EDS_241_Assignment_3

Marie Rivers

02/20/2022

In this assignment, the goal is to estimate the causal effect of maternal smoking during pregnancy on infant birth weight using the ignorability assumptions. The data are taken from the National Natality Detail Files and include a random sample of all births in Pennsylvania during 1989-1991. Each observation is a mother-infant pair.

Question 1: Application of estimateors based on treatment ignorability

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 mother consumed alcohol during pregnancy
- first = 1 if first child
- diabete = 1 if mother diabetic
- anemia = 1 if mother anemic

Read Data

```
data <- read_csv(here("data", "SMOKING_EDS241.csv"))
```

Question 1a

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 hypothesis.

```r
mean_smoking_birthwt <- data %>%
  filter(tobacco == 1) %>%
  summarise(mean(birthwgt)) %>%
  as.numeric()
mean_smoking_birthwt
```

```
## [1] 3185.747
```

```r
mean_nonsmoking_birthwt <- data %>%
  filter(tobacco == 0) %>%
  summarise(mean(birthwgt)) %>%
  as.numeric()
mean_nonsmoking_birthwt
```

```
## [1] 3430.286
```

```r
unadj_mean_dif <- round(mean_nonsmoking_birthwt - mean_smoking_birthwt, 2)
unadj_mean_dif
```

```
## [1] 244.54
```

```r
# another method of showing the same thing with the coefficient of a linear regression
model_1a <- lm_robust(formula = birthwgt ~ tobacco, data = data)
huxreg("infant birth weight" = model_1a)
```

	infant birth weight
(Intercept)	3430.286 ***
	(1.781)
tobacco	-244.539 ***
	(4.150)
N	94173
R2	0.037

*** p < 0.001; ** p < 0.01; * p < 0.05.

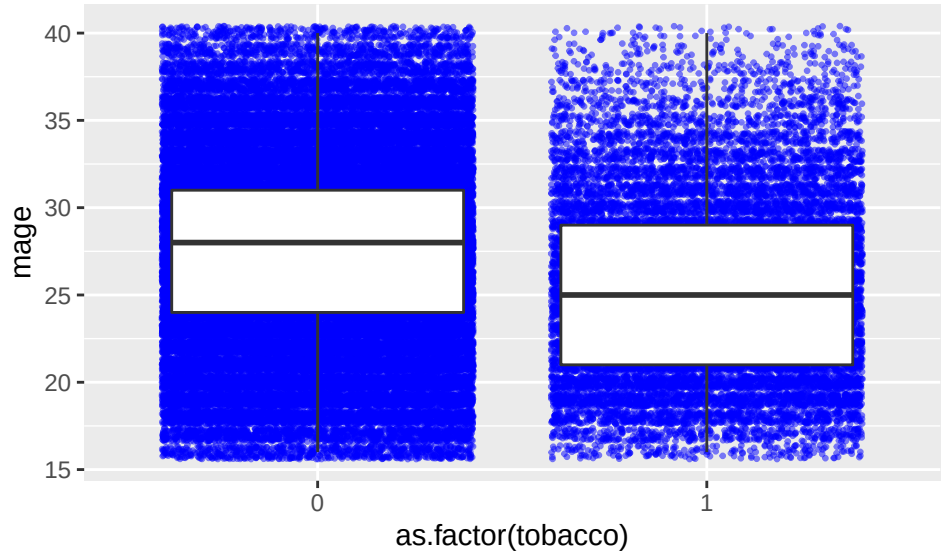
```r
model_1a_edu <- lm_robust(formula = meduc ~ tobacco, data = data)
model_1a_first <- lm_robust(formula = first ~ tobacco, data = data)
model_1a_age <- lm_robust(formula = mage ~ tobacco, data = data)

huxreg("mother's education level" = model_1a_edu, "first child" = model_1a_first, "mother's age" = model
```

	mother's education level	first child	mother's age
(Intercept)	13.239 ***	0.436 ***	27.453 ***
	(0.008)	(0.002)	(0.019)
tobacco	-1.318 ***	-0.072 ***	-1.915 ***
	(0.014)	(0.004)	(0.043)
N	94173	94173	94173
R2	0.061	0.003	0.020

