EDS241: Assignment 3

Desik Somasundaram

02/20/2022

This exercise asks you to implement some of the techniques presented in Lectures 6-7. The goal is to estimate the causal effect of maternal smoking during pregnancy on infant birth weight using the treatment ignorability assumptions. The data are taken from the National Natality Detail Files, and the extract "SMOKING_EDS241.csv"' is a random sample of all births in Pennsylvania during 1989-1991. Each observation is a mother-infant pair. The key variables are: The outcome and treatment variables are: birthwgt=birth weight of infant in grams tobacco=indicator for maternal smoking The control variables are: mage (mother's age), meduc (mother's education), mblack (=1 if mother black), alcohol (=1 if consumed alcohol during pregnancy), first (=1 if first child), diabete (=1 if mother diabetic), anemia (=1 if mother anemic)

1 Clean data

The following code loads and cleans the data.

```
# Load data
smokingdata <- read_csv("SMOKING_EDS241.csv")
# Clean data
smokingdata <-janitor::clean_names(smokingdata)
```

2 Unadjusted mean difference

(a) What is the unadjusted mean difference in birth weight of infants with smoking and nonsmoking mothers? Under what assumption does this correspond to the average treatment effect of maternal smoking during pregnancy on infant birth weight? Provide some simple empirical evidence for or against this assumption.

```
# smoking mothers mean
mu_nonsmoker = smokingdata %>%
  filter(tobacco == 0) %>%
  summarize(mean(birthwgt))

# non-smoking mothers mean
mu_smoker = smokingdata %>%
  filter(tobacco == 1) %>%
  summarize(mean(birthwgt))

# calculate mean  diff
mean_diff = as.numeric(mu_nonsmoker - mu_smoker)

# linear regression of choice covariate
```

```r
model_1 <- lm_robust(meduc ~ tobacco, data = smokingdata)
huxreg(model_1)
```

	(1)
(Intercept)	13.239 ***
	(0.008)
tobacco	-1.318 ***
	(0.014)
N	94173
R2	0.061

*** p < 0.001; ** p < 0.01; * p < 0.05.

The unadjusted mean difference in birth weight of infants 244.539 grams. Statistically different from zero. Under the "treatment ignorability assumption", this corresponds with the average treatment effect(ATE) of maternal smoking during pregnancy on infant birth weight. Assumption of "treatment ignorability" conditional on pre-treatment characteristics that allow us to assume that smoking mothers and nonsmoking mothers are good counterfactuals. The assumption of common support ensures that there is sufficient overlap in the characteristics of smoking mothers and nonsmoking mothers to find adequate matches so this aspect would need further analysis of the data. It's important to note that the treatment of smoking is not randomly assigned either. The treatment ignorability assumption says that conditional on observable covariates, the assignment to the treatment is independent of the outcome of infant birth weight. Observational regression bias can arise if the smoking mother and nonsmoking mothers are inherently different in a way which would affect their infant birth weight. For example, smoking mothers could be less health conscious which has a negative impact on their infant birth weights aside from the effect of smoking itself. The regression of tobacco one mother's education yields a statistically significant relationship which questions the validity of the assumption. There is omitted variable bias shown by the model_1 regression which prevents us from being able to interpret the unadjusted mean difference as a causal effect. Unconditional treatment ignorability is not holding true.

3 Introducing covariates

Assume that maternal smoking is randomly assigned conditional on the observable covariates listed above. Estimate the effect of maternal smoking on birth weight using a linear regression. Report the estimated coefficient on tobacco and its standard error.

```
model_2 <- lm_robust(formula =  birthwgt ~ tobacco + mage + meduc + as.factor(anemia)
                     + as.factor(diabete) + as.factor(alcohol) +
                       as.factor(mblack) + as.factor(first), data=smokingdata)

model_2_ht <- huxreg(model_2)
restack_across(model_2_ht,13)
```

	(1)		(1)
(Intercept)	3362.258 ***	as.factor(alcohol)1	-77.350 ***
	(12.076)		(14.039)
tobacco	-228.073 ***	as.factor(mblack)1	-240.030 ***
	(4.277)		(5.348)
mage	-0.694	as.factor(first)1	-96.944 ***
	(0.368)		(3.488)
meduc	11.688 ***	N	94173
	(0.862)	R2	0.072
as.factor(anemia)1	-4.796	*** p < 0.001; ** p < 0.01; * p < 0.05.	
	(17.874)		
as.factor(diabete)1	73.228 ***		
	(13.235)		

Table 1 shows the **estimated coefficients from the linear regression of effect of maternal smoking on birth weight. -228.073 is the estimated coefficient on tobacco with 4.277 being the standard error**.

