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)*

```
#================================================
#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)) %>%
  mutate(mean_diff_birth = mean_birth - dplyr::lead(mean_birth))
```

```r
#==================================
#  Create table for birth weight
#==================================
birth_table <- kable(mean_birth_smoking, format = "latex",
      col.names = c("Tobacco", "Mean Birth Weight", "Mean Birth Weight Difference"),
      caption = "Mean Difference in Birth Weight between Smokers and Non-smokers") %>%
  kable_styling(font_size = 7, latex_options = "hold_position")

birth_table
```

Table 1: Mean Difference in Birth Weight between Smokers and Non-smokers

Tobacco	Mean Birth Weight	Mean Birth Weight Difference
0	3430.286	244.5394
1	3185.747	NA

```r
#==================================
#find statistical difference between birth weight for those that do smoke and those that do not smoke
#==================================
mean_birth <- t.test(birthwgt ~ tobacco, data = smoking)
#print t.test
mean_birth
```

```
##
##  Welch Two Sample t-test
##
## data:  birthwgt by tobacco
## t = 58.932, df = 26945, p-value < 0.00000000000000022
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  236.4060 252.6727
## sample estimates:
## mean in group 0 mean in group 1
##        3430.286        3185.747
```

```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 ----
```

```r
#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]

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 Proportion and T- Test Results") %>%
```

```
kable_styling(font_size = 7, latex_options = "hold_position")
# Displaying the table
combined_results_table
```

Table 2: Treated and Untreated 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

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

Under what assumption does this correspond to the average treatment effect of maternal smoking during pregnancy on infant birth weight [0.5 pts]?

Answer: The assumption that this correlates to is ignorability. In comparing the mean difference between the control (no smoking) and treatment(smoking), we are seeing if there is a difference in the birth weight on average. When we calculate the mean difference, we get 244.539. Running a t.test on the mean of the birth weigth and on tobacco, we get a very low p-value. Just by ignoring the other covariants, we are able to get this result and conclude that mothers that do or do not smoke have an impact on their baby's birth weight when born.

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]

Answer: Looking at the table of the difference in the mean proportions, almost all covariates are statistically significant except diabetes. This would mean that almost all of these variables have an impact on the birth weight rather than just the smoking treatment variable. From this, we can reject our assumption since our original assumption was that we ignore all other covariates and assume that only tobacco has an impact on birth weight. Another quantitative evidence that can help us test our former assumption is the propensity score. Later on in this lab, we will be finding the propensity score to help assess our former assumption. However, what the propensity score does is it tells us the probability of receiving the treatment if the individual and the covariate as well.

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]

