Problem Set 1

Goal: Explore Treatment Effects & Randomization Inference
Application: The Abecedarian Project

In this problem set, we will look at the effects of The Abecedarian Project, a preschool intervention for children from low-income families implemented in North Carolina the 1960s. Please use R Markdown to create a html or pdf file with your solution. Please annotate all of your code thoroughly!

Part 1. Setup and Descriptive Statistics

1. Use `readRDS` to import the dataset, `earlytrainingproject_clean.rds`.
2. There are fourteen outcome variables that we'll be using: five child variable, three teen variables, and six adult variables. The five child variables are `iq5` (IQ age 5), `iq6` (IQ age 6), `iq12` (IQ age 12), `retn12` (whether or not the child was retained at age 12), and `iep12` (whether or not the child was receiving special education at age 12). The three teen variables are `iq15` (IQ at age 15), `hsgrad` (whether or not graduated high school at age 18), and `parent` (whether or not he/she was a parent by age 19). The six adult variables are `college`, `employed`, `convicted`, `felon`, `jailed`, and `marijuana` (all binary and self-explanatory). Create three vectors, `varsChild`, `varsTeen`, and `varsAdult`, that contain the names of the corresponding variables. Create a fourth vector, `vars`, that contains all the outcome variables, by concatenating the three vectors you just made.
3. Create a new data frame containing only the variables that we will use in the rest of our analysis: `SUBJECT` (subject ID number), `DC_TRT` (treatment indicator), `SEX`, and the outcome variables in `vars`.
4. Use `describe` to explore the data set. Notice that there are missing values for some variables.
5. Use `na.omit` to create a new data frame containing only those observations with no missing values. How many observations are dropped by removing the missing values?
6. What values does the variable `SEX` take?
7. Create a binary variable called `treat` (using `DC_TRT`) that is 1 if treated and 0 if control. Create a binary variable called `fem` (using `SEX`) that is 1 if female and 0 if male.
8. Use `Stargazer` to make tables of descriptive statistics for the variables in `vars`, separately for controls and treated observations.

Part 2. Estimates of Average Treatment Effects

1. Using `sapply` and `tapply`, create a table compute the mean of each variable in `vars` for the control observations, the corresponding means for treatment observations, and the difference in those means as estimates of the average treatment effects. Use `Stargazer` to create a table of the estimated average treatment effects.

2. Repeat question (1) above, but this time do the analysis separately by gender to produce a table of estimated average treatment effects conditional on gender.

Part 3. Randomization Tests

As we did in Lab 2, in this section we will use randomization inference to compute p-values for each of the 14 outcomes — first for the whole sample, then separately by gender. You might wish to first review Lab 2.

1. Start by setting the seed. Then, generate a matrix of permutation for treatment, where each column is a different permutation of the treatment vector. Use the "genperms" function for this step. Also, use "genprobexact" to compute the exact probability of treatment in the sample.

2. Next, create a function that take the observed outcomes as inputs and spits out a large vector of simulated estimates under the null hypothesis that the treatment effect is zero for each observation. The function should be composed of two lines of code: one using "genouts" and one using "gendist". "genouts" takes the sample outcomes and treatment vectors and generate a vector of hypothetical outcomes under the null. Note that the null is that there is zero treatment effect on each individual, so that $Y_1 = Y_0$ for all individuals and thus the average treatment effect is zero. "gendist" takes this hypothetical outcome, the permutation matrix, and the probability of treatment and create a large vector of simulated estimates of the average treatment effect. Even though the average treatment effect is zero in each simulation as we are imposing the null in each simulation, the estimated average treatment effect generally be non-zero and will vary across the simulations because of the permutations of the hypothetical assigned treatment vectors.

3. Then, use sapply function to apply the function to all 14 outcome variables. This step will give a large matrix, where each matrix is a simulated distribution of the estimates of the average treatment effects on each outcome under the null. Note that the absolute value of the estimate is the value of the test statistic.

4. In the last step, use apply function to evaluate whether the absolute value of the simulated estimate in the large matrix is larger than the absolute value of the estimate computed on the actual sample. Finally, calculate the mean of this true/false matrix to obtain the p-values. Create a table to present the resulting p-values.

5. Discuss your findings.

6. Repeat this process separately on the sub samples for male and female. Discuss your findings.