Problem set 11: week 15

1. Open a do file in Stata. All the commands we use in this problem set will be copied into here. This is so you can recall what we have done, and the analysis can be repeated. It will also be useful for you to annotate the do file as you go along. Load the wage5.dta dataset into Stata. Open a log file to record the output from Stata.

2. ALWAYS ALWAYS ALWAYS CLEAN THE DATA. To clean the data check the distributions of the variables (using the tabulate and summarize commands). Do all the values make sense? If the values don't make sense should you include those observations in your analysis? Are there any missing values? How many? Is the variable reliable? If not, should you use that variable in your analysis?

3. To estimate a regression you need to use the regress command. In this question your primary objective is to maximise the R^2 in the regression of log wage (lwage) on any of the independent variables in the data set. To do this you can enter the any other variables as controls (just add education, experience, hours, ... etc), you can also take polynomial transformations (for example square your variables such as experiences $\times$ experience, ... etc), interact any independent variables together (for example generate a new variable that equal to education times experience), ... etc. You can include as many variables as you like; the greater the number of variables the greater the likely value of the R^2. **Once you think you have finished building your model, interpret the estimated coefficient on education.**

4. Our primary goal in this question is to answer the research question: Does education impact wage? Run the following regressions and explain **why** you have entered the control variables you have and provide interpretations:

- $ln(wage_i) = \alpha + \beta_1 educ_i + \epsilon_i$.
- $ln(wage_i) = \alpha + \beta_1 educ_i + \beta_2 exp_i + \beta_3 exp_i^2 + \epsilon_i$.
- $ln(wage_i) = \alpha + \beta_1 educ_i + \beta_2 exp_i + \beta_3 exp_i^2 + \beta_4 IQ + \epsilon_i$.
- Out of the remaining variables in the data set, pick two to include in the model, providing reasons for why you do. Interpret the results.

5. Aside from controlling for observable characteristics, are there alternate ways/methods that may be more useful for finding the causal impact of education on wage?