EDS241: Assignment 2

Patty Park

Reminders: Make sure to read through the setup in markdown. Remember to fully report/interpret your results and estimates (in writing) + present them in tables/plots.

1 Part 1 Treatment Ignorability Assumption and Applying Matching Estimators (19 points):

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 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 identifies as Black), alcohol (=1 if consumed alcohol during pregnancy), first (=1 if first child), diabete (=1 if mother diabetic), anemia (=1 if mother anemic)

```
# Load data for Part 1
smoking <- read_csv("data/SMOKING_EDS241.csv")
```

1.1 Mean Differences, Assumptions, and Covariates *(3 pts)*

a) What is the mean difference in birth weight of infants with smoking and non-smoking mothers [1 pts:ANSWERED]?

Under what assumption does this correspond to the average treatment effect of maternal smoking during pregnancy on infant birth weight (its a theorethical assumption what is the counterfactual? test whether that assumption holds true) [0.5 pts]?

Calculate and create a table demonstrating the differences in the mean proportions/values of covariates observed in smokers and non-smokers (remember to report whether differences are statistically significant) and discuss whether this provides empirical evidence for or against this assumption. Remember that this is observational data. What other quantitative empirical evidence or test could help you assess the former assumption? [1.5 pts: 0.5 pts table, 1 pts discussion]

```r
#separate control and treatment from each other ----
m_smoking <- smoking %>% filter(tobacco == 1)
m_nonsmoking <- smoking %>% filter(tobacco == 0)

## Calculate mean difference. Remember to calculate a measure of statistical significance
mean_birth_smoking <- smoking %>%
  group_by(tobacco) %>%
  summarise(mean_birth = mean(birthwgt))

#find mean difference birth weight
mean_difference_birth <- mean_birth_smoking$mean_birth[1] - mean_birth_smoking$mean_birth[2]
print(mean_difference_birth)
```

```
## [1] 244.5394
```

```r
# Selecting binary and continuous variables from the dataset ----
#binary
pretreat_binary <- smoking %>%
select(anemia, diabete, alcohol, mblack, first, tobacco, birthwgt)

#continuous
pretreat_continuous <- smoking %>%
select(mage, meduc, tobacco, birthwgt)

#create empty dataframes
prop_test_results <- data.frame()
t_test_results <- data.frame()


## For continuous variables you can use the t-test ----
#t.test()
# Identifying continuous variables for t-tests
continuous_vars <- names(pretreat_continuous)[-3:-4]
      # comparing if age is grouped in a certain age
      # if it is, then it is probably impacting birth weigths

for (var in continuous_vars) {
# Dynamically creating the formula for the t-test
formula <- as.formula(paste(var, "~ tobacco"))
# Performing the t-test
t_test_result_cont <- t.test(formula, data = pretreat_continuous)
# Storing the tidy results of the t-test in the data frame
t_test_result_tidy <- broom::tidy(t_test_result_cont)
t_test_result_tidy$Variable <- var
t_test_results <- rbind(t_test_results, t_test_result_tidy)
}


## For binary variables you should use the proportions test
#prop.test()

binary_vars <- names(pretreat_binary)[-6:-7]
```

```r
for (var in binary_vars) {
# Splitting the data into treated and untreated groups for the current variable
treated <- pretreat_binary %>% filter(tobacco == 1) %>% pull(!!sym(var))
untreated <- pretreat_binary %>% filter(tobacco == 0) %>% pull(!!sym(var))
# Performing the proportion test
prop_test_result <- prop.test(x = c(sum(treated), sum(untreated)),
n = c(length(treated), length(untreated)),
correct = FALSE)
# Storing the tidy results of the proportion test in the data frame
prop_test_result_tidy <- broom::tidy(prop_test_result)
prop_test_result_tidy$Variable <- var
prop_test_results <- rbind(prop_test_results, prop_test_result_tidy)
}


# Covariate Calculations and Tables (code used from Assignment 1 key)

# Combining the results of proportion and t-tests into a single data frame
combined_results <- bind_rows(
prop_test_results %>% select(Variable, estimate1, estimate2, p.value),
t_test_results %>% select(Variable, estimate1, estimate2, p.value)
)
# Creating a table for output using kable and kableExtra
combined_results_table <- kable(combined_results, format = "latex",
col.names = c("Variable",
"Proportion or Mean Control",
"Proportion or Mean Treated", "P-Value"),
caption = "Treated and Untreated Pre-treatment Proportion and T- Test Results") %>%
kable_styling(font_size = 7, latex_options = "hold_position")
# Displaying the table
combined_results_table
```

