ARE 213 Problem Set 1B

Becky Cardinali, Yuen Ho, Sara Johns, and Jacob Lefler

Due 10/12/2020

Question 1

In Problem Set 1a, you used linear regression to relate infant health outcomes and maternal smoking during pregnancy. Please answer the following questions.

(a) Under the assumption of random assignment conditional on the observables, what are the sources of misspecification bias in the estimates generated by the linear model estimated in Problem Set 1a?

Question 2

Describe the propensity score approach to the problem of estimating the average causal effect of smoking when the treatment is randomly assigned conditional on the observables. How does it reduce the dimensionality problem of multivariate matching?

We know that if we condition on observables, we will get a consistent estimate of the ATE under the assumption. However, if the observables are high dimensional, it might be difficult to find a comparison unit with the same values of the observables. From lecture, we know that it is sufficient instead to condition on the propensity score. Using the propensity score allows us to compare treated and control units with the same probability of being treated. The propensity score does not require that all values of the observables be the same and so therefore avoids problems of multidimensionality.

Try a few ways to estimate the effects of maternal smoking on birthweight:

2a

a) First create the propensity score. For our purposes let's use a logit specification. First specify the logit using all of the "predetermined" covariates (don't include interactions). Next, include only those "predetermined" covariates that enter significantly in the first logit specification. How comparable are the propensity scores? If they are similar does this imply that we have the "correct" set of covariates in the logit specification used for our propensity score?

```
# get prop score using all predetermined variables
prop_all <- glm(tobacco ~ factor(stresfip) + dmage + factor(mrace3) + dmeduc +
                  dtotord + disllb + dfage + factor(birmon) + factor(orfath) +
                  factor(dmar) + dfeduc + dplural + factor(pre4000) + factor(preterm),
                family=binomial(link='logit'), data = mom_dt)
mom_dt[, prop_score_all := fitted(prop_all)]

# take a look at the output to see which are significant - omitting because takes up a lot of space
# summary(prop_all)
# only need to take out state of residence. birth month has a few months that are significant so will keep

# get prop score using significant variables from previous logit
```

```
prop_sig <- glm(tobacco ~ dmage + factor(mrace3) + dmeduc +
                  dtotord + disllb + dfage + factor(birmon) + factor(orfath) +
                  factor(dmar) + dfeduc + dplural + factor(pre4000) + factor(preterm),
                family=binomial(link='logit'), data = mom_dt)
mom_dt[, prop_score_sig := fitted(prop_sig)]

# how different are the prop scores?
prop_score_diff <- mom_dt$prop_score_all - mom_dt$prop_score_sig
summary(prop_score_diff)
```

```
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
## -0.2889930  0.0003105  0.0004940  0.0000000  0.0006952  0.2873649
```

```
# a few outliers but not very different
```

2b

Control directly for the estimated propensity scores using a regression analysis, and estimate an average treatment effect. State clearly the assumptions under which your estimate is correct.

```
# run regression with prop_score
prop_reg <- lm(dbrwt ~ tobacco + prop_score_sig, data = mom_dt)
```

Controlling directly for the propensity score, we get the ATE is -223.23 and is statistically significant at the 99.9% level.

2c

```
# create propensity weights
mom_dt[,prop_weights := ifelse(tobacco == 1,
                          1/prop_score_sig,
                          1/(1 - prop_score_sig))]
# normalize the weights
mom_dt[,norm_prop_weights := ifelse(tobacco == 1,
                               prop_weights/sum(mom_dt[tobacco == 1, prop_weights]),
                               prop_weights/sum(mom_dt[tobacco == 0, prop_weights]))]
# estimate tau
tau_ipw <- sum((mom_dt$tobacco*mom_dt$dbrwt)*(mom_dt$norm_prop_weights) - ((1 - mom_dt$tobacco)*mom_dt$dbrwt)
```

Using inverse propensity score weighting, we get that the ATE is now -225.29.