Answer: Looking at our table created by the `xBalance` function, we find that a number of covariates are imbalanced. Some examples of the covariates that are unbalanced are `meduc` which received a standardized difference of -0.64, `alcohol` which received a standardized difference of 0.32, and `mage` which received a standardized difference of -0.36. Because these numbers are more further away from zero, it means that these covariates are imbalanced. In other words, the mean average between these covariates between the treated and control are not close to each other, and there is a big difference between the two. From this, we need to dismiss the ignorability assumption because in this case, these covariates are too important to ignore and potentially have an impact on birth weight rather than just tobacco.

```
#============================
# ATE Regression univariate --
#============================

uni <- lm(birthwgt ~ tobacco, data = smoking)

#============================
#    ATE with covariates  ----
#============================

cov <- lm(birthwgt ~ tobacco + mage + meduc + anemia + diabete + alcohol + mblack + first, data = smoki

#============================
#    Tidied up results   ----
#============================

uni_table <- broom::tidy(uni)
cov_table <- broom::tidy(cov)

#============================
# Present Regression Results
#============================

uni_table_out <- kable(uni_table, format = "latex",
col.names = c(
"Variables",
"Estimates", "Standard error", "Statistic", "P-value"),
caption = "ATE for univariate") %>%
kable_styling(font_size = 7, latex_options = "hold_position")

cov_table_out <- kable(cov_table, format = "latex",
col.names = c(
"Variables",
"Estimates", "Standard error", "Statistic", "P-value"),
caption = "ATE for covariates") %>%
kable_styling(font_size = 7, latex_options = "hold_position")



#stargazer(cov, uni, type = "text")

#============================
#    Covariate balance   ----
#============================
cov_bal_ind <- xBalance(tobacco ~ mage + meduc + anemia + diabete + alcohol + mblack + first, data = sm
```

```
        report=c("std.diffs","chisquare.test", "p.values"))

#=================================================
# assign overall and results to individual variables
#=================================================
overall <- cov_bal_ind$overall
results <- cov_bal_ind$results

#============================
#   Create Balance Table ----
#============================
# for results
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")

#for overall
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 3: Balance Table results

Variable	Standard Difference unstratified	P-Value unstratified
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 4: Balance Table results

	Chi Squared	Degrees of Freedoms	P-value
unstrat	7642.691	7	0

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].

Answer: The covariate coefficients indicate the probability that if an individual is on the treatment end, in this case is smoking, they are x amount likely to also have other covariates associated with them. Using the propensity score, we are able to determine what type of bias may lie in the dataset and what we should be aware of in order to account for those bias. An example in this case is looking at the covariate `alcohol`. The covariate coefficient is around 2.03. This mean that if an individual smokes, they are 2.03 time likely to also drink. We would have to be more aware of this fact that there is more bias in alcohol and need to take that into account when doing other calculations, such as finding the ATE or ATT in order to minimize for the bias.

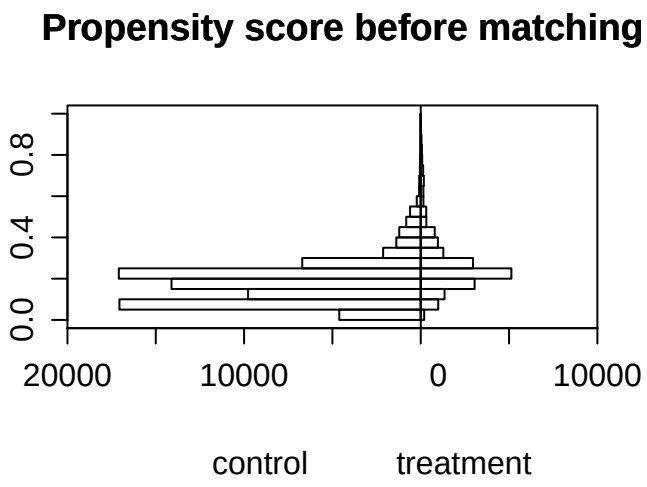
Create histograms of the propensity scores comparing the distributions of propensity scores for smokers ('treated') and non-smokers ('control'), discuss the overlap and what it means [1.5 pts].

Answer: This histogram, which is the propensity score before matching, shows how heavily imbalanced the covariates are. We see that on the left side, which is the control side (those that do not smoke), the histograms reach very far into the left side, with the maximums passing by the 15000 mark. One the right side, which is the treatment side (those that do smoke) the histograms don't go very far into the right, with maximum bar reaching around 5000. Visually, we can see that there is an imbalance in the score in general. Because of the imbalance we see on the histogram, there is bias in the dataset. This could mean that the other covariates, such as alcohol may also have an impact on the baby's birth weight rather than smoking on its own. In all, it gives us less confidence which we can state that smoking is the main reason that is causing the gap between the birth rates of those that are on the control (do not smoke) and those that are on the treatment (do smoke).

