Section 3 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.
- This assignment is due electronically through CANVAS on Sunday March 5 2023.
- 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).

Question 1: Why do the logistic, log, and identity link functions enable us to interpret coefficient as a log-odds ratios, log-risk ratios, and risk differences? Please show the math.

Question 2) Using the NHEFS dataset (available on CANVAS), please plot the unadjusted dose-response relation between smoking intensity and the *risk* of high blood pressure.

Question 3) Using the NHEFS data and an outcome regression model, estimate the conditionally and marginally adjusted risk ratio and risk difference for the association between quitting smoking and high blood pressure. Adjust for smokeintensity, sex, age, race, school, and marital status. Please use appropriate coding for all variables, but do not adjust for interaction effects. As always, use appropriate standard error eestimators.

Question 4) For the risk difference model used in question 3, are there any relevant interaction effects?

Question 5: Using the NHEFS data and a propensity score model with IP weighting, estimate the marginally adjusted risk ratio and risk difference for the association between quitting smoking and high blood pressure. Adjust for smokeintensity, sex, age, race, school, and marital status. Please use appropriate coding for all variables, but do not adjust for interaction effects. As always, use appropriate standard error eestimators.

Question 6: Is the positivity assumption met for the propensity score model fit in Question 5? Is the positivity assumption required to interpret the estimate in Question 5 AND Question 3 as a causal contrast of potential outcomes? Why?

Question 7: Consider the stem cell transplant data available on CANVAS. These data contain information on 177 patients who received stem cell transplants for leukemia treatment. The event of interest in these data is relapse, which occurred in 56 individuals. Competing events include transplant-related death, which occurred in 75 individuals. Finally, 46 individuals were censored in the study. The goal of this study is to evaluate the effect of transplant source (Type of Transplant) on relapse rates, adjusting for sex, disease type (lymphoblastic or myeloblastic leukemia, abbreviated as ALL and AML, respectively), phase at transplant (Relapse, CR1, CR2, CR3), source of stem cells (bone marrow and peripheral blood, coded as BM+PB, or peripheral blood, coded as PB), and age.

Please plot the cumulative sub-distribution risk of relapse (Status = 1) among those who received bone marrow and peripheral blood (BM+PB) transplants relative to those who received peripheral blood (PB) alone.