EDS241: Assignment 3

Halina Do-Linh

02/20/2022

For Assignment 3, we are implementing some techniques from Lectures 6-7. We want to estimate the causal effect of maternal smoking during pregnancy on infant birth weight using treatment ignorability assumptions. Data comes from the National Natality Detail Files, and the extract `SMOKE_EDS241.csv` is a random sample of all births in Pennsylvania during 1989-1991. Each observation is a mother-infant pair.

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)

Load Data

```
smoking_df <- read_csv(here("data/SMOKING_EDS241.csv")) # data is clean and tidy
```

Question A

What is the unadjusted mean difference in birth weight of infants with smoking and non-smoking 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 hypothesis.

```
# mean of smoking mothers
bw_smoke <- smoking_df %>%
  filter(tobacco == 1) %>%
  group_by(tobacco) %>%
  summarize(mean_bw_g = mean(birthwgt))
mean_bw_smoke <- bw_smoke$mean_bw_g
```

```r
# mean of non-smoking mothers
bw_nonsmoke <- smoking_df %>%
  filter(tobacco == 0) %>%
  group_by(tobacco) %>%
  summarize(mean_bw_non_g = mean(birthwgt))
mean_bw_nonsmoke <- bw_nonsmoke$mean_bw_non_g

# mean difference in birth weight of infants with smoking and non-smoking mothers
mean_diff = mean_bw_nonsmoke - mean_bw_smoke

# linear regression of other covariates regressed by tobacco to show OVB
mod_1 <- lm_robust(meduc ~ tobacco, data = smoking_df)
mod_2  <- lm_robust(mage ~ tobacco, data = smoking_df)

# created a table using huxtable
huxreg(mod_1, mod_2)
```

	(1)	(2)
(Intercept)	13.239 ***	27.453 ***
	(0.008)	(0.019)
tobacco	-1.318 ***	-1.915 ***
	(0.014)	(0.043)
N	94173	94173
R2	0.061	0.020

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

Answer: The unadjusted mean difference in birth weight of infants with smoking and non-smoking mothers is 244.54 grams.

Our assumption is that smoking has been randomly assigned and that the effect of smoking on infant birth weight between mothers who smoke and mothers who don't smoke are have different means, holding the other variables constant.

`mod_1` and `mod_2` provides some empirical evidence that the statement above is likely not true. Since both models show coefficients that are non-zero and are statistically significant, then there is a correlation between `meduc` and `tobacco` as well as between `mage` and `tobacco`. These variables were not considered when I found the mean difference of infant birth weight between mothers who smoke and mothers who don't smoke. This means that omitted variable bias could be occurring and that other variables are acting on infant birth weight in addition to smoking.

Question B

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 with all covariates
mod_3 <- lm_robust(birthwgt ~
                     tobacco +
                     anemia +
                     diabete +
                     alcohol +
                     mblack +
                     first +
                     mage +
                     meduc, data = smoking_df)

# creating a table using huxtable
mod3_ht <- huxreg(mod_3)
restack_across(mod3_ht, 13)
```

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

Answer: The estimated coefficient on tobacco is -228.07 and its standard error is 4.28.

Question C

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 2*2*2*2 = 16 cells.

Report the estimated average treatment effect of smoking on birthweight using the exact matching estimator and its linear regression analogue (Lecture 6, slides 12-14).

```
# creating indicators for mage and meduc
mother_df <- smoking_df %>%
  select("tobacco",
         "alcohol",
         "mblack",
         "mage",
         "meduc",
         "birthwgt") %>%
  mutate(mage_d = case_when(mage >= 34 ~ 1,
                            mage < 34 ~ 0)) %>%
  mutate(meduc_d = case_when(meduc >= 16 ~ 1,
                             meduc < 16 ~ 0)) %>%
  mutate(g = paste0(as.factor(mage_d),
                    as.factor(meduc_d),
                    as.factor(mblack),
                    as.factor(alcohol)))


