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

```r
# 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.

lm_robust - include all variables all in (age, education are numeric). The way to get cond't avg tx effect, have to have all factors - here add them linearly. If you add age, education as factors creates a number of bins. How do we control these as observables, the easiest way it to include them linearly.

```
# 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). *see TIA Table #ydiff = delta for x, w_ATE # of obs for rows, w_ATT = weights*

```
# mod3 <- lm_robust(data = data,
#                   birthwgt ~ tobacco + as.factor(alcohol) + as.factor()
#
#
#                   )
```

4 linear analogue - saturated model, exact matching estimator: calculating weights

-224

4.0.1 Answers

4.1 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.

4.1.1 Answers

4.2 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, glm is a logit

4.2.1 Answers