EDS 241: Assignment 3

Alexandra Yousefivand

03/25/2022

```
# load packages
packages=c("readxl", "stargazer", "here", "tidyr",
           "dplyr","stringr", "janitor",
           "cowplot", "ggplot2", "tinytex",
           "datasets", "tibble", "estimatr")

for (i in packages) {
  if (require(i,character.only=TRUE)==FALSE) {
    install.packages(i,repos='http://cran.us.r-project.org')
  }
  else {
    require(i,character.only=TRUE)
  }
}

#devtools::install_github('rstudio/rmarkdown')
options(scipen=999) # not scientific notation
```

```
# load data
data_raw <- read.csv("SMOKING_EDS241.csv")

data <- data_raw %>% clean_names()
```

Question (a) (a)

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

```
# subset smoking mothers
data_smoking <- data %>% filter(tobacco == 1)

# subset nonsmoking mothers
data_non_smoking <- data %>% filter(tobacco == 0)
```

```
unadjusted_mean_difference <-
  mean(data_non_smoking$birthwgt) - mean(data_smoking$birthwgt)
```

The unadjusted mean difference is 244.54 grams.

Under what hypothesis 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.

Treatment ignorability: all other confounding variables (education, race, alcohol, first, diabete, anemia) are held constant.

Question (b) (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 <- lm_robust(formula = birthwgt ~ tobacco,
                   data = data)
summary(model)
```

```
##
## Call:
## lm_robust(formula = birthwgt ~ tobacco, data = data)
##
## Standard error type:  HC2
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper    DF
## (Intercept)   3430.3      1.781 1926.11        0   3426.8   3433.8 94171
## tobacco       -244.5      4.150  -58.93        0   -252.7   -236.4 94171
##
## Multiple R-squared:  0.03676 ,   Adjusted R-squared:  0.03675
## F-statistic:  3473 on 1 and 94171 DF,  p-value: < 0.00000000000000022
```

```
tobacco_coefficient <- model$coefficients[[2]]
tobacco_coefficient
```

```
## [1] -244.5394
```

```
tobacco_std_error <- model$std.error[[2]]
tobacco_std_error
```

```
## [1] 4.149552
```

Question (c) (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).*

```r
data_indicator <- data %>%
  mutate(age_above_34 = case_when(mage >= 34 ~ 1,
                                  mage <  34 ~ 0),
         educ_above_16 = case_when(meduc >= 16 ~ 1,
                                   meduc <  16 ~ 0),
         g = as.factor(paste0(age_above_34,
                              educ_above_16,
                              mblack,
                              alcohol)))
```

```r
model <- lm(formula = birthwgt ~ age_above_34 + educ_above_16 + mblack + alcohol,
            data = data_indicator)
summary(model)
```

```
##
## Call:
## lm(formula = birthwgt ~ age_above_34 + educ_above_16 + mblack +
##     alcohol, data = data_indicator)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1987.59  -303.61    23.38   346.02  1346.02
##
## Coefficients:
##               Estimate Std. Error  t value             Pr(>|t|)
## (Intercept)   3393.612      2.001 1696.136 < 0.0000000000000002 ***
## age_above_34    13.969      4.992    2.798              0.00514 **
## educ_above_16   83.005      4.062   20.437 < 0.0000000000000002 ***
## mblack        -239.630      5.123  -46.776 < 0.0000000000000002 ***
## alcohol       -164.133     13.629  -12.043 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 494.9 on 94168 degrees of freedom
## Multiple R-squared:  0.03235,    Adjusted R-squared:  0.03231
## F-statistic: 787.1 on 4 and 94168 DF,  p-value: < 0.00000000000000022
```