Table 1: Treated and Untreated Pre-treatment Proportion and T- Test Results

Variable	Proportion or Mean Control	Proportion or Mean Treated	P-Value
anemia	0.0141031	0.0078005	0.0000000
diabete	0.0175187	0.0173636	0.8858005
alcohol	0.0441825	0.0071033	0.0000000
mblack	0.1354121	0.1086279	0.0000000
first	0.3645879	0.4360900	0.0000000
mage	27.4530853	25.5385632	0.0000000
meduc	13.2394207	11.9209454	0.0000000

Answer: The mean difference I got between the birth weights from mothers that did smoke and mothers that did not smoke was 244.539. The assumption that the ATE applies to is ignorability. there are no unobserved factors and they are not impacting the treatment(smoking) and control(no smoking).

1.2 ATE and Covariate Balance *(3 pts)*

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 an OLS regression with NO linear controls for the covariates [0.5 pts].

Perform the same estimate including the control variables [0.5 pts].

Next, compute indices of covariate imbalance between the treated and non-treated regarding these covariates (see example file from class). Present your results in a table [1 pts].

What do you find and what does it say regarding whether the assumption you mentioned responding to a) is fulfilled? [1 pts]

```
# ATE Regression univariate
uni_treat <- lm(birthwgt ~ tobacco, data = m_smoking)
uni_control <- lm(birthwgt ~ tobacco, data = m_nonsmoking)
uni <- lm(birthwgt ~ tobacco, data = smoking)

summary(uni_treat)
```

```
##
## Call:
## lm(formula = birthwgt ~ tobacco, data = m_smoking)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1685.75  -322.75    18.25   329.25  1314.25
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept) 3185.747      3.748     850 <0.0000000000000002 ***
## tobacco           NA         NA      NA                  NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 505 on 18151 degrees of freedom
```

```
summary(uni_control)
```

```
##
## Call:
## lm(formula = birthwgt ~ tobacco, data = m_nonsmoking)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1930.29  -290.29    28.71   340.71  1069.71
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept) 3430.286      1.781    1926 <0.0000000000000002 ***
## tobacco           NA         NA      NA                  NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 491 on 76020 degrees of freedom
```

```
# ATE with covariates
cov_treat <- lm(birthwgt ~ tobacco + mage+ meduc + anemia + diabete + alcohol + mblack + first, data = m
cov_control <- lm(birthwgt ~ tobacco + mage+ meduc + anemia + diabete + alcohol + mblack + first, data =
cov <- lm(birthwgt ~ tobacco + mage + meduc + anemia + diabete + alcohol + mblack + first, data = smoki
```

```
summary(cov_treat)
```

```
##
## Call:
## lm(formula = birthwgt ~ tobacco + mage + meduc + anemia + diabete +
##     alcohol + mblack + first, data = m_smoking)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1779.94  -307.14    15.27   330.78  1479.90
##
## Coefficients: (1 not defined because of singularities)
##              Estimate Std. Error t value             Pr(>|t|)
## (Intercept) 3074.5410    29.2546 105.096 < 0.0000000000000002 ***
## tobacco           NA         NA      NA                   NA
## mage          -4.2762     0.8225  -5.199        0.00000020264 ***
## meduc         22.8552     2.5315   9.028 < 0.0000000000000002 ***
## anemia         0.3030    31.3502   0.010               0.9923
## diabete       83.3709    28.1882   2.958               0.0031 **
## alcohol     -109.6314    18.1373  -6.045        0.00000000153 ***
## mblack      -226.0715    10.9567 -20.633 < 0.0000000000000002 ***
## first        -49.5090     8.2513  -6.000        0.00000000201 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 497 on 18144 degrees of freedom
## Multiple R-squared:  0.03154,    Adjusted R-squared:  0.03116
## F-statistic: 84.41 on 7 and 18144 DF,  p-value: < 0.00000000000000022
```

```
