EDS241: Assignment 03 - National Natality Detail

Julia Parish

2022/02/20

1 EDS241 Environmental Policy Evaluation Assignment 03

This statistical analysis was completed as an assignment for the course, Environmental Data Science 241: Environmental Policy Evaluation. It is an application of estimators based on treatment ignorability. The goal of this assignment was to 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 data files for this assignment 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 is Black
- `alcohol`: = 1 if consumed alcohol during pregnancy
- `first`: = 1 if first child
- `diabete`: = 1 if mother is diabetic
- `anemia`: = 1 if mother anemic

Note: This exercise asks you to implement some of the techniques presented in Lectures 6-7. This homework is a simple examination of these data. More research would be needed to obtain a more definitive assessment of the causal effect of smoking on infant health outcomes. Further, for this homework, you can ignore the adjustments to the standard errors that are necessary to reflect the fact that the propensity score is estimated. Just use heteroskedasticity robust standard errors in R. If you are interested, you can read Imbens and Wooldridge (2009) and Imbens (2014) for discussions of various approaches and issues with standard error estimations in models based on the propensity score.

2 Data

```
# read in the data
nn_data <- read_csv(here("hw03/data/smoking.csv")) %>%
  clean_names()
```

3 Homework Questions

3.1 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 assumption.

- For the last part of (a), regress your favorite covariate on the smoking status of mothers. For example, think of regressing meduc ~ tobacco. Is the mean difference in the education level of smoking and non-smoking mothers statistically different from zero? What does that say about the required assumption to interpret the unadjusted mean difference as causal?

```
# calculate the unadjusted mean difference in birth weight of infants with smoking and non-smoking moth

smoker <- nn_data %>% filter(tobacco == 1)
nonsmoker <- nn_data %>% filter(tobacco == 0)

smoker_mean <- round(mean(smoker$birthwgt), 3)
nonsmoker_mean <- round(mean(nonsmoker$birthwgt), 3)

unadj_diff <- smoker_mean - nonsmoker_mean
```

```
# Regress infant birth weight (birthwgt) in grams on the indicator for maternal smoking (tobacco)
mod_a1 <- lm_robust(birthwgt ~ tobacco, data = nn_data)
```

```
#create table with regression results
mod_a1_table <- tidy(mod_a1)

mod_a1_table %>%
  select(term, estimate, std.error, p.value, conf.low, conf.high) %>%
  kable()
```

term	estimate	std.error	p.value	conf.low	conf.high
(Intercept)	3430.2863	1.780943	0	3426.7957	3433.7769
tobacco	-244.5394	4.149552	0	-252.6725	-236.4063

```
mod_a1_table
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high	df	outcome
(Intercept)	3.43e+03	1.78	1.93e+03	0	3.43e+03	3.43e+03	9.42e+04	birthwgt
tobacco	-245	4.15	-58.9	0	-253	-236	9.42e+04	birthwgt

```
# Regress the education level (meduc) of the mother on the indicator for maternal smoking (tobacco)
mod_a2 <- lm_robust(meduc ~ tobacco, data = nn_data)
```

```
#create table with regression results
mod_a2_table <- tidy(mod_a2)

mod_a2_table %>%
  select(term, estimate, std.error, p.value, conf.low, conf.high) %>%
  kable()
```

term	estimate	std.error	p.value	conf.low	conf.high
(Intercept)	13.239421	0.0077600	0	13.224211	13.25463
tobacco	-1.318475	0.0142478	0	-1.346401	-1.29055

```
mod_a2_table
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high	df	outcome
(Intercept)	13.2	0.00776	1.71e+03	0	13.2	13.3	9.42e+04	meduc
tobacco	-1.32	0.0142	-92.5	0	-1.35	-1.29	9.42e+04	meduc

3.1.1 Answers

The mean infant birthweight for those born from mothers who do smoke is 3185.747. The mean infant birthweight for those born from mothers who do NOT smoke is 3430.286. The unadjusted mean difference in birth weight of infants with smoking and non-smoking mothers is -244.539 grams. This means that, on average, infants born to mothers who smoke weighed 244.54 grams less than infants born to mothers who do not smoke.

