

EDS 241 Assignment 3

Mia Forsline

2/20/2022

Set Up

Introduction

Goal: to estimate the causal effect of maternal smoking during pregnancy on infant birth weight using the treatment ignorability assumptions

- data come from the [Child & Family Data Archive's National Natality Detail Files](#)
- the data we will be using is a random sample of all births in Pennsylvania during 1989 - 1991
- each observation is a mother-infant pair

Variables

- Outcome variable: `birthwgt` = birth weight of infant in grams
- Treatment variable: `tobacco` = indicator for maternal smoking
- Control variables:
 - `mage` = mother's age
 - `meduc` = mother's education
 - `mblack` = 1 if the mother is Black
 - `alcohol` = 1 if the mother consumed alcohol during pregnancy
 - `first` = 1 if this is the mother's first child
 - `diabete` = 1 if the mother is diabetic
 - `anemia` = 1 if the mother is anemic

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

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

data <- data %>%
  mutate(mage = as.numeric(mage),
         meduc = as.numeric(meduc),
         mblack = as.factor(mblack),
         alcohol = as.factor(alcohol),
         first = as.factor(first),
         diabete = as.factor(diabete),
         anemia = as.factor(anemia)
  )
```

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

```
#Calculate the unadjusted mean difference by hand
data_smoke <- subset(data, tobacco == 1)
smoke_mean <- round(mean(data_smoke$birthwgt), digits = 2)

data_no_smoke <- subset(data, tobacco == 0)
no_smoke_mean <- round(mean(data_no_smoke$birthwgt), digits = 2)

diff <- no_smoke_mean - smoke_mean

#Calculate the unadjusted mean difference using a linear regression

mdl <- lm_robust(birthwgt ~ tobacco, data = data)
```

On average, mothers who smoked during pregnancy gave birth to babies that weighed 3185.75g. On average, mothers who did not smoke during pregnancy gave birth to babies that weighed 3430.29g. Thus, the unadjusted mean difference in infant birth weights between smoking and nonsmoking mothers is approximately 244.54g.

(a) 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

```
mdl2 <- lm_robust(mage ~ tobacco, data = data)
mage_diff <- round(mdl2$coefficients[[2]], digits = 2) * -1
```

Table 1 shows empirical evidence against the assumption that maternal smoking is randomly assigned among mothers who gave birth in Pennsylvania from 1989 - 1991.

Table 1: Maternal smoking predicts mothers' ages

	<i>Dependent variable:</i>
	Mother's Age (Years)
Maternal Smoking	-1.915*** (0.043)
Observations	94,173
R ²	0.020
<i>Note:</i>	*p<0.1; **p<0.05; ***p<0.01

This unadjusted mean value is valid if we assume that maternal smoking is randomly assigned among all mothers who gave birth in Pennsylvania during 1989 - 1991. However, that is an unrealistic assumption because this was not a randomized controlled treatment experiment and maternal smoking is not a treatment that can be randomly assigned. In other words, D is not independent of the potential outcomes $Y(1)$ and $Y(0)$.

Therefore, it is more likely that there are significant differences in the control variables (such as mother's age) depending on if mothers smoked during pregnancy or not. For example, on average, mothers who smoked during pregnancy are 1.91 years younger than mothers who did not smoke during pregnancy.

(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

```
mdl4 <- lm_robust(birthwgt ~ tobacco + mage + meduc + mblack + alcohol + first + diabete + anemia, data)
```

Table 1 shows the estimated coefficients and standard error of maternal smoking during pregnancy on infant birth weight (g).

Table 2: Maternal Smoking During Pregnancy Decreases Infant Birth Weights

	<i>Dependent variable:</i>
	Infant Birth Weight (g)
Maternal Smoking	-228.073*** (4.277)
Observations	94,173
R ²	0.072
<i>Note:</i>	*p<0.1; **p<0.05; ***p<0.01

(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 $2 \times 2 \times 2 \times 2 = 16$ 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).

```
#Create the covariate indicators
data_matching <- data %>%
  mutate(mage_ind = case_when(
    mage >= 34 ~ 1,
    mage < 34 ~ 0),
    meduc_ind = case_when(
    meduc >= 16 ~ 1,
    meduc < 16 ~ 0),
    g = paste0(mage_ind, meduc_ind, mblack, alcohol) #create all combinations of the 4 covariates in a
  )
```

Exact matching estimator

- Y = birthwgt
- X = g
- D = tobacco

```
TIA_table <- data_matching %>%
  group_by(g, tobacco)%>%
  summarise(n_obs = n(),
    Y_mean = mean(birthwgt, na.rm = T)) %>%
  gather(variables, values, n_obs:Y_mean) %>%
  mutate(variables = paste0(variables,"_",tobacco, sep="")) %>%
  pivot_wider(id_cols = g, names_from = variables, values_from = values) %>%
  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)
```

```
##
## =====
##      g    n_obs_0 n_obs_1 Y_mean_0 Y_mean_1 Y_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
## -----
```

Average treatment effect (ATE) estimation

