Econometrics – Problem Set #5

Junwoo Yang

May 8, 2021

Exercise 5.2. Suppose that a researcher, using wage data on 200 randomly selected male workers and 240 female workers, estimates the OLS regression

$$\widehat{Wage} = 10.73\,(0.16) + 1.78\,(0.29) \times Male, \quad R^2 = 0.09, \quad SER = 3.8,$$

where *Wage* is measured in dollars per hour and *Male* is a binary variable that is equal to 1 if the person is a male and 0 if the person is a female. Define the wage gender gap as the difference in mean earnings between men and women.

(a) What is the estimated gender gap?

(b) Is the estimated gender gap significantly different from 0? (Compute the p-value for testing the null hypothesis that there is no gender gap.)

(c) Construct a 95% confidence interval for the gender gap.

(d) In the sample, what is the mean wage of women? Of men?

(e) Another researcher uses these same data but regresses *Wages* on *Female*, a variable that is equal to 1 if the person is female and 0 if the person a male. What are the regression estimates calculated from this regression?

$$\widehat{Wage} = \underline{\qquad} + \underline{\qquad} \times Female, \quad R^2 = \underline{\qquad}, \quad SER = \underline{\qquad}.$$

Solution. (a) The estimated gender gap is $\hat{\beta}_1 = \$1.78$ per hour. ■

(b) The t-statistic and the p-value are

$$t^{\text{act}} = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} = \frac{1.78}{0.29} = 6.137931, \quad p\text{-value} = 2\Phi(-|t^{\text{act}}|) = 8.360316 \times 10^{-10}.$$

Thus, $H_0 \colon \beta_1 = 0$ can be rejected at very low significance level. ■

(c) The 95% confidence interval for the gender gap is

$$(\hat{\beta}_1 - z_{0.025}\, SE(\hat{\beta}_1), \hat{\beta}_1 + z_{0.025}\, SE(\hat{\beta}_1)) = (1.21161, 2.34839).$$ ■

(d) The sample mean wage of women is $\hat{\beta}_0 = \$10.73$ per hour, and the same of men is $\hat{\beta}_0 + \hat{\beta}_1 = \$12.51$ per hour. ■

(e) The constant term and the coefficient of *Female* should be

$$E(Wage_i | Female_i = 0) = E(Wage_i | Male_i = 1) = \hat{\beta}_0 + \hat{\beta}_1 = 12.51,$$
$$E(Wage_i | Female_i = 1) - E(Wage_i | Female_i = 0) = -\hat{\beta}_1 = -1.78,$$

respectively. Thus, $\widehat{Wage} = 12.51 - 1.78 \times Female$, and the regression R^2 and SER does not change. ■

Exercise 5.4. Read the box "The Economic Value of a Year of Education: Homoskedasticity or Heteroskedasticity?" in Section 5.4. Use the regression

$$\widehat{Earnings} = -12.12\,(1.36) + 2.37\,(0.10) \times \textit{Years Education}, \quad R^2 = 0.185, \quad SER = 11.24 \quad (5.23)$$

to answer the following.

(a) A randomly selected 30-year-old worker reports an education level of 16 years. What is the worker's expected average hourly earnings?

(b) A high school graduate (12 years of education) is contemplating going to a community college for a 2-year degree. How much are this worker's average hourly earnings expected to increase?

(c) A high school counselor tells a student that, on average, college graduates earn \$10 per hour more than high school graduates. Is this statement consistent with the regression evidence? What range of values is consistent with the regression evidence?

Solution. (a) $E(Earnings_i | \textit{Years Education}_i = 16) = -12.12 + 2.37 \times 16 = \$25.8$ per hour. ■

(b) The expected increase is $2\hat{\beta}_1 = 2 \times 2.37 = \$4.74$ per hour. ■

(c) The hypothesis testing for earning gap is $H_0\colon \beta_1 = 2.5$ vs. $H_1\colon \beta_1 \neq 2.5$. The t-statistic is

$$t^{\text{act}} = \frac{\hat{\beta}_1 - 2.5}{SE(\hat{\beta}_1)} = \frac{2.37 - 2.5}{0.1} = -1.3,$$

and p-value is 0.193601. Thus, the counselor's assertion is acceptable at 10% significance level. A 90% confidence interval for $4\beta_1$ is

$$4 \times (\hat{\beta}_1 - z_{0.05}\, SE(\hat{\beta}_1), \hat{\beta}_1 + z_{0.05}\, SE(\hat{\beta}_1)) = (8.822059, 10.13794).$$

Note that $4\beta_1 = \$10$ is already contained in the interval. ■

Exercise 5.10. Let X_i denote a binary variable, and consider the regression $Y_i = \beta_0 + \beta_1 X_i + u_i$. Let $\overline{Y}_0$ denote the sample mean for observations with $X = 0$, and let $\overline{Y}_1$ denote the sample mean for observations with $X = 1$. Show that $\hat{\beta}_0 = \overline{Y}_0$, $\hat{\beta}_0 + \hat{\beta}_1 = \overline{Y}_1$, and $\hat{\beta}_1 = \overline{Y}_1 - \overline{Y}_0$.

