EDS241: Assignment 2

Briana Barajas

02/06/2024

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")) %>%
  clean_names()
```

Table 1: Subtracted mean birthweight for smoking mothers from mean birthweight for non-smoking mothers to keep the mean difference positive.

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]***

```r
# calculate the mean difference in birthweight
birth_weight %>%
  group_by(tobacco) %>%
  summarise(mean_birthwgt = mean(birthwgt)) %>%
  mutate(mean_diff_bw = mean_birthwgt - dplyr::lead(mean_birthwgt)) %>%
  kbl(caption = "Subtracted mean birthweight for smoking mothers from mean birthweight for
      non-smoking mothers to keep the mean difference positive.") %>%
  kable_minimal()
```

```r
# calculate statistical significance of difference in mean birthweights
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. The mean values for smoking (`tobacco` = 1) and non-smoking (`tobacco` = 0) mothers, as well as the difference between these means (`mean_bw_diff`) is summarized in *Table 1*. The difference in birth weight is statistically significant ($p < 0.05$), so we can reject the null hypothesis that the difference in means is equal to 0.

|

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 maternal smoking or birth weight. 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.

```r
## ==========================================
##          subset control variables    ----
## ==========================================
# continuous variables in birth_weight df
cont_birth_df <- birth_weight %>%
  select(birthwgt, meduc, mage)

# categorical variables in birth_weight df
cat_birth_df <- birth_weight %>%
  select(-c(meduc, mage))

# create list of variable names
cont_birth_vars <- c("mage", "meduc")
cat_birth_vars <- c("anemia", "diabete", "tobacco", "alcohol", "mblack", "first")

## ==========================================
##          t-test on continuous vars ???   ----
## ==========================================
# initialize empty df for results
t_test_birth <- data.frame()

# use for-loop to apply t-test to each continuous variable and full the bw_t_test results table

for(i in cat_birth_vars){

  # create formula to iterate through variables in t-test
  formula <- as.formula(paste("birthwgt ~", i))

  # apply t-test, and fill results df
  t_test_birth <- t.test(formula, data = cat_birth_df)

  # tidy t-test results
  #t_test_birth <- broom::tidy(t_test_birth)

}
```

```r
continuous_vars <- names(pretreat_continuous)[grepl("97$", names(pretreat_continuous))]

for (var in continuous_vars) {
  # Dynamically creating the formula for the t-test
formula <- as.formula(paste(var, "~ treatment"))

# Performing the t-test
t_test_result <- 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) t_test_result_tidy$Variable <- var
t_test_results <- rbind(t_test_results, t_test_result_tidy)
}
```

ANS:

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]

```
# calculate the statistical significance of the difference in mean
```

```
## Calculate mean difference. Remember to calculate a measure of statistical significance

## For continuous variables you can use the t-test
#t.test()

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

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

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]*

```
# ATE Regression univariate


# ATE with covariates


# Present Regression Results


# Covariate balance


# Balance Table
```

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].*

```
## Propensity Scores


## PS Histogram Unmatched
```

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

## Covariate Imbalance post matching:


## Histogram of PS after matching
```

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)*

```
## Nearest Neighbor

## ATT
```

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.*

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


## Weighted least Squares (WLS) Estimates


## Present Results
```

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)?*

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)*

Load the new baseline data (pre-program) and the follow-up data (post-program, from Assignment 1) into R. Create a time denoting variable (with the same name) in BOTH datasets with a value of 0 for the pre-program dataset and 1 for the other one. 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)-=. Estimate a standard difference-in-differences (DiD) regression and interpret the results.

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

## Load the datasets
# progresa_pre <- read.csv() insert your filepath etc
# progresa_post <- read.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 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 ~ ...)
```

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

## Present Regression Results
```

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.*