```
#============================
##    Propensity Scores
#============================
#find Propensity score using glm function
ps  <- glm(tobacco ~  mage + meduc + anemia + diabete + alcohol + mblack + first, data = smoking,   fam

#============================
## PS Histogram Unmatched
#============================
#put propensity score back on original dataframe
smoking$psvalue <- predict(ps,  type    = "response")

#create histogram
ps_hist <- histbackback(split(smoking$psvalue,  smoking$tobacco),   main=
  "Propensity   score   before matching",   xlab=c("control",   "treatment"))
```




```
#print histogram
#print(ps_hist)

#============================
# create regression table
#============================
#setup for creating regression table
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",
"Coefficient Estimates",
"Standard Error", "z value", "P-Value"),
caption = "Propensity Score Regression Table") %>%
kable_styling(font_size = 7, latex_options = "hold_position")

#print regression table
regression_table
```

Table 5: Propensity Score Regression Table

Variable	Coefficient Estimates	Standard Error	z value	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

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].

Answer: In table 6 and 7, we can see a stark difference in the means control column. In the original dataset, as seen here and as well, we see how the means here are not similar to the means in the treated column. However, when looking at the matched dataset, the means in the treated and control columns are nearly identical. While they are not completely the same, it is much better than the ones in the original dataset. Looking at the balanced table, we see that the numbers have also gone down considerablly between the original and matched datset. For example, when looking at the covariate `meduc`, the unmatched balance table states that it is around -0.64. From this, we know that covariate is extremely unbalanced. But when looking at the matched dataset for `meduc`, we see it around 0.04. We can conclude that the match dataset is much more balanced and are comparing those that are similar in covariates, leading us to compare tobacco use. It should be noted that all of these numbers are not exact, meaning there is still a bit of bias in this dataset. However, the bias has gone down significantly from the original dataset, giving us more confidence in our matched dataset.

When we also look at the histogram, we can see a much more even distribution. In the histogram that we created for the data before we matched it, we saw it heavily skewed to the control side. In this histogram, however, we see both sides being basically a mirror of each other. The highest bar for both sides reaches up to over 5000. As with any matched dataset, the number of observations dropped significantly, in order to match with the amount there is for the treated variable. In the matched dataset, there is an equal number of control and treated variables, both 18512.

```
#================================
## Nearest-neighbor Matching
#================================
#find nearest neighbord
match_nn_reg    <- matchit(tobacco  ~  mage + meduc + anemia + diabete + alcohol + mblack + first, data
match_sum <- summary(match_nn_reg)

#create into dataframe
match.data = match.data(match_nn_reg)

#assign to individual variables
match_nn <- match_sum$nn
match_all <- match_sum$sum.all
match_matched <- match_sum$sum.matched

#create into table
match_all_table <- kable(match_all, format = "latex",
# col.names = c("Variable",
# "Coefficient Estimates",
# "Standard Error", "z value", "P-Value"),
caption = "Summary of Balance for All Data") %>%
kable_styling(font_size = 7, latex_options = "hold_position")

match_matched_table <- kable(match_matched, format = "latex",
# col.names = c("Variable",
# "Coefficient Estimates",
# "Standard Error", "z value", "P-Value"),
```

```
caption = "Summary of Balance for Matched Data") %>%
kable_styling(font_size = 7, latex_options = "hold_position")

nearest_neighbor_table <- kable(match_nn, format = "latex",
# col.names = c("Variable",
# "Coefficient Estimates",
# "Standard Error", "z value", "P-Value"),
caption = "Sample Sizes") %>%
kable_styling(font_size = 7, latex_options = "hold_position")

