

LI Econometrics

STATA ASSIGNMENT

In this assignment, you will explore correlates of earnings. You will use data on real individuals collected by the Office for National Statistics and published in the Quarterly Labour Force Survey. In the course of working on this assignment, you will:

- be introduced to an important source of data for research on the UK (the UK Data Service);
- become familiar with processing and analysing data using Stata;
- interpret regression output;
- develop critical thinking about economic phenomena and econometric analysis of them.

Practical details

Word limit: no more than 1000 words. Reasonable use of tables and figures does not count toward the word limit.

Submission file: one file (in `.pdf` or `.doc` format).

Figures and tables: should be numbered and titled appropriately. Tables should be formatted and presented as in standard economics journals (copy-pasted output from Stata is not acceptable).

Appendix: an appendix should be included at the end of the assignment, containing a copy of the Stata code used to obtain the results. This should be the exact code used to go from loading the dataset to generating the results presented (no more, no less). For convenience, you can simply copy and paste the `.do` file.

Preliminaries

You will have accessed the data and documentation you need to complete this assignment during Computer Workshop 5. Please refer to that workshop and the accompanying guide on how to access the dataset (“Computer Workshop 5 – UK Data Service Access Instructions”).

0. Loading data and defining the sample

- a) Load the main dataset (`lfsp_jm19_eul_pwt18.dta`) in Stata.
- b) We will focus on individuals reporting positive gross weekly earnings and not currently working towards a qualification. To keep only these observations, do:
`keep if GRSSWK > 0 & QULNOW == 2`
- c) Check: the resulting dataset should contain 8949 observations. If this is not the case, something has gone wrong somewhere.

1. Degrees and earnings [20 marks]

- Plot the distribution of weekly earnings (GRSSWK) in a histogram. What do you notice?
- Consider instead taking the logarithm of earnings. Compare the distribution of $\log(\text{GRSSWK})$ to the one above. When considering earnings in our regressions, why might it be preferable to take logarithms?
- Now let us think about education, in particular the correlation between having attained a degree and earnings. Using the variable detailing the highest degree earned (HIQUL15D), create a dummy variable taking a value of one if the respondent has attained a degree or equivalent. Then, estimate the regression:

$$\log(\text{earnings})_i = \alpha + \beta \text{Degree}_i + \epsilon_i \quad (1)$$

Report your findings in a table, and interpret the estimated coefficient on Degree_i , both in terms of econometrics and economic theory.

2. Men and women [20 marks]

- Create a dummy variable taking a value of one for female respondents (the relevant variable is SEX).
- Estimate equation 1 separately for men and women and report your findings in additional columns of the table you created above. What do you conclude about the returns to a degree for men and women?
- Instead of estimating the regression separately for men and women, estimate instead:

$$\log(\text{earnings})_i = \alpha + \beta_1 \text{Degree}_i + \beta_2 \text{Female}_i + \beta_3 \text{Degree}_i \times \text{Female}_i + \epsilon_i \quad (2)$$

How do the findings from this regression compare to the ones for men and women separately?

- What do we learn, on the basis of these regressions, about the earnings of men and women?

3. Other factors [30 marks]

- Pick an additional dimension on which information is available in the dataset (consult the user manual and codebook ([lfs_user_guide_vol3_variabledetails2019.pdf](#)) for details) and construct variable(s) which allow you to explore the relationship between this dimension and earnings. Without running additional regressions (yet), why does economic theory tell you to expect this dimension to matter for earnings? Can we expect these dimensions to interact with the “degree effect” or “gender effect” we explored already?
- Estimate additional regressions to test whether your theoretical predictions hold in the data. Present these in a second table and discuss.

4. Taking a step back [30 marks]

- Thinking about the first regression we estimated (equation 1), do you expect the estimated $\hat{\beta}$ to capture the true causal effect of attaining a degree on earnings? Why or why not?
- In this assignment, you have used labour force data from 2019. What differences, if any, do you expect to find if you were to re-do the assignment using data from 2022?