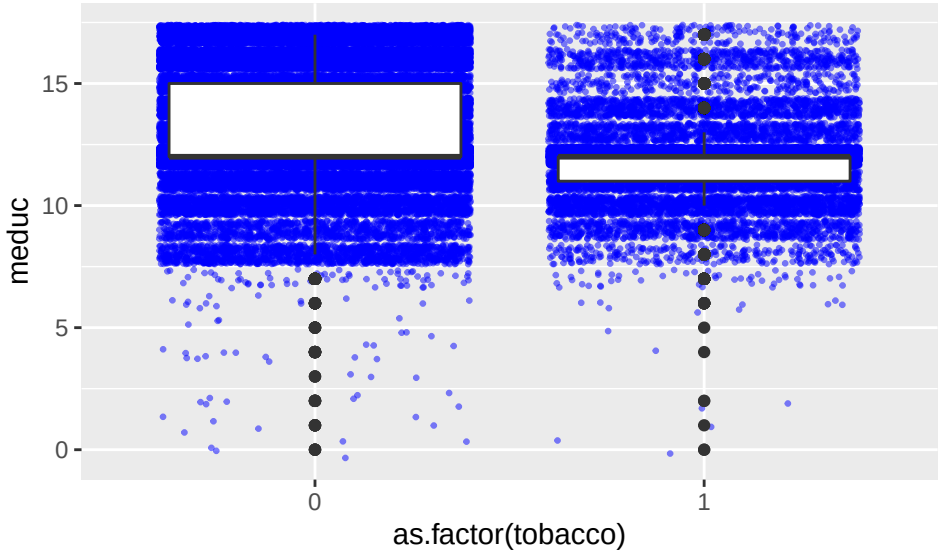
*** p < 0.001; ** p < 0.01; * p < 0.05.

The unadjusted mean difference in birth weight of infants with smoking and non-smoking mothers is 244.54 grams. Since this coefficient is negative, infant birth weight is, on average, lower for smoking mothers than for non-smoking mothers. This corresponds to the average treatment effect of maternal smoking during pregnancy on infant birth weight under the assumption that smoking status is randomly assigned and all else about the mothers' health and life characteristics (ie anemia, diabete, mblack, age...) is equal. In reality, there are variations between the smoking and non-smoking groups. For example, the mean difference in education and age for smoking and non-smoking mothers is statistically different from zero. The group of mothers who smoked, on average, had fewer years of education than the group of mothers who did not smoke. Additionally, the smoking mothers, on average, were also younger than the non-smoking mothers

```
# just to visualize the concepts discussed above
ggplot(data = data, aes(x = as.factor(tobacco), y = mage)) +
  geom_jitter(size = 0.5, color = "blue", alpha = 0.5) +
  geom_boxplot()
```



