HUDM 6026 - Computational Statistics
HW 04 – HW 03 Continued

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 04

This homework is a continuation of the group assignment you started last week. Last week you created a function that can be used to generate covariate data according to the correlation matrix specified in Setoguchi et al 2008 with some numeric and some dichotomous predictors, again, as specified in the papers related to this work. Now that we have had a chance to review how to generate data from linear and logistic regression models, you will complete the assignment from last week.

1. First, pick one of the propensity score (logistic) models described in the papers, but not the simplest, called scenario A. Then, augment your data generation function so that it now generates the full data set where the first ten columns are the ten predictors, $X_1, \ldots, X_{10}$, 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 A)?
4. 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) group participants. This is a naive method and will be biased. The second method is by OLS regression. Fit a multiple linear 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 naive 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.

5. Apply the estimation function to the data list of $r = 1000$ data sets and calculate and report the Monte Carlo estimated bias and MSE for both methods.

6. 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 and MSE of each of the two estimation methods.