summary(cov_control)
```

```
##
## Call:
## lm(formula = birthwgt ~ tobacco + mage + meduc + anemia + diabete +
##     alcohol + mblack + first, data = m_nonsmoking)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2025.91  -288.11    24.06   331.46  1357.29
##
## Coefficients: (1 not defined because of singularities)
##              Estimate Std. Error t value             Pr(>|t|)
## (Intercept) 3367.6958    12.9134 260.790 < 0.0000000000000002 ***
## tobacco           NA         NA      NA                   NA
## mage          0.1175     0.3967   0.296                0.767
## meduc         9.9178     0.9136  10.856 < 0.0000000000000002 ***
## anemia       -7.6021    19.8922  -0.382                0.702
## diabete      71.0170    13.3873   5.305          0.000000113 ***
## alcohol     -22.8825    20.8059  -1.100                0.271
## mblack      -240.1688     5.7990 -41.415 < 0.0000000000000002 ***
## first       -107.4647     3.7900 -28.355 < 0.0000000000000002 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 481.4 on 76013 degrees of freedom
## Multiple R-squared:  0.03911,    Adjusted R-squared:  0.03902
## F-statistic: 441.9 on 7 and 76013 DF,  p-value: < 0.00000000000000022
```

```
# Present Regression Results
stargazer(cov, uni, type = "LaTeX")


# Covariate balance (does birth weight need to be added? is it tobacco or birthwgt that is being regres
cov_bal_ind <- xBalance(tobacco ~ birthwgt + mage + meduc + anemia + diabete + alcohol + mblack + first
         report=c("std.diffs","chisquare.test", "p.values"))

overall <- cov_bal_ind$overall
results <- cov_bal_ind$results

overall
```

```
##         chisquare df p.value
## unstrat  10297.66  8       0
```

```
results
```

```
## , , strata = unstrat
##
##           stat
## vars           std.diff
##   birthwgt -0.495266796
##   mage     -0.361941990
##   meduc    -0.643735401
##   anemia    0.066702936
##   diabete   0.001186437
##   alcohol   0.315254487
##   mblack    0.084390427
##   first    -0.144997533
##           stat
## vars                                                                                                  p
##   birthwgt 0.00000000000000000000000000000000000000000000000000000000000000000000000000000000000000
##   mage     0.00000000000000000000000000000000000000000000000000000000000000000000000000000000000000
##   meduc    0.00000000000000000000000000000000000000000000000000000000000000000000000000000000000000
##   anemia   0.00000000000000069349363744652212303911548947161403147779553833227050851917
##   diabete  0.88580110424800984070259346481179818511009216308593750000000000000000000000
##   alcohol  0.00000000000000000000000000000000000000000000000000000000000000000000000000000000000000
##   mblack   0.00000000000000000000017907846465052483479371779507617472510484467293722 3
##   first    0.00000000000000000000000000000000000000000000000000000000000000000000000000948984
##
## attr(,"originals")
## [1] "birthwgt" "mage"     "meduc"    "anemia"   "diabete"  "alcohol"  "mblack"
## [8] "first"
```

```
# Balance Table

balance_results_table <- kable(results, format = "latex",
col.names = c("Variable",
"Standard Difference unstratified",
"P-Value unstratified"),
caption = "Balance Table results") %>%
kable_styling(font_size = 7, latex_options = "hold_position")

balance_overall_table <- kable(overall, format = "latex",
col.names = c(
"Chi Squared",
"Degrees of Freedoms", "P-value"),
caption = "Balance Table results") %>%
kable_styling(font_size = 7, latex_options = "hold_position")

#print table
balance_results_table
```

Table 2: Balance Table results

Variable	Standard Difference unstratified	P-Value unstratified
birthwgt	-0.4952668	0.0000000
mage	-0.3619420	0.0000000
meduc	-0.6437354	0.0000000
anemia	0.0667029	0.0000000
diabete	0.0011864	0.8858011
alcohol	0.3152545	0.0000000
mblack	0.0843904	0.0000000
first	-0.1449975	0.0000000

```
balance_overall_table
```

