Final Exam: EPI 560

Instructions:

- Submissions must include the .pdf or .html file and the reproducible .rmd file used to do the exam. R code for all applied questions must be provided and be executable in the .rmd file.
- This assignment is due electronically through CANVAS on Sunday April 30th 2023 at 11:59PM.

The Data: These data were collected from a US-based study of the effect of daily low-dose aspirin on pregnancy loss among women who've experienced a loss in the past. Women started taking aspirin prior to conception, and stopped taking it after 6 months of trying with no conception, or, if they conceived, at 36 weeks gestation. The primary outcome of interest was successful live birth. The primary exposure of interest was an indicator of taking 81 mg of aspirin at least 5 days per week during every week of follow-up. To account for potential confounding, the authors of the study collected the following information to be adjusted for:

Variable description (variable name, type)

```
- eligibility stratum (eligibility, binary)
- number of prior losses (loss_num, ordinal)
- age (continuous)
- number of pregnancy attempts (time_try_pregnant, ordinal)
- BMI (continuous)
- mean arterial pressure (meanAP, continuous)
```

Exposure and Outcome:

- exposure: taking 81 mg aspirin at least 5 days per week consistently over follow-up versus NOT taking 81 mg aspirin at least 5 days per week.
- live_birth: indicator of successful live birth occurring any time between the beginning and end of follow-up.

Question 1: Using the `aspirin.csv` data, please estimate the **marginally adjusted** average treatment effect using IP-weighting and marginal standardization (i.e. g-computation) for the relation between aspirin and live birth on the risk difference scale. Adjust for eligibility stratum, number of prior losses, age, number of prior pregnancy attempts, BMI, and mean arterial pressure. Check if the positivity assumption is met. Code these variables appropriately, and use the appropriate variance estimator to obtain 95% confidence intervals.

Question 2: Please interpret the results (i.e., risk difference) of the analyses in question 1.

Question 3: Suppose you fit an outcome regression model to answer question 1 that included only seven main effect terms (one for the exposure, and one for each confounder). In other words, suppose that there were no additional spline, polynomial, or dummy variables in the model (i.e., only seven terms). How many possible 2-way interactions could you have included in this model? How many possible k-way interactions could you have included in this model?

Question 4: The outcome in the aspirin.csv data is an indicator of whether live birth occurred at any point during follow-up. The data **do not** include a variable for precisely when this outcome occurred during follow-up. Additionally, the zero value for the live birth outcome indicates that administrative end of follow-up for an individual (i.e., there were no withdrawals from the study). Is the aspirin study outcome subject to left and/or right censoring? Why or why not?

Question 5: Suppose the aspirin study above provided an outcome variable with three categories: live birth (coded as 2), pregnancy loss (coded as 1), and administrative end of follow-up (coded as 0). Pregnancy loss, in this scenario, is a competing event for live birth. Suppose further that you used IP-weighting or marginal standardization to compute the risk difference for the effect of aspirin on live birth, and you would have censored pregnancy losses. Specifically what type of risk difference would you have estimated? What kinds of problems would arise if you wanted to interpret the real-world effects of aspirin on live birth?

Question 6: Consider a dataset with the following variables:

- Y, the outcome under study
- X, the exposure of interest
- $\{Z_1, Z_2, ...Z_p\}$, confounders of interest

One of the research goals is to fit the following regression model and interpret the coefficient for the exposure:

$$E(Y \mid X, Z) = \beta_0 + \beta_1 X + \beta Z$$

The researchers you will be collaborating with would like to do a complete case analysis, and are not certain about the amount of missing data, or which variables are missing observations. Describe all the scenarios where conducting a complete case analysis will yield a valid estimate of β_1.

Question 7: Suppose we were interested in estimating the **conditionally adjusted risk difference** for the relation between aspirin and live birth. Describe three strategies you can use (including variance estimation) to do this with the aspirin data.

Question 8: Pick one of the methods from question 8, and implement it in the aspirin data. Include the **conditionally adjusted risk difference** and appropriate 95% confidence intervals.