EDS241: Assignment 2

Erica Bishop

03/13/2023

```
# load libraries
library(tidyverse)
library(estimatr)
library(stargazer)
library(janitor)
library(here)
library(corrplot)
```

```
smoking_data <- read_csv(here("eds241_data/SMOKING_EDS241.csv")) |>
    clean_names()
```

1 Question 1: Application of estimators based on the "treatment of ignorability" assumption

The goal is to estimate the causal effect of maternal smoking during pregnancy on infant birth weight using the treatment ignorability assumptions (Lecture 6 & 7). 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)

1.1 (a)

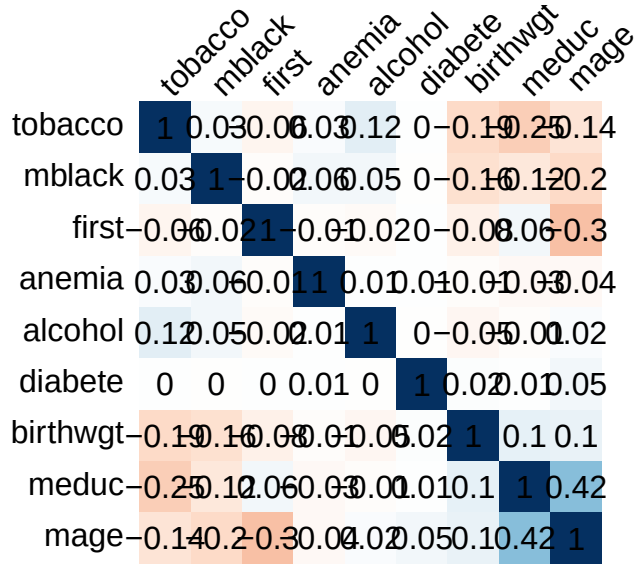
What is the unadjusted mean difference in birth weight of infants with smoking and nonsmoking mothers? Under what assumption does this correspond to the average treatment effect of maternal smoking during pregnancy on infant birth weight? Provide some simple empirical evidence for or against this assumption.

```
# calculate difference in means

means <- smoking_data |>
    group_by(tobacco == 1) |>
    summarise(mean(birthwgt, na.rm = TRUE))

mean_smoking <- as.numeric(means[2, 2])

mean_nonsmoking <- as.numeric(means[1, 2])

mean_diff_unadj <- mean_nonsmoking - mean_smoking
```

[1] "The unadjusted difference in average birthweights between babies born to mothers that smoked and mothers that did not smoke is 244.54 grams."

This unadjusted mean difference only corresponds to the average treatment effect under the treatment ignorability assumption, which states that there are no other variables that influence both the treatment (smoking and non-smoking) and the outcome (birth weight). Some evidence against this assumption is demonstrated in the correlation matrix below, as there are multiple variables that are correlated to both the treatment and the outcome, for example age has a 0.14 correlation with tobacco usage and a 0.1 correlation with birth weight. Education is also correlated with both birth weigth and tobacco usage.

```
# show correlations between other variables
cormatrix <- cor(smoking_data)

corrplot(cormatrix, method = "shade", shade.col = NA, tl.col = "black",
    tl.srt = 45, addCoef.col = "black", number.font = 8, cl.pos = "n",
    order = "FPC")
```



1.2 (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 an OLS regression with linear controls for the covariates. Report the estimated coefficient on tobacco and its standard error.

```
# run OLS
ols_mod <- lm(data = smoking_data, formula = birthwgt ~ tobacco +
    mage + meduc + mblack + alcohol + first + diabete + anemia)

ols_results <- summary(ols_mod)

ols_results
```

```
##
## Call:
## lm(formula = birthwgt ~ tobacco + mage + meduc + mblack + alcohol +
##     first + diabete + anemia, data = smoking_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -2024.45  -294.06    20.61   329.63  1502.19
##
## Coefficients:
##              Estimate Std. Error t value             Pr(>|t|)
## (Intercept) 3362.2582    11.9273 281.897 < 0.0000000000000002 ***
## tobacco     -228.0731     4.1775 -54.596 < 0.0000000000000002 ***
## mage          -0.6940     0.3566  -1.946               0.0516 .
## meduc         11.6883     0.8605  13.583 < 0.0000000000000002 ***
## mblack      -240.0303     5.1062 -47.007 < 0.0000000000000002 ***
## alcohol      -77.3497    13.4654  -5.744        0.00000000926 ***
## first        -96.9441     3.4466 -28.127 < 0.0000000000000002 ***
## diabete       73.2275    12.1035   6.050        0.00000000145 ***
## anemia        -4.7964    16.7544  -0.286               0.7747
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 484.7 on 94164 degrees of freedom
## Multiple R-squared:  0.0717, Adjusted R-squared:  0.07162
## F-statistic: 909.2 on 8 and 94164 DF,  p-value: < 0.00000000000000022
```

[1] "The coefficient on tobacco is -228.073. This means that if all other variables are held constant, the difference in infants born to smoking mothers would be about 228 grams lighter than babies born to mothers that didn't smoke." [1] "The standard error for the effect of tobacco in this model is 4.177."

