EDS 241: Assignment 3

Paloma Cartwright

2022-02-15

The goal is 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 extract "SMOKING_EDS241.csv"' is a random sample of all births in Pennsylvania during 1989-1991. Each observation is a mother-infant pair. The key variables are:

The outcome and treatment variables: `birthwgt` = birth weight of infant in grams `tobacco` = indicator for maternal smoking

The control variables: `mage` (mother's age) `meduc` (mother's education) `mblack` (=1 if mother black) `alcohol` (=1 if consumed alcohol during pregnancy) `first` (=1 if first child), `diabete` (=1 if mother diabetic) `anemia` (=1 if mother anemic)

Clean and plot data

The following code loads the data.

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

Question 1

(a) 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 assumption.

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

mean_smoker <- round(mean(smoker$birthwgt), 3)
mean_nonsmoker <- round(mean(nonsmoker$birthwgt), 3)

diff <- mean_nonsmoker - mean_smoker
```

The unadjusted mean difference in birth weight in grams for children whose mothers did not smoke versus those that did is 244.539 grams.

The assumption is that smoking status is independent of y(0) and y(1) implying unconditional treatment ignorability.

OH NOTES if you have income on lhs, whether you are a smoker or not will not be statistically different. Maybe the hypothesis is that you're testing whether the treatment status is independent of y1 y0 What

would need to be true for smoker vs nonsmoker to be the treatment effect? We have treatment ignorability. We should be working toward conditional treatment ignorability.

the hypothesis/assumpiton smoking status is independent of y(1) y(0) - unconditional treatment ignorability. We would need smoking status to be randomly assigned to mothers unconditionally.

Run a model to provide evidence run income on smoking status, education on smoking status, age on smoking status - provide someeee evidence that smoking status is correlated with these variables. You can do this with the mean differences between them but you want to know if they are statistically different. running the regression gives you a test statistic already. if it were true that smoking was randomly assigned then the regressions of other vars on smoking would not be statistically significant.

```
model1 <- lm_robust(mblack ~ tobacco, data = data)
summary(model1)
```

```
##
## Call:
## lm_robust(formula = mblack ~ tobacco, data = data)
##
## Standard error type:  HC2
##
## Coefficients:
##             Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper    DF
## (Intercept)  0.10863   0.001129  96.251 0.000e+00  0.10642  0.11084 94171
## tobacco      0.02678   0.002779   9.637 5.686e-22  0.02134  0.03223 94171
##
## Multiple R-squared:  0.001107 ,  Adjusted R-squared:  0.001096
## F-statistic: 92.88 on 1 and 94171 DF,  p-value: < 2.2e-16
```

The p-value for the impact of whether the mother is black on smoking during pregnancy is $5.6855795 \times 10^{-22}$. This is less than 0.5 meaning that it is statistically significant and the race of the mother is correlated with whether they used tobacco during pregnancy. This would contradict the assumption that smoking status is independent and that there is unconditional treatment ignorability.

Question 2

(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.

```
model2 <- lm_robust(birthwgt ~ tobacco + mage + meduc + mblack + alcohol + first + diabete + anemia,
                    data = data)
summary(model2)
```

```
##
## Call:
## lm_robust(formula = birthwgt ~ tobacco + mage + meduc + mblack +
##     alcohol + first + diabete + anemia, data = data)
##
## Standard error type:  HC2
##
## Coefficients:
##             Estimate Std. Error  t value   Pr(>|t|) CI Lower   CI Upper    DF
## (Intercept) 3362.258    12.0765 278.4133  0.000e+00 3338.588 3385.92805 94164
## tobacco     -228.073     4.2768 -53.3282  0.000e+00 -236.456 -219.69063 94164
```

```
## mage          -0.694     0.3682  -1.8849  5.944e-02   -1.416    0.02764 94164
## meduc         11.688     0.8618  13.5630  7.262e-42    9.999   13.37742 94164
## mblack      -240.030     5.3478 -44.8842  0.000e+00 -250.512 -229.54873 94164
## alcohol      -77.350    14.0392  -5.5096  3.607e-08 -104.866  -49.83312 94164
## first        -96.944     3.4880 -27.7934 2.528e-169 -103.781  -90.10763 94164
## diabete       73.228    13.2355   5.5327  3.162e-08   47.286   99.16895 94164
## anemia        -4.796    17.8739  -0.2683  7.884e-01  -39.829   30.23630 94164
##
## Multiple R-squared:  0.0717 ,    Adjusted R-squared:  0.07162
## F-statistic: 877.6 on 8 and 94164 DF,  p-value: < 2.2e-16
```

The coefficient for the average impact of tobacco on birth weight in grams is -228.0730765 and the standard error is 4.2767834.

(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).

```
data_matching <- 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),
    g = paste0(mage, meduc, mblack, alcohol)
    )
```

```
TIA_table <- data_matching %>%
  group_by(g, tobacco)%>%
  summarise(n_obs = n(),
            bwgt_mean= mean(birthwgt, na.rm = T)) %>% #Calculate number of observations and Y mean by X
  gather(variables, values, n_obs:bwgt_mean) %>% #Reshape data
  mutate(variables = paste0(variables, "_", tobacco, sep=""))%>% #Combine the treatment and variables f
  pivot_wider(id_cols = g, names_from = variables, values_from = values) %>% #Reshape data by treatment
  ungroup() %>%  #Ungroup from X values
  mutate(bwgt_diff = bwgt_mean_1 - bwgt_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 bwgt_mean_0 bwgt_mean_1 bwgt_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$w_ATE)*(TIA_table$bwgt_diff))
ATE
```

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

```
ATT=sum((TIA_table$w_ATT)*(TIA_table$bwgt_diff))
ATT
```

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

