

University of East Anglia: School of Economics

ECO-5006A: Introductory Econometrics

Autumn 2019

## Take-Home Data Analysis Assignment using Stata

**Please read the Take-Home Data Analysis Assignment Brief carefully before attempting the questions, provided by the LTS Hub, which provides some general information and further instructions. Please read the instructions below carefully as well.**

- Open data set `Stata.Exercise.dta` in Stata, which contains information on 9,664 university graduates in full-time employment, based on DLHE Survey of 2016/17. Please refer to the Assignment Brief for more information about this data set.
- In the first parts of this exercise, we are interested in estimating the salary gap between graduates who studied Economics and those who studied Business or Management studies.

We would like to investigate whether such a salary gap exists for recent graduates and whether it gets bigger/smaller once we control for differences in the graduates' characteristics, such as age, gender, tariff, degree classification, etc.

- In the last part, you are asked whether studying Economics or Business/Management is associated with the probability of graduating with getting into professional employment.
- Please have a good look at the available variables in Stata before attempting the questions below. Using commands `codebook`, `sum`, `tab` appropriately, will allow you to get a good understanding of the data.
- In questions that require you to use Stata commands to get your answer, make sure you **clearly show these Stata commands within your answers**, unless the instructions in the question state that you don't need to present these.
- For questions related to statistical significance, you don't need to set up the null and alternative hypothesis, unless this is specifically asked in the question. You just need to comment on the significance based on the  $p$ -values as found in Stata.
- Note that the marks associated with each individual part are given in squared brackets, summing up to 100.

## QUESTIONS

- (a) Provide descriptive statistics (i.e mean, standard deviation, minimum and maximum) for variables *salary*, *tariff*, *age*, *professional*, *male*, *degree\_class*, and *socecon\_class*, separately for Economics and Business/Management graduates (by adding to your commands “if Economics==1” for descriptive statistics of the Economics subject and “if Economics==0” for descriptive statistics of the Business/Management subject). Note that for the categorical variables *degree\_class* and *socecon\_class*, you need to provide the descriptive statistics for each category of these variables. Also, based on these descriptive statistics, describe in no more than 300 words the main similarities/differences between the two subjects. Clearly indicate the word count within your answer.

[13 marks]

- (b) Estimate the following log-linear regression model.

$$\log(\text{salary}_i) = \beta_0 + \beta_1 \text{Economics}_i + u_i \quad (1)$$

where  $\log(\cdot)$  represents the natural logarithm. Report your Stata estimated output and interpret the estimated coefficient associated with the *Economics* dummy variable. Based on this regression, is there statistical evidence that  $\log(\text{salary})$  differs between Economics and Business/Management graduates?

[9 marks]

- (c) Re-estimate your model of part (a), also adding variables for, *male*, *age*, *tariff*, *degree\_class*, and *socecon\_class*. Note that for *age*, you must also add the *age*-squared term. In addition, variables *degree\_class*, and *socecon\_class*, need to be included in the form of dummy variables appropriately. Report the estimated regression output.

[9 marks]

- (d) Compared to the regression in part (b), how does the estimated salary gap (in % terms) between Economics and Business/Management graduates change in terms of both magnitude and statistical significance, when adding the additional explanatory variables ?

[5 marks]

- (e) Using the results of your regression in part (c), provide an interpretation of the coefficients associated with the *socecon\_class* dummy variables, and lost test for their joint significance.

[9 marks]

- (f) Using the results of your regression in part (c), discuss the relationship between the age of the graduates and their salary.

[7 marks]

- (g) Using the results of your regression in part (c) and the command `margins` appropriately, calculate and report the predicted log of salary for a male Economics graduate, who is 22 year old, has a tariff score of 150, graduated with a first-class degree, and comes from “semi-routine” socioeconomic classification.

[5 marks]

- (h) Test whether your model estimated in part (c) suffers from a heteroskedasticity problem using a Breusch-Pagan test. Report and interpret the results of the test. In addition, re-estimate the models in part (c), but now using heteroskedasticity robust standard errors (RSEs), and report your results. In no more than 150 words, explain whether the use of the RSEs has changed the significance of your estimated coefficients. Please also clearly indicate the word count within your answer.

[5 marks]

- (i) Recode the `region` variable into a new variable that puts the 12 regions into 4 categories: North East, North West, Yorkshire and The Humber **in category 1**; East Midlands, West Midlands, East of England, and London **in category 2**; South East and South West **in category 3**, and Wales, Scotland and Northern Ireland **in category 4**. Do a twoway-tabulation between the original and the newly created variable, to check whether you have done the recoding correctly (report this tabulation with your results). Then, re-estimate your model in part (c) using RSEs, but now also include your newly created grouped regional variable in form of dummy variables, as well as interaction terms between these dummy variables and the *Economics* dummy variable. Based on this new regression model, calculate the estimated salary gap between Economics and Business/Management graduates, for each of the four grouped regions, and test whether these salary gaps are statistically different from zero.

[19 marks]

- (j) Estimate either a Probit or a Logit model that explains the probability of getting into professional employment (using variable *professional* as your dependent variable). Report the results of two models: (i) a Probit or Logit model that uses the explanatory variables as used in part (c) exactly in the same form; (ii) another Probit or Logit model where you have “improved” your specification used in part (c) in terms of explanatory variables used, for example, by using additional variables that are available in the data set, removing variables that you think are not relevant, using quadratics, interaction terms, etc. Then, in no more than 400 words (excluding Stata commands and tables) explain: (i) how did you arrive at your final model specification (e.g. why did you choose a Probit or a Logit model? Why did you include the variables in a certain form, e.g. dummies, quadratics, etc.? Why did you remove or didn’t include some of the variables available in the data set?); (ii) based on the results of your two models, discuss the relationship between studying Economics instead of Business/Management studies, and getting into professional employment.

[19 marks]

END OF QUESTIONS