1.3 (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.

```
#encode data into bins by specified thresholds
smoking_encoded <- smoking_data |>
  mutate(mage = as.factor(
    case_when(mage >= 34 ~ 1, #greater than 34 = 1
                          mage < 34 ~ 0)), #less than 34 = 0
         meduc = as.factor(
           case_when(meduc >= 16 ~ 1, #greater than 16 = 1
                           meduc < 16 ~ 0)),# less than 16 = 0
    mblack = as.factor(mblack),
    alcohol = as.factor(alcohol)
    ) %>%
  select(c(mage, meduc, mblack, alcohol, tobacco, birthwgt)) %>%  #drop other variables
  add_column(X = as.factor(case_when( #create a factor for 16 unique combinations of covariate
                 .$meduc == 0 & .$mage == 0 & .$mblack == 0 & .$alcohol == 0 ~ 1,
                 .$meduc == 1 & .$mage == 0 & .$mblack == 0 & .$alcohol == 0 ~ 2,
                 .$meduc == 1 & .$mage == 1 & .$mblack == 0 & .$alcohol == 0 ~ 3,
                 .$meduc == 1 & .$mage == 0 & .$mblack == 1 & .$alcohol == 0 ~ 4,
                 .$meduc == 1 & .$mage == 0 & .$mblack == 0 & .$alcohol == 1 ~ 5,
                 .$meduc == 0 & .$mage == 1 & .$mblack == 0 & .$alcohol == 0 ~ 6,
                 .$meduc == 0 & .$mage == 1 & .$mblack == 1 & .$alcohol == 0 ~ 7,
                 .$meduc == 0 & .$mage == 1 & .$mblack == 0 & .$alcohol == 1 ~ 8,
                 .$meduc == 0 & .$mage == 0 & .$mblack == 1 & .$alcohol == 0 ~ 9,
```

```
                  .$meduc == 0 & .$mage == 0 & .$mblack == 1 & .$alcohol == 1 ~ 10,
                  .$meduc == 0 & .$mage == 0 & .$mblack == 0 & .$alcohol == 1 ~ 11,
                  .$meduc == 1 & .$mage == 1 & .$mblack == 1 & .$alcohol == 1 ~ 12,
                  .$meduc == 0 & .$mage == 1 & .$mblack == 1 & .$alcohol == 1 ~ 13,
                  .$meduc == 1 & .$mage == 0 & .$mblack == 1 & .$alcohol == 1 ~ 14,
                  .$meduc == 1 & .$mage == 1 & .$mblack == 0 & .$alcohol == 1 ~ 15,
                  .$meduc == 1 & .$mage == 1 & .$mblack == 1 & .$alcohol == 0 ~ 16
                 )))


TIA_table <- smoking_encoded %>%
  group_by(X,tobacco)%>%
  summarise(n_obs = n(),
            Y_mean = mean(birthwgt, na.rm = TRUE))%>% #Calculate number of observations and Y mean by X
  ungroup()%>%
  mutate(total_obs = sum(n_obs))%>% #Calculate total number of observations
  group_by(tobacco)%>%
  mutate(total_obs_d = sum(n_obs))%>% #Calculate total number of observations by treatment cells
  group_by(X)%>%
  mutate(Y_diff = lead(Y_mean)-Y_mean,
         W_ATE = sum(n_obs)/total_obs,
         W_ATT = lead(n_obs)/lead(total_obs_d))%>% #Calculate difference in outcome and ATE and ATT wei
  ungroup()%>%
  mutate(ATE=sum(W_ATE*Y_diff, na.rm= TRUE),
         ATT=sum(W_ATT*Y_diff, na.rm= TRUE))%>% #Calculate ATE and ATT
  mutate_if(is.numeric, round, 2) #Round data

# TIA_table

gt::gt(TIA_table)
```