```
model3 <- lm_robust(birthwgt ~ tobacco +
                      mage + meduc + mblack + alcohol +
                      mage:meduc + mage:mblack + mage:alcohol +
                      meduc:mblack + meduc:alcohol + mblack:alcohol +
                      mage:meduc:mblack + mage:meduc:alcohol + meduc:mblack:alcohol +
                      mage:meduc:mblack:alcohol, data = data_matching)

summary(model3)
```

```
##
## Call:
## lm_robust(formula = birthwgt ~ tobacco + mage + meduc + mblack +
##     alcohol + mage:meduc + mage:mblack + mage:alcohol + meduc:mblack +
##     meduc:alcohol + mblack:alcohol + mage:meduc:mblack + mage:meduc:alcohol +
##     meduc:mblack:alcohol + mage:meduc:mblack:alcohol, data = data_matching)
##
## Standard error type:  HC2
##
## Coefficients:
##                            Estimate Std. Error   t value  Pr(>|t|) CI Lower
## (Intercept)                3445.873      2.232 1543.7974 0.000e+00 3441.498
## tobacco                    -226.245      4.220  -53.6114 0.000e+00 -234.516
## mage                         10.359      6.804    1.5225 1.279e-01   -2.977
## meduc                        37.809      4.535    8.3377 7.675e-17   28.921
## mblack1                    -241.839      5.733  -42.1852 0.000e+00 -253.075
## alcohol1                    -63.127     20.028   -3.1519 1.623e-03 -102.381
## mage:meduc                   -7.343     10.591   -0.6933 4.881e-01  -28.102
## mage:mblack1                -20.203     24.782   -0.8152 4.149e-01  -68.775
```

```
## mage:alcohol1                -50.068     43.319   -1.1558 2.478e-01 -134.973
## meduc:mblack1                 83.255     20.110    4.1399 3.478e-05   43.839
## meduc:alcohol1               113.829     43.439    2.6205 8.783e-03   28.690
## mblack1:alcohol1             -79.035     34.047   -2.3214 2.027e-02 -145.766
## mage:meduc:mblack1            -8.226     50.176   -0.1639 8.698e-01 -106.569
## mage:meduc:alcohol1          -14.721     80.388   -0.1831 8.547e-01 -172.281
## meduc:mblack1:alcohol1       -70.090    138.607   -0.5057 6.131e-01 -341.758
## mage:meduc:mblack1:alcohol1  123.650    249.369    0.4959 6.200e-01 -365.110
##                             CI Upper    DF
## (Intercept)                 3450.25 94157
## tobacco                     -217.97 94157
## mage                          23.69 94157
## meduc                         46.70 94157
## mblack1                     -230.60 94157
## alcohol1                     -23.87 94157
## mage:meduc                    13.42 94157
## mage:mblack1                  28.37 94157
## mage:alcohol1                 34.84 94157
## meduc:mblack1                122.67 94157
## meduc:alcohol1               198.97 94157
## mblack1:alcohol1             -12.30 94157
## mage:meduc:mblack1            90.12 94157
## mage:meduc:alcohol1          142.84 94157
## meduc:mblack1:alcohol1       201.58 94157
## mage:meduc:mblack1:alcohol1  612.41 94157
##
## Multiple R-squared:  0.06269 ,   Adjusted R-squared:  0.06254
## F-statistic:   400 on 15 and 94157 DF,  p-value: < 2.2e-16
```

(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.

```
# this is the model of the propensity score
ps_model <- glm(tobacco ~ mage + mage_sq + meduc + mblack + alcohol, family = binomial(), data = data_ma
summary(ps_model)
```

```
##
## Call:
## glm(formula = tobacco ~ mage + mage_sq + meduc + mblack + alcohol,
##     family = binomial(), data = data_matching)
##
## 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) -5.865e-01  2.833e-02 -20.700  < 2e-16 ***
## mage         2.383e-01  3.997e-02   5.962 2.49e-09 ***
## mage_sq     -9.450e-04  4.114e-05 -22.972  < 2e-16 ***
## meduc       -1.715e+00  3.683e-02 -46.570  < 2e-16 ***
## mblack1     -9.110e-02  2.595e-02  -3.510 0.000447 ***
## alcohol1     2.063e+00  6.055e-02  34.065  < 2e-16 ***
## ---
```

```
## 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
```

```
EPS <- predict(ps_model, type = "response")
PS_WGT <- (data_matching$tobacco/EPS) + ((1-data_matching$tobacco)/(1-EPS))
```

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

```
# WLS USING EPS WEIGHTS
wls1 <- lm(birthwgt ~ tobacco, data_matching, weights=PS_WGT)
summary(wls1)
```

```
##
## Call:
## lm(formula = birthwgt ~ tobacco, data = data_matching, weights = PS_WGT)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -9090.8  -370.4    34.9   412.0  7064.2
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3426.034      2.313 1481.47   <2e-16 ***
## tobacco     -227.062      3.264  -69.57   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 709.6 on 94171 degrees of freedom
## Multiple R-squared:  0.04888,    Adjusted R-squared:  0.04887
## F-statistic:  4840 on 1 and 94171 DF,  p-value: < 2.2e-16
```

```
se_model = starprep(wls1, stat = c("std.error"), se_type = "HC2", alpha = 0.05)
stargazer(wls1, se = se_model, type="text")
```

```
##
## ================================================
##                         Dependent variable:
##                     ---------------------------
##                               birthwgt
## ------------------------------------------------
## tobacco                     -227.062***
##                               (5.403)
##
## Constant                    3,426.034***
##                               (1.808)
##
## ------------------------------------------------
## 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
```