ECON 322: Econometric Analysis 1

Final Project

For this project, use the file dataXXX.rda that was sent to you by email, where XXX is your student ID. Make sure you received the file with the right student ID before starting. The file contains two dataset: Hours and Names. The first is for Part A and the second for Part B. Each part must be organized as a report. So even if I describe what you have to do using an enumeration, I do not want it to look like a question-answer type of document. You must provide all the codes that you used to obtain your results at the end of the document inside a separate section. You are evaluated on your ability to use as many concepts covered during the term as possible. Here are the concepts I want you to use for Part A and B:

- Interpretation of coefficients with significance tests.
- Prediction and partial effect (average or not) with their confidence intervals. Your models must include square variables and interactions.
- Inference with one and multiple restrictions. In that case, you can use the asymptotic distribution or the exact one, but you have to justify your choice.
- You must use dummy variables and interpret their coefficients.
- Test the equality of coefficients between two groups (Topic 7).
- Test heteroscedasticity and use robust tests and confidence intervals whenever the null hypothesis is rejected.
- Use weighted least squares when it is possible and justified (Topic 8). You can choose not to use this method, but I want to see a justification somewhere.
- Whenever you show a result in the form of numbers or graphs, I want you to explain what the result means and what it implies.
- J-test for non-nested models (Topic 9).
- Specification test (RESET) from Topic 9. This is used to select the functional form for your model and verify if the heteroscedasticity is not caused by a bad choice of function.
- Discuss whether measurement error is an issue and how it may affect the validity of your results.
- Discuss whether you have proxy variables that you can use. If you do, explain what is the effect of using them instead of the true variables (Topic 9).
- Using the Cook's distance to detect the presence of outliers (Topic 9).
- I want to see some graphs. There are many options from the term, so it is up to you to decide which ones are the most appropriate for your analysis.
- Anything else that is not listed and you think should be included.

I remind you that the assignment must be done individually. I expect all assignments to be very different from each other. For example, the choice of variables to include in your models and the functional forms to use is up to you. Given the number of variables, it is very unlikely to have two assignments with the same variables and functional forms.

Part A

The dataset Hours is a subset of the 1976 Panel Study of Income Dynamics (PSID) that contains the following 19 variables:

- **hours**: Wife's hours of work in 1975.
- **youngkids**: Number of children less than 6 years old in household.
- **oldkids**: Number of children between ages 6 and 18 in household.
- **age**: Wife's age in years.
- **education**: Wife's education in years.
- **wage**: Wife's average hourly wage, in 1975 dollars.
- **hhours**: Husband's hours worked in 1975.
- **hage**: Husband's age in years.
- **heducation**: Husband's education in years.
- **hwage**: Husband's wage, in 1975 dollars.
- **fincome**: Family income, in 1975 dollars.
- **tax**: Marginal tax rate facing the wife, and is taken from published federal tax tables. The taxable income on which this tax rate is calculated includes Social Security, if applicable to wife.
- **meducation**: Wife's mother's educational attainment, in years.
- **feducation**: Wife's father's educational attainment, in years.
- **unemp**: Unemployment rate in county of residence, in percentage points.
- **city**: Does the individual live in a large city?
- **experience**: Actual years of wife's previous labor market experience.
- **college**: Did the individual attend college?
- **hcollege**: Did the individual's husband attend college?

The purpose of this first part is to build a model that explains the number of hours worked by women in 1975. The dependent variable is therefore **hours**. Notice that hours and wage are equal to 0 when the women are not in the labour force. Models with dependent variables that contain many 0's are not usually estimated by OLS, but we ignore it for this project. You just have to be aware that you cannot take the log of variables that contain 0's. Alternatively, if you want to consider taking the log of these variables, you can restrict your sample to individuals with strictly positive hours. This is up to you. This part should be organized as follows:

- Explain which variables should be included in your model. At this stage, you should not use R. You have to base your decision on your intuition and on concepts covered during the term. Which variable do you think is likely to affect hours? Why do you think they are important? Do you think omitting them is likely to create a bias? Are there any variables you do not want to include? Why? Do you think that adding some variables will create some multicollinearity? The only rule that you have to respect is that at least one variable must be a dummy variable and at least one must be a continuous variable. Do not to select more than 6 independent variables.

- Once you have selected your variables, keep them until the end. Consider 3 or 4 different models. For example, you can try the log-lin, lin-lin, add square variables, interactions etc. The only rule you have to respect is that you are required to have at least one square variable and at least one interaction.
- Use R to select the best model. This is the ideal stage to use inference (f-test, t-test, RESET test, J-test) and by making sure your tests are robust to heteroscedasticity if the null is rejected. Show me that you know how to use the different concepts covered during the term.
- Once the model is selected, start your analysis: interpretations, detection of outliers, measures and interpretations of some average partial effects, predictions for different groups, tests equality of coefficients between two groups, graphs, etc.