```
ATE <- sum((TIA_table$w_ATE)*(TIA_table$Y_diff))
```

The exact matching estimator estimates an average treatment effect (ATE) of -224.2583.

Linear regression analogue

```
mdl6 <- lm_robust(birthwgt ~ tobacco +
  mage_ind + meduc_ind + mblack + alcohol +
  mage_ind:meduc_ind +
  mage_ind:mblack +
  mage_ind:alcohol +
  meduc_ind:mblack +
  meduc_ind:alcohol +
  mblack:alcohol +
  mage_ind:meduc_ind:mblack +
  mage_ind:meduc_ind:alcohol +
  meduc_ind:mblack:alcohol +
  tobacco:mage_ind:meduc_ind:mblack:alcohol, data = data_matching)

mdl7 <- lm(birthwgt ~ tobacco + as.factor(g), data = data_matching)

se_models <- starprep(mdl6, stat = c("std.error"), se_type = "HC2", alpha = 0.05)

stargazer(mdl7, se = se_models, type="text")
```

```
##
## =====
##               Dependent variable:
##               -----
##               birthwgt
## -----
## tobacco               -226.245
```

```

##
##
## as.factor(g)0001          -63.124
##
##
## as.factor(g)0010          -241.839
##
##
## as.factor(g)0011          -384.006
##
##
## as.factor(g)0100           37.809
##
##
## as.factor(g)0101           88.511
##
##
## as.factor(g)0110          -120.775
##
##
## as.factor(g)0111          -219.198
##
##
## as.factor(g)1000           10.359
##
##
## as.factor(g)1001          -102.853
##
##
## as.factor(g)1010          -251.686
##
##
## as.factor(g)1011          -443.862
##
##
## as.factor(g)1100           40.825
##
##
## as.factor(g)1101           26.737
##
##
## as.factor(g)1110          -146.188
##
##
## as.factor(g)1111          -185.751
##
##
## Constant                   3,445.873
##
## -----
## Observations                94,173
## R2                          0.063
## Adjusted R2                 0.063

```

```
## Residual Std. Error      487.101 (df = 94156)
## F Statistic              393.603*** (df = 16; 94156)
## =====
## Note:                    *p<0.1; **p<0.05; ***p<0.01
```

(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.

```
#Create dataframe that includes mother's age squared and necessary indicators
data_p <- data_matching %>%
  mutate(tobacco = as.integer(tobacco),
         mage_ind = as.integer(mage_ind),
         meduc_ind = as.integer(meduc_ind),
         mage_sq = as.integer(mage ** 2))

#create propensity score model
ps_model <- glm(tobacco ~ mage_ind + mage_sq + meduc_ind + mblack + alcohol,
               family = binomial(),
               data = data_p)
summary(ps_model)
```

```
##
## Call:
## glm(formula = tobacco ~ mage_ind + mage_sq + meduc_ind + mblack +
##      alcohol, family = binomial(), data = data_p)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.7121  -0.7330  -0.6362  -0.2762   2.7172
##
## Coefficients:
##              Estimate Std. Error z value      Pr(>|z|)
## (Intercept) -0.58645121  0.02833050 -20.700 < 0.0000000000000002 ***
## mage_ind      0.23829478  0.03996939   5.962  0.000000000249 ***
## mage_sq     -0.00094499  0.00004114 -22.972 < 0.0000000000000002 ***
## meduc_ind    -1.71521635  0.03683092 -46.570 < 0.0000000000000002 ***
## mblack1     -0.09109851  0.02595057  -3.510    0.000447 ***
## alcohol1      2.06253835  0.06054697  34.065 < 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: 86027  on 94167  degrees of freedom
## AIC: 86039
##
## Number of Fisher Scoring iterations: 5
```

```
#use estimated logistic equation to create EPS (estimated propensity score)
eps <- predict(ps_model, type = "response")

#create propensity score weights
ps_weight <- (data_p$tobacco/eps) + ((1 - data_p$tobacco) / (1 - eps)) #PS_WGT = p-score weight
```

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

```
#Exact answer changes when we apply the weights compared to the first question - what are those changes

# WLS USING EPS WEIGHTS
wls1 <- lm(formula = birthwgt ~ tobacco, data=data_p, weights=ps_weight)

se_models = starprep(wls1, stat = c("std.error"), se_type = "HC2", alpha = 0.05)

stargazer(wls1, se = se_models, type="text", omit = "Constant")
```

```
##
## =====
##                               Dependent variable:
##                               -----
##                               birthwgt
## -----
## tobacco                      -227.062***
##                               (5.403)
## -----
## Observations                  94,173
## R2                           0.049
## Adjusted R2                   0.049
## Residual Std. Error          709.623 (df = 94171)
## F Statistic                   4,839.785*** (df = 1; 94171)
## =====
## Note:                         *p<0.1; **p<0.05; ***p<0.01
```