Assignment 1
Applied Econometrics
ECON 5645 \ 6645

Due: January 24th, 2022 (by midnight)

Question 1 [5 marks] Please see the attached 'birth' dataset which consists of a hypothetical cross-section of regional data on mean birth weight (in grams) and GDP per capita (in US dollars).

(a) Run a regression of birth weight on GDP per capita. Present your work in a table and interpret the results, with emphasis on the magnitude and statistical significance of the parameter estimates. Are there concerns over an omitted variable bias? Explain. [Hint: to run a regression in Stata, type "reg y x" (inputting the correct variable names in lieu of "y" and "x")]

(b) For each observation, predict the mean level of birth weight using the results you found in part (a). Graph the fitted values, overlapping a scatter plot of the birth weight and GDP data. [Hint: To derive the fitted values, type "predict y_hat" after running your regression. Additionally the graph will require a command such as: "twoway (scatter y x) (line y_hat x)" - however, challenge yourself to present it in a more professional manner!]

Question 2 [5 marks] Please see the attached 'CCHS' dataset which consists of a subset of variables from the 2014 wave of the public-use Canadian Community Health Survey public-use.

(a) The Grossman model posits that an individual's stock of health is a function of income, education, and age. Test this hypothesis by running a regression of health utility on the above variables.[1] Include income in natural log terms. Based on this regression, present the results in a table and interpret.

(b) Some suggest that age may not have a linear relationship with health. Test this conjecture by including a quadratic term for age (i.e. age^2). Present the results in a table. What do they results suggest? [Hint: to derive age^2 in Stata, type "$gen\ age\_sq = age * age$"]

[1] The health utility index was derived by a team of researchers at McMaster University where a value of 1 denotes perfect health, and 0 indicates death; values below 0 suggest health scenarios worse than death.

(c) The "Healthy Immigrant Effect" posits that immigrants, upon arrival, tend to be healthier than those who are Canadian-born. However, as years since migration rise, their health falls to that of a typical Canadian-born individual. Test this hypothesis by including both the immigrant dummy variable (imm) and the years since migration variable (ysm) which equals 0 for Canadian-born respondents and counts the number of years since arrival for immigrants. Interpret your findings and present them in a table.

Question 3 [5 marks] Please see the attached 'SLID' dataset which consists of a subset of variables from the public-use 2010 wave of the Survey of Labour and Income Dynamics.

(a) In this question we will examine the returns to education using the Mincer Equation. Begin by deriving a set of dummy variables for level of education using the variable '*educ_level*'. That is, for each category in '*educ_level*', create a corresponding dummy variable (e.g., 'HS' = 1 if the respondent's highest level of education is high school completion; 0 otherwise). Next, do the same thing for region of residence using the '*region*' variable.

Now regress wage (in dollars) on your sets of dummy variables regarding education and region of residence while also including controls for years of experience, being in a union, having a managerial position, and working for a large firm. Given the survey is stratified, apply appropriate population weights. In terms of the reference categories for education and region, use a Bachelor's Degree and Ontario. Present your results in a table and interpret. [Hint: Population weights can be applied to a regression with the command '[pweight=wgt]' to the end of the regression syntax, where wgt is the hypothetical name of the weight variable.]

(b) Let's return to a more 'traditional' version of the Mincer equation, while adding a discrimination twist. Our explanatory variables of interest are: years of education, years of experience, being in a union, having a managerial position, and working for a large firm.

Interact each of these variables with the female dummy variable.[2] Regress wage on these explanatory variables along with the corresponding interaction terms.

Present your results in a table and interpret. Specifically, how do returns to education, experience, being unionized, having a managerial position, and working for a large firm differ based on gender?

[2] You may choose to do this by manually deriving interaction variables ($X_j \times female$) or by using the Stata interaction term ## in your regression equation. In doing the latter, you must tell Stata which variables are continuous and which are dummy variables. To do this, use the prefix "c." for education and experience (it stands for "continuous") and "i." for unionization, having a managerial position, working for a large firm, and being female (it stands for "indicator" - a synonym for "dummy"). Thus, a hypothetical regression in Stata, where X_1 is continuous and X_2 is a dummy variable, would look as such: $reg\ \ y\ \ c.X\_1\#\#i.female\ \ i.X\_2\#\#i.female$.