**Assignment 2: Setting up a Git repository**

**Due date: Friday May 14, 2020 by 12:00PM**

Gentzkow and Shapiro’s manual that you just read asserts that version control is an important part of the empirical workflow of any competent empiricists. The most popular form of version control is git. But git is for many people somewhat confusing, so I’d like to use this assignment as an opportunity to force you to set up your first git repository.

**Slack**

Install slack on your phone and desktop. Create an account. I will send out the work directory later.

**Git and Github**

Grant McDermott at University of Oregon has created a helpful deck of slides explaining what git is, how to set it up for your project, and how to integrate into your workflow. While his deck is primarily aimed at R users, it is not required. I use Stata and it incorporates with that software perfectly fine because git transfers move from a local directory of folders on my physical computer to the repository. I haven’t found a better soft introduction to git than Grant’s deck of slides, so I would like to use it to guide you through some steps and exercises.

<https://raw.githack.com/uo-ec607/lectures/master/02-git/02-Git.html#1>

1. Briefly write the difference between git and github with a simple explanation of what each is and a simple explanation of how each one is different and distinct from the other.
2. What is the benefit of using git to organize your empirical research? What about using git is challenging for you for right now? What steps can you take to minimize those challenges such that you can adopt git for this class?
3. What are the four main Git operations? What does each do? How are these different from one another?
4. The first step in your new empirical workflow is the creation of a Github repository (“repo”). You can either do this independently or do this through R functionality. You need to create a github account, then create your first repository called “Titanic”. Initialize with a Readme and create the separate folders that we discussed in class on Wednesday.
5. Post a link to your repository.

**OLS**

Next we will estimate a simple OLS model using a scripting file, either in Stata, R, python or your preferred software. The data is called titanic. It is the manifest of the Titanic ship with a marker stating whether the passenger died or not. You can download this data from my Github directory here:

<https://github.com/scunning1975/mixtape/raw/master/titanic.dta>

1. Download and place the above dataset into your /Titanic/Data subdirectory, commit and the push to Github.
2. Create a program called titanic (.do, .R, or equivalent in python, etc.) that does the following
   1. Has a header naming the file, the author of the file, a description of the file and the date last updated.
   2. Generate a dummy variable for gender (female=1)
   3. Generate a dummy if they’re seated in first class versus in other parts of the ship
   4. Estimate the following regression model twice – once without adjusting for heteroskedasticity and once adjusting for heteroskedasticity using the robust estimator of the variance we dicussed in class.
   5. Assume this is the short regression. That is, if individuals in first class were more likely to be women and children, and women and children were more likely to die, then we have an omitted variable bias problem. Decompose the regression parameter into the true parameter and any biases using the omitted variable formula we discussed in class.
   6. Given what you know from that decomposition, what do you think is likely to be true about ? Do you think it’s likely larger than the true effect of first class on surviving the Titanic sinking, the same as the true effect, or smaller than the true effect?
   7. Now estimate the following equation and interpret the sign on first class with and without adjusting for heteroskedasticity. How does it compare to the earlier regression?
   8. Create a “beautiful table” containing regression output from both equations. Your table should contain coefficients with standard errors in parentheses below each coefficient in each cell. Results should be organized vertically where each column is a regression model. You should report mean values of the dependent variable, the number of observations, and the constant.
   9. Interpret the coefficient on the constant and the first class variable for all four models. Assuming the second equation is the long regression, what is the causal effect of first class on surviving? Does your answer to (f) appear to have been true?
3. Create a separate table in which you implement the regression anatomy we discussed in class *manually*. That is estimate the following equations:

Where the first equation is the long regression, the second equation is an auxiliary regression of first class onto all other covariates, and the third equation is a regression of survivor onto the residuals from the second equation. Compare the coefficient on first class from the first equation to the coefficient in the third equation. Explain why you find this.