EDS241: Assignment 2 Template

Elliott Finn

02/02/2024

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
```

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

## 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 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() insert your filepath etc
# progresa_post_1999 <- read_csv()

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

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

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

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

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

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

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!)

```
## 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.