ECON-A4000 - Economics of Global Challenges
Instructor: Matti Liski
TA: Max Toikka
Return method: through mycourses by the deadline

Problem Set 2: Question 3

1. Recall the Autor (2014) article "Skills, education, and the rise of earnings inequality amonth the other 99 percent" (RA2). We will try to reproduce some of the graphs presented there using Finnish data (*(link to data)*). We work with a dataset from Statistics Finland. We have 15 observations for the years 1995-2009: for each we have mean earnings by gender and education level. The two types of education level are high school and college.

   (a) We want to reproduce Fig.1 (Autor 2014). Plot the graph showing the ratio of male university earnings to male high school earnings over the years 1995-2009. Do the same for female earnings.

   (b) Calculate the expected net present value of the gap for males and females separately when the discount rate is 4%, as we did in problem set 1 (question 4).

2. THIS PART IS NOT MANDATORY. However, I would like you to scrutinize the directions below and write back to me in your answer the following: do you have have enough technical background for doing this exercise in principle? If not, please let me know what part is unclear to you (software, coding, etc.). I'd like to learn how much you know about data analysis at this stage of your studies.

   In any empirical work, data cleaning constitutes probably 80% of the work, and actual empirical analysis is a much smaller part. Data cleaning is necessary in order to get the data to the state where you have the variables that you need so that you can run the regression you are interested in or construct specific graphs. In the first exercise above, you had a small dataset which had exactly the information you needed in order to construct the graph and calculate the earnings gap.
   However, the original dataset is quite a lot different. You can find it *here*.
   I will give you guidelines on how to clean this dataset. THE SESSION, WE WILL GO OVER THE CODE and you will get a chance to ask more questions about Stata commands. You have two ways of answering this question.

(a) Create a folder and save the .csv file in it.

(b) Open Stata and a new Do-file

(c) Load the dataset into Stata: **import delimited "fleed_puf_julk.csv", delimiters(",")**

(d) Drop all variables except for "vuosi sukup syntyv ktutk svatva tyotu"

(e) Generate "earnings", which is the sum of "tyotu" and "svatva"

(f) Replace earnings = 0 if it is missing

(g) "vuosi" is in numbers from 1 to 15. The baseline is 1994.

(h) The variable "syntyv" is birth year. Use it to obtain information on the person's age.

(i) Create a dummy variable for being of working age (aged 18-65)

(j) Drop individuals that are not of working age

(k) "sukup" takes value 1 for male and 2 for female. Transform it into a dummy variable called "female", so that it takes value 0 for male and 1 for female. This is not crucial in our case, but very important when running regressions.

(l) You can drop variables that you do not need any more as you go along.

(m) Here is a tricky bit. "ktutk" (korkein tutkinto) is a 2-digit code for the highest education attained, in which the first digit refers to the level of education (secondary, post-secondary, vocational, tertiary etc) and the second digit refers to a specialisation/subject area. In this case, we only want to extract the information referring to the level.

(n) We first transform the variable into a text (string) variable

(o) Next, we create a variable which takes the value of the first digit of "ktutk", call it "edu_level". Write **help substr** in the command window to see how to do it. Once done, you can make the variable a numerical one again.

(p) "edu_level" takes values 3, 4, 5, 6, 7, 8, missing. Create a dummy variable "tertiary" that takes value 0 for highschool ("edu_level"= 3, 4, missing) and 1 for college ("edu_level"= 5, 6, 7, 8).

(q) Now the aim is to calculate mean earnings by year for each of the four categories: male_highschool, male_college, female_highschool, female_college: **egen mean_earnings = mean(earnings), by(year female tertiary)**

(r) Now we have the same information multiplied many times, which is not very useful. If you drop the variable "earnings" and then input the command **duplicates drop**, you end up with a much smaller dataset of 60 observations, one for each combination of gender and education level for each of the 15 years.

(s) We now want to get to a simple dataset with 15 observations (one per year) and variables for mean earnings for each of the four categories. For my personal knowledge of stata code (I am sure some more sophisticated and less heavy code exists), this is done in a very tedious way.

(t) We first generate the four variables "earnings_f_college", "earnings_f_highschool", "earnings_m_college", "earnings_m_highschool".

(u) We will end up with three missing variables for each observation. What we want to do now is to copy the information from different observations into all the others. This can be done because you can specify the observation from which you want to copy the information. The easiest position from which you can copy the information is the first observation. So we can sort the dataset multiple times so that the information we want to copy is at the top of the group. For example, if we want to copy the information for "earnings_m_highschool", you can **sort year female tertiary"** and then **bys year: replace earnings_m_highschool = earnings_m_highschool[1]**: what this command is saying is to consider only observations with the same value for year, and then replace the value of the variable earnings_m_highschool (which for all the observations that are not tertiary=0 and female=0 is 0 or missing) with the value of the variable earnings_m_highschool in the observation at the top of group. You have to make sure that the data is sorted in a correct way to do this.

(v) Get the values for the other categories.

(w) Drop the variables "female", "tertiary" and "mean_earnings", so you can **duplicates drop** and get to your 15 observations.

(x) Finally, we **export excel using "earnings_edu_finland", firstrow(variables) replace**