Empirical Exercise - E7.2

Chi-Yuan Fang

2021-03-30

In the empirical exercises on earning and height in Chapters 4 and 5, you estimated a relatively large and statistically significant effect of a worker's height on his or her earnings. One explanation for this result is omitted variable bias: Height is correlated with an omitted factor that affects earnings. For example, Case and Paxson (2008) suggest that cognitive ability (or intelligence) is the omitted factor. The mechanism they describe is straightforward: Poor nutrition and other harmful environmental factors in utero and in early childhood have, on average, deleterious effects on both cognitive and physical development. Cognitive ability affects earnings later in life and thus is an omitted variable in the regression.

a. Suppose that the mechanism described above is correct. Explain how this leads to omitted variable bias in the OLS regression of *Earnings* on *Height*. Does the bias lead the estimated slope to be too large or too small? [Hint: Review Equation (6.1).]

Remark

Given true model

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + u. \tag{1}$$

Z is omitted variable if (1) $Cov(X, Z) \neq 0$, and (2) $\beta_2 \neq 0$. Fit the following models:

$$\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 X \tag{2}$$

$$\widehat{Y} = \widetilde{\beta}_0 + \widetilde{\beta}_1 X + \widetilde{\beta}_2 Z \tag{3}$$

$$\widehat{Z} = \widehat{\delta}_0 + \widehat{\delta}_1 X \tag{4}$$

where $\widetilde{\beta}_i$ is unbiased estimator of β_i, and $\widehat{\beta}_i$ is biased estimator of β_i. Thus,

$$\widehat{\beta}_1 = \widetilde{\beta}_1 + \widehat{\delta}_1 \times \widetilde{\beta}_2. \tag{5}$$

Solution

Because we know $Cov(height, cognitive\_ability) \neq 0$ and $Cov(cognitive\_ability, earning) \neq 0$, the simple regression model suffers from an omitted variables problem. Moreover, we think $Cov(height, cognitive\_ability) > 0$ and $Cov(cognitive\_ability, earning) > 0$, so $\widehat{\beta}_{height}$ is more positive than β_{height} (underestimated).

If the mechanism described above is correct, the estimated effect of height on earnings should disappear if a variable measuring cognitive ability is included in the regression. Unfortunately, there isn't a direct measure of cognitive ability in the data set, but the data set does include years of education for each individual. Because students with higher cognitive ability are more likely to attend school longer, years of education might serve as a control variable for cognitive ability; in this case, including education in the regression will eliminate, or at least attenuate, the omitted variable bias problem.

Use the years of education variable (educ) to construct four indicator variables for whether a worker has less than a high school diploma ($LT\_HS = 1$ if $educ < 12$, 0 otherwise), a high school

diploma ($HS = 1$ if $educ = 12$, 0 otherwise), some college ($Some\_Col = 1$ if $12 < educ < 16$, 0 otherwise), or a bachelor's degree or higher ($College = 1$ if $educ \geq 16$, 0 otherwise).

b. Focusing first on women only, run a regression of (1) *Earnings* on *Height* and (2) *Earnings* on *Height*, including LT_HS, HS, and $Some\_Col$ as control variables.
   i. Compare the estimated coefficient on *Height* in regressions (1) and (2). Is there a large change in the coefficient? Has it changed in a way consistent with the cognitive ability explanation? Explain.
   ii. The regression omits the control variable *College*. Why?
   iii. Test the joint null hypothesis that the coefficients on the education variables are equal to 0.
   iv. Discuss the values of the estimated coefficients on LT_HS, HS, and $Some\_Col$. (Each of the estimated coefficients is negative, and the coefficient on LT_HS is more negative than the coefficient on HS, which in turn is more negative than the coefficient on $Some\_Col$. Why? What do the coefficients measure?)

