Causality Recap Assignment

Dr. Soliman

Make sure you answer each question (and subsequestions!) to receive full credit, and use your lecture notes and any other resources to ensure these answers are complete.

Load the *STAR* data from here and assign it to an object called `star_df`. Read the help for the data here to understand what the variables correspond to. (Note: the data has been *reshaped* so don't mind the "k", "1", etc. in the variable names in the help.)

```
star_df <- read.csv("~/Library/CloudStorage/Dropbox/Clemson/Econometrics Course/data for tasks/star_data
```

1. Filter the STAR dataset to only keep first graders and the small class and regular class groups. Check the data, as there still may be missing information... Call this object `star_df_clean`. Hint: If you wanted just grade 2, you would use a pipe with `filter(grade == "2")`. In this case, you want to keep first graders in either small or regular class groups, and therefore you would need to change and extend the previous `filter()`.

```
star_df_clean <- star_df %>% filter(complete.cases(.)) %>%
    filter(star %in% c("small","regular") & grade == "1")
```

2. Compute the average math score for both groups, and the difference between the two. (Use base `R`.)

```
# there are many ways to answer this, these are just two
# one way, if you hadn't filtered NAs with complete cases
mean_small = mean(star_df_clean$math[star_df_clean$star == "small"], na.rm = TRUE)

# alternative way
mean_small = mean((star_df_clean %>% filter(star == "small"))$math)
mean_small
```

```
## [1] 539.0885
```

```
mean_regular <- mean((star_df_clean %>% filter(star == "regular"))$math)
mean_regular
```

```
## [1] 526.4434
```

```
ATE = mean_small - mean_regular
ATE
```

```
## [1] 12.64506
```

3. Create a dummy variable `treatment` equal to `TRUE` if student is in treatment group (i.e. small class size) and `FALSE` if in control group (i.e. regular class size). *Hint:* you can create the dummy variable with `treatment = (star == "small")`.

```
star_df_clean <- star_df_clean %>%
    mutate(treatment = (star == "small"))
table(star_df_clean$treatment)
```

```
##
## FALSE  TRUE
##  2359  1786
```

4. Regress math score on the treatment dummy variable. Are the results in line with question 2?

```
lm(math ~ treatment, star_df_clean)
```

```
##
## Call:
## lm(formula = math ~ treatment, data = star_df_clean)
##
## Coefficients:
##   (Intercept)  treatmentTRUE
##        526.44          12.65
```

Yes, the coefficient on treatmentTRUE, the treatment assignment indicator for being in a small class, is identical to what we found in question 2.

5. How do you interpret these coefficients? Is this a causal estimate? Why or why not?

The intercept represents the expected average math score for first graders in the control group, which consists of students in regular-sized classes. Specifically, the expected math score for this group is 526.44. This value aligns with the average computed in Question 2. The slope coefficient captures the difference in expected math scores between first graders in the treatment group (small classes) and those in the control group. In other words, students in small classes are expected to score, on average, 12.65 points higher than those in regular-sized classes. This difference can also be verified by comparing it to the average difference calculated in Question 2. Since the data comes from a randomized experiment, this estimate can be interpreted as causal.