Table 3: Balance Table results

	Chi Squared	Degrees of Freedoms	P-value
unstrat	10297.66	8	0

```
# kable(cov_bal_ind$results, format = "latex") %>%
#   kable_styling(font_size = 7, latex_options = "hold_position")
```

1.3 Propensity Score Estimation *(3 pts)*

c) Next, estimate propensity scores (i.e. probability of being treated) for the sample, using the provided covariates. Create a regression table reporting the results of the regression and discuss what the covariate coefficients indicate and interpret one coefficient [1.5 pts].

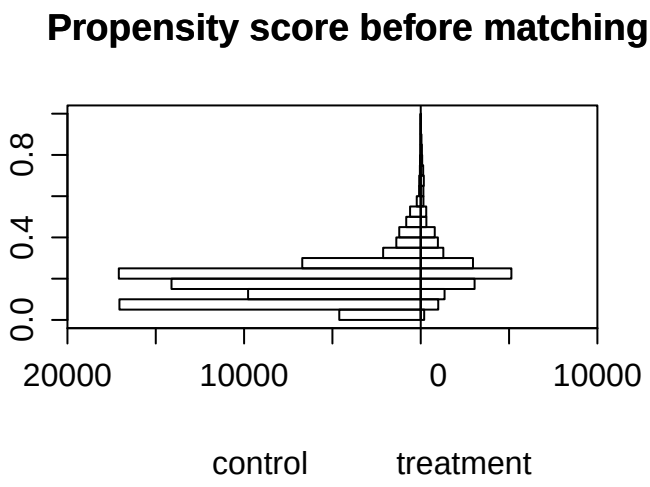
Create histograms of the propensity scores comparing the distributions of propensity scores for smokers ('treated') and non-smokers ('control'), discuss the overlap and what it means [1.5 pts].

```r
## Propensity Scores (less things are added in the other example)
ps  <- glm(tobacco ~  mage + meduc + anemia + diabete + alcohol + mblack + first, data = smoking,   fami
summary(ps)
```

```
##
## Call:
## glm(formula = tobacco ~ mage + meduc + anemia + diabete + alcohol +
##     mblack + first, family = binomial(), data = smoking)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4975  -0.6971  -0.5417  -0.3274   2.6985
##
## Coefficients:
##              Estimate Std. Error z value             Pr(>|z|)
## (Intercept)  3.493297   0.066613  52.442 < 0.0000000000000002 ***
## mage        -0.040562   0.001931 -21.007 < 0.0000000000000002 ***
## meduc       -0.297269   0.005152 -57.699 < 0.0000000000000002 ***
## anemia       0.333952   0.079367   4.208          0.000025797 ***
## diabete      0.159533   0.065861   2.422                0.0154 *
## alcohol      2.026641   0.060353  33.580 < 0.0000000000000002 ***
## mblack      -0.133447   0.026587  -5.019          0.000000519 ***
## first       -0.379167   0.019303 -19.643 < 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: 84459  on 94165  degrees of freedom
## AIC: 84475
##
## Number of Fisher Scoring iterations: 5
```

```r
## PS Histogram Unmatched
#attach the propensity scores to the dataframe
smoking$psvalue <- predict(ps,  type    = "response")

histbackback(split(smoking$psvalue, smoking$tobacco),   main=
  "Propensity   score   before  matching",  xlab=c("control",   "treatment"))
```




```
coef_table <- coef(summary(ps))
coef_table[, "Pr(>|z|)"] <- format.pval(coef_table[, "Pr(>|z|)"], digits = 2)



#create regression table
regression_table <- kable(coef_table, format = "latex",
col.names = c("Variable",
"Proportion or Mean Control",
"Proportion or Mean Treated", "P-Value"),
caption = "Treated and Untreated Pre-treatment Proportion and T- Test Results") %>%
kable_styling(font_size = 7, latex_options = "hold_position")

regression_table
```

