EDS241: Assignment 2

Guillermo Romero

Question 1:

Application of estimators based on the "treatment 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)

What is the unadjusted mean difference in birth weight of infants with smoking and nonsmoking mothers?

```
mean_weight_smk_nsmk <- df_smoking |>
  group_by(tobacco) |>
  summarise(mean_weight = mean(birthwgt))

round(mean_weight_smk_nsmk$mean_weight[1] - mean_weight_smk_nsmk$mean_weight[2],2 )
```

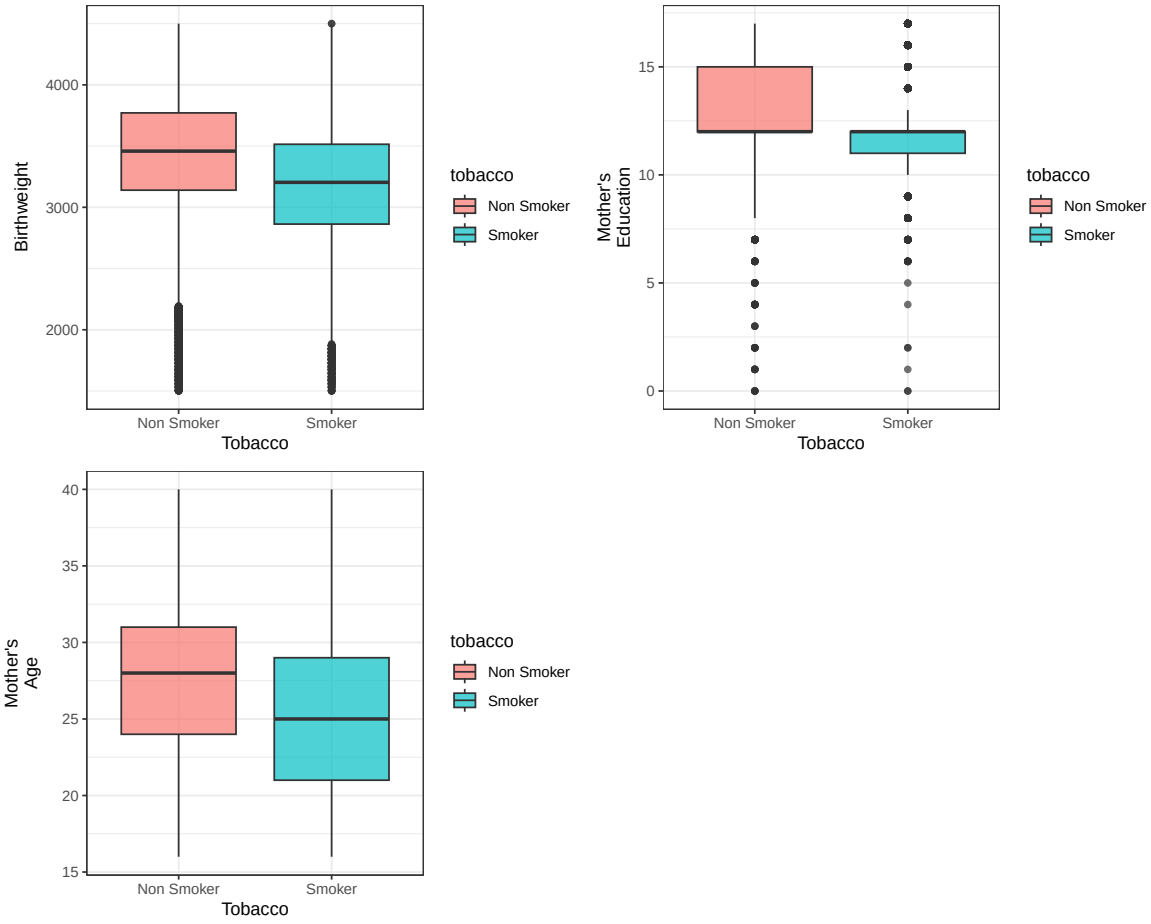
```
[1] 244.54
```

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

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.

