Final Exam: EPI 560

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:

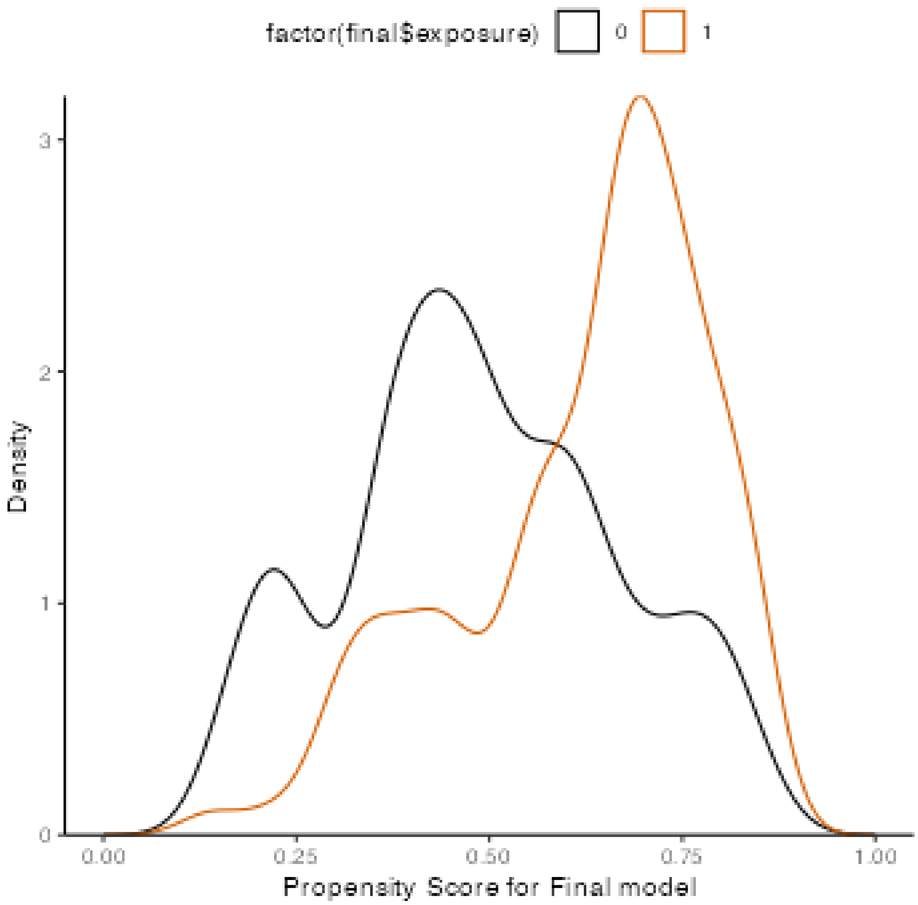
Variable description (variable name, type)

```
- 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.



The density plots do not precisely overlap, but we will also view a summary of the weights.

Min. 1st Qu. Median Mean 3rd Qu. Max. 0.5243 0.7378 0.8378 1.0058 1.0898 4.1050

The mean of the stabilized weights is essentially 1. The max weight is not large relative to the mean and min. This suggests that the weights are "well behaved." Thus, in this particular case even though the density plots do not precisely overlap, we are not concerned with violations of the positivity assumption.

Table 1: Marginally adjusted average treatment effect using IP-weighting and marginal standardization for the relation between aspirin and live birth on the risk difference scale

Method	Average Treatment Effect	95% CI
Inverse Probability Weighting	0.09	0.03-0.15
Marginal Standardization	0.09	0.04-0.14

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

There will be 9 more cases of successful live birth out of every 100 people in our sample if everyone took 81mg of aspirin at least 5 days per week consistently relative to if everyone did not.

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?

7 variables

We can have up to 120 k-way interactions.

$$2^7 - 7 - 1 = 120$$

Mathematically, there are 66 two way interactions based on the binomial coefficient.

$$\binom{7}{2} = \frac{7!}{2!(7-2)!} = 21$$

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?

The aspirin study outcome is subject to right censoring. Right Censoring occurs when an individual is enrolled in the study, but we don't know whether the individual experienced the event of interest (or not). The two primary causes of right censoring is due to participant withdrawal (not applicable for this study per the question) or, in the aspirin study's case, the study ends (also known as administrative censoring).

The aspirin study outcome is also subject to left censoring. Left censoring occurs when an individual is enrolled in the study and we know has experienced an event of interest (ie: live birth), but we have no information on when exactly the event occurred.

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?

When you censor competing risk, you estimate cause specific risk. Cause specific risks are interpreted as the risk of the outcome of interest if we completely prevented the competing event(s) from occurring. The problem with cause specific risks is that in the real world, we cannot prevent competing events to occur. In this case, we have to interpret our results in a way that illustrates how aspirin increase/decreased birth rates in a world where we completely prevented pregnancy loss. We could not know if aspirin had any association with pregnancy loss.

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.

There are three general classification of missing data patterns: **MCAR:** Data are considered missing completely at random (MCAR) if the probability of missingness does not depend on any other variables, either measured or unmeasured. **MAR:** Data are considered missing at random (MAR) if the probability of missingness depends on other variables, AND these other variables are measured and available in the data

MNAR: Data are considered not missing at random if the probability of missingness depends on other variables, AND these other variables are NOT measured, and thus not available in the data.

When we conduct complete case analysis, the validity of β_1 depends on the missingness pattern. When the probability of missingness does not depend on any other variables, measured or unmeasured, in the data set (MCAR), beta_1 will be valid after conducting a complete case analysis. Sometimes when data is MAR, we can still do a complete case analysis and get a valid beta 1 estimate. Specifically, if the missigness is dependent on variables in the data set that are confounders for the outcome and those confounders are accounted for in the mathematical model, the model itself accounts for bias introduced by doing a complete case analysis. We cannot have a valid beta 1 estimate if the missing data pattern is MNAR.

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.

We can first do **propensity score adjustment**. We use the propensity score to adjust the outcome regression model estimation. In our model we would essentially do

$$livebirth = exposure + propensityscore$$

and then use binomial distribution with identiy link which will provide us valid 95% CIs.

Using the propensity scores, we can also do a **propensity score stratification** where we create quantiles of the propensity score. We use these quantiles and quantify the effect of the exposure for each of the strata (the number of quantiles affects bias and variance. more quantiles = less bias and more variance, fewer quantiles = more bias and less variance). We would get point estimates for each of the strata and combine them and the standard errors using a weighted average in order to find the 95% CI.

GLM adjustment: We would fit the regression model to estimate risk difference of the effect of aspirin on our outcome (live birth). We would also adjust for confounding variables in our data set. We would calculate the risk difference by subtracting the predicted risk of live birth between our different exposure groups. We would use the sandwich variance estimate in this study's case.

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.

Regression model adjustment:

RD is 0.09 and the SE is 0.03. The 95% CI taken from the sandwich SE is (0.04,0.14)