EDS241: Assignment 2

Briana Barajas

02/09/2024

Reminders: Make sure to read through the setup in markdown. Remember to write out interpretations and report your results in writing (and table/plot etc) forms.

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:

continuous:

- `mage` (mother's age),
- `meduc` (mother's education),

categorical:

- `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
birth_weight <- read_csv(paste0(data_wd, "birthweight_simple.csv")) %>%
    janitor::clean_names()
```

Table 1: Mean birth weight (g) for treatment and control (1a)

tobacco	mean_birthwgt	mean_diff_bw
0	3430.286	244.5394
1	3185.747	NA

Question (a) 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 pt]**

```
## ======================================
##          Difference in Means        ----
## ======================================
# 1. Calculate difference in mean birth weight
birth_weight %>%
  # group by treatment
  group_by(tobacco) %>%

  #calculate means
  summarise(mean_birthwgt = mean(birthwgt)) %>%

  # calculate difference in means
  mutate(mean_diff_bw = mean_birthwgt - dplyr::lead(mean_birthwgt)) %>%

  # print results in table
  kbl(caption = "Mean birth weight (g) for treatment and control (1a)") %>%
  kable_minimal()
```

```
# 2. Calculate statistical significance of mean difference
t.test(birthwgt~tobacco, data = birth_weight)
```

```
##
##  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
```

ANS: As illustrated in **Table 1**, the mean difference in birth weight of infants with smoking versus non-smoking mothers is 244.54 grams. A t-test was done to assess the statistical significance of this difference. The p-value from this test was lower than the standard significance level ($\alpha = 0.05$), meaning we can reject the null hypothesis that the difference in means is equal to 0. In other words, the difference in mean birth weight when mothers did smoke, versus those that did not is statistically significant.

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

ANS: This mean difference corresponds with the average treatment effect (ATE) of maternal smoking, assuming that the treatment effect is constant across the entire population, and that the control variables (mother's age, mother's education, etc.) have no influence on whether or not a mother smokes. The latter point refers to the ignorability assumption.

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.

```
## =========================================
##       Prep subsets for testing      ----
## =========================================

# create df with tobacco and continuous variable
continuous <- birth_weight %>%
  select(tobacco, mage, meduc)

# create df with tobacco and categorical variables
binary <- birth_weight %>%
  select(tobacco, anemia, diabete, tobacco, alcohol, mblack, first)

# create list of variable names
binary_names <- binary %>% select(-tobacco) %>% names()
continuous_names <- names(continuous)[2:3]

# create empty data frame to store results
prop_test_results <- data.frame()
t_test_results <- data.frame()

## =========================================
##       prop test for binary vars     ----
## =========================================
# create propotion tests across all binary variables
for (i in binary_names) {

  # split data into treated and untreated
  treated <- binary %>% filter(tobacco == 1) %>% pull(!!sym(i))
  untreated <- binary %>% filter(tobacco == 0) %>% pull(!!sym(i))

  # perform the prop test
  prop_test_result <- prop.test(x = c(sum(treated),
                                      sum(untreated)),

                                n = c(length(treated),
                                      length(untreated)),correct = FALSE)

  prop_test_result_tidy <- broom::tidy(prop_test_result)
  prop_test_result_tidy$Variable <- i
  prop_test_results <- rbind(prop_test_results, prop_test_result_tidy)
}
```

```r
## ======================================
##       t-test for continuous vars    ----
## ======================================
# calculate t-test across all continuous variables
for (i in continuous_names) {
  # Dynamically creating the formula for the t-test
  formula <- as.formula(paste(i, "~ tobacco"))

  # Performing the t-test
  t_test_result <- t.test(formula, data = continuous)

  # Storing the tidy results of the t-test in the data frame
  t_test_result_tidy <- broom::tidy(t_test_result)
  t_test_result_tidy$Variable <- i
  t_test_results <- rbind(t_test_results, t_test_result_tidy)
}


## ======================================
##            display results          ----
## ======================================