This corresponds to the assumption of "treatment ignorability" conditional on pre treatment characteristics Xi (Rubin and Rosenbaum). There is evidence against this assumption as the distribution in the following boxplot is not balanced. Also, except for the control variable of the mother being diabetic, the F-statistic are large and the associated p-value is small, it suggests that the predictor variables in the model are jointly significant in explaining the variation in the outcome variable.

```
bw + meduc + mage + plot_layout(nrow = 2, byrow = TRUE)
```



```r
# EXAMINE BALANCE IN COVARIATES
# COVARIATE MEAN DIFFERENCES by tobacco
m1 <- lm(formula = birthwgt ~ tobacco, data = df_smoking)
m2 <- lm(formula = mblack ~ tobacco, data = df_smoking)
m3 <- lm(formula = alcohol ~ tobacco, data = df_smoking)
m4 <- lm(formula = first ~ tobacco, data = df_smoking)
m5 <- lm(formula = diabete ~ tobacco, data = df_smoking)
m6 <- lm(formula = anemia ~ tobacco, data = df_smoking)
m7 <- lm(formula = mage ~ tobacco, data = df_smoking)
m8 <- lm(formula = meduc ~ tobacco , data = df_smoking)


se_models = starprep(
  m1,
  m2,
  m3,
  m4,
  m5,
  m6,
  m7,
  m8,
  stat = c("std.error"),
  se_type = "HC2",
  alpha = 0.05
)
```

```r
stargazer(
  m1,
  m2,
  m3,
  m4,
  m5,
  m6,
  m7,
  m8,
  se = se_models,
  type = "latex",
  font.size = 'small',
  summary = FALSE,
  digits = 2,
  column.sep.width = '-8pt',
```

```
  no.space = TRUE
)
```

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: Mon, Mar 13, 2023 - 21:40:00

Table 1

	Dependent variable:							
	birthwgt	mblack	alcohol	first	diabete	anemia	mage	meduc
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
tobacco	−244.54***	0.03***	0.04***	−0.07***	0.0002	0.01***	−1.91***	−1.32***
	(4.15)	(0.003)	(0.002)	(0.004)	(0.001)	(0.001)	(0.04)	(0.01)
Constant	3,430.29***	0.11***	0.01***	0.44***	0.02***	0.01***	27.45***	13.24***
	(1.78)	(0.001)	(0.0003)	(0.002)	(0.0005)	(0.0003)	(0.02)	(0.01)
Observations	94,173	94,173	94,173	94,173	94,173	94,173	94,173	94,173
R^2	0.04	0.001	0.02	0.003	0.0000	0.001	0.02	0.06
Adjusted R^2	0.04	0.001	0.02	0.003	−0.0000	0.001	0.02	0.06
Residual Std. Error (df = 94171)	493.75	0.32	0.12	0.49	0.13	0.09	5.29	2.05
F Statistic (df = 1; 94171)	3,594.26***	104.36***	1,456.31***	308.07***	0.02	65.20***	1,919.59***	6,072.21***

Note: *p<0.1; **p<0.05; ***p<0.01

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

```
mod <-
  lm(
    formula = birthwgt ~ tobacco + as.factor(anemia) + as.factor(diabete) + as.factor(alco
      as.factor(mblack) + as.factor(first) + mage + meduc + birthwgt,
    data = df_smoking
  )

se_models = starprep(mod , stat = c("std.error"), se_type = "HC1", alpha = 0.05)

stargazer(mod, se = se_models, type="text", omit = "(LME)|(genus)|(species)")
```

```
================================================
                        Dependent variable:
                    ----------------------------
                              birthwgt
------------------------------------------------
tobacco                     -228.073***
                              (4.277)

as.factor(anemia)1             -4.796
                              (17.864)

as.factor(diabete)1            73.228***
                              (13.232)

as.factor(alcohol)1           -77.350***
                              (14.034)

as.factor(mblack)1            -240.030***
                              (5.348)

as.factor(first)1             -96.944***
                              (3.488)

mage                           -0.694*
                              (0.368)

meduc                          11.688***
                              (0.862)

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
```

The estimated coefficient on tobacco is -228.073 and the standar error is 4.277.

