ECOM20001: Econometrics 1

Tutorial 6: Single Linear Regression Testing

A. Getting Started

Please create a Tutorial6 folder on your computer, and then go to the LMS site for ECOM 20001 and download the following files into the Tutorial6 folder:

- tute6.R
- tute6_height.csv
- tute6_crime.csv

The first file is the R code for tutorial 6, the second two files are .csv files that contain two separate datasets for the tutorial.

The first (micro) dataset,[1] tute6_height.csv, has the following 5 variables:

- id: worker identifier
- earnings: annual labour earnings in $10,000's (in real terms, 2012=100)
- height: height without shoes in centimetres
- weight: weight without shoes in kilograms
- male: binary variable that equals 1 if worker is male and 0 otherwise
- age: age of the worker at time of survey

In total, the dataset contains this information for n=17,870 U.S. workers.

[1] The first dataset is the same as the one we used in tutorial 5, with one important difference that we discuss below. Recall this dataset is from Case, Anne and Christina Paxon (2008): "Stature and Status: Height, Ability, and Labor Market Outcomes," *Journal of Political Economy*, 116(3), pp. 499-532.

The second (county-level) dataset from England and Wales,[2] tute6_crime.csv , has the following 5 variables:

- county: county name
- police: number of police officers in 2012
- homicides: number of homicides in 2012

In total, the dataset contains this information for n=43 countries.

B. Go to the Code

With the R file downloaded into your Tutorial6 folder, you are ready to proceed with the tutorial. Please go to the tute6.R file to continue with the tutorial.

C. Questions

Having worked through the tute6.R code and graphs, please answer the following:

Earnings and Height

1. Estimate the following single linear regression model for worker *i*:

$$Earnings_i = \beta_0 + \beta_1 Height_i + u_i$$

Present the regression results, including discussion of statistical significance of the OLS regression estimate against a null of no relationship between earnings and height, for the change in earnings associated with a one-unit change in height. Also present the 95% confidence interval for the relationship between earnings and a one-unit change in height.

2. Present the regression results, including discussion of statistical significance of the OLS regression estimate against a null of no relationship between earnings and height, for the change in earnings associated with a 100 cm change in height. Also present the 95% confidence interval for the relationship between earnings and a 100 cm change in height.

[2] This dataset is from Machin, Stephen and Olivier Beck (2011): "Crime and Police Resources: The Street Crime Initiative," *Journal of the European Economic Association*, 9(4), pp. 678-701.

3. Present the regression results, including discussion of statistical significance of the OLS regression estimate against a null that a 10 cm increase in height has an associated $3000 increase in annual earnings.

Homicides and Police

1. Present a scatter plot with police on the horizontal axis and homicides on the vertical axis. What relationship do you see?
2. Using the scatter plot, examine whether there are any potential outliers.
   - If there are, discuss which data point(s) appear to be outliers and *why* they may be outliers.
   - Produce a second scatter plot with the outliers removed.
3. Estimate the following single linear regression model for county *i*:

$$Homicides_i = \beta_0 + \beta_1 Police_i + u_i$$

Present the regression results, including discussion of statistical significance of the OLS estimate against a null of no relationship between homicides and police, for the change in homicides associated with a one-unit change in police.

   - Also present the 95% confidence interval for the relationship between earnings and a one-unit change in police.
   - Does the OLS regression coefficient estimate and 95% CI have an easy or difficult numerical interpretation?

4. Rescale police from the raw data to being in terms of 1000s of police in a county. Using this rescaled variable, present a new set of regression results, including discussion of statistical significance of the OLS regression estimate against a null of no relationship between homicides and police, for the change in homicides associated with a one-unit change in the rescaled police variable.
   - Also present the 95% confidence interval for the relationship between earnings and a one-unit change in the re-scaled police variable.
   - Does the OLS regression coefficient estimate and 95% CI have an easier numerical interpretation than the interpretation in question 3. above?
5. Continuing to use the rescaled police variable, present a separate set of OLS regression results that omit any potential outliers you found in question 2. above. Present the a new set of regression results, including discussion of

statistical significance of the OLS estimate against a null of no relationship between homicides and police, for the change in homicides associated with a one-unit change in the rescaled police variable.

- Also present the 95% confidence interval for the relationship between earnings and a one-unit change in the re-scaled police variable.
- Compare the results in questions 4. and 5. and comment on the impact of any outliers on the regression results. Do you think the potential outliers should be omitted from the analysis?
- Present a scatter plot using the homicides and rescaled police variables as well as the estimated regression lines in questions 4. and 5. to visually highlight the impact of outliers on the regression results.

6. Finally, provide an *economic explanation* for the relationship between the number of homicides and the number of police in a county in 2012.
   - Did you originally find the scatter plot surprising? Why or why not?
   - Provide an economic explanation for why you might find a *positive* relationship between the number of homicides and the number of police.
   - Provide a separate economic explanation for why you might find a *negative* relationship between the number of homicides and the number of police.
   - Given your economic explanations for a positive and negative relationship, can you plausibly interpret your OLS estimates of the relationship between homicides and the number of police in questions 4. and 5. above as being "*causal*"? That is, can you interpret the estimated relationship as the causal impact of increasing the the number of police on the number of homicides?
     - If you think the OLS estimate of the relationship is causal explain why
     - If you do not think the OLS estimate of the relationship is causal, briefly describe an experiment that could be used to estimate the causal impact of increasing the number of police on the crime rate in a county. What would you expect the sign of the OLS estimate of the empirical relationship between the number of homicides and number of police with such experimental data to be?