# combine results to single df
combine_results <- bind_rows(
  prop_test_results %>% select(Variable, estimate1, estimate2, p.value),
  t_test_results %>% select(Variable, estimate1, estimate2, p.value)
)

# create output table
combined_results_table <- kable(combine_results, format = "latex",
                                col.names = c("Variable",
                                              "Proportion or Mean Treated",
                                              "Proportion or Mean Control", "P-Value"),
                                caption = "Compare Covarite Means Treated vs. Untreated (1a)") %>%
  kable_styling(font_size = 7, latex_options = "hold_position") %>%
  kable_minimal()

# print table
combined_results_table
```

Table 2: Compare Covarite Means Treated vs. Untreated (1a)

Variable	Proportion or Mean Treated	Proportion or Mean Control	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

ANS: The proportion and t-tests were used to compare the differences in control variables in those who smoked versus those who did not smoke and are summarized in **Table 2**. With the exception of the `diabete` variable, the difference in proportions and means were statistically significant (p-value < 0.05) for all covariates assessed. For example, the proportion of women who smoked and were anemic was 0.014,

compared to a proportion of 0.007 of women who did not smoke (control) and were anemic. Comparison of continuous variables such as mother's age were calculated using a t-test instead of proportion test. For the `mage` variable, we see the average age of mother's that smoked were 27, whereas the average age of mothers that did not smoke was 25. These results suggest potential violations of the ignorability assumption, as smoking status may not be independent of the excluded control variables.

Remember that this is observational data. What other quantitative empirical evidence or test could help you assess the former assumption? [1.5 pt: 0.5 pt table, 1 pt discussion]

ANS: In this case, propensity scores can be utilized to more accurately compare the difference in birth weight for the control (non-smoking) and treatment (smoking). By matching individuals using the values of the control variables (age, anemia, education,etc.) we can compare the effects of smoking without the impact of these external variables. Additional approaches include sensitivity analysis, or difference-in-differences (DiD) analysis.

Question (b) 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]**

```
## ======================================
##             run regressions         ----
## ======================================

# ATE Regression univariate
univariate <- lm(birthwgt~tobacco, data = birth_weight)

# ATE with covariates
multiple_reg <- lm(birthwgt~ anemia + diabete + tobacco +
                     alcohol + mblack + first + mage + meduc,
                   data = birth_weight)

## ======================================
##             plot results            ----
## ======================================

# tidy results
univariate <- broom::tidy(univariate)
multiple_reg <- broom::tidy(multiple_reg)

# print results
univariate %>% kbl(caption = "Univariate Linear Regression (1b)") %>%
  kable_minimal()
```

Table 3: Univariate Linear Regression (1b)

term	estimate	std.error	statistic	p.value
(Intercept)	3430.2863	1.790782	1915.52379	0
tobacco	-244.5394	4.078907	-59.95219	0

Table 4: Multiple Linear Regression (1b)

term	estimate	std.error	statistic	p.value
(Intercept)	3362.2582445	11.9272609	281.8969309	0.0000000
anemia	-4.7963916	16.7544421	-0.2862758	0.7746675
diabete	73.2275309	12.1035155	6.0501043	0.0000000
tobacco	-228.0730765	4.1774982	-54.5956133	0.0000000
alcohol	-77.3497487	13.4653594	-5.7443508	0.0000000
mblack	-240.0303000	5.1062333	-47.0073116	0.0000000
first	-96.9441154	3.4466349	-28.1271784	0.0000000
mage	-0.6940244	0.3565637	-1.9464246	0.0516067
meduc	11.6883416	0.8604935	13.5832997	0.0000000

```
multiple_reg %>% kbl(caption = "Multiple Linear Regression (1b)") %>%
  kable_minimal()
```

```
## ========================================
##         covariate balance table     ----
## ========================================

# compute balance for multiple linear regression
covariate_balance <- xBalance(tobacco ~ anemia + diabete +
                                alcohol + mblack + first + mage + meduc,
                              data = birth_weight,
                              report = c("std.diffs","chisquare.test", "p.values"))


