**POL 212**
**Winter 2024**
**Assignment 4**

1. (Re)load the ANES 2022 Pilot Study data in R (anes_pilot_2022_stata_20221214.dta). It will also be helpful to review the questionnaire and/or user guide to get a sense of which variables are included and how they are coded.

2. Collect the following variables: race, gender, age, income, education, feeling thermometer (Trump), and feeling thermometer (Biden).

    a. Make sure missing responses are coded as NA.

    b. Each of the seven variables should have classes (e.g., “numeric” or “factor”) corresponding to your choice for its level of measurement.

    c. Perform any variable rescaling or transformations that you think are appropriate.

3. Use tidyverse/“dplyr” functionality to report the group means of feeling thermometer (Biden) for the racial categories you created.

4. Estimate the following linear regression model using the lm() command in R and provide a table the results using “xtable”, “stargazer”, “jtools”, or another formatting package in R.

    y ~ XB + E

    y = (feeling thermometer Trump – feeling thermometer Biden)
    [difference between the two feeling thermometers]
    X_1 = race
    X_2 = gender
    X_3 = age
    X_4 = income
    X_5 = education

    *note: you may have more than five betas (regression coefficients) depending on how many of the X variables are coded as “factor.”

5. Repeat #4, but first combine education and income into a single summated rating scale called "SESclass". Use this as an explanatory X variable instead of education and income separately.

6. Write an R function that returns only the $R^2$ value from an "lm" object in R. (Hint: the command "names(summary(res))" will help you locate where this is stored.)

7. Use this function to extract the $R^2$ values from the two above models (#4 and #5). How much does $R^2$ decrease when we combine education and income in a single (index) variable?

**CHALLENGE QUESTION: ANSWER C1 or C2**

C1.) Design and employ a resampling strategy to simulate the underlying sampling distributions for the regression coefficients from any model above. Provide your results (preferably with histogram(s) or density plot(s)) and compare with the model estimates (of the model coefficients and standard errors) you obtained earlier (i.e., without resampling).

C2.) A new California lottery game gives you a 5 x 10 board, with sequential spaces numbered from 1 to 50. The zero space is marked "start", and your token is placed on it. You are handed a fair six-sided die and three coins. You are allowed to place the coins on three different (nonzero) spaces. Once placed, the coins may not be moved. After placing the three coins, you roll the die and move your token forward the appropriate number of spaces.

If, after moving the token, it lands on a space with a coin on it, you win. If not, you roll again and continue moving forward. If your token passes all three coins without landing on one, you lose.

Based on simulation results, on which three spaces should you place the coins to maximize your chances of winning?