - b_2X_{2i})^2}{\partial b_0} = -2\sum(Y_i - b_0 - b_1X_{1i} - b_2X_{2i}).$$

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) Substituting $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\overline{X}_1 - \hat{\beta}_2\overline{X}_2$ into the least squares objective function in (e) yields

$$\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - b_1X_{1i} - b_2X_{2i})^2 = \sum_{i=1}^{n}\Big((Y_i - \overline{Y}) - b_1(X_{1i} - \overline{X}_1) - b_2(X_{2i} - \overline{X}_2)\Big)^2,$$

which is identical to the least squares objective function in (a), except that all variables have been replaced with deviations from sample means. The result then follows as in (c). When

$$\sum_{i=1}^{n}((X_{1i} - \overline{X}_1)(X_{2i} - \overline{X}_2) = 0,$$

the estimated coefficient on X_1 in the OLS regression of Y on both X_1 and X_2 is the same as the estimated coefficient in the OLS regression of Y on X_1 only.

□

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)?

Answer. (a) $\hat{\beta}_1 = -253.228$

(b) (i)

(ii) The estimated slope is -253.228. The coefficient changes by a large amount. Thus, the regression in (a) suffer from omitted variable bias.

(iii) The prediction of the birth weight of Jane's child is

$$\hat{Y} = 3051.2486 - 217.5801 \times 1 - 30.4913 \times 0 + 34.0699 \times 8 = 3106.228.$$

(iv) $R^2 = 0.07285$ and $\overline{R}^2 = 0.07192$ are similar because of 3,000 samples which are large enough to say that there is no difference between $n - 1$ and $n - k - 1$.

(v) That coefficient is partial effect of *Nprevist* holding *Smoker* and *Alcohol* constant.

(c)

(d) (i)

(ii)

(iii)

(iv)

□

Appendix: R code

```r
# (a)
library(readxl)
bs <- read_excel("birthweight_smoking/birthweight_smoking.xlsx")
ols1 <- lm(birthweight ~ smoker, data = bs)
library(sandwich)
Cov1 <- vcovHC(ols1, type = "HC2")
library(lmtest)
coeftest(ols1, vcov. = Cov1)

##
## t test of coefficients:
##
##             Estimate Std. Error  t value  Pr(>|t|)
## (Intercept) 3432.060     11.889 288.6746 < 2.2e-16 ***
## smoker      -253.228     26.821  -9.4414 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

#coefci(ols1, vcov. = Cov1)

# (b)
ols2 <- lm(birthweight ~ smoker + alcohol + nprevist, data = bs)
Cov2 <- vcovHC(ols2, type = "HC2")
coeftest(ols2, vcov. = Cov2)

##
## t test of coefficients:
##
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3051.2486    43.7543 69.7360   <2e-16 ***
## smoker      -217.5801    26.1244 -8.3286   <2e-16 ***
## alcohol      -30.4913    73.2105 -0.4165   0.6771
## nprevist      34.0699     3.6123  9.4315   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

#coefci(ols2, vcov. = Cov2)
summary(ols2)

##
## Call:
## lm(formula = birthweight ~ smoker + alcohol + nprevist, data = bs)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2733.53  -307.57    21.42   358.09  2192.70
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3051.249     34.016  89.701  < 2e-16 ***
## smoker      -217.580     26.680  -8.155 5.07e-16 ***
## alcohol      -30.491     76.234  -0.400    0.689
## nprevist      34.070      2.855  11.933  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 570.5 on 2996 degrees of freedom
## Multiple R-squared:  0.07285,Adjusted R-squared:  0.07192
## F-statistic: 78.47 on 3 and 2996 DF,  p-value: < 2.2e-16

# (c)
```