```r
TIA <- data_indicator %>%
  group_by(g, tobacco) %>%
  summarise(n_obs = as.integer(n()),
            birthwgt_mean = mean(birthwgt, na.rm = TRUE)) %>%
  gather(variables, values, n_obs:birthwgt_mean) %>%
  mutate(variables = paste0(variables, "_", tobacco)) %>%
  pivot_wider(id_cols = g, names_from = variables, values_from = values) %>%
  ungroup() %>%
  mutate(diff = birthwgt_mean_1 - birthwgt_mean_0,
         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)) %>%
  mutate_if(is.numeric, round, 2)
```

```r
head(TIA)
```

```
## # A tibble: 6 x 8
```

```
##   g     n_obs_0 n_obs_1 birthwgt_mean_0 birthwgt_mean_1   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      214     448           3450.           3124. -326.   0.01  0.02
## 3 0010     7007    1980           3196.           3006. -190.   0.1   0.11
## 4 0011       71     226           3120.           2817. -303.   0     0.01
## 5 0100    13425     535           3483.           3274. -209.   0.15  0.03
## 6 0101      130      29           3511.           3413.  -97.7  0     0
```

```
# Average Treatment Effect
ATE = sum(TIA$w_ATE * TIA$diff)
ATE
```

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

```
# Average Treatment effect on the Treated
ATT = sum(TIA$w_ATT * TIA$diff)
ATT
```

```
## [1] -222.589
```

```
model <- lm_robust(formula = birthwgt ~ tobacco + g, data = data_indicator)
summary(model)
```

```
##
## Call:
## lm_robust(formula = birthwgt ~ tobacco + g, data = data_indicator)
##
## Standard error type:  HC2
##
## Coefficients:
##             Estimate Std. Error   t value
## (Intercept)  3445.87      2.232 1543.6015
## tobacco      -226.25      4.220  -53.6104
## g0001         -63.12     20.431   -3.0897
## g0010        -241.84      5.742  -42.1186
## g0011        -384.01     29.870  -12.8558
## g0100          37.81      4.535    8.3374
## g0101          88.51     38.413    2.3042
## g0110        -120.78     18.977   -6.3643
## g0111        -219.20    127.345   -1.7213
## g1000          10.36      6.819    1.5192
## g1001        -102.85     45.144   -2.2783
## g1010        -251.69     24.106  -10.4408
## g1011        -443.86     79.415   -5.5892
## g1100          40.82      7.404    5.5136
## g1101          26.74     55.254    0.4839
## g1110        -146.19     38.555   -3.7917
## g1111        -185.75    198.895   -0.9339
##                                                Pr(>|t|) CI Lower CI Upper    DF
## (Intercept) 0.00000000000000000000000000000000000000000 3441.498  3450.25 94156
## tobacco     0.00000000000000000000000000000000000000000 -234.517  -217.97 94156
```

```
## g0001       0.00200417530014243170863319498664623097284 -103.168   -23.08 94156
## g0010       0.00000000000000000000000000000000000000000 -253.093  -230.58 94156
## g0011       0.00000000000000000000000000000000000008586 -442.552  -325.46 94156
## g0100       0.00000000000000007691931962254939598305770   28.921    46.70 94156
## g0101       0.02121327353044220528910379641729377908632   13.222   163.80 94156
## g0110       0.00000000019707604983019180199844966328016 -157.970   -83.58 94156
## g0111       0.08520084240754143001250753286512917838991 -468.793    30.40 94156
## g1000       0.12871952077994658347748213600425515323877   -3.006    23.72 94156
## g1001       0.02270863724874112518636870561294927028939 -191.334   -14.37 94156
## g1010       0.00000000000000000000000016678365899005575 -298.934  -204.44 94156
## g1011       0.00000022877809705132968708107821242947 47 -599.513  -288.21 94156
## g1100       0.00000035254957046125877180749175199281 35   26.312    55.34 94156
## g1101       0.62845742779829927293633318186039105057716  -81.559   135.03 94156
## g1110       0.00014970515982212615922405940516171085619 -221.755   -70.62 94156
## g1111       0.35035048914475463988082992727868258953094 -575.582   204.08 94156
##
## Multiple R-squared:  0.06269 ,   Adjusted R-squared:  0.06253
## F-statistic:   375 on 16 and 94156 DF,  p-value: < 0.00000000000000022
```

```
tobacco_coefficient <- model$coefficients[["tobacco"]]
tobacco_coefficient
```

```
## [1] -226.245
```

Question (d) (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.

