Lab 5: Endogeneity and hypothesis testing

Julieth Santamaria

February 22, 2019

Summary[1]

Hypothesis testing

In economics, we usually want to test whether our estimates are significantly different from zero. In other words, the null hypothesis is $H_o : \hat{\beta} = 0$. We use t-statistics to assess our results. Two steps:

1. Calculate the t-statistic of your estimate (Often, $\beta_{null} = 0$)

$$t = \frac{\hat{\beta} - \beta_{null}}{SE(\hat{\beta})}$$

2. Then, compare the t-statistic you calculated to the critical value. The table below displays the critical value for each confidence level:

Level	Critical value
1%	2.58
5%	1.96
10%	1.65

Omitted variables

You decide to run this regression:	$Salary = \beta_0^{omit} + \beta_1^{omit} X_1 + u$
You forgot to include X_2:	$Salary = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$
The relation between X_1 and X_2 is:	$X_1 = \alpha + \delta X_2 + \epsilon$

Omitting a relevant variable will cause your estimates to be biased. The direction of the bias will be determined by the following formula

$$\beta_1^{omit} = \beta_1 + \beta_2 \delta$$

Measurement error

This happens when the observed value of a variable and the actual value of a variable differ. Suppose that instead of observing X_1^* you observe $X_1 = X_1^* + \eta$. Then your bias (which in this context is called attenuation bias) is the following:

$$plim\hat{\beta}_1 = \beta_1 \frac{\sigma^2_{X_1^*}}{\sigma^2_{X_1^*} + \sigma^2_\eta}$$

[1]Key terms are paraphrased or copied from Real Econometrics by Michael A. Bailey or from the class lectures by Joe Ritter

Application

1. Set the working directory and load nscg17.

```
setwd("G:/My Drive/U of M/TA/TA APEC3003/APEC 3003 - 2019/APEC 3003 R work/labs/")
load("../data/nscg17.rdata")
```

2. Run the following lines that we run last week.

```
nscg17 <- within(nscg17, {
  # Salary
  salary[salary >= 9999998 | salary==0] <- NA

  # Potential experience
  exper <- 2017-dgryr

  # Gender
  female <- NA
  female[gender=="F"]<-1
  female[gender=="M"]<-0
})
```

3. Create dummies for the variable education using the variable `dgrdg`

a. What type of variable is dgrdg?

b. Create the dummies

```
nscg17 <- within(nscg17, {
  bachelors<-NA
  bachelors<-as.numeric(dgrdg==1)
  masters<-NA
  masters<-as.numeric(dgrdg==2)
  phd<-NA
  phd<-as.numeric(dgrdg==3)
  professional<-NA
  professional<-as.numeric(dgrdg==4)
})
```

4. Create a dummy for whether the spouse works full time using the variable `spowk`

```
nscg17 <- within(nscg17, {
  spouse.works=NA
  spouse.works[spowk=="1"]<-1
  spouse.works[spowk %in% c("2","3")]<-0
  spouse.works[spowk=="L"]<-NA
})
```

5. Estimate the following equation:

$$Salary = \beta_0^{omit} + \beta_1^{omit} Experience + u$$

```
reg1<-lm(salary~exper,data=nscg17)
summary(reg1)
```

```
##
## Call:
## lm(formula = salary ~ exper, data = nscg17)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -124088  -38349  -11098   19940  953368
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 73322.17     555.04   132.1   <2e-16 ***
## exper         962.56      27.11    35.5   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 86460 on 69342 degrees of freedom
##   (14328 observations deleted due to missingness)
## Multiple R-squared:  0.01785,    Adjusted R-squared:  0.01784
## F-statistic:  1260 on 1 and 69342 DF,  p-value: < 2.2e-16
```

6. Calculate yourself the t statistic associated with the variable experience. Interpret the significance of the coefficient.

$$t = \frac{962.56}{27.11} = 35.5$$

It is significant at the 1% level.

7. You forgot to include the following set of variables.

a. For each one, determine the direction of the bias by filling out columns 2-3 of the table below

Omitted variable	β_2	δ	Direction ($\beta_2\delta$)	$\hat{\beta}_1$
Female	-	-	+ ($\beta_1^{omit} > \beta_1$)	822.86
Spouse works	-	-	+ ($\beta_1^{omit} > \beta_1$)	692.35
Education	+	-	- ($\beta_1^{omit} < \beta_1$)	1004.95
Weather	0	0	0	-

b. Run a regression that includes each of the omitted variables and record the coefficient associated with experience in the table (column 4). Check whether your hypothesis holds.

```
## Table 1: Regressions of salary on experience
## ================================================================
##                      [a]          [b]          [c]          [d]
##  (Intercept) 73322.17***  88185.96***  95653.05*** 134634.01***
##                (473.22)     (566.77)     (971.78)    (3136.58)
##        exper   962.56***    822.86***    692.35***   1004.95***
##                (30.73)      (29.93)      (36.61)      (30.23)
##       female             -28434.78***
##                            (613.78)
## spouse.works                          -19298.73***
##                                         (922.24)
##    bachelors                                       -73450.99***
##                                                     (3221.21)
##      masters                                       -59360.62***
##                                                     (3211.24)
##          phd                                       -42239.89***
##                                                     (3379.53)
##            N    69344        69344        51109        69344
```

```
##    r.squared    0.018        0.044        0.022        0.054
## ----------------------------------------------------------------
## Significance: 0.01=***, 0.05=**, 0.1=*
## Note: This table displays the summarized results of running a
## regression of salary on experienca and the omitted variables
## listed above.
```