Part B

The dataset Names is a subset of the dataset used by M. Bertrand and S. Mullainathan for their research paper "Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labour Market Discrimination", published in 2004 in the American Economic Review. It contains the following 27 variables:

- **name**: factor indicating applicant's first name.
- **gender**: factor indicating gender.
- **ethnicity**: factor indicating ethnicity (i.e., Caucasian-sounding vs. African-American sounding first name).
- **quality**: factor indicating quality of resume.
- **call**: Was the applicant called back?
- **city**: factor indicating city: Boston or Chicago.
- **jobs**: number of jobs listed on resume.
- **experience**: number of years of work experience on the resume.
- **honors**: Did the resume mention some honors?
- **volunteer**: Did the resume mention some volunteering experience?
- **military**: Does the applicant have military experience?
- **holes**: Does the resume have some employment holes?
- **school**: Does the resume mention some work experience while at school?
- **email**: Was the e-mail address on the applicant's resume?
- **computer**: Does the resume mention some computer skills?
- **special**: Does the resume mention some special skills?
- **college**: Does the applicant have a college degree or more?
- **minimum**: factor indicating minimum experience requirement of the employer.

- **equal**: Is the employer EOE (equal opportunity employment)?
- **wanted**: factor indicating type of position wanted by employer.
- **requirements**: Does the ad mention some requirement for the job?
- **reqexp**: Does the ad mention some experience requirement?
- **reqcomm**: Does the ad mention some communication skills requirement?
- **reqeduc**: Does the ad mention some educational requirement?
- **reqcomp**: Does the ad mention some computer skills requirement?
- **reqorg**: Does the ad mention some organizational skills requirement?
- **industry**: factor indicating type of employer industry.

This is a field experiment. It means that the authors are trying to reproduce the conditions of a lab experiment in which treatments are randomly assigned to individuals. This is done by generating CV's randomly and sending them to potential employers. As a result, CV's with African-American and Caucasian sounding first names have an equal chance to have high or low experience, have a college degree, etc. The purpose of this project is to test the presence of racial discrimination in the labour market. The dependent variable is **call** and the coefficient of interest is the one attached to **ethnicity**. Since call is a categorical variable, you need to convert it to a dummy variable (1 if call=yes and 0 otherwise). The model for this part is therefore a linear probability model (LPM), covered in Topic 7.

Because of the nature of the data, detecting discrimination on average is possible by simply running the regression:

$$call = \beta_0 + \beta_1 ethnicity + u\,.$$

However, this model only detects discrimination on average. It does not tell us, for example, if discrimination varies from industry, sex or education. That could be a starting point for your project to test if it exists on average, but I want you to dig a little deeper. Here is how that part should be organized.

- Formulate 4 different questions related to discrimination and build one model for each question. For example, (that example cannot be used), is discrimination different for male and female workers? A model that would allow to answer the question is:

  $$call = \beta_0 + \beta_1 ethnicity + \beta_2 male + \beta_3 male \times ethnicity + u\,.$$

  This model allows us to compare the effect of having an African-American sounding first name on the probability of being called back for male and female workers. You can compare it by city, by industry, by job wanted, etc. The rules are

  - All your models must include experience and experience^2.
  - You may also include other regressors if you think it is necessary. For example, if you add computer, you may also add reqcomm (the job requires computer skill). You just need to justify why you want to include them.
  - You have to interact some characteristics (e.g. gender:city or industry:gender) in at least 2 models.
  - At least one of your questions must be in the form: Is the effect of experience on the probability of being called back the same for group A and group B? For this question, you must interact experience and experience^2 with the appropriate dummy variable.

Formulate your questions, justify why do you think they are interesting questions, write down the models that you want to use to answer the questions and justify your choice of models.

- Estimate the models, interpret the coefficients and answer your questions by running the appropriate tests and by constructing confidence intervals. Notice that LPM's are heteroscedastic by construction, so all your tests and confidence intervals must be robust to heteroscedasticity.
- Using your model in which you interact experience and experience^2 with one or more dummy variables, measures and interpret some average partial effects and their confidence intervals.
- Using your model in which you interact experience and experience^2 with one or more dummy variables, compare the effect of experience on the probability of being called between two or more groups. To compare them, plot the predicted probabilities and their confidence intervals with respect to experience for each group. Interpret your results.
- Conclude by summarizing your main findings.