# print results in neat balance table
tidy.xbal(covariate_balance) %>%
  kbl(caption = "Covariate Model Balance Table (1b)") %>% kable_minimal()
```

ANS: The results demonstrate the regression coefficients for a univariate model only assessing birth

Table 5: Covariate Model Balance Table (1b)

vars	std.diff	p.value	NA.info
anemia	0.0667029	0.0000000	NA
diabete	0.0011864	0.8858011	NA
alcohol	0.3152545	0.0000000	NA
mblack	0.0843904	0.0000000	NA
first	-0.1449975	0.0000000	NA
mage	-0.3619420	0.0000000	NA
meduc	-0.6437354	0.0000000	NA

Table 6: Propensity Scores for Covariate Model (1c)

term	estimate	std.error	statistic	p.value
(Intercept)	3.4932967	0.0666128	52.441812	0.0000000
anemia	0.3339523	0.0793667	4.207713	0.0000258
diabete	0.1595334	0.0658605	2.422291	0.0154230
alcohol	2.0266407	0.0603530	33.579793	0.0000000
mblack	-0.1334468	0.0265866	-5.019324	0.0000005
first	-0.3791667	0.0193026	-19.643270	0.0000000
mage	-0.0405619	0.0019309	-21.007234	0.0000000
meduc	-0.2972694	0.0051521	-57.698672	0.0000000

weight and smoking (Table 3), compared to a multiple regression that adds additional variables (Table 4). The coefficient in the univariate model demonstrates the mean difference discussed above, and we see with the addition of smoking (control of 0 + 1), there is an average 244 gram decrease in birth weight. The coefficients for the multivariate model can be interpreted the same way. For example, for every one unit increase in mother's age (`mage`) there is an average -0.69 gram decrease in birth weight.

As for the covariate balance, the results in **Table 5** demonstrate how successful the propensity score matching was. For example, the `anemia` variable has a standardized different (std.diff) that's close to zero, and statistically significant (p-value < 0.05). This means, after matching, we can assess the impact of smoking on birth weight without the effect of anemia. The same is true for all other variables except diabetes.

Question (c) 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].**

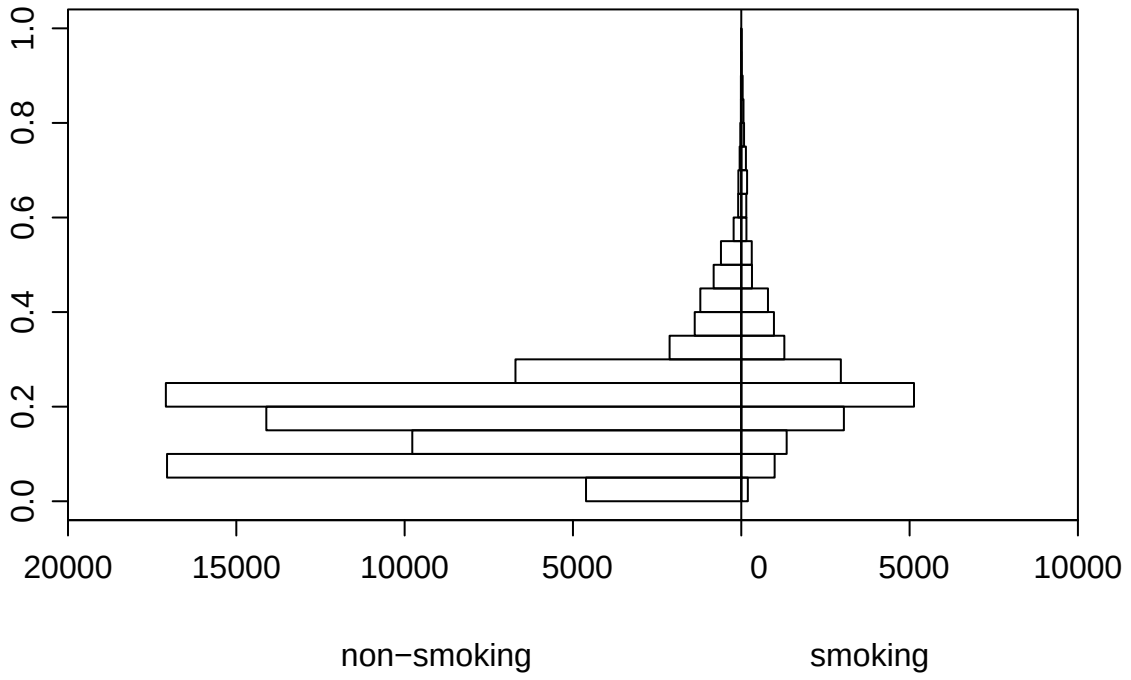
```
# calculate propensity score unmatched
propensity_scores <- glm(tobacco ~ anemia + diabete + alcohol +
    mblack + first + mage + meduc, data = birth_weight, family = binomial())

