GEB 6895: Business Intelligence
Department of Economics
College of Business Administration
University of Central Florida
Fall 2019

Assignment 5

Due Tuesday, October 22, 2019 at 11:59 PM
in *your* private mirror of the GEB6895F19 GitHub repo.

Instructions:

Complete this assignment within the space on your private mirror of the GEB6895F19 GitHub repo in the folder `assignment_05`. Create a folder called `my_answers` that will contain all of your work for this assignment, which can be in the form of a `docx` file, a `README.md` file or simply comments within in the relevant scripts.

Question 1:

Complete this exercise using the script `OLS_Omitted_Vars.R` in RStudio. This script generates simulated data for housing prices, which depended on income, location in California and the occurrence of earthquakes, which happen only in California. Specifically, the regression model is

$$HOUSE\_PRICE_i = \beta_0 + \beta_1 \times INCOME_i + \beta_2 \times IN\_CALI_i + \beta_3 \times EARTHQUAKE_i + \epsilon \quad (1)$$

where:

$HOUSE\_PRICE_i$	=	the average house price in zip code i
$INCOME_i$	=	the average income in zip code i
IN_CALI_i	=	whether or not zip code i is in California (i.e., $IN\_CALI_i = 1$ if zip code i is in Californa, zero otherwise)
$EARTHQUAKE_i$	=	whether or not zip code i experienced an earthquake (similarly, 1 or 0)

Run the entire script and compare the output from `summary(lm_full_model)`, which includes all variables, with that from `summary(lm_no_earthquakes)`, which omits the earthquake indicator. If there are no earthquakes in your simulation, run the script again to take another draw.

a) Compare the estimated coefficient for IN_CALI_i with and without the earthquake variable. How does this relate to the coefficient for $EARTHQUAKE_i$?

b) Compare the values of R^2 (labeled `Multiple R-squared`) and $\bar{R}^2$ (labeled `Adjusted R-squared`) for the two models. Which model do you recommend (pretending that you don't know the true model)?

Now consider the situation in which homeowners in California have perfect insurance against earthquakes. That is, any damages from earthquakes are fully repaired with no decrease in home value. To implement this in the script `OLS_Omitted_Vars.R`, change lines $62 - 63$ from `beta_earthquake <- - 0.50` to `beta_earthquake <- 0`. Run the entire script again and read the output.

c) For this new set of regressions, compare the estimated coefficient for IN_CALI_i with and without the earthquake variable. How does this relate to the new coefficient for $EARTHQUAKE_i$?

d) Compare the values of R^2 (labeled `Multiple R-squared`) and $\bar{R}^2$ (labeled `Adjusted R-squared`) for the two models. Now which model do you recommend (again, pretending that you don't know the true model)?

e) Re-run the script several times with `beta_earthquake <- - 0.50` (as in parts a-b) to see what happens over several realizations. What pattern emerges in the estimates from the output in `summary(lm_no_earthquakes)`? Can you relate this to the values of `beta_earthquake` and `prob_earthquake`?

Question 2:

Complete this exercise exercise using the script `OLS_On_Repeat.R` in RStudio. This script repeatedly generates simulated data for housing prices, which depended on income, location in California and the occurrence of earthquakes, which happen only in California. Specifically, the regression model is

$$HOUSE\_PRICE_i = \beta_0 + \beta_1 \times INCOME_i + \beta_2 \times IN\_CALI_i + \beta_3 \times EARTHQUAKE_i + \epsilon \quad (2)$$

where:

$HOUSE\_PRICE_i$	=	the average house price in zip code i
$INCOME_i$	=	the average income in zip code i
IN_CALI_i	=	whether or not zip code i is in California (i.e., $IN\_CALI_i = 1$ if zip code i is in Californa, zero otherwise)
$EARTHQUAKE_i$	=	whether or not zip code i experienced an earthquake (similarly, 1 or 0)

Run the entire script and observe the output from the simulation. In particular, observe the statistics printed at the bottom.

a) Consider the standard deviation of the each of the estimated coefficients. Order the variables from highest to lowest standard deviation. Which ones have highest and lowest variance? Compare this ordering with the variances of the explanatory variables themselves.

b) Now compare the average values of each of the estimated coefficients with their true values. Are they biased or unbiased? Keep in mind that a difference of less than 2 standard deviations could often happen by chance. Note also that an estimator is unbiased of the average value of the estimate is equal to the true value; it is unbiased if they are different.

Now consider the situation in which the true `income` variable is unobserved but another measurement `income_1` is observed. In other words, income is measured with error. To implement this in the script `OLS_On_Repeat.R`, change part of line $129 - 130$ from

```
list_of_variables <- c('income', 'in_cali', 'earthquake')
```

to

```
list_of_variables <- c('income_1', 'in_cali', 'earthquake')
```

Run the entire script again and observe the new output.

c) Now compare the average values of each of the estimates with their true values when `income` is unobserved but `income_1` is observed. Which ones are biased and which are unbiased? Again, keep in mind that a difference of less than 2 standard deviations could often happen by chance.

d) Re-run the script several times with `beta_income` set to a few different values. Did you notice anything unusual about the distribution of the coefficient for `income_1` realative to the true value? What pattern do you notice?