**DEPARTMENT OF ECONOMICS**
**ECON 4041H – RESEARCH METHODOLOGY**
**Winter 2020, Peterborough**

# Assignment #4

**Due date**: April 8, 2020

**Instructions**: You are required to analyze each question below, and provide a write-up of your findings. The write-up may be brief, but should be a few paragraphs. You should explain what you analyzed, comment directly on your statistical results, then sum up what your findings mean. Marks will be awarded based on depth of analysis. That means you should provide additional analysis where it assists in interpreting the results. In particular, you may find **emmeans()** useful in many cases.
Questions 1 and 2 are worth 30 marks. Question 3 is worth 15 marks.
You may work with others, but each of you is responsible for submitting your own problem set solution. Submit solution through SafeAssign.

1. Use the Labour Force Survey 2018 dataset "lfsA4.rds" to assess two effects: industry wage differentials, and the gender wage gap. The following variables are to be used:
   - *hrlyearn:* hourly earnings in $ (wage)
   - *educ:* educational attainment
   - *age_12:* five-year age categories (drop the top category "70 and over").
   - *sex*
   - *union:* three categories, recode to treat the two categories"Union member" and "Not a member but . . . " as one category
   - *cowmain:* class of worker: public, private. There is also a self-employed category which can be dropped because self-employed are not paid workers, so have no wages reported.
   - *firmsize:* self-explanatory
   - *naics_21*: industry classification

   a. Test for industry-wage differentials. All else being equal, do wages differ by industry? Estimate a model of wages (*hrlyearn*)—use log(*hrlyearn*) as your dependent variable. You will choose the appropriate specification, ie. the covariates to include. You will want to include covariates that would explain wage differences, and then see if there are still industry effects (*naics_21*) explaining wages.
   b. Test for the gender-wage gap. Use the same basic model as in part a., but now see if you can eliminate the gender-wage gap by inclusion of the following four covariates: education, union, class of worker and firmsize. You may run one model or several with subsets of the covariates to be tested; that is up to you. In particular, you are to determine if *sex* interacts with any or all of the four variables. What model you settle on and how you present results is up to you.

2. Canada is a country of immigrants. Use the 2016 Census file "cen16.rds" posted to Blackboard to see how immigration status affects labour market outcomes. Your task is to explore the effect of immigration status on income. Simply put, do immigrants' earnings differ from those born in Canada?

   Use $\log(totinc)$ as the dependent variable you are explaining. As covariates, at a minimum include a polynomial of age whose polynomial order you determine, educational attainment, sex, industry, and of course immigrant status. Use *citizen* to create a binary variable to capture immigrant status. Use a model to capture the effect of immigration status as you see appropriate. Discuss and defend your choice of model.

   The following variables are useful.

   - *citizen:* immigrants are those who are not "Canadian citizen by birth"
   - *totinc:* total income—numeric.
   - *cow:* class of worker
   - *hdgree:* educational attainment
   - *locstud:* location of study–where highest educational attainment was obtained
   - *agegrp:* age categories
   - *sex*
   - *naics:* industry classification
   - *pob/pobm/pobf:* place of birth of individual/their mother/their father
   - *vismin:* categories of visible minority status

   Restrict the sample to:

   - *agegrp:* ages $>$ 19 years and $<$ 85 years
   - *cow:* categories "Employee" and all "Self-employed"
   - *totinc:* values of *totinc* $> 0$.

   I have left this open to you to analyze as you see appropriate. Are there any characteristics available in the census dataset that might explain income differences (check documentation file for descriptions)? For example, those who obtained education in Canada may have a different experience from those whose education was obtained elsewhere (*locstud*). Also consider that some variables may interact with immigrant status. Try different variables, but do not report every single attempt. Report those that you find interesting.

   There is no correct answer. There are only more thorough investigations. Consider that you have been called in as a consultant to explain to someone how immigrant status affects income, and given your knowledge of economics in general and the methods from this course, use the census data to address that question. Try to illustrate the effects by using **emmeans()** and/or **margins()** where appropriate.

3. Explore further the example we did way back in the first week looking at the correlation between Ease of Doing Business and per capita GDP. Explore variables that explain per capita GDP in addition to the Ease of Doing Business index. Download your data from World Bank: World Development Indicators.
   To download, click on link above, then:
   a. Choose Country on left, click on "Countries" box to turn it blue. Click on checkmark below on left and all countries should be selected.
   b. Choose Series, and pick variables that you think are appropriate by clicking on their checkbox. Your dependent variable is "GDP per capita (constant 2010 US$)", and the test variable is "Ease of doing business score(0 = lowest performance to 100 = best performance". Select those two variables. Select at least 5 other variables that you think might cause countries' per capita GDP to differ. Things like "Manufactures exports" might be plausible. Or maybe "Agricultural raw materials exports" might be inversely proportional to income. Don't use other income variables, like GNI.
   c. Choose Time, and pick a recent year. I recommend 2018. Too recent and data coverage might be incomplete. Just pick one year. We did not get to panel data techniques, so we will only look at a cross-section. Once done, on right, click "Apply Changes". Then we need to organize the data for easy importing. On the left, choose the tab "Layout". For Time, choose "Row". For Series, choose "Column". For Country, choose "Page. Then click "Apply Changes".
   d. Now download your selection. Choose "Download options" and select "CSV". Your data will download in a zipped file. Extract the file that does not have the word meta-data in it. Then rename the extracted file to something easy. You need to do one more step, edit the data. Open the data in Excel or text editor or other spreadsheet. Page to the bottom. There will be two rows with "Data from database: .." and "Last Updated..". Delete those two lines. Page back up to top. The column titles won't work in R. Replace them with shorter names without spaces. So replace "GDP per capita (constant 2010US$)[NY.GDP.PCAP.KD]" with something like "gdpc". Replace all other variable names too. Wherever a column title has a space, rename it something simple. "Country Name" becomes "ctry". You can drop the year columns. You have only one year so you don't need them. Now "Export As" and export as a ".csv" file.
   e. To open the file 'wdi.csv' in R, use the R command
   wdi <- **read.csv(file =** 'wdi.csv', **header =** TRUE, **sep =** ',', **na.string =** '..')

   See if your set of explanatory variables affects the impact of the Ease of doing business index on explaining per capita GDP. In your regressions, use per capita GDP as the dependent variable. Test whether you should log() transform your variables by generating some scatterplots of per capita GDP against your other variable. Vary them so one is log() transformed, then the other, then both. Choose what seems to look best. Then progress to the next possible covariate. Rinse. Repeat.

   Note: when using log()-transformed variables, you need to drop all negative and missing values. An easy way to do that in a regression is to specify a subset of positive values only

   mod <- **lm**(log(y) $\sim \log(x_1) + \log(x_2)$, data = **subset**(wdi, $y > 0$ & $x_1 > 0$ & $x_2 > 0$))

   Write up your results in a few paragraphs explaining what you see. Include regression output.