```
data_logit <- data_indicator %>% mutate(mage_squared = mage^2)
```

```
model_logit <- glm(formula = tobacco ~ mage +
                                       mage_squared +
                                       meduc +
                                       mblack +
                                       alcohol,
                   family = binomial(link = "logit"),
                   data = data_logit)
summary(model_logit)
```

```
##
## Call:
## glm(formula = tobacco ~ mage + mage_squared + meduc + mblack +
##     alcohol, family = binomial(link = "logit"), data = data_logit)
##
## 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 < 0.0000000000000002 ***
## 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 < 0.0000000000000002 ***
## mblack       -0.059525   0.026506  -2.246               0.0247 *
## alcohol       2.022696   0.060358  33.511 < 0.0000000000000002 ***
## ---
## 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_logit, type = "response")
```

```
PS_WGT <- (data_logit$tobacco / EPS) + ((1 - data_logit$tobacco) / (1 - EPS))
```

```
head(PS_WGT)
```

```
##        1        2        3        4        5        6
## 1.081285 1.384348 1.286888 1.305889 1.566520 1.595757
```

Question (e) (e)

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

```
model_ps <- lm_robust(formula = birthwgt ~ tobacco +
                                           mage +
                                           mage_squared +
                                           meduc +
                                           mblack +
                                           alcohol,
                      weights = PS_WGT,
                      data = data_logit)
summary(model_ps)
```

```
##
## Call:
## lm_robust(formula = birthwgt ~ tobacco + mage + mage_squared +
##     meduc + mblack + alcohol, data = data_logit, weights = PS_WGT)
##
## Weighted, Standard error type:  HC2
##
## Coefficients:
```

```
##                Estimate Std. Error t value
## (Intercept)  2971.4438   57.06041  52.075
## tobacco      -220.2328    5.02912 -43.792
## mage           27.6275    4.58742   6.022
## mage_squared   -0.4777    0.08655  -5.519
## meduc           7.4716    1.58370   4.718
## mblack       -220.9902    8.24519 -26.802
## alcohol       -71.9137   16.73365  -4.298
##
## (Intercept)  0.00000000000000000000000000000000000000000000000000000000000000000000000000000000000000
## tobacco      0.00000000000000000000000000000000000000000000000000000000000000000000000000000000000000
## mage         0.0000000017243504421538613121094724034421119540638756006956100463867187500000000000000000
## mage_squared 0.0000000342278957546733582340513224462341668186127208173274993896484375000000000000000000
## meduc        0.0000023876968997998810587297946472773446657811291515827178955078125000000000000000000
## mblack       0.00000000000000000000000000000000000000000000000000000000000000000000000000000000000000
## alcohol      0.0000172871301575458120067078915393921079157735221087932586669921875000000000000000000
##                CI Lower CI Upper    DF
## (Intercept)  2859.6060 3083.282 94166
## tobacco      -230.0899 -210.376 94166
## mage           18.6362   36.619 94166
## mage_squared   -0.6473   -0.308 94166
## meduc           4.3675   10.576 94166
## mblack       -237.1507 -204.830 94166
## alcohol      -104.7114  -39.116 94166
##
## Multiple R-squared:  0.0736 ,    Adjusted R-squared:  0.07354
## F-statistic: 764.2 on 6 and 94166 DF,  p-value: < 0.00000000000000022
```

```r
tobacco_coeff_ps <- model_ps$coefficients[["tobacco"]]
tobacco_coeff_ps
```

```
## [1] -220.2328
```