4 Exact matching

Use the exact matching estimator to estimate the effect of maternal smoking on birth weight. For simplicity, consider the following covariates in your matching estimator: create a 0-1 indicator for mother's age (=1 if mage>=34), and a 0-1 indicator for mother's education (1 if meduc>=16), mother's race (mblack), and alcohol consumption indicator (alcohol). These 4 covariates will create 222*2 = 16 cells. Report the estimated average treatment effect of smoking on birthweight using the exact matching estimator and its linear regression analogue.

```r
# create indicator based on mother's age (=1 if mage>=34)
smokingdata<- smokingdata %>%
              mutate(age_indicator = case_when(
                      mage < 34 ~ 0,
                      mage >= 34 ~ 1))

# create indicator based on mother's education (1 if meduc>=16)
smokingdata<- smokingdata %>%
              mutate(educ_indicator = case_when(
                      meduc < 16 ~ 0,
                      meduc >= 16 ~ 1))

# create group variable to capture all interactions
smokingdata<- smokingdata %>%
              mutate(g = paste0(age_indicator,educ_indicator,mblack,alcohol))

# regression on including tobacco and the 4 grouped indicators (mother age, mother education, mother ra
model_3 <- lm_robust(birthwgt ~ tobacco +  as.factor(g), data= smokingdata)
```

```r
# exact matching table
TIA_table <- smokingdata %>%
  group_by(g,tobacco)%>%
  summarise(n_obs = n(),
            birthwgt_mean= mean(birthwgt, na.rm = T))%>% #Calculate number of observations and Y mean b
  gather(variables, values, n_obs:birthwgt_mean)%>% #Reshape data
  mutate(variables = paste0(variables,"_",tobacco, sep=""))%>% #Combine the treatment and variables for
  pivot_wider(id_cols = g, names_from = variables,values_from = values)%>% #Reshape data by treatment a
  ungroup()%>%  #Ungroup from X values
  mutate(birthwgt_mean_diff = birthwgt_mean_1 - birthwgt_mean_0, #calculate Y_diff
         w_ATE = (n_obs_0+n_obs_1)/(sum(n_obs_0)+sum(n_obs_1)),
         w_ATT = n_obs_1/sum(n_obs_1))%>% #calculate weights
  mutate_if(is.numeric, round, 2) #Round data


stargazer(TIA_table, type= "text", summary = FALSE, digits = 2)
```

```
##
## =====================================================================================
##      g   n_obs_0 n_obs_1 birthwgt_mean_0 birthwgt_mean_1 birthwgt_mean_diff w_ATE w_ATT
## -------------------------------------------------------------------------------------
## 1  0000  44274   13443      3445.69         3220.25          -225.44        0.61  0.74
## 2  0001   214     448       3450.28         3124.25          -326.03        0.01  0.02
## 3  0010  7007    1980       3195.97         3006.31          -189.66         0.1  0.11
## 4  0011   71      226       3120.07         2817.34          -302.73          0   0.01
## 5  0100  13425    535       3483.02         3273.94          -209.08        0.15  0.03
## 6  0101   130     29        3510.95         3413.21           -97.74          0     0
## 7  0110   625     61        3319.22         3159.05          -160.17        0.01    0
## 8  0111    4      10        2983.5          3097.7            114.2           0     0
## 9  1000  5115     976       3467.41         3171.42          -295.98        0.06  0.05
## 10 1001   56      45        3358.32         3097.73          -260.59          0     0
## 11 1010   396     135       3185.08         2994.67          -190.41        0.01  0.01
## 12 1011    7      26        2739.71         2846.38           106.67          0     0
## 13 1100  4492     201       3487.19         3249.45          -237.74        0.05  0.01
```

```
## 14 1101   57      17        3534.91         3037.47         -497.44         0     0
## 15 1110   147     19        3328.29         2852.16         -476.13         0     0
## 16 1111    1       1          3459            2835            -624          0     0
## --------------------------------------------------------------------------------------
```

```
# Multivariate matching estimates of ATE
ATE=sum((TIA_table$w_ATE)*(TIA_table$birthwgt_mean_diff))
ATE
```

```
## [1] -224.2583
```

Table 2 shows the exact matching table used to estimate ATE.

The exact matching estimator estimates the average treatment effect(ATE) of smoking on birthweight as -224.258. The linear analogue estimates the average treatment effect(ATE) of smoking on birthweight as -226.245.

