EDS241: Assignment 2 Template

Hope Hahn

02/09/2024

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_data <- read_csv(here::here("assignment2", "data", "SMOKING_EDS241.csv"))
```

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

```
## Calculate mean difference. Remember to calculate a
## measure of statistical significance subset non smokers
non_smoking <- smoking_data %>%
    filter(tobacco == 0)
```

```
# susbset smokers
yes_smoking <- smoking_data %>%
    filter(tobacco == 1)

# calculate the difference in means between this group
mean_difference <- round((mean(non_smoking$birthwgt) - mean(yes_smoking$birthwgt)),
    2)
```

- **The mean difference in birth weight of infants between smoking and non-smoking mothers is 244.54. This corresponds to the average treatment effect under the ignorability assumption, assuming there is an equal distribution/mean of covariates among smoking and non smoking mothers, and covariates are not influencing treatment effects.**

```
## Covariate Calculations and Tables (feel free to use code
## from Assignment 1 key)

# Selecting binary and continuous variables from the
# dataset
binary_vars <- smoking_data %>%
    select(mblack, alcohol, first, diabete, anemia, tobacco)

continuous_vars <- smoking_data %>%
    select(mage, meduc, birthwgt, tobacco)

# Initialize empty data frames to store results of tests
prop_test_results <- data.frame()
t_test_results <- data.frame()

# prop tests
# --------------------------------------------------------------------
binary_names <- names(binary_vars)
for (var in binary_names) {

    # Splitting the data into smoking and nonsmoking groups
    # for the current variable
    smoking_binary <- binary_vars %>%
        filter(tobacco == 1) %>%
        pull(!!sym(var))

    nonsmoking_binary <- binary_vars %>%
        filter(tobacco == 0) %>%
        pull(!!sym(var))

    # Performing the proportion test
    prop_test_result <- prop.test(x = c(sum(smoking_binary),
        sum(nonsmoking_binary)), n = c(length(smoking_binary),
        length(nonsmoking_binary)), 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)
```

```r
}

# t-tests
# ------------------------------------------------------------------------
continuous_names <- names(continuous_vars)[1:3]
for (var in continuous_names) {

    formula <- as.formula(paste(var, "~ tobacco"))

    # t-test
    t_test_result <- t.test(formula, data = continuous_vars)

    # store tidy results of t-test in data frame
    t_test_result_tidy <- broom::tidy(t_test_result)
    t_test_result_tidy$Variable <- var
    t_test_results <- rbind(t_test_results, t_test_result_tidy)

}

# 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 Non-smoking",
        "Proportion or Mean Smoking", "P-Value"), caption = "Smoking and Non-smoking Proportion and T- 
    kable_styling(font_size = 7, latex_options = "hold_position")

# Displaying the table
combined_results_table
```

Table 1: Smoking and Non-smoking Proportion and T- Test Results Summary

Variable	Proportion or Mean Non-smoking	Proportion or Mean Smoking	P-Value
mblack	0.1354121	0.1086279	0.0000000
alcohol	0.0441825	0.0071033	0.0000000
first	0.3645879	0.4360900	0.0000000
diabete	0.0175187	0.0173636	0.8858005
anemia	0.0141031	0.0078005	0.0000000
tobacco	1.0000000	0.0000000	0.0000000
mage	27.4530853	25.5385632	0.0000000
meduc	13.2394207	11.9209454	0.0000000
birthwgt	3430.2863025	3185.7469149	0.0000000

- **Differences in covariates between smoking and non-smoking mothers are statistically significant for all covariates except for diabetes (using a significance level of alpha = 0.05). All covariates have p-values of approximately 0 while diabetes has a p-value of approximately 0.886. Because most of the covariates are statistically significant, this provides empirical evidence against our assumption, and we would reject the null hypothesis that there is no difference in distribution/means between covariates in smoking and non-smoking mothers, meaning that there are differences in baseline characteristics between**

the two groups. To test the differences in distributions of covariates among the two groups, we could also calculate propensity scores and use histograms to visualize the differences in distribution.

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]