# view model summary
prop_scores_tidy <- broom::tidy(propensity_scores)
prop_scores_tidy %>%
    kbl(caption = "Propensity Scores for Covariate Model (1c)") %>%
    kable_minimal()
```

```
# add psvalue for predicted propensity score
birth_weight$psvalue <- predict(propensity_scores, type = "response")

# plot unmatched, propensity histogram
histbackback(split(birth_weight$psvalue, birth_weight$tobacco),
    main = "Fig 1. Unmatched Propensity Scores", xlab = c("non-smoking",
        "smoking"))
```

Fig 1. Unmatched Propensity Scores



ANS: The coefficients in the regression table (**Table 6**) represent the estimate propoensity score for each covariate. For example, there is a 0.133 decrease in the odds of being in the treated group (smoking), if the `mblack` variable equals 1. The unmatched propensity scores were also visualized using histograms, displayed in **Figure 1**. The right-tail skew in the non-smoking group demonstrates differences in the likelihood of being assigned to the control group, this can be corrected using matching. There is a reasonable amount of overlap, so there should be suitable matches between individuals in the control and treatment groups.

Question (d) 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].**

```
## ======================================
##        nearest neighbor matching     ----
## ======================================

# match using nearest neighbors method
match_bw <- matchit(tobacco ~ anemia + diabete + alcohol + mblack + first +
                      mage + meduc, data = birth_weight,
                    method = "nearest", ratio = 1)
```

Table 7: Covariate Balance Table - Matched (1d)

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

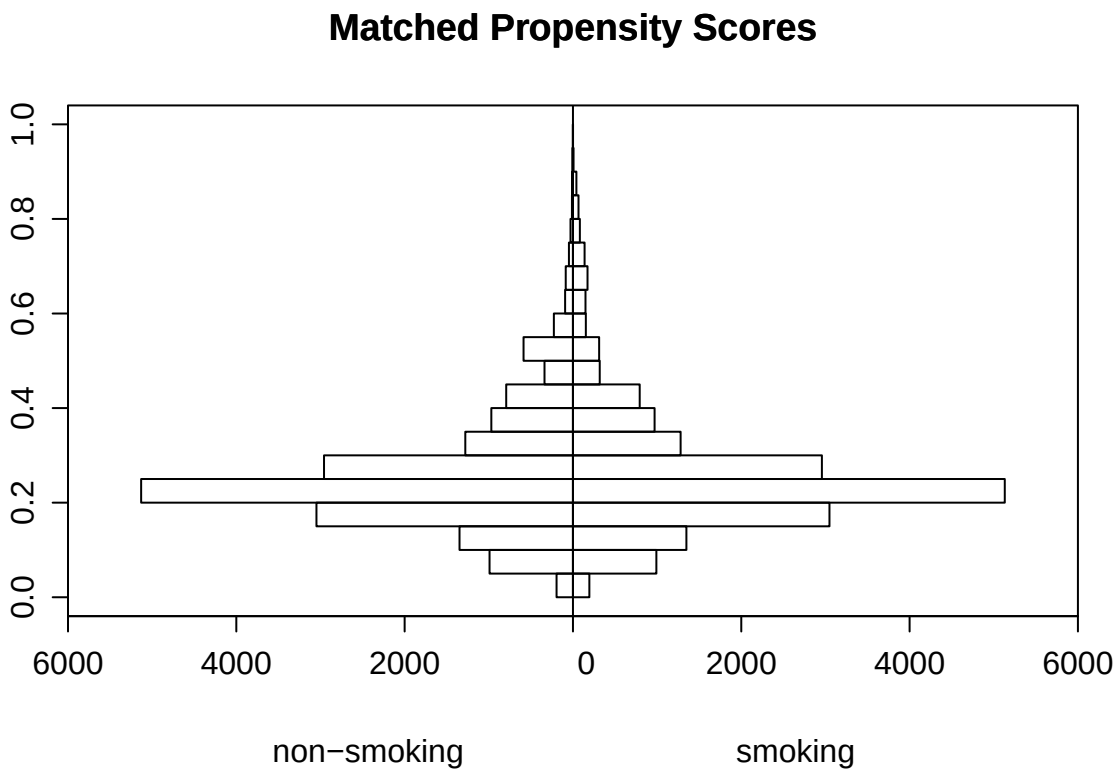
```
# store match data
match_bw_df <- match.data(match_bw)

