HUDM 6026 - Computational Statistics
HW 03 – Monte Carlo Simulation Studies

Instructions.

- You may use whatever software you like to write up your answers to the hw. For example, you might choose to use Rmarkdown for assignments that are primarily code-based; whereas, you might choose to use Word or LaTeX for more writing-heavy assignments. It's up to you.
- No matter which program you use, you should turn in a pdf or html version.
- **This is a group assignment that is meant to be completed by groups of size 2-4. Only one assignment will be turned in per group. You are responsible for finding one or more classmates to group up with. If you are having trouble doing so, please let me know and I'll try to find a group for you.**

HW 03

Lee, Lessler, and Stuart (2011) ran a Monte Carlo simulation study to look at the impact of weight trimming on weighting approaches to propensity score estimation on bias and SE of causal estimators. The data generation process they used is based on the data generation process described by Setoguchi et al (2008) with some modifications. Both these papers can be found in the 'Files' section on Canvas in the 'Readings' folder.

Looking at the Lee et al paper, focus on the Methods section where the simulation setup is discussed. The following important facts are noted:

- Each simulated data set had a sample size of $N = 500$ observations.
- The variables in each data set include
  - A binary exposure variable (more on how this was generated later).
  - A continuous outcome variable (more on how this was generated later).
  - 10 covariates (4 associated with both exposure and outcome, 3 associated only with the exposure, and 3 associated only with the outcome).
- The covariates were generated as $N(0, 1)$ but some were pairwise correlated. This information comes from Table 1 in the Setoguchi et al paper. In particular, $Cor(X_1, X_5) = Cor(X_3, X_8) = 0.2$ and $Cor(X_2, X_6) = Cor(X_4, X_9) = 0.9$.
- After being generated with the given correlations, covariates $X_1, X_3, X_5, X_6, X_8, X_9$ were dichotomized (i.e., made 0/1 binary) by cutting them off at their respective means.

The generation of the 0/1 exposure variable is the focus of much of this work. To generate the exposure variable Setoguchi et al used 7 different logistic regression models. You can find them spelled out in the Appendix of the Setoguchi paper (p. 555). These logistic models are called 'propensity score models' because they model the probability (or propensity) of assignment to treatment (=1) or control (=0) conditions.

Once the exposure variable, A, has been generated from one of the 7 scenarios above, the outcome is generated based on A and the covariates. The regression coefficients for the exposure and outcome models are also given on p. 555 of Setoguchi. One final detail to note is that although Setoguchi et al use a dichotomous outcome, Lee et al modify this and generate a numeric outcome variable as follows.

$$Y_i < -\alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + \alpha_3 X_{3i} + \alpha_4 X_{4i} + \alpha_5 X_{8i} + \alpha_6 X_{9i} + \alpha_7 X_{10i} + \gamma_1 A_i.$$

Note that the model was specified by Lee et al with no random error term. My guess is that this was an oversight, but in any case it shouldn't have a noticeable effect on the results. Your tasks for this assignment are as follows.

1. Pick one of the seven propensity score models (but not the simplest one - scenario A) and create a function that can generate a data frame with n rows (n should be an argument to the function with a default value of 500) and columns from that scenario. The first ten columns are the ten covariates, $X_1, \ldots, X_{10}$, the 11th column is the probability of exposure (i.e., the propensity score, ps) based on the scenario you pick, the 12th column is the dichotomous exposure variable A, and the 13th and final column is the outcome variable Y.

2. Next, replicate your function to create a list of $R = 1000$ data frames randomly generated from the DGP you selected.

3. What is the value of the true average treatment effect here (i.e., the simulated effect of exposure)? Write a function to estimate the average treatment effect by two methods. The first method is to simply take the mean outcome score for the treated (A = 1) minus the mean outcome score for the control (A = 0). This is a naive method and will be biased. The second method is by OLS regression. Fit a regression model of Y regressed on the covariates $X_1, \ldots, X_{10}$ and the exposure A. This method should be unbiased for the true average treatment effect. Note that another way to implement the first method of mean difference by treatment group is to regress the outcome on the exposure without controlling for any other covariates. The coefficient on A in that regression will represent the mean difference on the outcome by treatment group.

4. Conclude with a brief discussion and reflection on the process where you discuss (a) any challenges your group encountered and how you solved them and (b) what you take away from the results of the simulation study in terms of the bias of each of the two methods discussed in (3).