INFO371 Problem set 1: Estimating causal effect with BA and CS method

April 2, 2022

Introduction

Your first real task is to estimate the impact of the *Progresa* program, a government social assistance program in Mexico, using real data. This program, as well as the details of its impact, are described in Schultz (2004) (available on Canvas). The data (*progresa-sample.csv*) is available on canvas in files/data.

Please read the paper to familiarize yourself with the PROGRESA program before beginning this problem set, so you have a rough sense of where the data are coming from and how they were generated.

The goal of this problem set is to make you familiar with the simple estimators that you are learning in class (cross-sectional and before-after), and to use those to measure the impact of Progresa on secondary school enrollment rates. Your task is to estimate the impact of progresa subsidies on the school attendance. Note: this means to estimate the causal effect.

Please submit a) your code (as rmd) and b) the compiled output (html). 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. While some of the intermediate output may be informative, please don't include too many lines of it in your solutions!

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 all your collaborators below:

1.

2. ...

About *Progresa* program

In 1990s, Mexican government decided to improve the school attendance of poor rural children by introducing a cash subsidy to families. However, the families were only able to claim the money if a) they were considered poor, and b) if their children attended school. Most importantly in the current context, the subsidy was introduced in a randomized manner where initially only certain villages were eligible for subsidies. In this problem set we analyze this time period where the subsidies formed essentially a randomized control trial. Read more in Schultz (2004).

The timeline of the program was:

variable name	description
year	year in which data is collected
sex	male = 1
indig	indigenous = 1
dist_sec	nearest distance to a secondary school
sc	enrolled in school in year of survey (=1)
grc	grade enrolled
fam_n	family size
min_dist	min distance to an urban center
dist_cap	min distance to the capital
poor	poor = "pobre", not poor = "no pobre"
progresa	treatment = "basal", control = "0"
hohedu	years of schooling of head of household
hohwag	monthly wages of head of household
welfare_index	welfare index used to classify poor
hohsex	gender of head of household (male=1)
hohage	age of head of household
age	years old
folnum	individual id
village	village id
sc97	enrolled in school in 1997 (=1)

Table 1: Variables in the data, collected for each child each year (1997, 1998).

- Baseline survey conducted in 1997
- Intervention–subsidies for *poor households* in *treatment villages* begins in 1998, wave 1 data collected in 1998
- wave 2 data collected in 1999
- Evaluation ends in 2000, at which point all villages become eligible to the subsidy.

Note that:

- Progresa program was only available for poor families, so in the analysis below we only consider poor household.
- The central variable here is *sc*, the dummy telling if the child did attend the school or not.

When you are ready, download the progresa-sample.csv data from Canvas. The data are actual data collected to evaluate the impact of the Progresa program. In this file, each row corresponds to an observation taken for a given child for a given year. There are two years of data (1997 and 1998), and just under 40,000 children who are surveyed in both years. Table 1 describes the variables in the dataset.

1 Graphical exploration (20 pt)

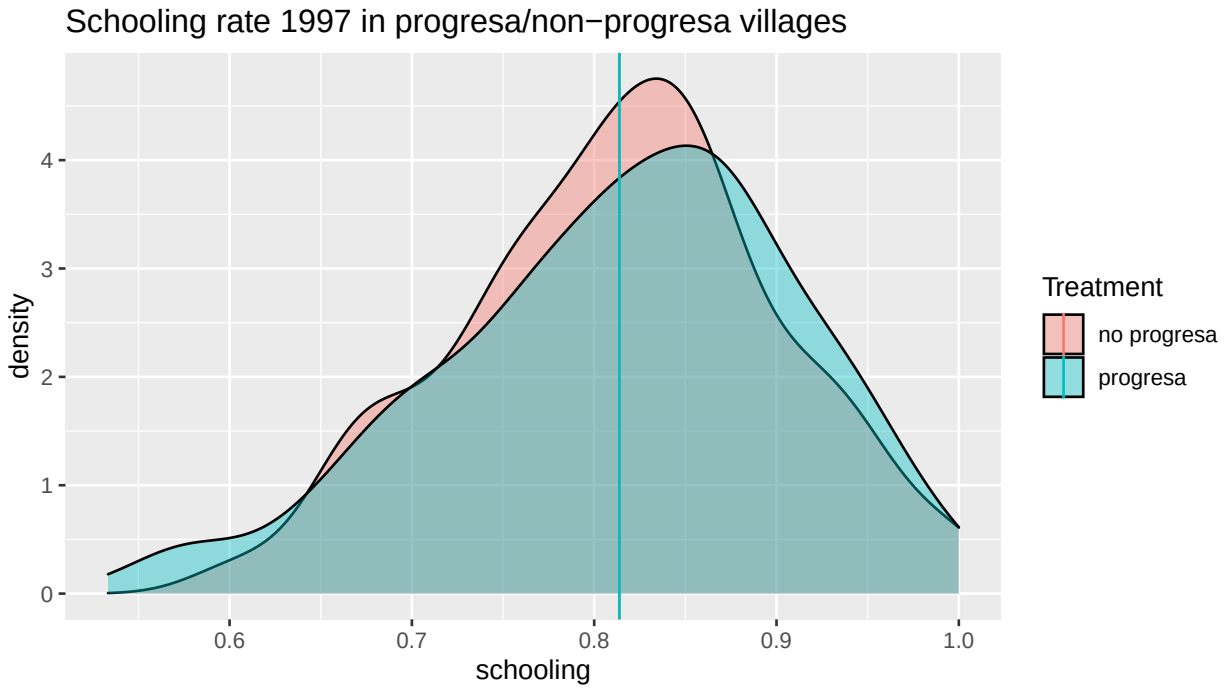
Before we get into regression, it is worthwhile to have visual image of the data.

