Introduction to Econometrics September/October 2018
Practical tutorial

Empirical problem set 1

Problem 1. E4.1

Start R Studio. From the menu choose 'Session/Set Working Directory/...' and set it to a convenient folder, e.g. `H:/R/C_exercises`.Open our R script file under the name 'c_exerc1.r'. The idea this time is to run most commands from this script file. Of course: it there are errors, correct them in the script file.
On Canvas you will find a Stata data file **CPS08** that contains a limited version of the data set used in Table 3.1 for 2008. Just to be sure we also included the Rdata file **CPS08** in case you experience problems importing Stata files (on the VU computers some of you experienced problems). It contains data for full-time, full-year workers, age 25-34, with a high school diploma or B.A./B.S. as their highest degree. A detailed description is given in **CPS08_Description**, also available on Canvas. (These are the same data as in **CPS92_08** but are limited to the year 2008.) Download the data files to your default R folder. In this exercise, you will investigate the relationship between a worker's age and earnings. (Generally, older workers have more job experience, leading to higher productivity and earnings.)
Only this very first time we run the commands from the scrip file one by one. It would even be better if you try yourself first

(a) Compute summary statistics for the variables average hourly earnings (`AHE`) and `Age` and draw histograms of the data. Also make a boxplot of the variable AHE. Is it reasonable to assume that the variable AHE is normally distributed?

(b) Make a scatterplot of AHE against Age.

(c) Split the sample in data corresponding to females and data corresponding to males. (Hint: create a logical vector being true if female=1 and being false for female=0. See also Vector-Indices.R). Are the hourly average earnings for Females and Males equal? Did you assume equal variances or did you not? Can you simply conclude that payment is unfair, or is a different explanation possible (e.g. if you consider the Bachelor degree)?

(d) Find a 99% confidence interval for the AHE for females. (Hint: using the (1 or 2-sample) t-test you get confidence intervals as well.)

(e) Run a regression of average hourly earnings (`AHE`) on age (`Age`) for the model

$$Y_i = \beta_0 + \beta_1 X_i + u_i, \quad i = 1, \dots n.$$

What is the estimated intercept? What is the estimated slope? Use the estimated regression to answer this question: How much do earnings increase as workers age by 1 year?
Note: you may study the example scripts in `OLS-Regression.R` and
`OLS-Regression-extra.R`. During the course we will adapt those script file; suggestions are welcome (commands you want to use often, but that perhaps are not in the current version).
Also use the commands `summary` to get more output and the command `names` to produce

the names of the items in the output list of the regression, produced by R.
How can you get the estimated coefficients in a vector?
Find the value of R^2. Also compute this value yourself using AHE, the fitted values and the residuals.

(f) Comment on the size of the regression's slope. Is the estimated effect of `Age` on `AHE` large or small? Explain what you mean by "large" and "small".

(g) Bob is a 26-year-old worker. Predict by hand Bob's earnings using the estimated regression.
Alexis is a 30-year-old worker. Predict Alexis's earnings using the estimated regression.
Also let R do the computations, and use the estimated coefficients produced by R.

(h) Does age account for a large fraction of the variance in earnings across individuals? Explain.

(i) From the lecture, we know that, the larger the variance of the regressor variable, the smaller the variance (and the standard error) of $\widehat{\beta}_1$. Illustrate this feature for the Californian schools data set by splitting the sample into two parts with equally many observations, where one subsample has low variance in age and the other is taken randomly from the entire sample. Estimate the model parameters seperately for both samples. Visualize the different outcomes in a diagram.

Problem 2. E4.2: Do-it-yourself

On Canvas you will find a Stata data file **TeachingRatings**. Just to be sure we also included the Rdata file **TeachingRatings** in case you experience problems importing Stata files. The data file contains data on course evatuations, course characteristics, and professor characteristics for 463 courses at the University of Texas at Austin. A detailed description is given in **TeachingRatings_Description**, also available on Canvas. One of the characteristics is an index of the professor's beauty as rated by a panel of six judges. In this exercise, you will investigate how course evaluations are related to the professor's "beauty".

(a) Construct a scatterplot of average course evaluations (`Course_Eval`) on the professor's beauty (`Beauty`). Does there appear to be a relationship between the variables?

(b) Run a regression of average course evaluation (`Course_Eval`) on the professor's beauty (`Beauty`). What is the estimated intercept? What is the estimated slope? Explain why the estimated intercept is equal to the sample mean of `Course_Eval`. (`Hint:` What is the sample mean of `Beauty`?)

(c) Professor Watson has an average value of `Beauty`, while Professor Stock's value of `Beauty` is one standard deviation above the average. Predict Professor Stock's and Professor Watson's course evaluations.

(d) Comment on the size of the regression's slope. Is the estimated effect of `Beauty` on `Course_Eval` large or small? Explain what you mean by "large" and "small".

(e) Does `Beauty` explain a large fraction of the variance in evaluations across courses? Explain.