#print all tables
match_all_table
```

Table 6: Summary of Balance for All Data

	Means Treated	Means Control	Std. Mean Diff.	Var. Ratio	eCDF Mean	eCDF Max	Std. Pair Dist.
distance	0.2600121	0.1766914	0.6189832	1.6360736	0.1405622	0.2945192	NA
mage	25.5385632	27.4530853	-0.3662918	0.9709174	0.0765809	0.1676368	NA
meduc	11.9209454	13.2394207	-0.8189780	0.5661696	0.0734004	0.2773124	NA
anemia	0.0141031	0.0078005	0.0534503	NA	0.0063027	0.0063027	NA
diabete	0.0175187	0.0173636	0.0011823	NA	0.0001551	0.0001551	NA
alcohol	0.0441825	0.0071033	0.1804336	NA	0.0370792	0.0370792	NA
mblack	0.1354121	0.1086279	0.0782790	NA	0.0267842	0.0267842	NA
first	0.3645879	0.4360900	-0.1485559	NA	0.0715021	0.0715021	NA

```
match_matched_table
```

Table 7: Summary of Balance for Matched Data

	Means Treated	Means Control	Std. Mean Diff.	Var. Ratio	eCDF Mean	eCDF Max	Std. Pair Dist.
distance	0.2600121	0.2559875	0.0298986	1.2066380	0.0024168	0.0217607	0.0300139
mage	25.5385632	25.4881556	0.0096441	1.0003986	0.0020516	0.0070516	0.1029867
meduc	11.9209454	11.8534046	0.0419533	0.8306459	0.0050622	0.0255619	0.1443385
anemia	0.0141031	0.0133870	0.0060736	NA	0.0007162	0.0007162	0.1069886
diabete	0.0175187	0.0172433	0.0020996	NA	0.0002755	0.0002755	0.0718056
alcohol	0.0441825	0.0239643	0.0983851	NA	0.0202182	0.0202182	0.1670133
mblack	0.1354121	0.1318863	0.0103044	NA	0.0035258	0.0035258	0.0982137
first	0.3645879	0.3641472	0.0009157	NA	0.0004407	0.0004407	0.0988918

```
nearest_neighbor_table
```

Table 8: Sample Sizes

	Control	Treated
All (ESS)	76021	18152
All	76021	18152
Matched (ESS)	18152	18152
Matched	18152	18152
Unmatched	57869	0
Discarded	0	0

```
#====================================
## Covariate Imbalance post matching:
#====================================
```

```
cov_bal_ind_2 <- xBalance(tobacco   ~  mage + meduc + anemia + diabete + alcohol + mblack + first,  data
         report=c("std.diffs","chisquare.test", "p.values"))

#convert to individual variables
overall_2 <- cov_bal_ind_2$overall
results_2 <- cov_bal_ind_2$results

cov_table_2 <- kable(overall_2, format = "latex",
# col.names = c("Variable",
# "Coefficient Estimates",
# "Standard Error", "z value", "P-Value"),
caption = "Balance Table results for Matched data") %>%
kable_styling(font_size = 7, latex_options = "hold_position")

cov_table_3 <- kable(results_2, format = "latex",
# col.names = c("Variable",
# "Coefficient Estimates",
# "Standard Error", "z value", "P-Value"),
caption = "Balance Table results for Matched data") %>%
kable_styling(font_size = 7, latex_options = "hold_position")

#print tables
cov_table_2
```

Table 9: Balance Table results for Matched data

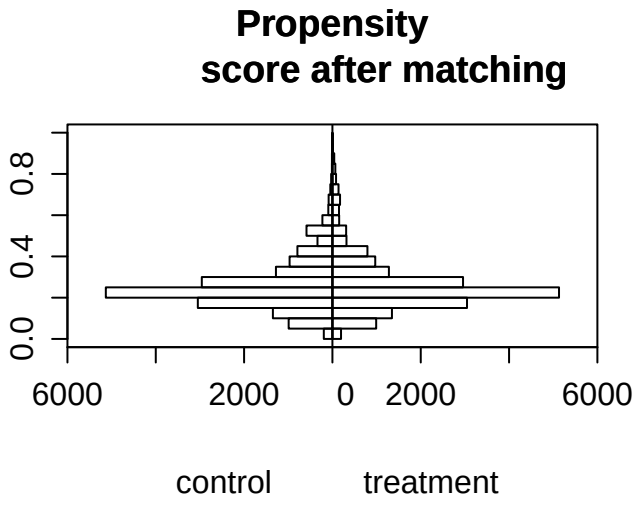
	chisquare	df	p.value
unstrat	124.2445	7	0