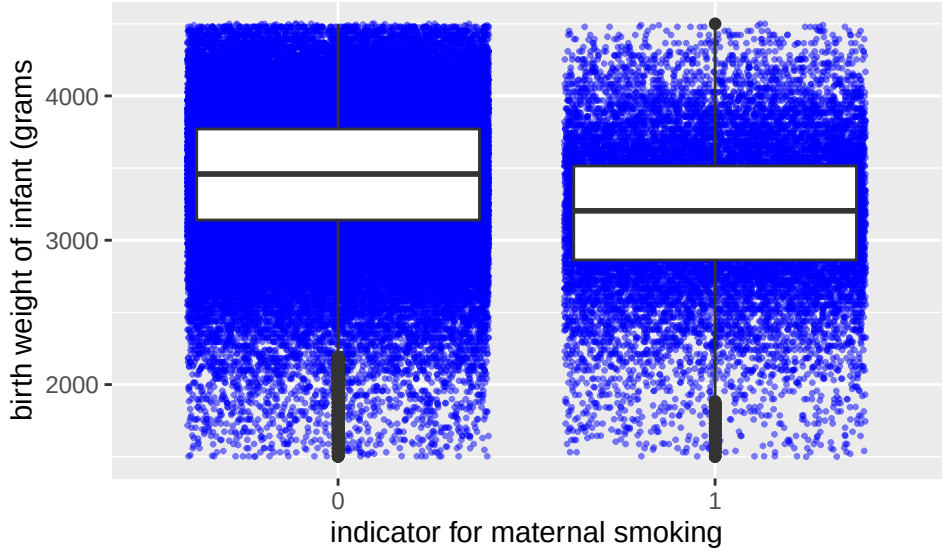

```
# just to visualize the concepts discussed above
ggplot(data = data, aes(x = as.factor(tobacco), y = meduc)) +
  geom_jitter(size = 0.5, color = "blue", alpha = 0.5) +
  geom_boxplot()
```



Question 1b

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.

```
# a visualization of the relationship
ggplot(data = data, aes(x = as.factor(tobacco), y = birthwgt)) +
  geom_jitter(size = 0.5, color = "blue", alpha = 0.5) +
  geom_boxplot() +
  labs(x = "indicator for maternal smoking", y = "birth weight of infant (grams")
```



```
model_1b <- lm_robust(formula = birthwgt ~ tobacco + mage + meduc + anemia + diabete + alcohol + mblack
huxreg("infant birth weight" = model_1b)
```

	infant birth weight
(Intercept)	3362.258 ***
	(12.076)
tobacco	-228.073 ***
	(4.277)
mage	-0.694
	(0.368)
meduc	11.688 ***
	(0.862)
anemia	-4.796
	(17.874)
diabete	73.228 ***
	(13.235)
alcohol	-77.350 ***
	(14.039)
mblack	-240.030 ***
	(5.348)
first	-96.944 ***
	(3.488)
N	94173
R2	0.072

*** p < 0.001; ** p < 0.01; * p < 0.05.

```
tobacco_coef_1b <- abs(model_1b$coefficients[2])
tobacco_se <- round(model_1b[[2]][2], 2)
```

Holding all other covariates equal, an infant born to a mother who smoked during pregnancy will weigh, on average, 228.07 grams less than an infant born to a mother who did not smoke. The robust standard error for this estimated coefficient is 4.28. While this model controls for a number of corvariates, it does not include covariates such as mother's drug use and mother's size.

Question 1c

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 2x2x2x2 = 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).

age_indicator edu_indicator mblack alcohol

```
data_1c <- data %>%
  mutate(age_indicator = if_else(mage >= 34, true = 1, false = 0)) %>%
  mutate(edu_indicator = if_else(meduc >= 16, true = 1, false = 0)) %>%
  mutate(g = paste0(age_indicator, edu_indicator, mblack, alcohol)) %>%
  dplyr::select(tobacco, age_indicator, edu_indicator, mblack, alcohol, birthwgt, g)
```

```r
# from TIA.Rmd
TIA_table <- data_1c %>%
  group_by(g,tobacco)%>%
  summarise(n_obs = n(),
            birthwgt_mean= mean(birthwgt, na.rm = T))%>% #Calculate number of observations and birthwgt
  gather(variables, values, n_obs:birthwgt_mean)%>% #Reshape data
  mutate(variables = paste0(variables,"_",tobacco, sep=""))%>% #Combine the treatment and variables for
  pivot_wider(id_cols = g, names_from = variables,values_from = values)%>% #Reshape data by treatment a
  ungroup()%>%  #Ungroup from X values
  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)),
         w_ATT = n_obs_1/sum(n_obs_1))%>% #calculate weights
  mutate_if(is.numeric, round, 2) #Round data


huxtable(TIA_table)
```