```r
# ATE Regression univariate Regression of birthweight on
# tobacco
model_uni <- lm(birthwgt ~ tobacco, data = smoking_data)

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

# Present Regression Results
se_models = starprep(model_uni, model_cov, stat = c("std.error"),
    se_type = "HC2", alpha = 0.05)

stargazer(model_uni, model_cov, se = se_models, type = "latex",
    ci = FALSE, no.space = TRUE, header = FALSE, omit = c("Constant"),
    omit.stat = c("adj.rsq", "ser", "f"), covariate.labels = c(""),
    dep.var.labels = c("Birthwgt"), dep.var.caption = c(""),
    title = "Average Treatment Effect", table.placement = "H",
    notes = "Robust standard errors in parantheses", notes.align = "l")
```

Table 2: Average Treatment Effect

	Birthwgt	
	(1)	(2)
	−244.539***	−228.073***
	(4.150)	(4.277)
mage		−0.694*
		(0.368)
meduc		11.688***
		(0.862)
first		−96.944***
		(3.488)
mblack		−240.030***
		(5.348)
alcohol		−77.350***
		(14.039)
diabete		73.228***
		(13.235)
anemia		−4.796
		(17.874)
Observations	94,173	94,173
R^2	0.037	0.072
Note:	*p<0.1; **p<0.05; ***p<0.01 Robust standard errors in parantheses	

- **The regression results show that the ATE with no linear controls is -244.54 and is statistically significant. This means that we reject the null hypothesis that there is no mean difference in birthweight between smoking and non-smoking mothers, and on average, smoking mother's birthweight is -244.54 grams lower than non-smoking mothers. The ATE with controls added is -228.07, and is also statistically significant. Using this ATE, we also reject the null hypothesis that there is no mean difference in birthweight between smoking and non-smoking mothers, and on average, non-smoking mother's birthweight is -228.07 grams lower than non-smoking mothers. The ATE calculated using controls and without controls are different, which means that the baseline characteristics are different between the treatment groups and are influencing the effects of treatment.**

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

# Balance Table
balance_table <- kable(cov_balance, format = "latex", caption = "Covariate Imbalance before matching")
    kable_styling(font_size = 7, latex_options = "hold_position")

# Displaying the table
balance_table
```

- **These results show that there are imbalances among the covariates between the smoking and non-smoking groups, this means that the ignorability assumption is not met, and**

Table 3: Covariate Imbalance before matching

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

	chisquare	df	p.value
unstrat	7642.691	7	0

there is not an equal distribution/means in covariates and baseline characteristics between the two 'treatment' (smoking/non-smoking) groups. The chi squared value tells us to reject the null hypothesis that there are no differences in baseline characteristics between the two groups, as well as the p-values for all covariates except for diabetes.

Question (c) Propensity Score Estimation *(3 pts)*

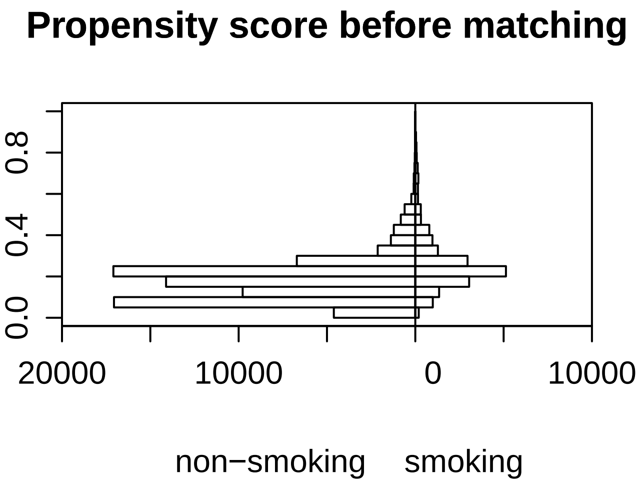
c) Next, estimate propensity scores (i.e. probability of being treated) for the sample, using the provided covariates. Create a regression table reporting the results of the regression and discuss what the covariate coefficients indicate and interpret one coefficient [1.5 pts]. Create histograms of the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

```r
## Propensity Scores
ps <- glm(tobacco ~ mage + meduc + first + mblack + alcohol +
    diabete + anemia, data = smoking_data, family = binomial())

# use gtsummary to make table
library(gtsummary)
# print the table in a nice format
ps %>%
    tbl_regression()
```

