

University of East Anglia: School of Economics

ECO-5006A: Introductory Econometrics

Autumn 2020

Stata Project

NOTE THAT MORE QUESTIONS WILL BE ADDED IN DUE COURSE. THIS IS POSTED NOW, SO THAT YOU CAN START WORKING ON IT, IF YOU WISH, AND GET A GOOD FEEL FOR THE DATA.

Please read the Project Assignment Brief carefully before attempting any of the questions. This Brief is available on the ‘Project’ section on the module’s Blackboard and provides some general information and further instructions. Please read the instruction below carefully as well.

- All the analysis needs to be done using the Econometrics software package, Stata.
- The data set `PROJECT_2020.dta` contains information on 28,424 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.
- The questions of the Project are based on the following main topic:

**“Does studying Economics pay off, relative to other subjects in Social Sciences?
Evidence using data of recent UK graduates”**

- In particular, is there evidence that Economics graduates ‘do better/worse’ in the graduate market, relative to graduates who studied the other subjects available in the data? And how much better/worse do they do? By ‘doing better/worse’, we mean whether:
 - they earn more/less based on self-reported salaries of graduates (before any deductions)
 - they are more/less likely to secure a managerial / professional position after graduation

Thus, in your analysis, you need to use both outcome variables ‘salary’ and ‘professional’.

- In questions that require you to use Stata commands to get your answer, make sure you **clearly show these Stata commands within your answers**.
- Presentation of your answers matters. Thus, please, all graphs, equations, results and discussions need to be well-presented.
- For the main text, you need to use ‘calibri’ font of size 11, and allow 1.15 line spacing. For text within tables, you can use smaller font, up to size 9. You are also allowed to change the size of your graphs, as long as the graphs are still clear to read (i.e. clear legends, titles, etc.)
- Note that the marks associated with each individual part of this project will be revealed when all questions are made available.

QUESTIONS

- (a) Investigate the main question of the Project by using descriptive statistics only; e.g. by appropriate use of means, medians, variances, graphs, etc. Don't forget that you need to investigate this in terms of both outcome variables 'salary' and 'professional'. There is no word limit for this question, but your answer needs to be presented within two A4 sides (so, all tables, graphs and discussions need to be presented within two A4 sides, i.e. one full page). The assigned weight of this part will be revealed in due course, but this will be worth around **10-15% of the overall mark** of the Project.
- (b) In this part you need to investigate the main question of the Project by using **regression analysis** (i.e. appropriate MLR models). Don't forget that you need to investigate this in terms of both outcome variables, 'salary' and 'professional'. That is, you will need two separate MLR models, one using 'salary' as the dependent variable, and one using 'professional' as dependent variable. So, in this part, you need to investigate whether Economics graduates are expected to 'earn more/less' relative to each of the other subjects, and whether Economics graduates are more/less likely 'to get into professional roles', holding other variables fixed (i.e. if Economics graduates had the same tariff scores, the same socio-economic background, etc. with the graduates of the other subjects).

Here are some important instructions/notes. Please read these very carefully:

- The dependent variable *salary* must be used in a **logarithmic form** (i.e. the natural log of *salary*). Note that the *professional* dependent variable is binary (taking value 0 for 'non-professional' and 1 for 'professional'). We still haven't seen examples of using a binary dependent variable, but we will see an example on the live lecture of **Week 11**, to explain how to interpret the coefficients in such models.
- Your main explanatory variable (i.e. *subject*) is **categorical**, so it needs to be added in the MLR model in the form of dummy variables. We will see examples of how to include categorical variables in the MLR model using dummy variables in the material of **Week 9** (both on the Asynchronous lecture notes/videos and on the Synchronous sessions).
- In your discussion of your results, you need to provide an appropriate interpretation of the coefficients of the *subject*-related dummy variables. You also need to conduct hypothesis testing, to test: (i) whether there is statistical evidence that the mean salary of Economics graduates differs from the mean salary of graduates of each of the other subject, holding the other variables fixed (you can do this by using the *p*-values); (ii) whether there is evidence of joint significance of the subject dummy variables (this is done by an *F*-test, which will be covered in the material of **Week 10**).
- It is up to you to decide which other explanatory variables you add to your model. Note that categorical variables (such as *degree_class* or *region*, need to be added in form of dummies variables). For each explanatory variable that you add, you need to offer a short justification on why it is important for these variables to be included in the model (about 100-150 words for each). Also, for the variables that you decide **not** to add, you also need to provide justification as to why these were not added (about 50-100 words for each).
- Note that variables *tariff* and *age* must be included in the model.
 - For *tariff*, you need to decide whether you use it in its linear form, or whether you include a quadratic term / replace it by the natural log of *tariff*. Your choice needs to be justified within your justifications above.
 - For the variable representing the graduates' age, it must be included in the model as a quadratic function (i.e. add both *age* and *age*²). You also need to provide two graphs, one for predicted salary against *age*, and one for the predicted probability of getting into professional employment against *age*. Then, based on these graphs, you need to discuss the relationship between age and salary/professional (in about 200-250 words overall).

- Your ‘salary’ regression model needs to be tested for violation of MLR5 (i.e. whether there is a heteroskedasticity problem). If there is evidence of heteroskedasticity, then the standard errors presented in your regressions must be made ‘robust to heteroskedasticity’. Please note that testing for heteroskedasticity and correcting the standard errors, will be covered in the material of **Week 12**. Also note that in your ‘professional’ regression model, heteroskedasticity robust standard errors must be used.
- Note that all your regression results need to be presented in **one or two tables**. There are Stata commands that create such tables automatically. I will prepare a video, showing how to do this, using the example of the Experience/Wage relationship.
- The assigned weight of this part will be revealed in due course, but you can expect that this will be worth around 50% of the overall mark of the Project.

The exact form of the remaining questions will be added in due course, but here is what they will cover:

- Providing theoretical justification of your main findings in this project. This discussion needs to focus on the main topic of the project (i.e. ‘does study economics pay off?’). You also need to provide two academic references as part of your justification.
- Identifying the 3 most important problems/limitations in your models/results and explain how these could be addressed.
- Presenting your most important findings of your regression analysis within a single graph (it can be a ‘combined graph’), also providing a discussion/summary of these findings in non-technical language.
- Only for the ‘salary’ regression model, as additional analysis, pick one of the explanatory variables and create interaction terms between the subject dummy variables and this explanatory explanatory variables. Then, add these interaction terms to your model (the main model you used in part (b)) and provide a discussion of the results. Your choices needs to be justified (e.g. why adding such interaction terms between these variables is important?)