g	n_obs_0	n_obs_1	birthwgt_mean_0	birthwgt_mean_1	birthwgt_diff	w_ATE	w_ATT
0000	4.43e+04	1.34e+04	3.45e+03	3.22e+03	-225	0.61	0.74
0001	214	448	3.45e+03	3.12e+03	-326	0.01	0.02
0010	7.01e+03	1.98e+03	3.2e+03	3.01e+03	-190	0.1	0.11
0011	71	226	3.12e+03	2.82e+03	-303	0	0.01
0100	1.34e+04	535	3.48e+03	3.27e+03	-209	0.15	0.03
0101	130	29	3.51e+03	3.41e+03	-97.7	0	0
0110	625	61	3.32e+03	3.16e+03	-160	0.01	0
0111	4	10	2.98e+03	3.1e+03	114	0	0
1000	5.12e+03	976	3.47e+03	3.17e+03	-296	0.06	0.05
1001	56	45	3.36e+03	3.1e+03	-261	0	0
1010	396	135	3.19e+03	2.99e+03	-190	0.01	0.01
1011	7	26	2.74e+03	2.85e+03	107	0	0
1100	4.49e+03	201	3.49e+03	3.25e+03	-238	0.05	0.01
1101	57	17	3.53e+03	3.04e+03	-497	0	0
1110	147	19	3.33e+03	2.85e+03	-476	0	0
1111	1	1	3.46e+03	2.84e+03	-624	0	0

```r
# 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
```

The estimated average treatment effect of smoking on birth weight using the exact matching estimator is -224.26 grams. Since this value is negative, smoking results in decreased birth weight. The magnitude of the effect using the exact matching method is smaller than the effect when estimated using linear regression and controlling for other covariates (part 1b). The magnitude of the effect is largest when using uadjusted mean difference or linear regression without controling for other covariates (part 1a).

Question 1d

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_1d <- data %>%
  mutate(mage_squared = mage^2)
```

```
model_1d <- glm(tobacco ~ mage + mage_squared + meduc + mblack + alcohol, family = binomial(), data = da
huxreg("tobacco" = model_1d)
```

	tobacco
(Intercept)	1.930 ***
	(0.192)
mage	0.078 ***
	(0.015)
mage_squared	-0.002 ***
	(0.000)
meduc	-0.322 ***
	(0.005)
mblack	-0.060 *
	(0.027)
alcohol	2.023 ***
	(0.060)
N	94173
logLik	-42412.664
AIC	84837.329

*** p < 0.001; ** p < 0.01; * p < 0.05.

```
# EPS = estimated propensity score
EPS <- predict(model_1d, type = "response")

ps_weight <- (data_1d$tobacco / EPS) + ((1 - data_1d$tobacco) / (1 - EPS))
```

The `ps_weight` dataframe includes propensity scores for each observation.

Question 1e

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

```
model_wls <- lm(birthwgt ~ tobacco + mage + mage_squared + meduc + mblack + alcohol, weights = ps_weigh

huxreg("infant birth weight" = model_wls)
```

	infant birth weight
(Intercept)	2971.444 ***
	(36.122)
tobacco	-220.233 ***
	(3.223)
mage	27.627 ***
	(2.693)
mage_squared	-0.478 ***
	(0.049)
meduc	7.472 ***
	(0.849)
mblack	-220.990 ***
	(4.994)
alcohol	-71.914 ***
	(13.709)
N	94173
R2	0.074
logLik	-728569.509
AIC	1457155.018

*** p < 0.001; ** p < 0.01; * p < 0.05.

```
tobacco_coef_1e <- round(model_wls$coefficients[2], 2)
```

The estimated effect of maternal smoking on birth weight using a propensity score weighted regression (WLS) is -220.23 grams. As with the models above, the model estimates that smoking results in decreased birth weight. This method results in the smallest magnitude effect of maternal smoking on birth weight.

Note: 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.