Solution. Let n_0 be the number of observation with $X = 0$ and n_1 be the same with $X = 1$. Then,

- $\sum_{i=1}^n X_i = \sum_{i=1}^n X_i^2 = n_1, \quad \overline{X} = \frac{n_1}{n} \quad \overline{Y}_1 = \frac{1}{n_1} \sum_{i=1}^n X_i Y_i$
- $\sum_{i=1}^n (X_i - \overline{X})^2 = \sum_{i=1}^n X_i^2 - n\overline{X}^2 = n_1 - \frac{n_1^2}{n} = \frac{n_0 n_1}{n}$
- $\overline{Y} = \frac{1}{n} \sum_{i=1}^n Y_i = \frac{1}{n}(n_0 \overline{Y}_0 + n_1 \overline{Y}_1) = \frac{n_0}{n} \overline{Y}_0 + \frac{n_1}{n} \overline{Y}_1$

Thus,

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^n X_i (Y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2} = \frac{\sum_{i=1}^n X_i Y_i - n_1 \overline{Y}}{n_0 n_1 / n} \\
&= \frac{n}{n_0}(\overline{Y}_1 - \overline{Y}) = \frac{n}{n_0}\left(\overline{Y}_1 - \frac{n_1}{n}\overline{Y}_1 - \frac{n_0}{n}\overline{Y}_0\right) = \overline{Y}_1 - \overline{Y}_0 \\
\hat{\beta}_0 &= \overline{Y} - \hat{\beta}_1 \overline{X} = \left(\frac{n_0}{n}\overline{Y}_0 + \frac{n_1}{n}\overline{Y}_1\right) - (\overline{Y}_1 - \overline{Y}_0)\frac{n_1}{n} = \overline{Y}_0.
\end{aligned}$$

■

Exercise 5.11. A random sample of workers contains $n_m = 100$ men and $n_w = 150$ women. The sample average of men's weekly earnings $\overline{Y}_m$ is €565.89, and the standard deviation s_m is €75.62. The corresponding values for women are $\overline{Y}_w =$ €502.37 and $s_w =$ €53.40. Let *Women* denote an indicator variable that is equal to 1 for women and 0 for men, and suppose that all of 250 observations are used in the regression $Y_i = \beta_0 + \beta_1 Women + u_i$. Find the OLS estimates of β_0 and β_1 and their corresponding standard errors.

Solution. By Exercise 5.10, $\hat{\beta}_1 = \overline{Y}_w - \overline{Y}_m = -63.52$ and $\hat{\beta}_0 = \overline{Y}_m = 565.89$. The standard errors are

$$SE(\hat{\beta}_1) = SE(\overline{Y}_w - \overline{Y}_m) = \sqrt{\frac{s_w^2}{n_w} + \frac{s_m^2}{n_m}} = 8.728931, \quad SE(\hat{\beta}_0) = SE(\overline{Y}_m) = \frac{s_m}{\sqrt{n_m}} = 7.562. \quad \blacksquare$$

Empirical exercise 5.3. On the text website, `http://www.pearsonglobaleditions.com`, you will find the data file `Birthweight_Smoking`, which contains data for a random sample of babies born in Pennsylvania in 1989. The data include the baby's birth weight together with various characteristics of the mother, including whether she smoked during the pregnancy. A detailed description is given in `Birthweight_Smoking_Description`, also available on the website. In this exercise, you will investigate the relationship between birth weight and smoking during pregnancy.

(a) In the sample:

  (i) What is the average value of *Birthweight* for all mothers?

  (ii) For mothers who smoke?

  (iii) For mothers who do not smoke?

(b) (i) Use the data in the sample to estimate the difference in average birth weight for smoking and nonsmoking mothers.

  (ii) What is the standard error for the estimated difference in (i)?

  (iii) Construct a 95% confidence interval for the difference in the average birth weight for smoking and nonsmoking mothers.

