Lab 4: Regression analysis

Julieth Santamaria

February 15, 2019

Summary

Key Terms[1]

- Predicted value or fitted value ($\hat{Y}$): Is the value of Y predicted by the estimated equation.
- Endogeneity: An independent variable is endogenous if changes in it are related to factors in the error term. Three reasons:
  - Omitted variable: Leaving out a variable that affects the dependent variable and is correlated with the independent variable
  - Measurement error: X is measured inaccurately
  - Reverse causality: X explains Y, and Y explains X
- Exogeneity: The opposite of endogeneity. An independent variable is exogenous if changes in it are unrelated to factors in the error term.
- R-squared: Is a measure of goodness of fit. It ranges from 0 to 1. A high R^2 means the predicted values are close to the actual ones. But be careful, a high R^2 is neither neccessary not sufficient condition for an analysis to be successul.

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. Select the sample of social scientists

a. Open the codebook and look for the codes of social scientists using the variable `n2ocpr`. Work with the people at your table to find 6 codes.

b. Create a column vector with those codes called "'soc.sci.list""

c. Create a subset of nscg17 that only contains social scientists

```
soc.sci.list <- c("412320","422350","432360","442310","442370","452380") # b
nscg17.soc.sci <- subset(nscg17, n2ocpr %in% soc.sci.list) # c
```

3. Run the following lines, what do they do? Discuss in groups.

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

  # Potential experience
  exper <- 2017-dgryr

  # Gender
  female <- NA
  female[gender=="F"]<-1
```

[1] Key terms are paraphrased or copied from Real Econometrics by Michael A. Bailey

```r
  female[gender=="M"]<-0
})
```

4. Make a regression of salary on experience

```r
reg1 <-lm(salary~exper,data=nscg17.soc.sci)
summary(reg1)
```

```
##
## Call:
## lm(formula = salary ~ exper, data = nscg17.soc.sci)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -107424  -34551  -11827   17974  953581
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  73277.5     3089.1  23.722  < 2e-16 ***
## exper          794.7      160.0   4.968 7.35e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 82990 on 1948 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.01251,    Adjusted R-squared:  0.01201
## F-statistic: 24.68 on 1 and 1948 DF,  p-value: 7.349e-07
```

a. Write down the estimated model. Identify the intercept and the slope. Interpret.
   $Salary = \hat{\beta}_0 + \hat{\beta}_1 Experience + \hat{u}$ or $\hat{Salary} = \hat{\beta}_0 + \hat{\beta}_1 Experience$

Elements in the output:

- Dependent variable: salary
- Independent variable: exper
- $\hat{\beta}_0 = 73277.5$
- $SE(\hat{\beta}_0) = 3089.1$
- $\hat{\beta}_1 = 794.7$
- $SE(\hat{\beta}_1) = 160.0$

For each additional year of experience, annual salary increases by 794.7 dollars.

b. Is experience exogenous? Why or why no?

Think of the three forms of endogeneity: - Omitted variables - Measurement error - Reverse causality

c. Include a dummy for being female in your regression. Compare the coefficient associated with experience. Why did it change?

```r
reg2 <-lm(salary~exper+female,data=nscg17.soc.sci)
summary(reg2)
```

```
##
## Call:
## lm(formula = salary ~ exper + female, data = nscg17.soc.sci)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -117615  -35470  -11026   18572  959848
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  90595.5     4145.1  21.856  < 2e-16 ***
## exper          646.2      160.3   4.033 5.73e-05 ***
## female      -24082.7     3888.4  -6.193 7.16e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 82210 on 1947 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.03159,    Adjusted R-squared:  0.0306
## F-statistic: 31.76 on 2 and 1947 DF,  p-value: 2.68e-14
```

The coefficient associated with experience is smaller in regression 2 as compared to regression by, possibly because of **omitted variable bias** (OVB). The logic to understand OVB is the following:

1. Suppose that the true regression is: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$.
2. Suppose that X_1 and X_2 are correlated: $X_2 = \alpha_0 + \gamma X_1 + u$
3. Instead of running that regression in 1., you decide to run: $Y = \beta_0^{omit} + \beta_1^{omit} X_1 + e$.

By running the regression in 3 (that does not include X_2), then your estimate associated with X_1 will be:

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

So, the omitted variable bias will depend on two effects: first, the relation between Y and X_2 (or β_2), and second, the relation between X_1 and X_2. In our case, the first regression we estimated did not include `female`. Therefore, the first regression estimates the relationship described by equation 3. The second regression we made includes the variable female. In other words: $\beta_1^{omit} = 1170.1$ and $\beta_1 = 1011.11$.

Not including the variable `female` biases our coefficients **upwards** because

- Women on average earn less than men ($\beta_2 < 0$)
- Women on average have less experience than men ($\gamma < 0$)

Thus, $\beta_2\gamma > 0$ and omitting the variable will cause our estimate to be biased upward.