ECON 123 Spring 2023 – Guidelines for the Empirical Project

There is a required empirical project for the course. You may work in groups of up to four students. If you don't have a group, but would like us to help match you to other students, please let the instructor or the TA know.

This project is your opportunity to use the tools you learn to answer a question you come up with and that you care about, though we will give suggestions and guidance. Your project needs to include you applying methods from the course to analyze real data. You must make it clear the relationship between your project and the previous literature or any previous work by yourself or others in your group. I encourage you to continue a research project started in Econ 117, though you are also welcome to start a new project. Replicating (part of) an existing study is acceptable, as is expanding on one of your problems sets or labs. If you use a previous empirical paper/project as a starting point, you must also submit that paper as well, and your project will be evaluated based on the incremental work beyond that starting point.

Research Proposal (due April 4)

The initial proposal of 2 to 3 pages, which must include:

1. A clearly stated research question, with motivation for why this is an interesting research question within economics.
2. A short description of the data you plan to use to answer this question, and why it is appropriate given your research question.
3. A brief description of what econometric methods you plan to implement, and why they are appropriate given your research question.
4. If you are doing a replication study or starting with a previous paper by yourself, you must state that that is the case, and describe what you will replicate or do differently from the past work.
5. Your proposal needs to contain a statement confirming that you have access to the data set.

Final Project (due May 10th)

The final project is due on **May 10th**. There is no length requirement, but most projects will be around 15 to 25 pages long, including all tables and figures. Final projects must include:

- An introduction (which should clearly motivate and lay out the research question).
- A short literature review (which briefly touches on key passed work on this topic). This should be 1-2 pages max.
- A data description section (2-5 pages).
- A methods section which describes what analysis you performed.
- A result section discussing your findings.
- A short conclusion.

ADDITIONAL ADVICE FOR THE FINAL PROJECT

Creating Your Sample

Before writing your data description, think through which variables from your data set you plan to use in your analysis, and whether you plan to use all observations or a subset of observations. Some datasets contain hundreds (or thousands) of variables, while you will only focus on a much smaller

number of variables. The variables you plan to analyze might need to be constructed from the original raw variables in the data set. Depending on your project, you may wish to only use observations that satisfy certain restrictions, while excluding other observations.

For example, suppose you are interested in examining the effect of schooling on wages, and plan to do a regression of log wage on schooling and other control variables. In that context, you may wish to include a variable for hourly wage, a variable for years of schooling, a variable for years of work experience (if that is available in your data set), and other control variables that you might wish to include in a regression analysis (perhaps indicator variables for race and sex, for example). The variables you wish to use might have to be constructed from the original variables in the data set. For example, you might need to use a variable for total annual earnings and annual number of hours worked to create a variable for hourly wage, and you might need to recode categorical variables to create indicator variables for race. You might wish to limit your analysis to working-age adults who worked at least a certain number of hours in the previous year (say adults aged 22 to 55 who worked at least 1000 hours in the previous calendar year). You may additionally wish to exclude observations from your analysis with missing values of key variables. For example, you might restrict your sample to exclude observations with missing values of schooling. You may additionally wish to exclude from your analysis observations with values of variables that are implausible or are extreme outliers. For example, if any observation has an hourly wage of one cent per hour, or a million dollars per hour, you may suspect that such values are misreported and thus exclude such observations. Thus, you might, for example, wish to limit our analysis to individuals reporting an hourly wage between the minimum wage and $1000 per hour.

From your raw data, create the data set that you will analyze. Keep the variables that you need for your analysis (possibly creating any variables that you need from the original variables), while dropping any variables that are not needed. Keep the observations that you will analyze, while dropping any observations that are not relevant for your analysis or which you wish to exclude because of missing values or implausible values.

Exploring your Data

For your dataset, examine whether there are missing values for any of your variables, and, if so, how they are coded. For continuous variables, you should examine the sample minimum, maximum, mean, and standard deviation of each variable. You should understand the units of each variable. You should ensure that the results are sensible, and extreme or nonsensible values should cause you to further inspect that data. For example, you might find that the maximum years of schooling is 99, and further investigation might reveal that missing years of schooling is coded as 99 in the dataset which would motivate you to exclude those observations with schooling equal to 99 from your analysis. Examining the mins and maxs of the variables might cause you to change your sample exclusion restrictions. You should examine histograms for key continuous variables. You should examine sample means for indicator variables (equivalently, fraction of observations for which the variable equals one). For categorical (discrete, but taking more than two values) random variables that you do not wish to recode as indicator variables, you should report tabulations (in other words, fraction of the sample in each category). Depending on your project, you may wish to examine summary statistics for the variables conditional on values of indicator variables. For example, if you are interested in male vs female differences in wage equations, you might look at the sample minimum/maximum/mean/standard deviations of wages and of schooling separately for men and women.

Writing Your Data Description

Your empirical project must contain a detailed data description. The data description section should be a few pages long, perhaps 2-5 pages. You should clearly describe the source of the data, the unit of observation, and the number of observations. You should explain which variables from the original dataset you are using, and how any variable that you constructed from that original dataset was constructed. You should state both the population sampled by the original dataset (for example, all U.S. residents) as well as which subsample of the data set you will use (for example, individuals between the ages of 22 and 55 years old, who worked at least 1000 hours last calendar year, with nonmissing values of wages and of schooling). You should create a table(s) of descriptive statistics, reporting the means and standard deviations of each variable, and you may wish to include other descriptive statistics such as the minimum, maximum, or median for each variable. Depending on your project, you may wish to report tables of descriptive statistics separately by subgroup (such as for men and women separately). You may look at Tables 1 and 2 of the Chetty et al (2017) "Mobility Report Cards: The Role of Colleges in Intergenerational Mobility", or Table 3 of Bertrand and Mullainathan (200) "Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination", for inspiration. You may wish to include histograms of any important continuous variables. Your data description (and your tables and figures) should talk about the variables themselves rather than the name or number they might have in the survey (e.g., don't talk about q55 but instead talk about income or years of schooling). The units of the variables should be made clear. The writing should be in clear, English sentences, with writing in standard prose format along with the relevant figures and tables. Your tables should look professional and aesthetically pleasing and should not simply be copied and pasted from raw R output. See your stargazer handout for more advice on using stargazer to produce professional tables of regression output.