EDS241: Assignment 3

Charles Hendrickson

02/20/2022

Load the SMOKING_EDS241.csv data

```r
# load data
smoking_data <- read.csv("SMOKING_EDS241.csv")
```

(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 hypothesis.

The unadjusted mean difference in birth weight of infants with smoking and nonsmoking mothers is -244.5 grams.

Under the "treatment ignorability" assumption, this corresponds to the average treatment effect of maternal smoking during pregnancy on infant birth weight

When regressing the mother's education level (meduc) on the indicator for maternal smoking (tobacco), the mean difference in the education level of smoking and non-smoking mothers is -1.318 units, which is statistically significant because the p-value ($< 2.2e-16$) is much lower than the 5% significance level. Therefore, we cannot interpret the unadjusted mean difference as causal because maternal smoking is not randomly assigned due to a mother's education level being statistically different from zero on the maternal smoking treatment variable.

```r
# Regress birth weight of infant in grams (birthwgt) on the indicator for maternal smoking (tobacco)
summary(lm_robust(birthwgt ~ tobacco, data = smoking_data))
```

```
##
## Call:
## lm_robust(formula = birthwgt ~ tobacco, data = smoking_data)
##
## Standard error type:  HC2
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper    DF
## (Intercept)   3430.3      1.781 1926.11        0   3426.8   3433.8 94171
## tobacco       -244.5      4.150  -58.93        0   -252.7   -236.4 94171
##
## Multiple R-squared:  0.03676 ,   Adjusted R-squared:  0.03675
## F-statistic:  3473 on 1 and 94171 DF,  p-value: < 2.2e-16
```

```r
# Regress the mother's education level (meduc) on the indicator for maternal smoking (tobacco)
summary(lm_robust(meduc ~ tobacco, data = smoking_data))
```

```
##
## Call:
## lm_robust(formula = meduc ~ tobacco, data = smoking_data)
##
```

```
## Standard error type:  HC2
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper    DF
## (Intercept)   13.239    0.00776 1706.12        0   13.224   13.255 94171
## tobacco       -1.318    0.01425  -92.54        0   -1.346   -1.291 94171
##
## Multiple R-squared:  0.06057 ,   Adjusted R-squared:  0.06056
## F-statistic:  8563 on 1 and 94171 DF,  p-value: < 2.2e-16
```

(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.

The estimated coefficient on tobacco is -228.073 grams and the standard error is 4.2768 grams.

```
summary(lm_robust(birthwgt ~ ., data = smoking_data))
```

```
##
## Call:
## lm_robust(formula = birthwgt ~ ., data = smoking_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
## anemia        -4.796    17.8739  -0.2683  7.884e-01  -39.829   30.23630 94164
## diabete       73.228    13.2355   5.5327  3.162e-08   47.286   99.16895 94164
## tobacco     -228.073     4.2768 -53.3282  0.000e+00 -236.456 -219.69063 94164
## alcohol      -77.350    14.0392  -5.5096  3.607e-08 -104.866  -49.83312 94164
## mblack      -240.030     5.3478 -44.8842  0.000e+00 -250.512 -229.54873 94164
## first        -96.944     3.4880 -27.7934 2.528e-169 -103.781  -90.10763 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
##
## Multiple R-squared:  0.0717 ,    Adjusted R-squared:  0.07162
## F-statistic: 877.6 on 8 and 94164 DF,  p-value: < 2.2e-16
```

(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).

