EDS241: Assignment template/example

Shale Hunter

02/17/2022

1 Load Data

```
smoke = read_csv(here("HW3/SMOKING_EDS241.csv"))
```

2 Questions

a) What is the unadjusted mean difference in birth weight of infants with smoking and non- smoking mothers? Under what hypothesis 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 non- smoking mothers is -244.5393875grams. This corresponds to the ATE under the assumption that mothers are assigned randomly to the smoking/non-smoking group. This is probably not a strong assumption because it is likely that there are external factors that effect a mother's likelihood to smoke during pregnancy, such as (only thinking about other variables in our dataset) whether or not this is the mother's first child (mothers might know less about the harmful effects of smoking when they are on their first child) or mother's education level (women with more education may be less likely to smoke because of a greater awareness of the harmful effects of smoking). Both these effects are shown to be significant in the models below, showing that smoking is not random across these other conditions.

```
summary(estimatr::lm_robust(data = smoke, tobacco ~ first))
```

```
##
## Call:
## estimatr::lm_robust(formula = tobacco ~ first, data = smoke)
##
## Standard error type:  HC2
##
## Coefficients:
##             Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper    DF
## (Intercept)   0.2120   0.001752  120.98 0.000e+00  0.20858  0.21545 94171
## first        -0.0456   0.002561  -17.81 8.202e-71 -0.05062 -0.04058 94171
##
## Multiple R-squared:  0.003261 ,  Adjusted R-squared:  0.00325
## F-statistic: 317.1 on 1 and 94171 DF,  p-value: < 2.2e-16
```

```
summary(estimatr::lm_robust(data = smoke, tobacco ~ meduc))
```

```
##
## Call:
```

```
## estimatr::lm_robust(formula = tobacco ~ meduc, data = smoke)
##
## Standard error type:  HC2
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper    DF
## (Intercept)  0.78933  0.0077083  102.40        0  0.77423  0.80444 94171
## meduc       -0.04594  0.0005421  -84.75        0 -0.04701 -0.04488 94171
##
## Multiple R-squared:  0.06057 ,   Adjusted R-squared:  0.06056
## F-statistic:  7183 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.

```
tbco_lm = summary(estimatr::lm_robust(data = smoke, birthwgt ~ tobacco))
tbco_lm
```

```
##
## Call:
## estimatr::lm_robust(formula = birthwgt ~ tobacco, data = smoke)
##
## 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
```

The model above predicts a coefficient of -244.5 (grams) for tobacco, with a standard error of 4.15 grams.

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).

```
# dummy variables
age = as.numeric(as.logical(smoke$mage >= 34))
edu = as.numeric(as.logical(smoke$meduc >= 16))
blk = smoke$mblack
alc = smoke$alcohol

# grouped variable
grp = paste0(age, edu, blk, alc)

smoke = smoke %>% mutate(grp = grp)

analogue_lm = estimatr::lm_robust(data = smoke, birthwgt ~ tobacco + factor(grp))
EMcompare = summary(analogue_lm)
```

2.0.1 Exact Matching

```
treatment_table <- smoke %>%
  group_by(grp,tobacco)%>%
  # Calculate number of observations and Y mean by X by treatment cells:
  summarise(n_obs  = n(),
            Y_mean = mean(birthwgt, na.rm = T))%>%
#old way to pivot_longer: gather(variables, values, n_obs:Y_mean)
  pivot_longer(names_to = "variables", values_to = "values", n_obs:Y_mean) %>%
  # Combine the treatment and variables for re-reshaping
  mutate(variables = paste0(variables, "_", tobacco))%>%
  # Reshape data by treatment and X cell
  pivot_wider(id_cols = grp, 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(treatment_table, type= "text", summary = FALSE, digits = 2)
```