```
cov_table_3
```

Table 10: Balance Table results for Matched data

	std.diff.unstrat	p.unstrat
mage	0.0096451	0.3581627
meduc	0.0399657	0.0001408
anemia	0.0061509	0.5578802
diabete	0.0021077	0.8408575
alcohol	0.1116157	0.0000000
mblack	0.0103614	0.3235874
first	0.0009158	0.9304778

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



1.5 ATT with Nearest Neighbor *(3 pts)*

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

## ATT
NT=sum(smoking$tobacco) #number of treated
sumdiff<-sum(sumdiff_data$diff)/2
ATT_m_nn = 1/NT * sumdiff
ATT_m_nn
```

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

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

Answer: Here, the ATT is -222.789. This means that this is the difference between the means of the control and the treated. This difference of means is smaller than the original mean difference. While it is smaller than the mean difference from the original dataset, it is still a pretty large number. From this, we have more confidence that smoking does have an impact on birth weight. In order to solidify this theory, we would need to run other tests that can give us the standard error or p value. A t-test would be an appropriate test to run in this instance.

1.6 ATE with WLS Matching *(3 pts)*

```r
#=========================================================
## Weighted least Squares (WLS) estimator Preparation
#=========================================================
#find weights and ATE on weighted scores
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
#with controls
reg_wls_c   <-lm(birthwgt   ~ tobacco + mage + meduc + anemia + diabete + alcohol + mblack + first,
          data = smoking, weights = wgt)

#setup for creating regression table
coef_table_2 <- coef(summary(reg_wls_c))
# coef_table_2[, "Pr(>|z|)"] <- format.pval(coef_table_2[, "Pr(>|z|)"], digits = 2)

## Present Results
weighted_table <- kable(coef_table_2, format = "latex",
# col.names = c("Variable",
# "Coefficient Estimates",
# "Standard Error", "z value", "P-Value"),
caption = "Weighted Least Squares Estimation table") %>%
kable_styling(font_size = 7, latex_options = "hold_position")

#print table
weighted_table
```

Table 11: Weighted Least Squares Estimation table

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3384.224583	11.4872927	294.6059332	0.0000000
tobacco	-224.854183	3.2180177	-69.8735070	0.0000000
mage	-2.162627	0.3536926	-6.1144251	0.0000000
meduc	12.865499	0.8690655	14.8038317	0.0000000
anemia	10.811567	16.6819881	0.6480982	0.5169230
diabete	63.306452	12.1117043	5.2268823	0.0000002
alcohol	-69.215047	13.6868845	-5.0570345	0.0000004
mblack	-238.022495	4.9713029	-47.8792981	0.0000000
first	-89.986297	3.4817872	-25.8448585	0.0000000

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.

Answer: In this example, our weighted ATE is `r(paste(round(ATE_IPW))`. As this is a different value from the ATT, we can be confident that this dataset has strong heterogeneity. When looking at the weighted regression table, it shows us the strength of relationship between that particule covariate and the observed variable. We see that covariates such as `tobacco` and `mblack` have a big size impact on the baby's weight. If the individual smokes and is black, then the baby's weight goes down considerably, in this case, more than 200 grams. When looking at the precision of these numbers, they all look pretty precise, with variables `anemia`, `diabete`, and `alcohol` having higher standard errors. For these variables, this could mean that they may not play a big part on influencing the baby's weight verses the other variables.