(c) Run a regression of *Birthweight* on the binary variable *Smoker*.

  (i) Explain how the estimated slope and intercept are related to your answers in parts (a) and (b).

  (ii) Explain how the $SE(\hat{\beta}_1)$ is related to your answer in (ii) of (b).

  (iii) Construct a 95% confidence interval for the effect of smoking on birth weight.

(d) Do you think smoking is uncorrelated with other factors that cause low birth weight? That is, do you think that the regression error term—say, u_i—has a conditional mean of 0 given Smoking (X_i)?

Solution. (a)

```
library(readxl)
bs <- read_excel("birthweight_smoking/birthweight_smoking.xlsx")
mean(bs$birthweight)

## [1] 3382.934

library(dplyr)
avgs <- bs %>%
  group_by(smoker) %>%
  summarise(mean=mean(birthweight), sd=sd(birthweight), n=n())
print(avgs)

## # A tibble: 2 x 4
##   smoker  mean    sd     n
##    <dbl> <dbl> <dbl> <int>
## 1      0 3432.  585.  2418
## 2      1 3179.  580.   582
```

(i) The average value of *Birthweight* for all mothers, say $\overline{Y}$, is 3382.9336667. ■

(ii) $\overline{Y}_1 = 3178.83161512027$. ■

(iii) $\overline{Y}_0 = 3432.05996691481$. ■

(b)

```
avgs_1 <- avgs %>% filter(smoker==1)
avgs_0 <- avgs %>% filter(smoker==0)
gap <- avgs_1$mean - avgs_0$mean
gap_se <- sqrt(avgs_1$sd^2/avgs_1$n + avgs_0$sd^2/avgs_0$n)
gap_cil <- gap - qnorm(0.975)*gap_se
gap_ciu <- gap + qnorm(0.975)*gap_se
result <- cbind(gap, gap_se, gap_cil, gap_ciu)
print(result, digits = 5)

##          gap gap_se gap_cil gap_ciu
## [1,] -253.23 26.821  -305.8 -200.66
```

(i) $\overline{Y}_1 - \overline{Y}_0 = -253.2283518$. ■

(ii) $SE(\overline{Y}_1 - \overline{Y}_0) = 26.8210626$. ■

(iii) The 95% confidence interval for difference is $(-305.7966686, -200.660035)$. ■

(c)

```
ols <- lm(birthweight ~ smoker, data = bs)
library(sandwich)
Cov <- vcovHC(ols, type = "HC2")
library(lmtest)
coeftest(ols, vcov. = Cov)

##
## 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(ols, vcov. = Cov)

##                 2.5 %    97.5 %
## (Intercept) 3408.7485 3455.3714
## smoker      -305.8179 -200.6388
```

(i) $\hat{\beta}_0 = 3432.0599669$ is exactly same with $\overline{Y}_0$ in (iii) of (a), and $\hat{\beta}_1 = -253.2283518$ is exactly same with $\overline{Y}_1 - \overline{Y}_0$ in (i) of (b). ■

(ii) Since $\hat{\beta}_1 = \overline{Y}_1 - \overline{Y}_0$, the heteroskedasticity-robust standard error $SE(\hat{\beta}_1) = 26.821$ is almost same with $SE(\overline{Y}_1 - \overline{Y}_0)$ in (ii) of (b). ■

(iii) The 95% confidence interval generated by `coefci` is $(-305.8179001, -200.6388035)$. ■

(d)
```r
ols_bt <- lm(birthweight ~ tripre0, data = bs)
Cov_bt <- vcovHC(ols_bt, type = "HC2")
coeftest(ols_bt, vcov. = Cov_bt)
```

```
##
## t test of coefficients:
##
##             Estimate Std. Error  t value  Pr(>|t|)
## (Intercept) 3390.282     10.733 315.8746 < 2.2e-16 ***
## tripre0     -734.882    151.159  -4.8617 1.224e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
t.test(subset(bs, tripre0==1)$birthweight, # no prenatal visits
       subset(bs, tripre0==0)$birthweight, var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  subset(bs, tripre0 == 1)$birthweight and subset(bs, tripre0 == 0)$birthweight
## t = -4.8617, df = 29.295, p-value = 3.639e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1043.9017  -425.8632
## sample estimates:
## mean of x mean of y
##  2655.400  3390.282
```

No, birth weight is also related to whether the mother visits prenatal care. Mothers who didn't receive prenatal care have a lower average birth weight than mothers who received. Thus, there are omitted variable bias and the first least squares assumption $E(u_i|X_i) = 0$ doesn't hold. ■