The unadjusted mean difference corresponds to the average treatment effect of mother's smoking during pregnancy on infant birth weight, assuming that the treatment of whether a mother is a smoker or not is randomly assigned and statistically significant. Smoking status of mothers during pregnancy is independent of `Y(1)` and `Y(0)`.

Empirical evidence against the assumption that smoking treatment is randomly assigned to mothers during pregnancy is that another variable, education, is significantly correlated with the indicator for maternal smoking as shown in the linear regression model, `mod_a2`.

3.2 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.

```
# Regress infant birth weight (birthwgt) in grams conditional on all variables in the data set

mod_b <- lm_robust(birthwgt ~ ., data = nn_data)
```

```
#create table with regression results
mod_b_table <- tidy(mod_b)

mod_b_table %>%
  select(term, estimate, std.error, p.value, conf.low, conf.high) %>%
  kable()
```

term	estimate	std.error	p.value	conf.low	conf.high
(Intercept)	3362.2582445	12.0764983	0.0000000	3338.588438	3385.9280506
anemia	-4.7963916	17.8739216	0.7884338	-39.829085	30.2363013
diabete	73.2275309	13.2354917	0.0000000	47.286110	99.1689514
tobacco	-228.0730765	4.2767834	0.0000000	-236.455526	-219.6906273
alcohol	-77.3497487	14.0391720	0.0000000	-104.866374	-49.8331235
mblack	-240.0303000	5.3477693	0.0000000	-250.511870	-229.5487301
first	-96.9441154	3.4880224	0.0000000	-103.780602	-90.1076293
mage	-0.6940244	0.3681995	0.0594445	-1.415691	0.0276425
meduc	11.6883416	0.8617788	0.0000000	9.999265	13.3774186

```
mod_b_table
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high	df	outcome
(Intercept)	3.36e+03	12.1	278	0	3.34e+03	3.39e+03	9.42e+04	birthwgt
anemia	-4.8	17.9	-0.268	0.788	-39.8	30.2	9.42e+04	birthwgt
diabete	73.2	13.2	5.53	3.16e-08	47.3	99.2	9.42e+04	birthwgt
tobacco	-228	4.28	-53.3	0	-236	-220	9.42e+04	birthwgt
alcohol	-77.3	14	-5.51	3.61e-08	-105	-49.8	9.42e+04	birthwgt
mblack	-240	5.35	-44.9	0	-251	-230	9.42e+04	birthwgt
first	-96.9	3.49	-27.8	2.53e-169	-104	-90.1	9.42e+04	birthwgt
mage	-0.694	0.368	-1.88	0.0594	-1.42	0.0276	9.42e+04	birthwgt
meduc	11.7	0.862	13.6	7.26e-42	10	13.4	9.42e+04	birthwgt

3.2.1 Answers

The estimated effect of maternal smoking on birth weight is a decrease of 228.07 grams on average. The standard error is 4.28.

3.3 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). Once you have your 4 dummy variables, you can create a group variable `g` using the paste0() function. For example, `mutate(g = paste0(d1,d2,d3,d4))` . The resulting `g` will include all potential observed combinations of the 4 dummy variables in the data. You can then control for `factor(g)` in the regression model. To calculate the exact matching estimator, use this `g` grouping variable and the code from lines 76 to 97 in the `TIA.R` script on gauchospace. In this case, Y = birthwgt, X = g, D = tobacco. Since we observe `Y1` or `Y0`, you can ignore line 77. *see TIA Table #ydiff = delta for x, w_ATE # of obs for rows, w_ATT = weights*

3.3.1 Section 1: Exact matching estimator

Use the exact matching estimator to estimate the effect of maternal smoking on birth weight. Consider the following covariates in your matching estimator:

- mother's age (=1 if mage>=34),
- 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.