Table 4: Treated and Untreated Pre-treatment Proportion and T- Test Results

	Variable	Proportion or Mean Control	Proportion or Mean Treated	P-Value
(Intercept)	3.49329673960041	0.0666128153596041	52.4418119958167	<0.0000000000000002
mage	-0.0405619249450694	0.00193085515144564	-21.0072334606252	<0.0000000000000002
meduc	-0.297269382261615	0.00515210095176822	-57.698671870859	<0.0000000000000002
anemia	0.333952290590898	0.0793666904085223	4.20771344844988	0
diabete	0.159533411674843	0.0658605440326628	2.42229113072198	0.02
alcohol	2.02664067000151	0.0603529830525079	33.5797928701934	<0.0000000000000002
mblack	-0.133446847724597	0.0265866156610708	-5.01932436327335	0
first	-0.379166672610414	0.0193026250975255	-19.6432697985219	<0.0000000000000002

```
#gtsummary()
```

1.4 Matching Balance *(3 pts)*

(d) Next, match treated/control mothers using your estimated propensity scores and nearest neighbor matching. Compare the balancing of pretreatment characteristics (covariates) between treated and non-treated units in the original dataset (from c) with the matched dataset (think about comparing histograms/regressions) [2 pts]. Make sure to report and discuss the balance statistics [1 pts].

```r
## Nearest-neighbor Matching

m.nn    <- matchit(tobacco  ~  mage + meduc + anemia + diabete + alcohol + mblack + first, data = smoki
summary(m.nn)
```

```
##
## Call:
## matchit(formula = tobacco ~ mage + meduc + anemia + diabete +
##     alcohol + mblack + first, data = smoking, method = "nearest",
##     ratio = 1)
##
## Summary of Balance for All Data:
##          Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mean
## distance        0.2600        0.1767          0.6190     1.6361    0.1406
## mage           25.5386       27.4531         -0.3663     0.9709    0.0766
## meduc          11.9209       13.2394         -0.8190     0.5662    0.0734
## anemia          0.0141        0.0078          0.0535          .    0.0063
## diabete         0.0175        0.0174          0.0012          .    0.0002
## alcohol         0.0442        0.0071          0.1804          .    0.0371
## mblack          0.1354        0.1086          0.0783          .    0.0268
## first           0.3646        0.4361         -0.1486          .    0.0715
##          eCDF Max
## distance   0.2945
## mage       0.1676
## meduc      0.2773
## anemia     0.0063
## diabete    0.0002
## alcohol    0.0371
## mblack     0.0268
## first      0.0715
##
## Summary of Balance for Matched Data:
##          Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mean
## distance        0.2600        0.2560          0.0299     1.2066    0.0024
## mage           25.5386       25.4882          0.0096     1.0004    0.0021
## meduc          11.9209       11.8534          0.0420     0.8306    0.0051
## anemia          0.0141        0.0134          0.0061          .    0.0007
## diabete         0.0175        0.0172          0.0021          .    0.0003
## alcohol         0.0442        0.0240          0.0984          .    0.0202
## mblack          0.1354        0.1319          0.0103          .    0.0035
## first           0.3646        0.3641          0.0009          .    0.0004
##          eCDF Max Std. Pair Dist.
## distance   0.0218          0.0300
## mage       0.0071          0.1030
## meduc      0.0256          0.1443
## anemia     0.0007          0.1070
## diabete    0.0003          0.0718
## alcohol    0.0202          0.1670
## mblack     0.0035          0.0982
## first      0.0004          0.0989
##
## Sample Sizes:
##           Control Treated
```

```
## All         76021   18152
## Matched     18152   18152
## Unmatched   57869       0
## Discarded       0       0
```