## ========================================
##      covariate imbalance post-match ----
## ========================================
# compute post-match covariate imbalance
covariate_balance_match <- xBalance(tobacco ~ anemia + diabete + alcohol +
                                      mblack + first + mage + meduc,
                                    data = match_bw_df,
                                    report = c("std.diffs","chisquare.test", "p.values"))

# print results in balance table
tidy.xbal(covariate_balance_match) %>%
  kbl(caption = "Covariate Balance Table - Matched (1d)") %>% kable_minimal()
```

```
## ========================================
##        prop scores of matched       ----
## ========================================


# plot matched, propensity histogram
histbackback(split(match_bw_df$psvalue, match_bw_df$tobacco),
             main = "Matched Propensity Scores",
             xlab = c("non-smoking", "smoking"))
```



ANS: After conducting nearest neighbor matching, the standardized difference values (std.diff) have moved closer to 0 (compare matched **Table 7** to unmatched Table 5). The change in standardized difference values indicates that the matching was successful, and the additional covariates were matching in a way that will allow us to isolate the impact that the treatment has on birth weight. The updated histograms in **Figure 2** displays a similar result. The improvement in the histograms indicates that the matching process led to changes in the propensity scores that improved balance between groups.

Question (e) ATT 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)**

```
# calculate difference by subgroup
sumdiff_data <- match_bw_df %>%
    group_by(subclass) %>%
    mutate(pair_diff = birthwgt[tobacco == 1] - birthwgt[tobacco ==
        0])

# calculate sum of treatment column (NT)
NT <- sum(birth_weight$tobacco)

# calculate ATT
sumdiff <- sum(sumdiff_data$pair_diff)/2
ATT_nn <- 1/NT * sumdiff
```

Table 8: Weighted Least Squares Regression (1f)

term	estimate	std.error	statistic	p.value
(Intercept)	3384.224583	11.4872927	294.6059332	0.0000000
tobacco	-224.854183	3.2180177	-69.8735070	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
mage	-2.162627	0.3536926	-6.1144251	0.0000000
meduc	12.865499	0.8690655	14.8038317	0.0000000

```
# print results
paste("Estimate of ATT using matched dataset:", ATT_nn)
```

```
## [1] "Estimate of ATT using matched dataset: -222.936480828559"
```

ANS: The average treatment effect on the treated (ATT) is similar to the average treatment effect (ATE) that was conducted in Question 1a. In this case, the ATT demonstrates that when mother's smoke, their babies weigh 222 grams less on average. This value is only slightly smaller than the mean difference in birth weight of unmatched mothers. Since the ATT was calculated using matched data, we can see that smoking has a large impact on birth weight even when other factors (education, first, alcohol, etc.) are held constant.

Question (f) 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.**

```
# preliminary calculations for weighted least squares (WLS)
D <- birth_weight$tobacco
PS <- birth_weight$psvalue