```
##
## ============================================================
##    grp  n_obs_0 Y_mean_0 n_obs_1 Y_mean_1 Y_diff  w_ATE w_ATT
## ------------------------------------------------------------
## 1  0000  44274  3445.69   13443  3220.25  -225.44 0.61  0.74
## 2  0001   214   3450.28    448   3124.25  -326.03 0.01  0.02
## 3  0010  7007   3195.97   1980   3006.31  -189.66  0.1  0.11
## 4  0011   71    3120.07    226   2817.34  -302.73   0   0.01
## 5  0100  13425  3483.02    535   3273.94  -209.08 0.15  0.03
## 6  0101   130   3510.95    29    3413.21  -97.74    0     0
## 7  0110   625   3319.22    61    3159.05  -160.17 0.01    0
## 8  0111    4     2983.5    10     3097.7   114.2    0     0
## 9  1000  5115   3467.41    976   3171.42  -295.98 0.06  0.05
## 10 1001   56    3358.32    45    3097.73  -260.59   0     0
## 11 1010   396   3185.08    135   2994.67  -190.41 0.01  0.01
## 12 1011    7    2739.71    26    2846.38  106.67    0     0
## 13 1100  4492   3487.19    201   3249.45  -237.74 0.05  0.01
## 14 1101   57    3534.91    17    3037.47  -497.44   0     0
## 15 1110   147   3328.29    19    2852.16  -476.13   0     0
## 16 1111    1      3459      1      2835    -624     0     0
## ------------------------------------------------------------
```

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

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

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

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

```r
coefs = as.matrix(EMcompare$coefficients[,1] + EMcompare$coefficients[1,1])
lm_coefs = coefs[-1, ]


# This section made sure that lm() and lm_robust gave the same output:

# logue_lm = lm(data = smoke, birthwgt ~ tobacco + factor(grp))
# compare = summary(logue_lm)
# coefs = as.matrix(compare$coefficients[,1] + compare$coefficients[1,1])
# base2 = coefs[-1, ]
# verify2 = data_frame(base_coefs, base2)


# comparison
verify_df = data.frame(lm_coefs, treatment_table$Y_mean_1)
verify_df
```

```
##                 lm_coefs treatment_table.Y_mean_1
## tobacco         3219.628                  3220.25
## factor(grp)0001 3382.749                  3124.25
## factor(grp)0010 3204.034                  3006.31
## factor(grp)0011 3061.867                  2817.34
## factor(grp)0100 3483.682                  3273.94
## factor(grp)0101 3534.384                  3413.21
## factor(grp)0110 3325.098                  3159.05
## factor(grp)0111 3226.675                  3097.70
## factor(grp)1000 3456.232                  3171.42
## factor(grp)1001 3343.020                  3097.73
## factor(grp)1010 3194.187                  2994.67
## factor(grp)1011 3002.011                  2846.38
## factor(grp)1100 3486.698                  3249.45
## factor(grp)1101 3472.610                  3037.47
## factor(grp)1110 3299.685                  2852.16
## factor(grp)1111 3260.123                  2835.00
```

```r
coe = data_frame(EMcompare$coefficients[-1,1], treatment_table$Y_diff)
coe
```

```
## # A tibble: 16 x 2
##    `EMcompare$coefficients[-1, 1]` `treatment_table$Y_diff`
##                              <dbl>                    <dbl>
##  1                          -226.                    -225.
##  2                           -63.1                   -326.
##  3                          -242.                    -190.
##  4                          -384.                    -303.
##  5                            37.8                   -209.
##  6                            88.5                    -97.7
##  7                          -121.                    -160.
##  8                          -219.                     114.
##  9                            10.4                   -296.
## 10                          -103.                    -261.
## 11                          -252.                    -190.
## 12                          -444.                     107.
## 13                            40.8                   -238.
## 14                            26.7                   -497.
```

```
## 15                          -146.                    -476.
## 16                          -186.                    -624
```

DOES THIS COMPARISON MAKE SENSE???

The estimated average treatment effect of smoking on birthweight using the exact matching estimator is -225.44 grams, while the comparable linear regression gives an average treatment effect of c(tobacco = -226.245032864622) grams. These values are quite close.

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.

```
ps_model = glm(data = smoke, formula = tobacco ~ mage + (mage * mage) + meduc + mblack + alcohol, family
propensity = predict(ps_model, type = "response")
```

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

```
ps_wt = (smoke$tobacco / propensity) + ((1-smoke$tobacco) / (1-propensity))

wps_model = estimatr::lm_robust(data = smoke, formula = birthwgt ~ tobacco, weights = ps_wt)
summary(wps_model)
```

```
##
## Call:
## estimatr::lm_robust(formula = birthwgt ~ tobacco, data = smoke,
##     weights = ps_wt)
##
## Weighted, Standard error type:  HC2
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper    DF
## (Intercept)     3426       1.85 1851.90        0   3422.1   3429.4 94171
## tobacco         -225       4.92  -45.73        0   -234.7   -215.4 94171
##
## Multiple R-squared:  0.04814 ,   Adjusted R-squared:  0.04813
## F-statistic:  2092 on 1 and 94171 DF,  p-value: < 2.2e-16
```