5 Propensity Score

Estimate the propensity score for maternal smoking using a logit estimator and based on the following specification: mother's age, mother's age squared, mother's education, and indicators for mother's race, and alcohol consumption.

```
# add mother's age variable squared
smokingdata <- smokingdata %>%
                mutate(mage_squared = mage * mage)

model_4 <- glm(tobacco ~ mage + mage_squared + meduc + as.factor(mblack) + as.factor(alcohol), family =
summary(model_4)
```

```
##
## Call:
## glm(formula = tobacco ~ mage + mage_squared + meduc + as.factor(mblack) +
##     as.factor(alcohol), family = binomial(), data = smokingdata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5482  -0.7182  -0.5461  -0.3214   2.6709
##
## Coefficients:
##                       Estimate Std. Error z value         Pr(>|z|)
## (Intercept)           1.929611   0.191814  10.060          < 2e-16 ***
## mage                  0.077636   0.014915   5.205 0.00000019355476 ***
## mage_squared         -0.001941   0.000278  -6.983 0.00000000000288 ***
## meduc                -0.321597   0.005144 -62.520          < 2e-16 ***
## as.factor(mblack)1   -0.059525   0.026506  -2.246           0.0247 *
## as.factor(alcohol)1   2.022696   0.060358  33.511          < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 92325  on 94172  degrees of freedom
## Residual deviance: 84825  on 94167  degrees of freedom
## AIC: 84837
##
## Number of Fisher Scoring iterations: 5
```

```
EPS <- predict(model_4, type = "response") # estimated propensity score (EPS)
PS_weighted <- (smokingdata$tobacco / EPS) + ((1 - smokingdata$tobacco)/(1 - EPS)) # weight EPS
```

Use the propensity score weighted regression (WLS) to estimate the effect of maternal smoking on birth weight (Lecture 7, slide 12).

```
# add propensity scores and weighted propensity scores to data as columns
smokingdata <- smokingdata %>%
  mutate("EPS" = EPS, "PS_weighted" = PS_weighted)


model_5 <- lm_robust(birthwgt ~ tobacco + mage + mage_squared + meduc +
                     as.factor(mblack) + as.factor(alcohol), data=smokingdata, weights=PS_weighted)

wls_ATE <- model_5$coefficients[2]
```

The weighted OLS regression estimates the average treatment effect(ATE) of smoking on birthweight as -220.233

5.1 Appendix

```
model_3_ht <- huxreg(model_3)
restack_across(model_3_ht,20)
```

	(1)		(1)
(Intercept)	3445.873 ***		(6.819)
	(2.232)	as.factor(g)1001	-102.853 *
tobacco	-226.245 ***		(45.144)
	(4.220)	as.factor(g)1010	-251.686 ***
as.factor(g)0001	-63.124 **		(24.106)
	(20.431)	as.factor(g)1011	-443.862 ***
as.factor(g)0010	-241.839 ***		(79.415)
	(5.742)	as.factor(g)1100	40.825 ***
as.factor(g)0011	-384.006 ***		(7.404)
	(29.870)	as.factor(g)1101	26.737
as.factor(g)0100	37.809 ***		(55.254)
	(4.535)	as.factor(g)1110	-146.188 ***
as.factor(g)0101	88.511 *		(38.555)
	(38.413)	as.factor(g)1111	-185.751
as.factor(g)0110	-120.775 ***		(198.895)
	(18.977)	N	94173
as.factor(g)0111	-219.198	R2	0.063
	(127.345)	*** p < 0.001; ** p < 0.01; * p < 0.05.	
as.factor(g)1000	10.359		

Table A1 shows the estimated coefficients from the linear regression of effect of maternal smoking on birth weight with the inclusion of the following covariates: mother age, mother education, mother race and alcohol.

```
model_5_ht <- huxreg(model_5)
restack_across(model_5_ht,13)
```

Table A2 shows the estimated coefficients from weighted OLS estimating the effect of maternal smoking on birth weight with the inclusion of the following covariates: mother age, mother age squared, mother education, mother race and alcohol.

	(1)
(Intercept)	2971.444 ***
	(57.060)
tobacco	-220.233 ***
	(5.029)
mage	27.627 ***
	(4.587)
mage_squared	-0.478 ***
	(0.087)
meduc	7.472 ***
	(1.584)
as.factor(mblack)1	-220.990 ***
	(8.245)
as.factor(alcohol)1	-71.914 ***
	(16.734)
N	94173
R2	0.074

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.