1. (4pt) Load data. How many cases do we have? How many different villages? How many cases of poor in progresa villages?

2. (4pt) Compute average schooling rate of poor household by villages (you can use village id as the grouping variable) for 1997 and 1998. Compare it between progresa villages, and in non-progresa villages in 1997 and 1998. Here just report the averages, you'll do a graphical comparison of distributions below.

   Note: this asks you to compare the schooling rate *by village*, i.e. you need a single number (avg schooling rate) for each village. Thereafter, you should compare *averages of village averages*.

3. (4pt) Display the average schooling rate before the program (1997) separately for progresa/non-progresa villages. Mark sample average rate (separately for progresa/non-progresa villages) on the figure. Attempt to overlay these density estimates. You can try to replicate this example.



   Hint: ggplot's `geom_density` makes such density plots, you can add transparency by `alpha`.

4. (4pt) Repeat for the program year (1998)

5. (4pt) Comment the results. Do the distributions look similar? Do you see the schooling rate in progresa villages increasing over that of the control villages?

---

2 Measuring impact

Next, we measure the impact of Progresa. We do it in two ways: first using the cross-sectional estimator, and thereafter by before-after estimator. Both estimators we implement in turn in three ways: a) just table of averages; b) simple regression where we only introduce control/treatment group (or time in case of before-after estimator); c) multiple regression.

2.1 Cross-sectional (CS) estimator (40pt)

CS estimator compares data for treated (poor in progresa villages) and non-treated controls (poor in non-progresa villages) after the treatment (i.e. 1998). We start with a simple table.

1. (3pt) What is the identifying assumption behind this CS estimator? Do you think these are satisfied here? Explain!

   Hint: see lecture notes Ch 5.5.1 "Counterfactual and Identifying Assumption" and 5.5.2 "A Few Popular Estimators".

2. (3pt) Why do we look at only poor households, and only year 1998?

3. (4pt) compute average schooling rate (variable *sc*) for treated and non-treated controls after the program. Compare these means. How big effect do you find?

   Hint: it should be 3.88 pct points.

4. (5pt) Based on this number, can you claim progresa was effective (i.e. it increased schooling rate)? Interpret the number (in terms of percent points increase or decrease).

Reading the result from the table is an easy and intuitive approach but it does not provide any standard errors and statistical significance estimates. It is also hard to include other relevant characteristics that may influence the effect size. Linear regression helps here.

5. (5pt) Implement the CS estimator using linear regression: regress the outcome after treatment on the treatment indicator. Do not include any other controls (except the intercept).

   If you know how to do it the go ahead in your own way. But if you need a little help then you can follow these steps:

   (a) Ensure you are only comparing the relevant groups: the control group that was not treated, and the treatment group that was actually treated.

   (b) Create a dummy variable T that tells if someone is in the treatment or control group.

   (c) Regress the outcome on T.

6. (3pt) Compare the results. You should get exactly the same number as when just comparing the group means.

7. (2pt) Is the effect statistically significant?

So far we ignored the other relevant covariates. If the experiment was conducted correctly, those should not matter. But if randomization was imperfect, it may not be the case.

8. (5pt) Estimate the multiple regression model. Include all covariates, such as education, family size and whatever else you consider relevant for the current case.

9. (5pt) Compare the results. Do other covariates substantially change the results?

2.2 Before-After Estimator (40pt)

(5pt each, except question 5)

Instead of comparing treatment and control villages in 1998, we can also compare just treatment villages after (1998) and before (1997) the program was introduced. We follow fairly similar steps as what you did above.

1. (3pt) What is the identifying assumption behind this estimator? Do you think they are fullfilled? Explain!

2. (3pt) Why do we have to select only progresa villages and only poor for this task?

3. (4pt) compute average schooling rate (variable *sc*) for the poor for the treated villages before and after the program. Compare these means. How big effect do you find?

   Hint: it should be 2.38 pct points.

4. (5pt) Based on this number, can you claim progresa was effective (i.e. it increased schooling rate)? Interpret the number (in terms of percent points increase or decrease).

Next, do the same with linear regression:

5. (5pt) Implement the BA estimator using linear regression: regress the outcome for the treated group on the after-program indicator. Do not include any other controls (except the intercept).

   If you know how to do it the go ahead in your own way. But if you need a little help then you can follow these steps:

   (a) Ensure you are only comparing the relevant groups: the control group is before and treatment group is after the policy was implemented.

   (b) Create a dummy variable *After* that tells if we are looking the period were the policy is already there.

   (c) Regress the outcome on *After*.

6. (2pt) Compare the results. You should get exactly the same number as when just comparing the group means.

7. (3pt) Is the effect statistically significant?

So far we ignored other relevant covariates. If the identifying assumptions were correct, those should not matter. But if not, this may not be the case.

8. (5pt) Estimate the multiple regression model. Include all covariates, such as education, family size and whatever else you consider relevant for the current case.

9. (5pt) Compare the results. Do other covariates substantially change the results?

10. (5pt) Comment the identifying assumptions behind the CS and BA models. Which one do you find more convincing?

---

3 Finally

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

References

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.