```r
# create 0-1 indicators for mother's education and age.

matching_nn_data <- nn_data %>%
  mutate(
    mage_sq = (mage*mage),
    mage = case_when(
      mage >= 34 ~ 1,
      mage <34 ~ 0),
    meduc = case_when(
      meduc >= 16 ~ 1,
      meduc < 16 ~ 0),
    mblack = as.factor(mblack),
    alcohol = as.factor(alcohol),
    covariates = paste0(mage, meduc, mblack, alcohol)
    )
```

```r
# create average treatment estimate of smoking on birth weight using exact matching estimator

tia_table <- matching_nn_data %>%
  group_by(covariates, tobacco) %>%
  summarise(n_obs = n(), # number of observations
            birthwgt_mean = mean(birthwgt, na.rm = TRUE)) %>% # calculate birthwgt mean by X by treatme
  gather(variables, values, n_obs:birthwgt_mean) %>% # reshape the dataframe

  mutate(variables = paste0(variables, "_", tobacco, sep = "")) %>%  # combine the treatment and variab
  pivot_wider(id_cols = covariates,  # reshape data by treatment and X cell
              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)), # calculate ATE
         w_ATT = n_obs_1 / sum(n_obs_1)) %>%  # calculate ATT weights
  mutate_if(is.numeric, round, 2)
```

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

```
##
## ===================================================================================
##    covariates 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 AND ATT
ate = sum((tia_table$w_ATE)*(tia_table$birthwgt_diff))
ate
```

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

```
att = sum((tia_table$w_ATT)*(tia_table$birthwgt_diff))
att
```

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

3.3.1.1 Answers

The average treatment effect of smoking on birthweight using the exact matching estimator is -224.26 grams.

6.1

3.4 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. glm(formula, family = binomial(), data) is a logit model.

```
# create a new dataframse and add a new column transforming the age variable by squaring it

propensity_data <- matching_nn_data %>%
  mutate(mage_sq = mage^2) %>%
  select(tobacco,
         mage,
         mage_sq,
         meduc,
         mblack,
         birthwgt,
         alcohol)
```

```
# ESTIMATE PROPENSITY SCORE MODEL

propensity_model <- glm(tobacco ~ mage + mage_sq + meduc + mblack + alcohol,
                        family = binomial(),
                        data = propensity_data)
```

```
# create new EPS variable for the estimated propensity score
eps <- predict(propensity_model, type = "response")

eps_sample <- head(eps, 5) # sample eps
```

3.4.1 Answers

A sample (n = 5) of the estimated propensity score for maternal smoking during pregnancy using a logit estimator (glm) are 0.0447602, 0.2299876, 0.2299876, 0.2299876, 0.1757896.

3.5 Question E:

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

```
# create new variable for the weighted propensity score
ps_wgt <- (propensity_data$tobacco / eps) +
  ((1 - propensity_data$tobacco) / (1 - eps))

wgt_sample <- head(eps, 5) # sample weighted propensity score
```

```
# propensity score weighted regression (WLS) lm(formula = Y ~ D + X1 ..., data=DF, weights=wgt)

mod_wgt <- lm_robust(birthwgt ~ tobacco, data = propensity_data, weights = ps_wgt)
summary(mod_wgt)
```

```
##
## Call:
## lm_robust(formula = birthwgt ~ tobacco, data = propensity_data,
##     weights = ps_wgt)
##
## Weighted, Standard error type:  HC2
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper    DF
## (Intercept)   3427.2      1.804 1899.51        0   3423.7     3431 94171
## tobacco       -227.6      5.366  -42.41        0   -238.1     -217 94171
##
## Multiple R-squared:  0.04915 ,   Adjusted R-squared:  0.04914
## F-statistic:  1798 on 1 and 94171 DF,  p-value: < 0.00000000000000022
```

7.1

```
# create propensity score weighted regression table
mod_wgt_table <- tidy(mod_wgt)

mod_wgt_table %>%
  select(term, estimate, std.error, p.value, conf.low, conf.high) %>%
  kable()
```

term	estimate	std.error	p.value	conf.low	conf.high
(Intercept)	3427.2250	1.804268	0	3423.689	3430.7613
tobacco	-227.5555	5.366082	0	-238.073	-217.0381

3.5.1 Answers

To create a weighted propensity score, weights were assigned as shown in variable, `ps_wgt`. A sample (n = 5) of the weighted propensity score for maternal smoking during pregnancy are 0.0447602, 0.2299876, 0.2299876, 0.2299876, 0.1757896.

The estimated effect of maternal smoking on birth weight using `WLS` is a decrease in infant birth weight of 227.56 grams on average compared to infants born to mothers who do not smoke.

Index of comments

6.1 where is the saturated model equivalent?

7.1 you made a slight mistake somewhere as your estimate is off. See solution key.