ECON 490: *Current Population Survey (CPS) Class Activity*

For this activity, we will explore the CPS capstone data set posted to Dropbox. You will also get some practice working with data in R in a less structured context than what we've seen previously in the R coding activities.

To get started:

- Download the CPS data set as well as the PDF documentation from Dropbox – there is a link on the "Capstone Data Resources" page on Canvas.
- Open the data in R using either 1) `File > Open File...` or 2) `setwd() + readRDS()`
- Create a new R script file and include "ECON 490 CPS Activity: Your Name" at the top as a comment.

As you work through the questions below, write out your answers (with comments to indicate each question) in your code file. Everyone will answer the first set of questions below; you'll then answer the questions corresponding to your group number in the second section.

Questions for Everyone

The following questions are the kinds of basic questions you should ask whenever opening a new data set. Check each of the questions below in your data set (use a comment to separate each question, e.g., `# Q1`, etc.).

Q1) What years are covered in this data set?

Q2) How many observations are included in your data set? From the documentation, what is each row of the data?

Q3) What states do we observe in the data? What Census regions do we observe?

Q4) What proportion of observations have a missing value for the `hourly.earnings` variable? From the documentation, for whom do we observe hourly earnings?

Group Questions

Everyone should have a group number – you only have to complete the questions below corresponding to your specific group number!

GROUP 1: *Poverty Status and Migration*

- We want to explore how poverty status and migration are related – are people living in poverty more or less likely to move than those who are not?
- Start by exploring relevant variables in the data set.
    - How are poverty and migration status measured?
    - What is the relevant time span for both variables?
- Next, calculate the proportion of people who have moved at all in the past year (i.e., create a dummy variable using migration status).
    - Using this dummy variable, compare the average migration rate for those who did and did not have family incomes below the poverty line over the last year.
- Finally, compare how longer moves (i.e., across state-lines and international) vary by poverty status.

GROUP 2: *Average Wages across Industries*

- We want to explore how average wages vary across industry. To get started, use the Dropbox documentation to explore the industry variable.
    - How is this variable coded in our CPS data set?
    - How are unemployed folks included in this variable?
- Note that we just get numeric values of industry codes. To see what these codes actually mean, Google "IPUMS CPS industry 2017" to get definitions of the industry codes included in the CPS data set.
    - Using `tidyverse`, find the most common non-missing values of industry code.
    - Using the IPUMS definitions, what are the 5 most common occupations in our data set?
- Create a subsample of the CPS data set including individuals are ***currently employed*** in any of the 5 industries you've identified above.
    - Using this data set, run a regression with `inc.wage` as the outcome variable, and industry as the explanatory variable to see how average earnings vary across these categories.
    - Remember to make industry a factor variable!
    - What industry has the highest average earnings? What industry has the lowest?

GROUP 3: *Health Status and Employment*

- We want to explore how health status covaries with labor market outcomes – are healthier people more likely to be employed? Are they more likely to have health insurance?
- To get started, explore the Dropbox documentation for the health status variable.
    - How is health status defined? How is the data collected (i.e., who is reporting this variable – individual respondents or their doctors?)
    - Check the coding for the employment indicator variable and the health insurance variables to confirm how they are coded.
- Next, calculate average employment and insurance coverage rates across each level of health status.
    - Do the patterns in the data follow your expectations?
- Suppose you wanted to include health status as an outcome variable in a regression. Could you use the current version of this variable as currently coded?
    - Transform health status into something that you can include as an outcome variable and run a regression with health status as the outcome with the employment indicator variable from above and age as explanatory variables.
    - How do these results compare to the summary statistics above? How can we interpret each coefficient?