Empirical Assignment #1 (3 pages)
Labor Economics CERGE-EI, Spring 2015, Daniel Münich and Mariola Pytlíková

The purpose of this assignment is to:
A) make you familiar with standard data work preceding almost every empirical research,
B) upgrade your experience as empirical researcher using real empirical data.

The better and the more carefully you get familiar with the data and their possible deficiencies, the easier, effective will be your follow-up analysis and the higher will be the credibility of your research findings. Below is a sequence of steps you should follow and questions you should answer. You should document findings for each step.

The length of your written output should not exceed 10 one-sided pages! You are not expected to include motivation and literature review (as you would include into regular paper). But the rest of your output should try resembling at least a raw version of a regular paper, meaning a balanced combination of explanatory text and results (tables/graphs).

Keep in mind that in a real empirical analysis like this one, unique "right" answers and unique solutions are rather exceptions. Sometimes, you will have to make explicit assumptions and simplifications plus arbitrary decisions to assure tractability of a problem you face.

You are not required to account for the phenomena of *self-selection* because the data you are provided are not rich enough, but your concluding comments on possible impact of this phenomenon on results will be appreciated.

You are given a random sample *RQdata.dta* of Czech men and women being surveyed in December 1996. The data coding is based on enclosed questionnaire and variables labels. The file is saved in Stata 8 format. Variable *identa* identifies households and variable *pers* identifies individuals within household. If something is really unclear, do not hesitate to ask. Your tasks are:

1) Merge your dataset with *ur95.xls* data containing district specific unemployment and vacancy rates (shares on the labor force) and check whether data were merged properly. Before you merge, you may want to save the xls file in ASCI (tab delimited) file and retrieve it into Stata using '`. infile`' or '`. insheet`' command. Alternatively, you can use *Stata Transfer* software (available in the CERGE-EI lab – or consult RA) to transfer from *.xls* format into Stata *.dta* format.

2) Make sure you understand the meaning of all variables, their coding, and occurrence of missing or strange values. Use both graphical and statistical tools (Stata recommended but not mandatory) to check data quality and possible mistakes/errors in data. Do not forget to explore the scope of missing values and outliers. If data allow, explore whether missing values of some variables have some systematic pattern and if so, at the end of your overall analysis, discuss their possible impact on your results. Mention at least briefly, if you decide to remove some observations from your empirical analysis. Note, that the sample and variables are not perfect – common in empirical research. Note that you do not know more

about the data than is described here, in the questionnaire, and by data labels.

3) Use descriptive statistics (explore how in papers published for example in journals like *Labor Economics*, or *The Review of Economics and Statistics* for typical approach – not too detailed & not too brief) to outline key and interesting patterns in the data. Think carefully which statistics and graphs are important to be presented. Make sure that you explain clearly how you created your working variables. Make sure you present basic statistics for the working sample in case you remove some observations.

4) Use simple OLS procedure to estimate the *basic* Mincerian model for **men only(!)**

a) $\ln Y_i = \alpha + \beta_1 EDU_i + \beta_2 EXP_i + \beta_3 EXP_i^2 + \varepsilon_i$

b) Choose what will be your left-hand-side variable (earnings, hourly wage, etc.) and explain. It is recommended that you perform the whole analysis on a subsample of individuals who work about full-time (6-10 hours per day). But use basic statistics to check if these people are different than those excluded from your analysis and at the end briefly comment on possible impacts of this exclusion.

5) Consider additional explanatory variables of your choice to estimate *extended model*. Briefly explain economic reasons to include additional variables and possible implications of their exclusion. Choose your *favorite extended model* specification (explanatory variables and their functional form) and defend your choice by economic/econometric arguments.

a) Compare estimates of the *basic* and *extended* model.
b) Should you control for inflation in earnings? What happens if you do not?
c) Test for presence of heteroskedasticity and use robust estimates if proper.
d) Test for omitted variables.
e) What are the returns to experience of labor market entrants?
f) Compute point estimate of labor market experience of maximum earnings.
g) Instead of polynomial shape of the earnings – experience profile, estimate specification using 3-segment connected linear spline of the profile.
h) Many similar studies do not have data on actual years of education as reported in "A09" and have to impute years of education from school level attained as in "A05". Note that primary school in the CR lasts for 8-9 years because educational system had changed over decades. To create *years_of_education* variable, do you prefer using A05 or A09 and explain why the choice affects your estimates (if it does). Note that some discrepancies between information in A05 and A09 could be due to unknown features of the educational system over time and other features. Do you have some in mind?

6) Assume that different school level attainments result in different annual returns. Estimate *basic* model using attainment dummies instead of years of education. Describe clearly how you define your categorical variables. Compute return to a year of education for each specific school type/level. *Consult P. Kennnedy's textbook on the treatment of coefficients on dummy variables, pp.248-258.*

7) In the following, what is your estimate of the expected percentage difference in earnings (or wages) between (explain how you reach your results):

a) Someone who has children and somebody who does not?
b) Somebody who has one child and somebody who does not have any?

c) Test for statistical differences.

8) Specifications above assumed that coefficients on education and experience variables are identical in all locations. Now assume that the effect of human capital variables is different in Prague.
   a) Propose proper specification and estimate.
   b) Test Ho: no difference due to Prague.
   c) Your only available variable describing district labor market conditions is unemployment rate. Does it help explaining variation in earning/wage? If yes, provide plausible economic interpretation.

9) Use **both samples of men and women** (after inspecting data quality for women). Consider *basic* Mincerian model, allow for gender specific coefficients, perform Oaxaca's decomposition of the raw wage/earning gap. Describe, explain, comment and interpret your results (consider Stata command `.decompose etc.`). Finally, test i) Ho: returns to education of both genders are equal, ii) Ho: experience profiles of both genders are equal. Summarize key differences between genders.

10) By Gini coefficient and 90/10 decile ratio (the top decile compared with the bottom decile), describe inequality in earnings within subsamples of men and women. Do the same considering households as earning units.

HINT:
You will be tempted to cut & paste large number of Stata regression printouts into the text or table appendices. You should not do it! You should present only those results, statistics etc. which are really important for what your goal is. Consider using `.outreg` and `.outreg2` or `.estout` and `.estimates store` Stata commands. Or simply cut & paste coefficient estimates and clean it as follows to save your time and space (in a scientific paper, you present only Std. Err. or t- or P-values, but not all!). Use `"Courier New" font 9` to make fit tables well in MS Word.

```
Number of obs =    3720, R-squared     =  0.0880, Adj R-squared =  0.0867
------------------------------------------------------------------------------
         c13 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         c14 |   .0001014   7.61e-06    13.34   0.000     .0000865    .0001163
         e03 |   .0763062   .0334354     2.28   0.023     .0107527    .1418598
         a09 |  -.0214649   .0127147    -1.69   0.091    -.0463934    .0034636
        e02a |   -.496976   .0611714    -8.12   0.000    -.6169088   -.3770433
        e02e |   .0132556   .0184894     0.72   0.473    -.0229949     .049506
       _cons |   8.838834   .1857443    47.59   0.000     8.474663    9.203005
------------------------------------------------------------------------------
```