Lab 6: Midterm review

Julieth Santamaria

March 1, 2019

Basic linear prediction model[1]

$$Y = \beta_0 + \beta_1 X_1 + u$$

, where Y is the outcome, response variable or dependent variable; β_0 is the intercept, or the predicted value when $X_1 = 0$; β_1 is the slope; X_1 is the independent or explanatory variable; and u is the error term or the residual.

- Predicted value or fitted value ($\hat{Y}$): Is the value of Y predicted by the estimated equation, i.e., $\hat{Y} = \hat{\beta_0} + \hat{\beta_1} X_1$
- Biasedness: $\hat{\beta_1}$ is unbiased if X_1 is exogenous
- Consistency: $\hat{\beta_1}$ is consistent if, as N increases, the sampling distribution of $\hat{\beta_1}$ gets tighter, and the mean of the sampling distribution gets closer to β_1
- Standard error: In this context we will refer to it as the precision of $\hat{\beta}$, i.e.,

$$SE(\hat{\beta}) = \sqrt{\frac{\hat{\sigma}_u^2}{N \times \hat{\sigma}_X^2 \times (1 - R^2)}}$$

- 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.
- Heteroskedasticity:
  - $\hat{\beta_1}$ will *not* be biased
  - SEs will be biased towards zero
  - It will affect your t-test and the p-value.
- Multicollinearity: The independent variables are highly correlated. It will cause your SE to be very large.

Endogeneity

An independent variable is endogenous if it is correlated to factors in the error term. It will cause bias in your estimates. Three reasons for endogeneity:

- Omitting a *relevant* variable:

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

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

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

- 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:

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

- Reverse causality: X explains Y, and Y explains X

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

- Power: Probability of rejecting a false H_o
- Confidence interval at 5% level

$$\hat{\beta} \pm 1.96 \ SE(\hat{\beta})$$

Interpreting regressions

- $Y = \beta_0 + \beta_1 X_1 + u$
  - A change in 1 *(unit)* of X_1 will change Y in *(units)*
- $\log(Y) = \beta_0 + \beta_1 X_1 + u$
  - A change in 1 *(unit)* of X_1 will change Y in $\beta_1 \times 100$ percent
- $\log(Y) = \beta_0 + \beta_1 \log(X_1) + u$
  - A change of 1 percent in X_1 will change Y in β_1 percent

Note: Notice that in this summary does not include *all* the key concepts that you need to review for the midterm. One example of a term I did not include and that you need to review is the *Central Limit Theorem*.

Application (measurement error)

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. Assume that experience is observed with error. Let's check two options:

- You observe experience + two years
- You observe experience with a random error of mean two and S.D.=5

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

  # Potential experience
  exper <- 2017-dgryr

  # Measurement errors
  N=nrow(nscg17)
  exper.error1 <- exper+2
  exper.error2 <- exper+ rnorm(N,2,5)
})
```

```
## Table 1: Regressions of salary on experience
## ------------------------------------------------
##                      [a]          [b]          [c]
##  (Intercept) 73322.17*** 71397.05*** 73657.75***
##                (473.22)     (522.46)     (482.02)
##        exper   962.56***
##                 (30.73)
## exper.error1                962.56***
##                             (30.73)
## exper.error2                             838.83***
##                                          (27.66)
##            N    69344        69344        69344
##    r.squared    0.018        0.018        0.016
## ------------------------------------------------
## Significance: 0.01=***, 0.05=**, 0.1=*
## Note: This table displays the summarized results of running a
## regression of salary on experience [a], and experienced measured
## with two types of error.
```

The first case does not affect the estimate because the variance of the error is zero. However, the second type of error makes the coefficient be biased towards zero. This is called **Attenuation Bias**