Characteristic	**log(OR)**	**95% CI**	**p-value**
mage	-0.04	-0.04, -0.04	<0.001
meduc	-0.30	-0.31, -0.29	<0.001
first	-0.38	-0.42, -0.34	<0.001
mblack	-0.13	-0.19, -0.08	<0.001
alcohol	2.0	1.9, 2.1	<0.001
diabete	0.16	0.03, 0.29	0.015
anemia	0.33	0.18, 0.49	<0.001

- **The covariate coefficients indicate the probability/odds that an individual will be a part of a specific 'treatment' group based off of baseline characteristics. For example, the coefficient for meduc is is -0.30, this means that with a 1 unit increase in mother education level, the log(odds) of the mother being a smoker decreases by 0.30.**

```r
## PS Histogram Unmatched
smoking_data$psvalue <- predict(ps, type = "response")

histbackback(split(smoking_data$psvalue, smoking_data$tobacco),
    main = "Propensity score before matching", xlab = c("non-smoking",
        "smoking"))
```



- **There is partial overlap in the histograms between the non-smoking and smoking groups. This means that the range of distributions within the smoking group is only part of the non-smoking group, and there is an imbalance among the different groups (they are not evenly matched). The control group does not contain good counterfactuals for the treated group.**

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].

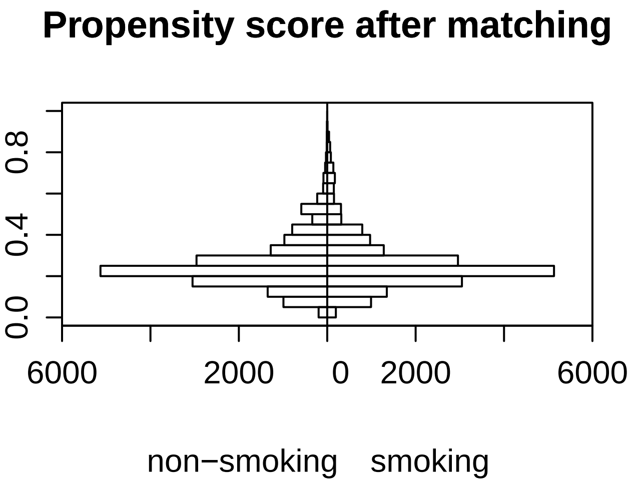
```r
## Nearest-neighbor Matching
m.nn <- matchit(tobacco ~ mage + meduc + first + mblack + alcohol +
    diabete + anemia, data = smoking_data, method = "nearest",
    ratio = 1)
match.data = match.data(m.nn)

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

# Balance Table
matching_balance_table <- kable(matching_balance, format = "latex",
    caption = "Covariate Imbalance after matching") %>%
    kable_styling(font_size = 7, latex_options = "hold_position")

# Displaying the table
matching_balance_table
```

Table 4: Covariate Imbalance after matching

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

	chisquare	df	p.value
unstrat	124.2445	7	0

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



- **Compared to the previous propensity score histogram before matching, there is a higher overlap and the distributions are much more balanced after matching. Matching allowed the counterfactuals to match the treatment group to eliminate the effects of covariates on the estimated treatment effects. The chi square value and p-scores of most variables (all but meduc and alcohol) show that the differences between baseline characteristics are not statistically significant.**

Question (e) 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)

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

sumdiff <- sum(sumdiff_data$diff)/2
```

```r
## ATT
NT <- sum(smoking_data$tobacco)
ATT_m_nn = 1/NT * sumdiff
ATT_m_nn
```

```
## [1] -222.9365
```

```r
att_match_table <- kable(ATT_m_nn, format = "latex", caption = "ATT after matching") %>%
    kable_styling(font_size = 7, latex_options = "hold_position")
```

- **The ATT using the matching dataset is -222.9364808. This means that the average effect of treatment among the treated group is -222.9364808, so smoking mothers birthweight is, on average, -222.9364808 grams lower than the non-treated. This estimate is the ATT and not ATE because when we used nearest neighbors matching, we matched for the smoking (treated) group, and we did not perform matching for the non-treated group (non-smoking).**

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.