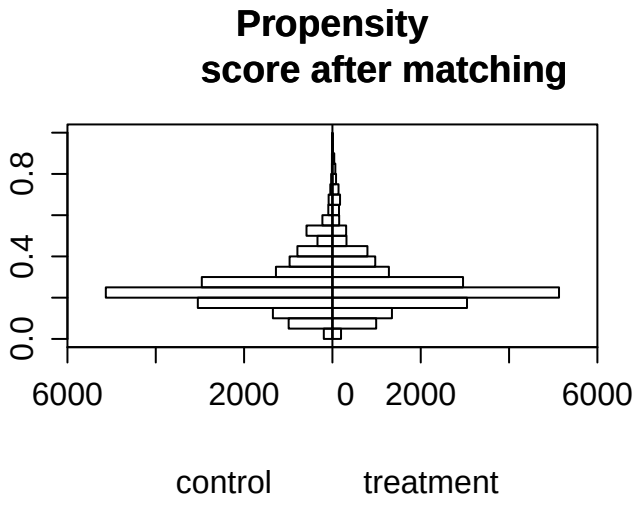
```
match.data  = match.data(m.nn)

## Covariate Imbalance post matching:
xBalance(   tobacco ~  mage + meduc + anemia + diabete + alcohol + mblack + first,  data=match.data,
         report=c("std.diffs","chisquare.test", "p.values"))
```

```
##         strata():  unstrat
##         stat      std.diff
## vars
## mage                  0.01
## meduc                 0.04 ***
## anemia                0.01
## diabete               0.00
## alcohol               0.11 ***
## mblack                0.01
## first                 0.00
## ---Overall Test---
##         chisquare df                        p.value
## unstrat       124  7 0.000000000000000000000000999
## ---
## Signif. codes:  0 '***' 0.001 '** ' 0.01 '*  ' 0.05 '.  ' 0.1 '    ' 1
```

```
# ps_2  <- glm(tobacco  ~  mage + meduc + anemia + diabete + alcohol + mblack + first,
#         data  =match.data,    family  = binomial())
# summary(ps_2)
#
# match.data$psvalue    <- predict(ps_2,    type    = "response")

## Histogram of PS after matching
histbackback(split(match.data$psvalue,  match.data$tobacco),    main= "Propensity
        score   after   matching",  xlab=c("control",   "treatment"))
```



1.5 ATE with Nearest Neighbor *(3 pts)*

(e) Estimate the ATT using the matched dataset. Report and interpret your result (Note: no standard error or significance test is required here)

```
## Nearest Neighbor
sumdiff_data<-match.data%>%
  group_by(subclass)%>%
  mutate(diff=birthwgt[tobacco==1]-birthwgt[tobacco==0])

## ATT

N=length(smoking$tobacco)
NT=sum(smoking$tobacco)
NC=N-NT

sumdiff<-sum(sumdiff_data$diff)/2
ATT_m_nn = 1/NT * sumdiff
ATT_m_nn
```

```
## [1] -222.7886
```

```
#mean(sumdiff_data$diff)
```

1.6 ATE with WLS Matching *(3 pts)*

f) Last, use the original dataset and perform the weighted least squares estimation of the ATE using the propensity scores (including controls). Report and interpret your results, here include both size and precision of estimate in reporting and interpretation.

```r
## Weighted least Squares (WLS) estimator Preparation

PS <- smoking$psvalue
Y <- smoking$birthwgt
D <- smoking$tobacco
EY1 <- D*Y/PS / sum(D/PS)
EY0 <- (1-D)*Y/(1-PS) / sum((1-D) / (1-PS))
ATE_IPW = sum(EY1) - sum(EY0)
ATE_IPW
```

```
## [1] -231.3871
```

```r
smoking$wgt = (D/PS + (1-D)/(1-PS))

## Weighted least Squares (WLS) Estimates
#without controls
reg_wls <-lm(birthwgt   ~ tobacco,
             data = smoking, weights = wgt)


#with controls
reg_wls_c   <-lm(tobacco    ~ birthwgt + mage + meduc + anemia + diabete + alcohol + mblack + first,
          data = smoking, weights = wgt)


## Present Results
summary(reg_wls)
```

```
##
## Call:
## lm(formula = birthwgt ~ tobacco, data = smoking, weights = wgt)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -8572.3  -366.3    33.4   413.0  6328.2
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept) 3427.178      2.292 1495.08 <0.0000000000000002 ***
## tobacco     -231.387      3.267  -70.82 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 704.7 on 94171 degrees of freedom
## Multiple R-squared:  0.05056,    Adjusted R-squared:  0.05055
## F-statistic:  5015 on 1 and 94171 DF,  p-value: < 0.00000000000000022
```

```r
summary(reg_wls_c)
```

```
##
## Call:
## lm(formula = tobacco ~ birthwgt + mage + meduc + anemia + diabete +
```

```
##     alcohol + mblack + first, data = smoking, weights = wgt)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7445 -0.5688 -0.4692 -0.3380  3.9900
##
## Coefficients:
##                Estimate   Std. Error t value             Pr(>|t|)
## (Intercept)  1.265548631  0.015173207  83.407 < 0.0000000000000002 ***
## birthwgt    -0.000219223  0.000003137 -69.874 < 0.0000000000000002 ***
## mage         0.002296647  0.000349225   6.576      0.0000000000484 ***
## meduc       -0.008365729  0.000858679  -9.743 < 0.0000000000000002 ***
## anemia       0.007097817  0.016471789   0.431              0.66654
## diabete      0.025347420  0.011960527   2.119              0.03407 *
## alcohol     -0.015465455  0.013516149  -1.144              0.25254
## mblack      -0.014067489  0.004967837  -2.832              0.00463 **
## first       -0.000247804  0.003450083  -0.072              0.94274
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6844 on 94164 degrees of freedom
## Multiple R-squared:  0.05184,    Adjusted R-squared:  0.05176
## F-statistic: 643.5 on 8 and 94164 DF,  p-value: < 0.00000000000000022
```

