SUMMER 2019/20 TERM 1 EPSE 581C: ASSIGNMENT 2
Due: Wednesday, June 19th

- Please make sure you write your answers to these questions in your own words. Even if you work with a group to formulate your responses, do not just copy someone else's sentences/words.
- There is no need to record more than 2 decimal places for any of these problems; HOWEVER, do not round numbers until you obtain a final answer.
- All problem data are available online in .csv format.

Question 1: (Find the best specified model) Load the dataset 'datq1' for Q1 from the webpage. These data contain information on a response variable y and two covariates x and w. These data have been collected in a non-experimental framework, so although we know that x and w likely relate to the response y, we can also be quite sure that we are missing important covariates that relate all these variables together. Adapt the R code from Q2 of HW1 to find a best specified model that relates x and w to the response y (this should mostly amount to replacing the data frame 'dat2' from HW1, Q2 with this problem's data frame 'datq1'). You will likely want to start by plotting some raw data, then proposing a simple model and performing some residual diagnostics. Based on these diagnostics, you then may want to respecify the model, perhaps adding interactions or higher order terms into the model. What model seems to be best specified, given the covariates you have to work with? Why?

Question 2: (Propensity scores) In this question, we will perform a propensity score adjustment to estimate the (partially unconfounded) effect of a graduate student tutoring intervention offered to students of STAT 302: Introduction to Probability on final course grade. As in the example in class, this tutoring service was offered free to all students one week before the final exam. Students were free to self-enrol in the tutoring service to receive up to one hour of one-on-one tutoring with a trained graduate student in the Statistics Department.

The dataset (called 'datps', available on the webpage) contains information on all 112 students from one section of STAT 302. In addition to their final course grade and a variable indicating whether or not they opted to take advantage of the tutoring service, we have information on the Sex and Major of each student, as well as their course participation mark (measured 1 to 6), and their midterm exam score (measured out of 40 points).

Here, we will mimic an applied setting where we do not know if we have accounted for all confounders, and we do not know the correct functional specification for our measured covariate(s).

(a) To begin, estimate the average causal effect naively; that is, by assuming that student self-enrolment in the tutoring service is completely unconfounded with any student characterisitics (measured or unmeasured). Here, we can perform a simple t-test on the course Final Grade between the Tutor and Non-Tutor groups:

```
t.test(datps$Final~datps$Tutor)
```

Is there evidence of a mean group difference? What is the mean of the Final Grade in the Tutored group? In the Non-Tutored group?

(b) Of course, the above estimate is likely a very poor estimate of the average causal effect of treatment, since there are likely important reasons why students choose to enrol in the tutoring service or not! Considering the measured covariates, which do you think might be most related or least related to whether or not a student chooses to enrol in the tutoring service? Why?

(c) Now we will construct propensity scores for each of our sample units, and then perform a matching operation to pair a student in the Tutor group with one in the Non-Tutor group that have approximately the same propensity score. Recall that to do this we need to perform a logistic regression to estimate the probability of assignment to treatment, given the observed covariates. Suppose we fit a full first-order model:

$$\text{logit}(Final) = \beta_0 + \beta_1 Sex + \beta_2 Year + \beta_3 Major + \beta_4 Participation + \beta_5 Midterm + \varepsilon$$

Fit this model using R's 'glm' command and store the resulting estimated propensity scores:

```
psmod1 <- glm(formula=Tutor~Sex+Year+Major+Participation+Midterm_out_of_40, family=bi
datps$ps1 <- exp(fitted(psmod1))/(1+exp(fitted(psmod1)))
```

Plot histograms of the propensity scores in the Tutor and Non-Tutor groups:

```
par(mfrow=c(1,2))
hist(datps$ps1[datps$Tutor==0])
hist(datps$ps1[datps$Tutor==1])
```

What is the estimated range of the propensity scores? Over what range do the propensity scores not overlap so well between the two groups? What implication(s) does this have for eventual propensity score matching?

(d) Now we will create matched Tutor and Non-Tutor units according to their propensity scores using R's 'MatchIt' package. Download and install this package into your RStudio workspace. Create the matches as follows:

```
mod_match <- matchit(formula = Tutor ~ Midterm_out_of_40 + Year+Sex+Major+Participatio
dat.m <- match.data(mod_match)
summary(mod_match)
```

The last line of code will produce a bunch of statistics summarizing the balance of covariates between Tutor and Non-Tutor groups, first for the raw data, and second for the matched data. For which covariates does the mean difference between treatment groups decrease? How many matches are created (final bit of output)?

(e) Estimate the average causal effect using the propensity score matched sample via another *t*-test on the course Final Grade between the Tutor and Non-Tutor groups:

```
with(dat.m, t.test(Final ~ Tutor))
```

Is there evidence of a mean group difference? What is the mean of the Final Grade in the Tutored group? In the Non-Tutored group? How have these estimates and this inference changed from the naive, unadjusted estimate in part (a)?

(f) We only did very rough checking of covariate balance in part (d) (examining mean differences between treatment groups). Let's look a bit closer by plotting the actual distributions of some covariates and comparing between the two treatment groups after matching.

```
par(mfrow=c(1,2))
hist(dat.m$Midterm_out_of_40[dat.m$Tutor==0])
hist(dat.m$Midterm_out_of_40[dat.m$Tutor==1])
```

How good does the balance look between matched treatment groups for the Midterm covariate?

```
par(mfrow=c(1,2))
hist(dat.m$Participation[dat.m$Tutor==0])
hist(dat.m$Participation[dat.m$Tutor==1])
```

What about for the Participation covariate?

```
table(dat.m$Sex,dat.m$Tutor)
table(dat.m$Year,dat.m$Tutor)
table(dat.m$Major,dat.m$Tutor)
```

What about for the Sex, Year, and Major covariates?

(g) We also didn't check the adequacy of our logistic regression model in part (c) by performing any diagnostics. Examine the residuals vs. fitted plot and qq-plot for this model with the following command (only pay attention to these first two plots; ignore the final two):

```
plot(psmod1)
```

Does the average value of the residuals seem to always be zero (look at the red line of best-fit: this is an estimate of the "average residual" for a given fitted value)? Do the residuals appear normally distributed? Confirm the shape of the residuals by plotting a histogram.

```
par(mfrow=c(1,1))
hist(residuals(psmod1))
```

Why might some poor regression diagnostics be of concern here?