(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.

```
df_matching <- df_smoking |>
  mutate(mage_i = case_when(mage >= 34 ~ 1,
                            TRUE ~ 0)) |>
  mutate(meduc_i = case_when(meduc >= 16 ~ 1,
                             TRUE ~0)) |>
  select(birthwgt,tobacco, alcohol, mblack, mage_i, meduc_i)

df_matching
```

```
# A tibble: 94,173 x 6
   birthwgt tobacco alcohol mblack mage_i meduc_i
      <dbl>   <dbl>   <dbl>  <dbl>  <dbl>   <dbl>
 1     4129       0       0      0      0       1
 2     3638       0       0      0      0       0
 3     3694       0       0      0      0       0
 4     3799       0       0      0      0       0
 5     3175       0       0      0      1       0
 6     2892       0       0      0      0       0
 7     3572       0       0      0      0       0
 8     3232       0       0      0      0       0
 9     3572       0       0      0      0       0
10     2820       0       0      1      1       0
# ... with 94,163 more rows
```

```
linear_est <-
  lm(formula = birthwgt ~ tobacco + alcohol + mblack + mage_i  + meduc_i,
     data = df_matching)

se_models = starprep(
  linear_est,
  stat = c("std.error"),
  se_type = "HC2",
  alpha = 0.05
```

```
)

stargazer(linear_est, se = se_models, type = "text")
```

```
================================================
                        Dependent variable:
                    ----------------------------
                               birthwgt
------------------------------------------------
tobacco                      -226.769***
                               (4.213)

alcohol                       -71.854***
                               (14.073)

mblack                       -238.643***
                               (5.258)

mage_i                         5.104
                               (5.034)

meduc_i                       42.064***
                               (4.003)

Constant                     3,445.583***
                               (2.176)

------------------------------------------------
Observations                    94,173
R2                              0.062
Adjusted R2                     0.062
Residual Std. Error      487.182 (df = 94167)
F Statistic          1,250.663*** (df = 5; 94167)
================================================
Note:                *p<0.1; **p<0.05; ***p<0.01
```

```
TIA_table <- df_smoking %>%
  mutate(
    mage_indicator = case_when(mage >= 34 ~ 1,
                               mage < 34 ~ 0),
```

```
    meduc_indicator = case_when(meduc >= 16 ~ 1,
                                meduc < 16 ~ 0)
  ) %>% #Create observed Y variable
  mutate(factors = as.factor(paste0(
    mblack, alcohol, mage_indicator, meduc_indicator
  ))) %>%
  group_by(factors, tobacco) %>%
  summarise(n_obs = n(),
            Y_mean = mean(birthwgt, na.rm = T)) %>% #Calculate number of observations and
  gather(variables, values, n_obs:Y_mean) %>% #Reshape data
  mutate(variables = paste0(variables, "_", tobacco, sep = "")) %>% #Combine the treatment
  pivot_wider(id_cols = factors,
              names_from = variables,
              values_from = values) %>% #Reshape data by treatment and X cell
  ungroup() %>%  #Ungroup from X values
  mutate(
    Y_diff = Y_mean_1 - Y_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)
```

```
=================================================================
   factors n_obs_0 n_obs_1 Y_mean_0 Y_mean_1 Y_diff  w_ATE w_ATT
-----------------------------------------------------------------
1      1     44274   13443  3445.69  3220.25  -225.44 0.61  0.74
2      2     13425    535   3483.02  3273.94  -209.08 0.15  0.03
3      3     5115     976   3467.41  3171.42  -295.98 0.06  0.05
4      4     4492     201   3487.19  3249.45  -237.74 0.05  0.01
5      5      214     448   3450.28  3124.25  -326.03 0.01  0.02
6      6      130     29    3510.95  3413.21  -97.74    0     0
7      7      56      45    3358.32  3097.73  -260.59   0     0
8      8      57      17    3534.91  3037.47  -497.44   0     0
9      9     7007    1980   3195.97  3006.31  -189.66  0.1  0.11
10    10      625     61    3319.22  3159.05  -160.17 0.01    0
```

```
11   11      396     135   3185.08  2994.67  -190.41 0.01  0.01
12   12      147     19    3328.29  2852.16  -476.13   0     0
13   13      71      226   3120.07  2817.34  -302.73   0   0.01
14   14       4      10     2983.5   3097.7   114.2    0     0
15   15       7      26    2739.71  2846.38  106.67    0     0
16   16       1       1      3459     2835    -624     0     0
-----------------------------------------------------------------
```

```
TIA_table
```

```
# A tibble: 16 x 8
   factors n_obs_0 n_obs_1 Y_mean_0 Y_mean_1 Y_diff w_ATE w_ATT
   <fct>     <dbl>   <dbl>    <dbl>    <dbl>  <dbl> <dbl> <dbl>
 1 0000      44274   13443    3446.    3220. -225.   0.61  0.74
 2 0001      13425     535    3483.    3274. -209.   0.15  0.03
 3 0010       5115     976    3467.    3171. -296.   0.06  0.05
 4 0011       4492     201    3487.    3249. -238.   0.05  0.01
 5 0100        214     448    3450.    3124. -326.   0.01  0.02
 6 0101        130      29    3511.    3413.  -97.7  0     0
 7 0110         56      45    3358.    3098. -261.   0     0
 8 0111         57      17    3535.    3037. -497.   0     0
 9 1000       7007    1980    3196.    3006. -190.   0.1   0.11
10 1001        625      61    3319.    3159. -160.   0.01  0
11 1010        396     135    3185.    2995. -190.   0.01  0.01
12 1011        147      19    3328.    2852. -476.   0     0
13 1100         71     226    3120.    2817. -303.   0     0.01
14 1101          4      10    2984.    3098.  114.   0     0
15 1110          7      26    2740.    2846.  107.   0     0
16 1111          1       1    3459     2835  -624    0     0
```

```
#MULTIVARIATE MATCHING ESTIMATES OF ATE AND ATT
ATE = sum((TIA_table$w_ATE) * (TIA_table$Y_diff))
ATE
```

```
[1] -224.2583
```

```
ATT = sum((TIA_table$w_ATT) * (TIA_table$Y_diff))
ATT
```

```
[1] -222.589
```

ATE of exact matching estimator is -224.2583 and its linear regression analogue is -226.769.