IMT574 Problem Set 3: Causality (100pt)

Your name:

Deadline: Feb 5th

1. Please write clearly! Answer each question in a way that if the code chunks are removed from your document, the result is still readable!
2. Please keep data file in the same folder as your code, and read these w/o any path like `"data.csv"` (or `"./data.csv"`). This makes it much easier to check your code!
3. If you use notebooks, upload both the .ipynb and html/pdf file.

Instructions

The goal of this problem set is to get experience with estimation of causal effects, in particular using the differences-in-differences (DiD) method. This is a very common task in economics/government/business analytics. Your task is to estimate the impact of progresa subsidies on the school attendance using the actual data.

Progresa was a a government social assistance program in Mexico. This program, as well as the details of its impact, are described in the paper "School subsidies for the poor: evaluating the Mexican Progresa poverty program", by Paul Shultz (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 come from and how they were generated. If you just proceed into the problem set without understanding Progresa or the data, it will be very difficult! I also recommend you to consult the section 3.3 in lecture notes.

The timeline of the program was:

- Baseline survey conducted in 1997
- Intervention begins in 1998, "Wave 1" of data collected in 1998
- "Wave 2 of data" collected in 1999
- Evaluation ends in 2000, at which point the control villages were treated.

The data are the 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. For each child-year observation, the following variables are collected:

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)

You may load the following packages

```
import pandas as pd
import numpy as np
import scipy.stats as stats
import statsmodels.formula.api as smf
```

1 Descriptive analysis (30pt)

1.1 Summary statistics (10pt)

First, learn about data.

1. Report summary statistics (mean, standard deviation, and number of missings) for all of the demographic variables in the dataset (i.e., everything except year, folnum, village). A central variable, *progresa* is coded in a rather unintuitive way. Find it's actual coding scheme. Does this fit with the documentation above?

   Present these in a single table alphabetized by variable name. Do NOT simply expect the grader to scroll through your output!

1.2 Differences at baseline? (20pt)

Now let's investigate the differences in baseline. Are the baseline (1997) demographic characteristics for the poor different in treatment and control villages?

1. (4pt) Use t-test to determine whether there is a statistically significant difference in the average values of each of the variables in the dataset. Focus only on the data from 1997 for poor.

2. (4pt) Present your results in a single table with the following columns and 14 (or so) rows:

Variable	Avg (Treatment villages)	Avg (control)	difference	p-value
sex	0.5193	0.5051	0.0143	0.0122
...				

3. (4pt) Do you find any statistically significant differences between treatment and control villages as baseline?

4. (4pt) Why does it matter if there are differences at baseline?

5. (4pt) What does this imply about how to measure the impact of the treatment?

2 Measuring Impact

Our goal is to estimate the causal impact of the Progresa program on the schooling outcomes of individuals in Mexico. We will focus on the impact of the program on the poor, since only the poor were eligible to receive the Progresa assistance.

2.1 Before-after estimator (10pt)

First, implement the before-after estimator. Compare the schooling rate of poor households in progresa villages before (i.e. 1997) and after (i.e. 1998) the program.

1. (2pt) compute the estimator by just comparing the average schooling rates for these villages.

2. (3pt) now re-compute the estimator using linear regression, and individual schooling rates. Do not include other regressors.

3. (3pt) finally, estimate a multiple regression model that includes other covariates.

4. (2pt) compare all the estimators. Are your estimates statistically significant? What do they suggest about the efficacy of the progresa program?

2.2 Cross-sectional estimator (10pt)

Now let's implement the cross-sectional estimator. Proceed along the same lines as what you did above.

1. (2pt) Begin by estimating the impact of Progresa by compring the average enrollment rate among poor households in the treatment villages and the average enrollment rate among poor households in the control villages. What do you find?

2. (3pt) Now repeat the estimator using simple regression.

3. (3pt) Third, use multiple regression to get the same estimate.

4. (2pt) Finally, as above, compare your three estimators. What do you find? Are the effects statistically significant?

2.3 Differences-in-differences estimator (30pt)

Now we are ready for DiD estimator. Proceed along the same lines as above.

1. (6pt) Start with the simple table. However, DiD requires 4-way comparison. So compare the average enrollment rate among poor households in the treatment villages and the average enrollment rate among poor households in the control villages, both 1997 and 1998. What do you find?

2. (10pt) Now repeat the estimator using simple regression.
3. (8pt) And as above, use multiple regression to get the same estimate.
4. (6pt) Finally, as above, compare your three estimators. What do you find? Are the effects statistically significant?

2.4 Compare the estimators (20pt)

Now you have used three estimators to assess the effect of Progresa program.

1. (10pt) List the identifying assumptions (counterfactual assumptions) behind all three models. Which ones do you find more/less plausible?
2. (10pt) Compare the estimates of all three models. Do your analysis suggest that progresa program had a positive impact on schooling rates?