```r
## Weighted least Squares (WLS) estimator Preparation
PS <- smoking_data$psvalue
Y <- smoking_data$birthwgt
D <- smoking_data$tobacco

smoking_data$wgt = (D/PS + (1 - D)/(1 - PS))

## Weighted least Squares (WLS) Estimates
reg_wls_c <- lm(birthwgt ~ tobacco + mage + meduc + first + mblack +
    alcohol + diabete + anemia, data = smoking_data, weights = wgt)


## Present Results Present Regression Results
wls_table <- kable(summary(reg_wls_c)$coefficients, format = "latex",
    caption = "ATE with WLS matching") %>%
    kable_styling(font_size = 7, latex_options = "hold_position")

wls_table
```

- **The WLS estimate of ATE is -224.854. This estimate tells us that the average birthweight is -224.854 grams less among smoking mothers than non-smoking mothers. The standard error is approximately 3, which is low, meaning that this estimate is relatively precise.**

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)?

Table 5: ATE with WLS matching

	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
first	-89.986297	3.4817872	-25.8448585	0.0000000
mblack	-238.022495	4.9713029	-47.8792981	0.0000000
alcohol	-69.215047	13.6868845	-5.0570345	0.0000004
diabete	63.306452	12.1117043	5.2268823	0.0000002
anemia	10.811567	16.6819881	0.6480982	0.5169230

- **In e, we calculated ATT which was the average treatment effect among the treated group because we calculated counterfactuals based on the characteristics of the treatment group. However, in f, we calculated the average treatment effect among the whole population with WLS, which runs the regression while controlling for covariates. Because of this, the average treatment effects were calculated using different methods, and the resulting effects are different**

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.

Question (a) 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(here::here("assignment2", "data",
    "progresa_pre_1997.csv"))
progresa_post_1999 <- read_csv(here::here("assignment2", "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!) *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 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)

# make table
fd_table <- kable(summary(fd_manual)$coefficients, format = "latex",
    caption = "Manual first difference") %>%
```

```r
    kable_styling(font_size = 7, latex_options = "hold_position")

fd_table
```

Table 6: Manual first difference

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1156.752	64.4938	-17.935859	0.0000000
treatment	287.905	85.6020	3.363297	0.0007723

- **The manual first different calculation results show that a post treatment, the value of animal holdings of a household is 287.9 higher than the pre treatment, on average. However, the standard error is 85.6, which means that these results are not incredibly precise.**

Question (b) 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!)

```r
## Fixed Effects Regression
within1 <- plm(vani ~ treatment, index = c("state", "year"),
    model = "within", effect = "twoways", data = progresa)

## Present Regression Results Calculate standard errors
## (note slightly different procedure with plm package)
se_within1 <- coeftest(within1, vcov = vcovHC(within1, type = "HC2",
    method = "white1"))[, "Std. Error"]
# Reformat standard errors for stargazer()
se_within1 <- list(se_within1)
# Output results with stargazer
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
```

- **With the fixed effects estimate, the value of animal holdings of a household is 231.844 lower when treated versus untreated. The standard error for this estimate is 56.66, meaning that this estimation is slightly more precise than the FD estimation, but is still not very precise. The differences between the FD and FE estimation are drastically different as the FE estimator estimates a decrease in value with treatment while the FD estimator estimates an increase.**

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 omitted 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.

- **The FD estimator solves a specific omitted variable problem because if an omitted variable does not change over time, then any changes in Y over time cannot be caused by the omitted variable. The FE estimator holds the parameters constant to account for omitted variables. An example of a potential omitted variable for this example could be quality or size of homes/land in certain areas (maybe larger houses/land can accommodate more animals.) In this case FD would be better because these variables would differ per area, but would be constant over time.**