INFO371 Problem Set 2: Diff-in-Diff Estimator

Your name:

April 2, 2022

Introduction

This week your task is (again!) to estimate the impact of the *Progresa* program, a government social assistance program in Mexico, using real data. However, this time you will use differences-in-differences estimator.

This program, as well as the details of its impact, are described in Schultz (2004) (available on Canvas). The data (*progresa-sample.csv.gz*, the same dataset as last week) is available on canvas in files/data. Please consult the explanations from the last week problem set for description of the program and variables. To put it briefly: from beginning of 1998, the families who were considered poor in certain villages (progresa villages) received a government subsidy given their kids attended school. The progresa/non-progresa villages were chosen randomly.

The goal of this problem set is to

- refresh your t-test;
- implement and use the double-differences estimator

Your task is to estimate the impact of progresa subsidies on the school attendance.

Please submit a) your code (rmd) and b) the lab in a final output form (html or pdf).

Always explain and comment your results, *do not expect* the grader is able to pick the correct number out of many with no further explanations.

Working together is fun and useful but you have to submit your own work. Discussing the solutions and problems with your classmates is all right but do not copy-paste their solution! Please list the students you collaborated with.

1 Was the randomization done correctly? (30pt)

Your first task is to analyze whether randomization was performed correctly. Perfect randomization ensures that the treatment group and the control group are similar. This is less important in terms of observable characteristics, but very-very important for unobservables. Obviously, we can only analyze the observables: are the pre-treatment (1997) demographic and village-related characteristics for the poor equal (in average) in treatment and control villages?

1. (10pt) Present your results in a single table with the following columns and 14 (or so) rows. The table should look something like this:

Variable name	Average (T)	Average (C)	Difference (T − C)	p-value
sex	0.519	0.505	0.014	0.012
indig	...	...	...	...
...				

You can see a very similar table in Adams-Prassl *et al.* (2020), Table A8, page 33 (just for males/females, not for villages).

Suggestion: use t-test to determine whether the difference between T and C villages is statistically significant for each of the variables in the dataset. Focus only on the data from 1997 for poor. Ignore variables such as *folnum* and *village* that do not carry social significance. t-test can be done with `t.test` in R. For instance `t.test(x, y)` compares two unpaired vectors `x` and `y`, and outputs confidence intervals, p-value, and other things.

Suggestion 2: There are many ways you can create this table, here is one suggestion you may follow:

(a) create an empty data frame that contains the values you need: variable name, average for T, average for C, their difference, and p-value from t-test. In R, you can also just create a `NULL`-object:

```
df <- NULL
```

This will be your final data frame, the one you will print.

(b) do a for-loop over all the variable names, data for which you want to compare. Inside the loop:

(c) extract the variable (your loop variable) values for T and C group as separate vectors

(d) compute averages of these two vectors

(e) compute p-value for t-test between these two vectors

(f) create a new one-line data frame (one row of the required table) that contains the values for this variable only.

(g) attach this new one-line data frame to the final data frame (the one you created as empty). You can use `rbind` as

```
df <- rbind(df, newDF)
```

As a result, your data frame will contain all the required values for all the variables. You can use `knitr::kable` or `xtable` library to print it in a nicer way. You can use argument `check.names=FALSE` to create a data frame with non-syntactic variable names, e.g. ones that contain space like "p value".

2. (4pt) Did you find any statistically significant differences between treatment and control villages?

3. (8pt) Why do we focus on 1997 differences only?

4. (8pt) Why does it matter if treatment and control villages differ?

2 Measuring impact (45pt)

Next, we measure the impact of Progresa. You already did it with CS and BA estimator, so the only thing to do is essentially to repeat the previous with DiD estimator. DiD relaxes the identifying assumptions but you pay the price in the form of stricter data requirements and more complex analysis.

1. (2pt) First, let's just compare group averages. Now you need four groups: treated and control, before and after treatment. DiD is the difference in the trends for treated and control groups. Compute these group averages and the corresponding DiD estimator.

   Hint: it should be 0.0313.

Now it's time to introduce regression. You should regress the schooling outcome on pre/post reform indicator (i.e. year), treatment/control indicator (i.e. progresa), and their corresponding interaction effect. You should only include the treatment and control groups (i.e. poor), not "rich" families.

2. (4pt) Estimate the effect using DiD simple regression (no other covariates).
3. (4pt) Interpret all the coefficients.
4. (4pt) Report the result: it should be the same as above. Is it statistically significant?
5. (3pt) Now estimate the effect using multiple regression–include all relevant control variables.
6. (3pt) Compare the results. Is the multiple-regression version similar? Is it statistically significant?
7. (5pt) What are 95% confidence intervals for this estimator? Does this encompass all other estimates you received in this and in the previous PS?

   Hint: the upper boundary of 95% CI should be 0.0409.

8. (8pt) What is the identifying assumption behind this DiD estimator? Would you be able to test it to a certain extent using the dataset here? Explain!

   Hint: what do you expect to see when comparing different villages? The same villages over time? Do you have this information in these data?

9. (8pt) Compare this assumption with the assumptions behind CS and BA estimator. Which ones do you find more plausible? Why?

   Base your claims in the institutional settings: it is possibly imperfect randomized experiment in poor rural villages. Do you think some assumptions are more likely satisfied that others?

10. (4pt) Based on all your work you did above–what is your conclusion about the efficacy of the Progresa program?

---

3 Leidner *et al.* (2021) (25pt)

Leidner *et al.* (2021) (available on canvas in files/readings) analyse the effect of online versus in-person instruction from COVID-19 perspective; they find that online instructions help to avoid a substantial number of cases.

1. Read the paper.
2. (3pt) Let us focus on the unmatched analysis. What kind of counties are the authors comparing in the paper? How many counties do they have? How many of these had remote, how many in-person instructions?
3. (2pt) Which time period are they looking at?
4. (4pt) What is treatment in this case? (You can define it in different ways).
5. (3pt) What is the main outcome measure the authors discuss?
6. (4pt) Why do authors analyze percentage positive results in testing?
7. (4pt) The authors provide the figures for COVID-19 incidence (in table). Take the numbers and use those to do DiD youself! Show the calculations and the answer.

   Note: You don't really need computers, the table provides the four averages, so you only have to compute the differences of those.

   Hint: effect should be 11.8 (or -11.8, depending which way you define it).
8. (5pt) What are the identifying assumptions? Do you find these credible?

   Hint: read about the limitations of the study and think which limitations are about identification.

Finally

. . . tell how much time (hours) did you spend on this PS. Feel free to add other feedback.

References

Adams-Prassl, A., Callison-Burch, C., Hara, K., Milland, K., Savage, S. and Bigham, J. (2020) The gender wage gap in an online labour market: The cost of interruptions, university of Oxford.

Leidner, A. J., Barry, V., Bowen, V. B., Silver, R., Musial, T., Kang, G. J., Ritchey, M. D., Fletcher, K., Barrios, L. and Pevzner, E. (2021) Opening of large institutions of higher education and county-level COVID-19 incidence United States, july 6september 17, 2020, *Morbidity and Mortalilty Weekly Report*, **70**, 14–19.

Schultz, T. P. (2004) School subsidies for the poor: evaluating the Mexican Progresa poverty program, *Journal of Development Economics*, **74**, 199 – 250, new Research on Education in Developing Economies.