1.7 Differences in Estimates *(1 pts)*

g) Explain why it was to be expected given your analysis above that there is a difference between your estimates in e) and f)?

Answer: e is with ATT, looking only those that were treated. ATE is looking at all treated and untreated.

2 Part 2 Panel model and fixed effects (6 points)

We will use the progresa data from last time as well as a new dataset. In the original dataset, treatment households had been receiving the transfer for a year. Now, you get an additional dataset with information on the same households from before the program was implemented, establishing a baseline study (from 1997), and the same data we worked with last time (from 1999). *Note: You will need to install the packages plm and dplyr (included in template preamble). Again, you can find a description of the variables at the bottom of PDF and HERE.

2.1 Estimating Effect with First Difference *(3 pts: 1.5 pts estimate, 1.5 pts interpretation)*

Setup: Load the new baseline data (progresa_pre_1997.csv) and the follow-up data (progresa_post_1999.csv) into R. Note that we created a time denoting variable (with the same name, 'year') in BOTH datasets. Then, create a panel dataset by appending the data (i.e. binding the dataset row-wise together creating a single dataset). We want to examine the same outcome variable as before, value of animal holdings (vani).

```
rm(list=ls()) # clean environment

## Load the datasets
# progresa_pre_1997 <- read_csv() insert your filepath etc
# progresa_post_1999 <- read_csv()

## Append post to pre dataset
#progresa <- rbind(progresa_pre_1997, progresa_post_1999)
```

a) Estimate a first-difference (FD) regression manually, interpret the results briefly (size of coefficient and precision!) *Note: Calculate the difference between pre- and post- program outcomes for each family. To do that, follow these steps and the code given in the R-template:

```
### Code included to help get you started
## i. Sort the panel data in the order in which you want to take differences, i.e. by household and tim

## Create first differences of variables
# progresa <- progresa %>%
#   arrange(hhid, year) %>%
#   group_by(hhid)

## ii. Calculate the first difference using the lag function from the dplyr package.
#     mutate(vani_fd = vani - dplyr::lag(vani))

## iii. Estimate manual first-difference regression (Estimate the regression using the newly created va
# fd_manual <- lm(vani_fd ~ ...)
```

2.2 Fixed Effects Estimates *(2 pts: 1 pts estimate, 1.5 interpretation)*

b) Now also run a fixed effects (FE or 'within') regression and compare the results. Interpret the estimated treatment effects briefly (size of coefficient and precision!)

```
## Fixed Effects Regression

## Present Regression Results
```

2.3 First Difference and Fixed Effects and Omitted Variable Problems *(1 pts)*

c) Explain briefly how the FD and FE estimator solves a specific omitted variable problem? Look at the example on beer tax and traffic fatalities from class to start thinking about ommitted variables. Give an example of a potential omitted variable for the example we are working with here that might confound our results? For that omitted variable, is a FE or FD estimator better? One example is enough.