Solution

```
# https://cran.r-project.org/web/packages/stargazer/stargazer.pdf
# https://www.jakeruss.com/cheatsheets/stargazer/
# https://cran.r-project.org/web/packages/stargazer/vignettes/stargazer.pdf

library(sandwich); library(lmtest)
library(stargazer)
library(car)
```

```
# import data
library(readxl)
Earnings_and_Height <- read_xlsx("Earnings_and_Height/Earnings_and_Height.xlsx")

# create dummay variables
LT_HS <- c()

for (i in 1:length(Earnings_and_Height$educ)){
  if (Earnings_and_Height$educ[i] < 12){
    LT_HS[i] <- c(1)
  } else {
    LT_HS[i] <- c(0)
  }
}

HS <- c()

for (i in 1:length(Earnings_and_Height$educ)){
  if (Earnings_and_Height$educ[i] == 12){
    HS[i] <- c(1)
  } else {
    HS[i] <- c(0)
  }
}

Some_Col <- c()

for (i in 1:length(Earnings_and_Height$educ)){
  if (Earnings_and_Height$educ[i] > 12 & Earnings_and_Height$educ[i] < 16){
    Some_Col[i] <- c(1)
```

```
  } else {
    Some_Col[i] <- c(0)
  }
}


College <- c()

for (i in 1:length(Earnings_and_Height$educ)){
  if (Earnings_and_Height$educ[i] >= 16){
    College[i] <- c(1)
  } else {
    College[i] <- c(0)
  }
}

Earnings_and_Height <- cbind(Earnings_and_Height, LT_HS, HS, Some_Col, College)

# correlation matrix
Earnings_and_Height_interest <- Earnings_and_Height[,c("earnings", "height", "LT_HS", "HS", "Some_Col")]

cor(Earnings_and_Height_interest)
```

```
##              earnings      height       LT_HS          HS    Some_Col
## earnings  1.00000000  0.10428470 -0.22769147 -0.18970620  0.02474663
## height    0.10428470  1.00000000 -0.06625784 -0.04820749  0.01787849
## LT_HS    -0.22769147 -0.06625784  1.00000000 -0.25273987 -0.19240154
## HS       -0.18970620 -0.04820749 -0.25273987  1.00000000 -0.43467287
## Some_Col  0.02474663  0.01787849 -0.19240154 -0.43467287  1.00000000
```

```
# data # women
Earnings_and_Height0 <- Earnings_and_Height[Earnings_and_Height$sex == 0,]

# Model 1
E72_M1 <- lm(formula = earnings ~ height,
             data = Earnings_and_Height0)

# Mdoel 2
E72_M2 <- lm(formula = earnings ~ height + LT_HS + HS + Some_Col,
             data = Earnings_and_Height0)

# Adjust standard errors
cov1 <- vcovHC(E72_M1, type = "HC1")
rb_se1 <- sqrt(diag(cov1))

cov2 <- vcovHC(E72_M2, type = "HC1")
rb_se2 <- sqrt(diag(cov2))

# output table
stargazer(E72_M1, E72_M2,
          type = "text",
          column.labels = c("Women", "Women"),
          se = list(rb_se1, rb_se2),
          digits = 4)
```

```
##
## ========================================================================
##                                     Dependent variable:
##                         ---------------------------------------------
##                                           earnings
##                                  Women                    Women
##                                   (1)                      (2)
## ------------------------------------------------------------------------
## height                         511.2222***               135.1421
##                                 (97.5846)                (92.3164)
##
## LT_HS                                                  -31,857.8100***
##                                                           (834.9586)
##
## HS                                                     -20,417.8900***
##                                                           (637.8055)
##
## Some_Col                                               -12,649.0700***
##                                                           (716.5866)
##
## Constant                      12,650.8600**             50,749.5200***
##                               (6,299.1510)               (6,003.8190)
##
## ------------------------------------------------------------------------
## Observations                     9,974                     9,974
## R2                               0.0027                    0.1382
## Adjusted R2                      0.0026                    0.1378
## Residual Std. Error  26,800.9000 (df = 9972)   24,917.3800 (df = 9969)
## F Statistic         26.7214*** (df = 1; 9972) 399.6163*** (df = 4; 9969)
## ========================================================================
## Note:                                        *p<0.1; **p<0.05; ***p<0.01
```

i. The estimated coefficient on height falls by approximately 73.5649% when the education variables are added as control variables in the regression. This result coincides with our expectation.

```
(E72_M2$coefficients[2] - E72_M1$coefficients[2]) / E72_M1$coefficients[2]
```

```
##    height
## -0.735649
```

ii. *College* is perfectly collinear with other education regressors and the constant regressor.

iii.

- **Prepare**

  $H_0 : \beta_{LT\_HS} = \beta_{HS} = \beta_{Some\_Col} = 0$ v.s. H_1 : not H_0

  Let the significance level be 0.05.