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: The main reason why the two estimates are different is because we are looking at a different set of population. ATT is looking at individuals that are nearly identical, with tobacco being the only variable that is changed. ATE is looking at the whole dataset as all observations are kept. In this smoking scenario, the treated were not randomly assigned. This would mean that ATT would have more leverage as it accounts for this and is able to create a fair compared group between the treated and controlled. In the ATE example, we are putting more weight on those that have lower variance and make the control and treated more identical. However, there are some issues doing a weighed estimation, such as having trouble with probabilities close to one or zero, the weight may not add up to 1, and standard errors are hard to compute.

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("data/progresa_pre_1997.csv")
progresa_post_1999 <- read_csv("data/progresa_post_1999.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!)

Answer: In our first difference regression results, the treatment variable we get is around 288. What this tells us is that from year 1997 (before the treatment) to 1999 (after the treatment), this is the effect that the treatment had on, in this case, values on animals households owned. We can see that in year 1999, the value of animals households owned increased. Looking at the standard error, the error for treatment is around 85. While this is a somewhat high number, it is still much lower that the treatment coefficient and the intercept coefficient. We also know that this variable is significant since the p value for treatment is less than 0.05.

*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.
# progresa <- progresa %>%
#   group_by(year) %>%
  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 ~ treatment, data = progresa)

   stargazer(fd_manual, type="text")
```

```
##
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                                vani_fd
## ------------------------------------------------
## treatment                    287.905***
##                               (85.602)
##
## Constant                    -1,156.752***
##                               (64.494)
##
## ------------------------------------------------
## Observations                   13,514
## R2                              0.001
## Adjusted R2                     0.001
## Residual Std. Error    4,929.875 (df = 13512)
## F Statistic           11.312*** (df = 1; 13512)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

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 (it got bigger, or it got smaller). Interpret the estimated treatment effects briefly (size of coefficient and precision!)

Answer: Running the fixed effects regression, our coefficient becomes much less than before. In other words, the treatment coefficient became smaller. In this instance, we are controlling for the state and year, meaning that whatever changes that happened in the state or what changes happened becuase of the year change is fixed. (eg. change in policy for the state, or change in technology for year). In this fixed effects case, the treatment coefficient became smaller, as well as the standard error. From this, we can interpret it as so: with the treatment, the value of animal holdings in household has gone down. Since the standard error also shrank, it means that this result is more presice in its accuracy. I would conclude that the fixed effects is a better indicator of the effects of the treatment on values of animal holdings in households.

```
## Fixed Effects Regression

within1 <- plm(vani ~ treatment, index = c("state", "year"), model = "within", effect = "twoways", data

se_within1 <- coeftest(within1, vcov = vcovHC(within1, type = "HC2", method="white1"))[, "Std. Error"]

# Reformat standard errors for stargazer()
```

```r
se_within1 <- list(se_within1)


## Present Regression Results
stargazer(within1, keep=c("treatment"), se = se_within1, type="text")
```

```
##
## ========================================
##                   Dependent variable:
##               ---------------------------
##                           vani
## ----------------------------------------
## treatment             -231.844***
##                        (56.662)
##
## ----------------------------------------
## Observations            27,996
## R2                       0.001
## Adjusted R2              0.0003
## F Statistic    17.206*** (df = 1; 27987)
## ========================================
## Note:         *p<0.1; **p<0.05; ***p<0.01
```

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.

Answer: First difference and Fixed Effects solves the issue on time-invariant variables. First difference is able to get rid of the time-invariant variables by being able to be help constant over time. Because we are looking at the difference bewteen the dependent and independent variable, it is able to hold it constant. In the Fixed Effects model, it is able to account for the unobserved variables and take that into account into the regression. One omitted variable that may impact our results is inflation. During that time period, inflation grew dramatically. This may have caused the household that received the cash transfer to not only hold on to the cash transfer that they received, but also impact the value of animal holding during that time. To account for this, I would use the first difference model. First difference model is holding time sensitive variables constant. It will find the difference of inflation over this time period and become constant. An issue that may arise is that there are too few years to find a constant difference, meaning this number may not be as strong holding back the influence time has on it.