**Questions week 4**

**Please download from CANVAS the file “catholic\_schools.csv”. The file contains information about 5,671 students and their twelfth-grade mathematics achievement. We also observe several covariates for each student, such as socio-economic characteristics and pre-treatment academic characteristics from the academic base year of grade 8. We are interested in whether attending a Catholic (versus a public) school has a causal effect on math achievements.**

|  |  |
| --- | --- |
| **Variable name** | **Description** |
| math12 | 12th grade standardized mathematics score – out outcome variable |
| catholic | attended catholic high school (vs. public one) – our treatment variable |
| read8 | 8th grade standardized reading score |
| math8 | 8th grade standardized mathematics score |
| female | student is female |
| race | Student’s race/ethnic background (categories – see below) |
| parmar8 | parents marital status in 8th grade |
| faminc8 | total annual family income in 8th grade |
| fathed8 | Father’s highest level of education in 8th grade |
| mothed8 | Mother’s highest level of education in 8th grade |
| fight8 | student got into fight with another student in 8th grade |
| nohw8 | student rarely completed homework in 8th grade |
| disrupt8 | student frequently disruptive in 8th grade |
| riskdrop8 | # of risk factors for later dropout in 8th grade |

1. **Draw a causal diagram for our scenario and justify your choices about the arrows.**
2. **Let’s start with some descriptive analysis:**
   1. **Should we log-transform our outcome variable (math12)? No need**
   2. **Check the balance (proportions or frequencies) of our treatment variable ‘catholic’ for our sample.**
   3. **Check the frequency distribution across treatment and control of all categorical covariates (to check balance).**
   4. **Check the density plots of the variables math8, read8, faminc8, mothed8 and fathed8 across the levels of our treatment variable. Is there sufficient common support in your view and how balanced do treatment and control appear?**
   5. **Check whether the covariates really affect the treatment using logistic regression.**
3. **Let’s estimate the ATT (average treatment effect on the treated) using stratification only on the two dummy variables of nohw8 and disrupt8. Provide both an estimate and standard error.**
4. **Let’s now check whether it’s possible to estimate any estimands using stratification for** 
   1. **The dummy female, the categorical variable ‘race’ and ‘income’ (which we treat as discrete).**
   2. **.. for all variables (realising exact matching)?**
5. **Let’s now estimate a propensity score based only on the nohw8 and disrupt8 dummies (as in C) and using logistic regression for our nuisance function.** 
   1. **What is the ATT based on the inverse probability weighting estimator (IPWE) without normalised weights?**
   2. **What is the ATT based on the inverse probability weighting estimator (IPWE) with normalised weights?**

***Tip: no need to provide standard errors***

1. **Let’s also use OLS to estimate..**
   1. **.. the simple mean difference (the naïve estimate) of our treatment**
   2. **.. the ATE estimate based on traditional regression adjustment in which you control for all variables according to your causal diagram (compare with and without log-transformations and with education being treated as discrete vs. not).**
2. **Let’s now estimate a propensity score based on different nuisance functions and all available variables. You will need to decide which input variables to include in our nuisance functions.** 
   1. **Create a trimmed data set where we remove the 1% most extreme PS values. Check the overlap and balance of the propensity scores (PS) based on logistic regression using a histogram and love plot (for propensity scores of ATE and ATT) for both the trimmed and original data sets. Any preference for any of the two data sets?**
   2. **Let’s use logistic regression and the IPWE for the ATE and ATT. Report the estimates and standard errors. Does it matter whether we use the trimmed or non-trimmed data set?**
   3. **Check the overlap and balance of the PS based on gradient boosting regression models (GBM) using a histogram and love plot (for propensity scores of ATE and ATT) for the data set you preferred (trimmed or raw).**
   4. **Let’s use GBM and the IPWE for the ATE and ATT. Report the estimates and standard errors. Check both the analytics and bootstrapped SE (100 replications).**

***Tip: For this exercise, recall the PS Weight uses normalised weights (the default settings of PSWeight). Recall what normalised weights do.***

1. **How much do you trust your final model results?**