X	tobacco	n_obs	Y_mean	total_obs	total_obs_d	Y_diff	W_ATE	W_ATT	ATE	ATT
1	0	44274	3445.69	94173	76021	-225.44	0.61	0.74	-225.37	-227.71
1	1	13443	3220.25	94173	18152	NA	0.61	NA	-225.37	-227.71
2	0	13425	3483.02	94173	76021	-209.08	0.15	0.03	-225.37	-227.71
2	1	535	3273.94	94173	18152	NA	0.15	NA	-225.37	-227.71
3	0	4492	3487.19	94173	76021	-237.74	0.05	0.01	-225.37	-227.71
3	1	201	3249.45	94173	18152	NA	0.05	NA	-225.37	-227.71
4	0	625	3319.22	94173	76021	-160.17	0.01	0.00	-225.37	-227.71
4	1	61	3159.05	94173	18152	NA	0.01	NA	-225.37	-227.71
5	0	130	3510.95	94173	76021	-97.74	0.00	0.00	-225.37	-227.71
5	1	29	3413.21	94173	18152	NA	0.00	NA	-225.37	-227.71
6	0	5115	3467.41	94173	76021	-295.98	0.06	0.05	-225.37	-227.71
6	1	976	3171.42	94173	18152	NA	0.06	NA	-225.37	-227.71
7	0	396	3185.08	94173	76021	-190.41	0.01	0.01	-225.37	-227.71
7	1	135	2994.67	94173	18152	NA	0.01	NA	-225.37	-227.71
8	0	56	3358.32	94173	76021	-260.59	0.00	0.00	-225.37	-227.71
8	1	45	3097.73	94173	18152	NA	0.00	NA	-225.37	-227.71
9	0	7007	3195.97	94173	76021	-189.66	0.10	0.11	-225.37	-227.71
9	1	1980	3006.31	94173	18152	NA	0.10	NA	-225.37	-227.71
10	0	71	3120.07	94173	76021	-302.73	0.00	0.01	-225.37	-227.71
10	1	226	2817.34	94173	18152	NA	0.00	NA	-225.37	-227.71
11	0	214	3450.28	94173	76021	-326.03	0.01	0.02	-225.37	-227.71

11	1	448	3124.25	94173	18152	NA	0.01	NA	-225.37	-227.71
12	0	1	3459.00	94173	76021	-624.00	0.00	0.00	-225.37	-227.71
12	1	1	2835.00	94173	18152	NA	0.00	NA	-225.37	-227.71
13	0	7	2739.71	94173	76021	106.67	0.00	0.00	-225.37	-227.71
13	1	26	2846.38	94173	18152	NA	0.00	NA	-225.37	-227.71
14	0	4	2983.50	94173	76021	114.20	0.00	0.00	-225.37	-227.71
14	1	10	3097.70	94173	18152	NA	0.00	NA	-225.37	-227.71
15	0	57	3534.91	94173	76021	-497.44	0.00	0.00	-225.37	-227.71
15	1	17	3037.47	94173	18152	NA	0.00	NA	-225.37	-227.71
16	0	147	3328.29	94173	76021	-476.13	0.00	0.00	-225.37	-227.71
16	1	19	2852.16	94173	18152	NA	0.00	NA	-225.37	-227.71

```
# MULTIVARIATE MATCHING AS REGRESSION ESTIMATOR
se_models = starprep(ols_mod, stat = c("std.error"), se_type = "HC2",
    alpha = 0.05)
stargazer(ols_mod, se = se_models, type = "text")
```

```
##
## ================================================
##                         Dependent variable:
##                     ---------------------------
##                               birthwgt
## ------------------------------------------------
## tobacco                     -228.073***
##                               (4.277)
##
## mage                          -0.694*
##                               (0.368)
##
## meduc                         11.688***
##                               (0.862)
##
## mblack                      -240.030***
##                               (5.348)
##
## alcohol                     -77.350***
##                               (14.039)
##
## first                       -96.944***
##                               (3.488)
##
## diabete                       73.228***
##                               (13.235)
##
## anemia                        -4.796
##                               (17.874)
##
## Constant                   3,362.258***
##                               (12.076)
##
## ------------------------------------------------
## Observations                   94,173
## R2                              0.072
## Adjusted R2                     0.072
```

```
## Residual Std. Error    484.733 (df = 94164)
## F Statistic         909.176*** (df = 8; 94164)
## ================================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

[1] "The estimated average treatment effect of smoking on birthweight is -225.37 grams, using the exact matching method." [1] "This differs only slightly (by about 3 grams) from the estimated effect of smoking on infant birthweigth using the linear regression analogue, which was -228.073 grams."