

SOC 114
SOCIAL DATA SCIENCE

University of California – Los Angeles
Spring Quarter 2021

Exercise 1

Instructions:

Your completed exercise must be submitted through Turnitin to your T.A. no later than 5:00 p.m., **Tuesday, April 20, 2021**. This quiz is worth 10 points (i.e., 10 percent of your final grade).

1. Describe your working directory structure. Give an accounting of level 0 to level 3 (or whatever level makes sense in your case) directories, as I outlined in class using my directory structure. Use the project you will complete for this class as an example directory. Then describe some of your mnemonic conventions for naming files for that project.
2. Using data from the 2018 GSS, tabulate the distribution of schooling (VAR: degree). What do you notice about the distribution? Now create a binary variable of no college completion (i.e., anything less than a 4-year college-degree) versus college completion (a 4-year college degree or more). Tabulate your college completion variable. What percentage of respondents in 2018 had a college degree? Cross-tabulate respondent's highest degree variable to respondent's highest year of school completed (VAR: educ). What do you notice?
3. Using data from the 2018 GSS, consider the relationship between mothers' education (VAR: maeduc) and respondents' education (VAR: educ). What is the independent variable in this relationship and what is the dependent variable? Simplify the two variables into 4 categories: (1) less than high school (< 12 years); (2) high school degree (12 + years); (3) some college (13-15 years); and (4) 4-year college degree or more (16 + years). Cross-tabulate the two newly created variables. Among respondents whose mothers did not complete high school, what percentage obtain a 4-year college degree? Among respondents whose mothers completed college, what percentage obtain a 4-year college degree? What other sociologically interesting patterns do you observe? Use a χ^2 test to test the relationship between these two variables. What does it show?

4. Using data from the 2018 GSS, consider the relationship between number of siblings (VAR: sibs) and respondents' 4-category education created above. What is the independent variable in this relationship and what is the dependent variable? What is the mean and median number of siblings for each level of educational attainment? What is the correlation between number of siblings and the continuous measure of years of schooling? What does this tell you about the relationship between family size and educational attainment? Now consider the correlation between number of siblings and educational attainment 'adjusting' for mothers' education. What do you observe now?
5. Include your log file of all the calculations for this assignment.