Section 4 Assignment

Instructions (read carefully):

- Each student must submit their own (independent) work.
- Assignments must be done using RMarkdown.
- Submissions must include the .pdf file and the reproducible .rmd file used to do the homework. R code for all applied questions must be provided and be executable in the .rmd file.
- Students should use the practices covered in Section 0 of the course to organize their folder structure and code (i.e., using RStudio Projects with the `here` package).
- This assignment is due electronically through CANVAS on Sunday April 9 2022, unless otherwise noted.

Question 1: You have been provided with a dataset from an observational cohort study to evaluate the effect of daily supplementation with 2000 mg of Omega-3 Fatty Acids (1200 mg EPA + 800 mg DHA) on the number of severe asthma attacks over a three month follow-up period. The outcome is a binary [0,1] indicator of severe asthma attacks (any versus none). The exposure is an indicator of whether the person took their supplement or not. There is also information on four strong baseline confounding variables: age (in years), body mass index (in kg/mˆ2), asthma medication use (regular versus irregular use), and a measure of overall diet quality. All of the variables in the dataset are complete (i.e., no missing data).

You've been asked to demonstrate to your clinical colleagues the types of missingness, and how they might affect the confounder adjusted association between Omega-3 supplementation and asthma attacks. Specifically, you've been asked to set a proportion of the observations for the Omega-3 supplementation exposure variable to missing under MCAR, MAR, and MNAR. The exact proportion of missingness doesn't matter.

Describe three different mechanism you would use to set a proportion of the exposure variable to missing if the missingness was MCAR (mechanism 1), MAR (mechanism 2), and MNAR (mechanism 3).

Question 2: Define non-monotone missingness. Describe the main complication introduced by the non-monotone missingness pattern. What type of imputation techniques can you use with non-monotone missing data.

Question 3: You are asked to estimate the ATE, and the outcome variable is missing 17% of it's observations. You and your colleagues are confident that this missingness is MAR. Your clinical colleague states that they want you to do a complete case analysis because you "should never impute the outcome." Describe an analytic scenario where a complete case analysis will still be valid, even though your missing outcome data are MAR. Describe an analytic scenario where a complete case analysis will NOT be valid.

Question 4: Use the `section4_q4_data.csv` data to estimate the average treatment effect between quitting smoking and the difference in weight between 1971 and 1982. Use marginal standardization AND inverse probability of exposure weighting to estimate this effect, and adjust for sex, age, and race using main effects terms only (i.e., no splines, no polynomials, no interactions). Use the `mice` package to impute any missing data.

Impute missing outcome data using all other variables in the dataset.

Impute missing age data using only `sex` and `qsmk`.

Impute missing race data using `age`, `sex`, and `wt82_71`.

Use the bootstrap to obtain correct 95% confidence intervals for the estimated ATE. Be sure to set appropriate seeds for reproducibility. Set the number of resamples to 200. It may take a few minutes for your code to run.