# calculate weights
birth_weight$wgt <- (D/PS + (1-D)/(1-PS))

# calculate WLS estimate
wls <- lm(birthwgt ~ tobacco + anemia + diabete + alcohol + mblack + first + mage + meduc,
          data = birth_weight, weights = wgt)

# print results
broom::tidy(summary(wls)) %>%
  kbl(caption = "Weighted Least Squares Regression (1f)") %>% kable_minimal()
```

ANS: The coefficient estimates in **Table 8** represent the estimated effects of each variable on birth weight, while holding all other variables constant. This is different than the previous regressions because the coefficients are influenced by the weights assigned to each observation. For example, after accounting for differences in the weighting of observations, it is estimated that smoking causes an estimated decreases of

-244 grams when all other variables are held constant. The `tobacco` variable also has a low standard error, indicating high certainty for this estimate.

Question (g) 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)?**

ANS: The estimated coefficient for tobacco is slightly more negative for the weighted least square estimate (f) compared to the ATT (1). This occurs because the WLS model makes adjustments for all covariates included in the model. The more negative value indicates that smoking is more strongly associated with decrease in birth weight once the covariates are weighted.

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

We will use the progresa data (progresa.csv) from last time as well as a new dataset, progresa_pre.csv. 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 (year 1997). **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.*

Question (a) Estimating Effect with First Difference (3 pts)

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).

```
## Load the datasets
progresa_pre <- read_csv(paste0(data_wd, "progresa_pre_1997.csv"))
progresa_post <- read.csv(paste0(data_wd, "progresa_pre_1999.csv"))

## Append post to pre dataset
progresa <- rbind(progresa_pre, progresa_post)
```

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 for each individual and for each variable used (i.e the outcome and the independent variables).[3 pts] 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 time.

## Create first differences of variables
progresa <- progresa %>%
    arrange(hhid, year) %>%
    group_by(hhid) %>%  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
## variables.)
fd_manual <- lm(vani_fd ~ treatment, data = progresa)

# print results
broom::tidy(fd_manual) %>%
    kbl(caption = "Q2, pt(a)") %>%
    kable_minimal()
```

Table 9: Q2, pt(a)

term	estimate	std.error	statistic	p.value
(Intercept)	-1156.752	64.4938	-17.935859	0.0000000
treatment	287.905	85.6020	3.363297	0.0007723

ANS: The results of the first difference regression and displayed in **Table 9**. On average, households that received the treatment experienced a 287-unit difference in their `vani_fd` values compared to households that did not receive the subsidy. `vani_fd` represents the difference in `vani` between 1997 and 1999. The p-value for the coefficient is statistically significant (p-value < 0.05), meaning the difference in `vani_fd` between treated and non-treated households is also significant.

Question (b) Fixed Effects Estimates (2 pts)

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 remove NAs in treatment colum
within <- plm(vani ~ treatment, index = c("state", "year"), effect = "twoways",
    data = progresa)

# present regression results
broom::tidy(within) %>%
    kbl() %>%
    kable_minimal()
```

ANS: The coefficient calculated in the fixed effects regressions is likely negative because the order of subtraction was switched, with this in mind I will be comparing a positive 231 difference to the estimate of 287 calculated before (2a). For the fixed effects model, we see a smaller change in the difference in `vani` based on treatment. Although the value of the coefficient is different, it is still statistically significant. This means that the average difference in `vani` between 1997 and 1999 was 231 for treated groups.

Question (c) 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.

ANS: The first difference and fixed effects models work by preventing time from causing omitted variable bias. Between two different years, the values of specific animals may very greatly depending on supply and demand, disease outbreaks, or other cultural factors. Any factor that changes the value of the animals over time has the potential to cause omitted variable bias. By holding time constant, we can indirectly account for variables such as market changes that affect `vani`. Let's say two households have nearly equivalent values for `vani` in 1997, but household A has more sheep, and household B has more cows. Say there is a cold-spell,

and wool sales increase well into 1999, increasing the value of `vani` for household A. In this example, the fixed effects estimator considers each household separately, so it accounts for the fact that A and B have different types of animals. The FD estimator can be useful for looking in changes over time, but would not properly account for the livestock differences between household A and B.