# linear regression analogue
mod_4 <- lm_robust(birthwgt ~  tobacco +
                     mage_d +
                     meduc_d +
                     mblack +
                     alcohol +
                     mage_d:meduc_d +
                     mage_d:mblack +
                     mage_d:alcohol +
                     meduc_d:mblack +
                     meduc_d:alcohol +
                     mblack:alcohol +
                     mage_d:meduc_d:mblack +
                     mage_d:meduc_d:alcohol +
                     meduc_d:mblack:alcohol +
                     mage_d:meduc_d:mblack:alcohol,
                   data = mother_df)

mod4_ht <- huxreg(mod_4)
restack_across(mod4_ht, 17)
```

	(1)		(1)		(1)
(Intercept)	3445.873 ***	mage_d:alcohol	-50.068	N	94173
	(2.232)		(43.319)	R2	0.063
tobacco	-226.245 ***	meduc_d:mblack	83.255 ***	*** p < 0.001; ** p < 0.01; * p < 0.05.	
	(4.220)		(20.110)		
mage_d	10.359	meduc_d:alcohol	113.829 **		
	(6.804)		(43.439)		
meduc_d	37.809 ***	mblack:alcohol	-79.035 *		
	(4.535)		(34.047)		
mblack	-241.839 ***	mage_d:meduc_d:mblack	-8.226		
	(5.733)		(50.176)		
alcohol	-63.127 **	mage_d:meduc_d:alcohol	-14.721		
	(20.028)		(80.388)		
mage_d:meduc_d	-7.343	meduc_d:mblack:alcohol	-70.090		
	(10.591)		(138.607)		
mage_d:mblack	-20.203	mage_d:meduc_d:mblack:alcohol	123.650		
	(24.782)		(249.369)		

```
# exact matching estimator
TIA_table <- mother_df %>%
  group_by(g, tobacco) %>%
  summarise(n_obs = n(),
            birthwgt_mean = mean(birthwgt, na.rm = T)) %>%
  gather(variables, values, n_obs:birthwgt_mean) %>% # reshape data
  mutate(variables = paste0(variables,"_", tobacco, sep="")) %>%
  pivot_wider(id_cols = g, names_from = variables, values_from = values) %>%
  ungroup() %>%
  mutate(birthwgt_diff = birthwgt_mean_1 - birthwgt_mean_0, # calculate birthwgt_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) # makes table
```

```
##
## ================================================================================
##     g    n_obs_0 n_obs_1 birthwgt_mean_0 birthwgt_mean_1 birthwgt_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_diff))
```

Answer: The estimated average treatment effect of smoking on low birth weight using the exact matching estimator is -224.26 and the estimated coefficient on tobacco from the linear regression analogue is -226.25.

Question D

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.

```
mother_df <- mother_df %>%
  mutate(mage_2 = mage * mage)

mod_5 <- glm(tobacco ~ mage + mage_2 + meduc + mblack + alcohol,
             family = binomial(),
             data = mother_df)

EPS <- predict(mod_5, type = "response") # estimated propensity score (EPS)
EPS_weighted <- (mother_df$tobacco / EPS) + ((1 - mother_df$tobacco) / (1 - EPS)) # weighted EPS
```

Question E

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

```
# regression with weights
WLS <- lm_robust(formula = birthwgt ~ tobacco, data = mother_df, weights = EPS_weighted)
# regression with weights with covariates
WLS_2 <- lm_robust(formula = birthwgt ~ tobacco + mage + mage_2 + meduc + mblack + alcohol, data = moth

huxreg(WLS, WLS_2)
```

	(1)	(2)
(Intercept)	3425.994 ***	2971.444 ***
	(1.854)	(57.060)
tobacco	-225.475 ***	-220.233 ***
	(5.025)	(5.029)
mage		27.627 ***
		(4.587)
mage_2		-0.478 ***
		(0.087)
meduc		7.472 ***
		(1.584)
mblack		-220.990 ***
		(8.245)
alcohol		-71.914 ***
		(16.734)
N	94173	94173
R2	0.048	0.074

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