The estimated average treatment effect of smoking on birthweight using the exact matching estimator and its linear regression analogue is -224.2583 grams.

```
# Create a new variable called "D_mage" by using "ifelse", which takes a condition, followed by what th

smoking_data_dummy <- smoking_data %>%
  mutate(D_mage = ifelse(mage >= 34, 1, 0)) %>%   # Conditional on Mother's Age
  mutate(D_meduc = ifelse(meduc>=16, 1, 0)) %>%   # Conditional on Mother's Education
  mutate(D_mblack = ifelse(mblack == 1, 1, 0)) %>%
  mutate(D_alcohol = ifelse(alcohol == 1, 1, 0)) %>%
```

```r
  mutate(g = paste0(D_mage,D_meduc,D_mblack,D_alcohol))

summary(lm_robust(birthwgt ~ tobacco + factor(g), data = smoking_data_dummy))
```

```
##
## Call:
## lm_robust(formula = birthwgt ~ tobacco + factor(g), data = smoking_data_dummy)
##
## Standard error type:  HC2
##
## Coefficients:
##                Estimate Std. Error   t value  Pr(>|t|) CI Lower CI Upper    DF
## (Intercept)    3445.87      2.232 1543.6015 0.000e+00 3441.498  3450.25 94156
## tobacco        -226.25      4.220  -53.6104 0.000e+00 -234.517  -217.97 94156
## factor(g)0001   -63.12     20.431   -3.0897 2.004e-03 -103.168   -23.08 94156
## factor(g)0010  -241.84      5.742  -42.1186 0.000e+00 -253.093  -230.58 94156
## factor(g)0011  -384.01     29.870  -12.8558 8.586e-38 -442.552  -325.46 94156
## factor(g)0100    37.81      4.535    8.3374 7.692e-17   28.921    46.70 94156
## factor(g)0101    88.51     38.413    2.3042 2.121e-02   13.222   163.80 94156
## factor(g)0110  -120.78     18.977   -6.3643 1.971e-10 -157.970   -83.58 94156
## factor(g)0111  -219.20    127.345   -1.7213 8.520e-02 -468.793    30.40 94156
## factor(g)1000    10.36      6.819    1.5192 1.287e-01   -3.006    23.72 94156
## factor(g)1001  -102.85     45.144   -2.2783 2.271e-02 -191.334   -14.37 94156
## factor(g)1010  -251.69     24.106  -10.4408 1.668e-25 -298.934  -204.44 94156
## factor(g)1011  -443.86     79.415   -5.5892 2.288e-08 -599.513  -288.21 94156
## factor(g)1100    40.82      7.404    5.5136 3.525e-08   26.312    55.34 94156
## factor(g)1101    26.74     55.254    0.4839 6.285e-01  -81.559   135.03 94156
## factor(g)1110  -146.19     38.555   -3.7917 1.497e-04 -221.755   -70.62 94156
## factor(g)1111  -185.75    198.895   -0.9339 3.504e-01 -575.582   204.08 94156
##
## Multiple R-squared:  0.06269 ,   Adjusted R-squared:  0.06253
## F-statistic:   375 on 16 and 94156 DF,  p-value: < 2.2e-16
```

```r
TIA_table <- smoking_data_dummy %>%
  group_by(g, tobacco)%>%
  summarise(n_obs = n(),
            Y_mean= mean(birthwgt, na.rm = T))%>% #Calculate number of observations and Y mean by X by
  gather(variables, values, n_obs:Y_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(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
## -------------------------------------------------------------
```

```
# MULTIVARIATE MATCHING ESTIMATES OF ATE AND ATT
ATE=sum((TIA_table$w_ATE)*(TIA_table$Y_diff))
ATE
```

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

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

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

(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 a 'mother's age squared' variable in the smoking_data_dummy
smoking_data_dummy <- smoking_data_dummy %>%
  mutate(mage_squared = mage*mage)
```

```
# Estimate the propensity score model and predict (EPS)
ps_model <- glm(formula = tobacco ~ mage + mage_squared + meduc + D_mblack + D_alcohol, family = binomia

# Table of the ps_model
summary(ps_model)
```

```
##
## Call:
## glm(formula = tobacco ~ mage + mage_squared + meduc + D_mblack +
##     D_alcohol, family = binomial(), data = smoking_data_dummy)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5482  -0.7182  -0.5461  -0.3214   2.6709
##
## Coefficients:
##               Estimate Std. Error z value          Pr(>|z|)
## (Intercept)   1.929611   0.191814  10.060           < 2e-16 ***
## mage          0.077636   0.014915   5.205 0.00000019355476 ***
## mage_squared -0.001941   0.000278  -6.983 0.00000000000288 ***
## meduc        -0.321597   0.005144 -62.520           < 2e-16 ***
```

```
## D_mblack     -0.059525   0.026506  -2.246           0.0247 *
## D_alcohol     2.022696   0.060358  33.511          < 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: 84825  on 94167  degrees of freedom
## AIC: 84837
##
## Number of Fisher Scoring iterations: 5
```

```r
# First 5 propensity scores for maternal smoking
EPS <- predict(ps_model, type = "response")
EPS_5 <- head(EPS, 5)
EPS_5
```

```
##          1          2          3          4          5
## 0.07517473 0.27763810 0.22293173 0.23423784 0.36164225
```

```r
PS_WGT <- (smoking_data_dummy$tobacco/EPS) + ((1-smoking_data_dummy$tobacco)/(1-EPS))
```

The first five estimated propensity scores for maternal smoking are the outputs from 'EPS_5 <- head(EPS, 5)'.

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

The estimated effect of maternal smoking on birth weight is -225.475 grams using the propensity score weighted regression (WLS).

```r
# propensity score weighted regression (WLS)
summary(lm(formula = birthwgt ~ tobacco, data = smoking_data_dummy, weights = PS_WGT))
```

```
##
## Call:
## lm(formula = birthwgt ~ tobacco, data = smoking_data_dummy, weights = PS_WGT)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -8124.3  -367.4    34.5   413.7  5931.1
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3425.994      2.288 1497.28   <2e-16 ***
## tobacco     -225.475      3.263  -69.11   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 703.6 on 94171 degrees of freedom
## Multiple R-squared:  0.04827,    Adjusted R-squared:  0.04826
## F-statistic:  4776 on 1 and 94171 DF,  p-value: < 2.2e-16
```