- **Calculate**

```
linearHypothesis(E72_M2,
                 c("LT_HS = 0", "HS = 0", "Some_Col = 0"),
                 test = "F")
```

```
## Linear hypothesis test
##
## Hypothesis:
```

```
## LT_HS = 0
## HS = 0
## Some_Col = 0
##
## Model 1: restricted model
## Model 2: earnings ~ height + LT_HS + HS + Some_Col
##
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1   9972 7.1628e+12
## 2   9969 6.1895e+12  3 9.7326e+11 522.52 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- **Conclusion**

  Because $p - value < \alpha = 5\%$, we reject H_0. There is significant evidence that educational differences exist.

iv.

- Other things being equal, workers with less than a high school education on average earn $31,857.81 less per year than a college graduate on average.
- Other things being equal, workers with a high school education on average earns $20,417.89 less per year than a college graduate on average.
- Other things being equal, workers with a some college on average earns $12,649.07 less per year than a college graduate on average.

c. Repeat (b), using data for men.

Solution

```
# data # men
Earnings_and_Height1 <- Earnings_and_Height[Earnings_and_Height$sex == 1,]

# Model 3
E72_M3 <- lm(formula = earnings ~ height,
             data = Earnings_and_Height1)

# Mdoel 4
E72_M4 <- lm(formula = earnings ~ height + LT_HS + HS + Some_Col,
             data = Earnings_and_Height1)

# Adjust standard errors
cov3 <- vcovHC(E72_M3, type = "HC1")
rb_se3 <- sqrt(diag(cov3))

cov4 <- vcovHC(E72_M4, type = "HC1")
rb_se4 <- sqrt(diag(cov4))

# output table
stargazer(E72_M3, E72_M4,
          type = "text",
          column.labels = c("Men", "Men"),
          se = list(rb_se3, rb_se4),
          digits = 4)
```

```
##
```

```
## ================================================================================
##                                       Dependent variable:
##                          ---------------------------------------------------
##                                             earnings
##                                    Men                          Men
##                                    (1)                          (2)
## --------------------------------------------------------------------------------
## height                         1,306.8600***                 744.6809***
##                                  (98.8569)                    (92.2615)
##
## LT_HS                                                      -31,400.4900***
##                                                               (869.6952)
##
## HS                                                         -20,345.8500***
##                                                               (701.6438)
##
## Some_Col                                                   -12,610.9200***
##                                                               (797.8023)
##
## Constant                      -43,130.3400***                9,862.7400
##                                 (6,925.0110)                (6,541.3220)
##
## --------------------------------------------------------------------------------
## Observations                       7,896                        7,896
## R2                                 0.0209                       0.1658
## Adjusted R2                        0.0207                       0.1654
## Residual Std. Error  26,671.2900 (df = 7894)    24,623.2200 (df = 7891)
## F Statistic         168.2010*** (df = 1; 7894) 392.0380*** (df = 4; 7891)
## ================================================================================
## Note:                                          *p<0.1; **p<0.05; ***p<0.01
```

i. The estimated coefficient on height falls by approximately 43.0175% when the education variables are added as control variables in the regression. This result coincides with our expectation.

```
(E72_M4$coefficients[2] - E72_M3$coefficients[2]) / E72_M3$coefficients[2]
```

```
##     height
## -0.4301754
```

ii. *College* is perfectly collinear with other education regressors and the constant regressor.

iii.

- **Prepare**

  $H_0 : \beta_{LT\_HS} = \beta_{HS} = \beta_{Some\_Col} = 0$ v.s. H_1 : not H_0

  Let the significance level be 0.05.

- **Calculate**

```
linearHypothesis(E72_M4,
                 c("LT_HS = 0", "HS = 0", "Some_Col = 0"),
                 test = "F")
```

```
## Linear hypothesis test
##
## Hypothesis:
## LT_HS = 0
```

```
## HS = 0
## Some_Col = 0
##
## Model 1: restricted model
## Model 2: earnings ~ height + LT_HS + HS + Some_Col
##
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1   7894 5.6155e+12
## 2   7891 4.7843e+12  3 8.3112e+11 456.94 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- **Conclusion**

  Because $p - value < \alpha = 5\%$, we reject H_0. There is significant evidence that educational differences exist.

iv.

- Other things being equal, workers with less than a high school education on average earn $31,400.49 less per year than a college graduate on average.
- Other things being equal, workers with a high school education on average earns $20,345.85 less per year than a college graduate on average.
- Other things being equal, workers with a some college on average earns $12,610.92 less per year than a college graduate on average.