MATH 598/782 – PROJECT 4

Please submit your solutions by 11.59 pm (EDT) on Friday 12th April by uploading a single pdf to myCourses.

The First Steps data

https://www.math.mcgill.ca/dstephens/598-Causal-2024/Exercises/Project4/FirstSteps_data.csv

records the effect of a program aimed (amongst other things) at increasing the birthweight of a newborn in a low income household (variable `bwt`). The treatment (exposure) is household participation in the program (binary variable `firststep`. The data are observational, that is, allocation to the program is not performed using an experimental intervention. The data are studied in the in the following two papers

- Alam S, Moodie EEM, Stephens DA. Should a propensity score model be super? The utility of ensemble procedures for causal adjustment. *Statistics in Medicine*. 2019; 38: 1690–1702.

  A full description of the data is given in this paper.

- Moodie EEM, Saarela O, Stephens DA. A doubly robust weighting estimator of the average treatment effect on the treated. *Stat.* 2018;7:e205.

(a) The Alam et al. (2019) paper uses "*plasmode*" simulation method to compare various ensemble and other machine learning approaches. The plasmode approach takes an existing data set and simulates outcome data using the same covariate distribution as the real data set and a specified model. One simple simulation is given below

```
fn<-"https://www.math.mcgill.ca/dstephens/598-Causal-2024/Exercises/Project4/FirstSteps_data.csv"
FS.data<-read.csv(fn)
n<-nrow(FS.data)

#Make a copy of data set
Copy.dat<-FS.data

#Code the discrete covariates
Copy.dat$sex_coded[Copy.dat$sex=="F"]<- 0
Copy.dat$sex_coded[Copy.dat$sex=="M"]<- 1

Copy.dat$race_coded[Copy.dat$race=="white"]<- 0
Copy.dat$race_coded[Copy.dat$race=="asian"]<- 1
Copy.dat$race_coded[Copy.dat$race=="hispanic" | Copy.dat$race=="other"
                  | Copy.dat$race=="black"]<- 3

Copy.dat$smoker_coded[Copy.dat$smoker=="N"]<- 0
Copy.dat$smoker_coded[Copy.dat$smoker=="Y"]<- 1

Copy.dat$drinker_coded[Copy.dat$drinker=="N"]<- 0
Copy.dat$drinker_coded[Copy.dat$drinker=="Y"]<- 1

Copy.dat$parity_coded[Copy.dat$parity==0]<- 0
Copy.dat$parity_coded[Copy.dat$parity==1]<- 1
Copy.dat$parity_coded[Copy.dat$parity>=2]<- 2

#Carry out a fit to the original data to get some realistic parameters.
fit.orig<-lm(bwt~firstep+sex+smoker,data=Copy.dat)
summary(fit.orig)$coef

+               Estimate Std. Error    t value     Pr(>|t|)
+ (Intercept) 3374.73996   16.77632 201.160909 0.000000e+00
+ firstep      -51.51902   30.18712  -1.706656 8.801045e-02
+ sexM         124.50608   22.10743   5.631866 1.982352e-08
+ smokerY     -238.73186   43.50390  -5.487596 4.485028e-08
```

```r
be<-coef(fit.orig)
be[2]<-50          #Fix the treatment effect in an additive model at 50

set.seed(23)
Y<-be[1]+be[2]*Copy.dat$firstep+be[3]*Copy.dat$sex_coded+
         be[4]*Copy.dat$smoker_coded+rnorm(n)*summary(fit.orig)$sigma

Synth.data<- data.frame(sex=as.factor(Copy.dat$sex_coded),
                      age=Copy.dat$age,
                      race=as.factor(Copy.dat$race_coded),
                      parity=as.factor(Copy.dat$parity_coded),
                      married=Copy.dat$married, welfare=Copy.dat$welfare,
                      smoker=as.factor(Copy.dat$smoker_coded),
                      drinker=as.factor(Copy.dat$drinker_coded),
                      wpre=Copy.dat$wpre, edu=Copy.dat$edu,
                      firstep=Copy.dat$firstep, bwt=Y)

#Carry out a fit to the synthetic data
fit.synth<-lm(bwt~firstep+sex+smoker,data=Synth.data)
summary(fit.synth)$coef

+               Estimate Std. Error    t value     Pr(>|t|)
+ (Intercept) 3387.29414   16.78133 201.848947 0.000000e+00
+ firstep       63.81977   30.19614   2.113508 3.465630e-02
+ sex1         125.35802   22.11403   5.668709 1.604250e-08
+ smoker1     -336.99662   43.51690  -7.744041 1.389154e-14
```

Data sets can also be resampled (with replacement) from the original data to make a replicated analysis more realistic.

Using a plasmode approach, carry out a replicated (Monte Carlo) simulation study that compares the performance of propensity score regression and inverse probability weighting approaches when, in the synthetic data generating model, **five** of the variables appear in the synthetic outcome model, and the treatment effect is presumed additive. Assess whether these models can recover the true ATE from the synthetic data sets, and assess the variability of the resulting estimates. 10 Marks

(b) For the **original** data, use the causal methods described through the course to estimate the ATE associated with the `firstep` treatment intervention. Report the sampling variability (standard errors) of the estimates that you compute using the different causal adjustment methods. 10 Marks