VERDICKT_JULIA-PS2

February 5, 2024

1 Problem Set 2

1.1 Introduction to the assignment

For this assignment, you will be using data from the Progresa program, 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 bCourses). Please 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!

The goal of this problem set is to implement some of the basic econometric techniques that you are learning in class to measure the impact of Progresa on secondary school enrollment rates. The timeline of the program was:

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

When you are ready, download the progresa_sample.csv data from bCourses. 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 each year. For each child-year observation, the following variables are collected:

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
grc	grade enrolled
fam_n	family size
min_dist	min distance to an urban center
dist_cap	min distance to the capital
poor	poor = 1
progresa	treatment =1

Variable name	Description
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	schooling in 1997
grc97	grade enrolled in 1997

1.2 Part 1: Descriptive analysis

1.2.1 1.1 Summary Statistics

Present summary statistics (mean, median and standard deviation) for all of the demographic variables in the dataset (i.e., everything except year, folnum, village). Present these in a single table alphabetized by variable name. Do NOT simply expect the grader to scroll through your output!

Note: For this and subsequent problems, you will need to be careful in how you deal with missing (NULL) values. You should not blindly drop rows and columns where any data field is missing. For instance, in calculating the average `hohwag`, you should average the `hohwag` values from all households that report a value (even if the household does not have a recorded `age` value, for example).

[ ]:
```
# your code here
import pandas as pd
import numpy as np
import scipy.stats as stats
import warnings
from IPython.display import Image
import dataframe_image as dfi

warnings.filterwarnings('ignore')

progresa_df = pd.read_csv("progresa_sample.csv")
progresa_df['progresa_num'] = progresa_df['progresa'].replace({'basal': 1, '0':␣
↪0})
progresa_df['poor_num'] = progresa_df['poor'].replace({'pobre': 1, 'no pobre':␣
↪0}).astype(int)

summ_stat = progresa_df.drop(['year','folnum','village'], axis = 1)
summ_stat = summ_stat.describe().transpose().reset_index(names = "variable")
```

```
summ_stat = summ_stat[['variable',
                       'mean',
                       '50%',
                       'std']].rename(columns = {'50%': 'median',
                                                 'std': 'standard deviation'})

summ_stat = summ_stat.sort_values(by = 'variable').reset_index(drop = True)
dfi.export(summ_stat, 'summ_stat.jpeg')
Image('summ_stat.jpeg')
```

[ ]:

	variable	mean	median	standard deviation
0	age	11.366460	11.000000	3.167744
1	dist_cap	147.674452	132.001494	76.063134
2	dist_sec	2.418910	2.279000	2.234109
3	fam_n	7.215715	7.000000	2.352900
4	grc	3.963537	4.000000	2.499063
5	grc97	3.705372	4.000000	2.572387
6	hohage	44.436717	43.000000	11.620372
7	hohedu	2.768104	2.000000	2.656106
8	hohsex	0.925185	1.000000	0.263095
9	hohwag	586.985312	500.000000	788.133664
10	indig	0.298324	0.000000	0.457525
11	min_dist	103.447520	111.228612	42.089441
12	poor_num	0.846498	1.000000	0.360473
13	progresa_num	0.615663	1.000000	0.486441
14	sc	0.819818	1.000000	0.384342
15	sc97	0.813922	1.000000	0.389172
16	sex	0.512211	1.000000	0.499854
17	welfare_index	690.346564	685.000000	139.491130

1.2.2 1.2 Differences at baseline?

Are the baseline (1997) demographic characteristics **for the poor** different in treatment and control villages? Present your results in a single table with the following columns and 14 (or so) rows (alphabetized by variable name):

Variable name	Average value (Treatment villages)	Average value (Control villages)	Difference (Treat - Control)	p-value
Male	?	?	?	?

Hint: Use a 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 from poor households (i.e., poor==‘pobre’).

[ ]:
```
# your code here

df12 = progresa_df[(progresa_df['poor'] == 'pobre') &
                   (progresa_df['year'] == 97)].drop(['year',
                                                      'folnum',
                                                      'village'],
                                                     axis = 1)
df12 = df12.groupby('progresa').mean(numeric_only = True).transpose().
 ↪reset_index(names = "Variable Name")
df12.columns.name = None
df12 = df12.rename(columns={"0": "Average value (Control villages)",
                      "basal":  "Average value (Treatment villages)" })
cols = df12.columns.tolist()
cols[-1], cols[-2] = cols[-2], cols[-1]
df12 = df12[cols]
df12['Difference (Treat - Control)'] = df12['Average value (Treatment␣
 ↪villages)'] - df12['Average value (Control villages)']

treatment_df97 = progresa_df[(progresa_df['year'] == 97) &
                             (progresa_df['poor'] == 'pobre') &
                             (progresa_df['progresa'] == 'basal')]
control_df97 = progresa_df[(progresa_df['year'] == 97) &
                           (progresa_df['poor'] == 'pobre') &
                           (progresa_df['progresa'] == '0')]

variables = df12['Variable Name'].tolist()
t_values = []
p_values = []
significance = []
for variable in variables:
    t_stat, p_value = stats.ttest_ind(treatment_df97[variable],␣
 ↪control_df97[variable], equal_var=False, nan_policy = 'omit')
```

```
    t_values.append(t_stat)
    p_values.append(p_value)
    if p_value < 0.05:
        significance.append(True)
    else:
        significance.append(False)

df12['p-value'] = p_values
df12['significant'] = significance
df12 = df12.head(-2).sort_values(by = "Variable Name")
dfi.export(df12, 'df12.jpeg')
Image('df12.jpeg')
```

[ ]:

	Variable Name	Average value (Treatment villages)	Average value (Control villages)	Difference (Treat - Control)	p-value	significant
13	age	10.716991	10.742023	-0.025032	4.783633e-01	False
7	dist_cap	150.829074	153.769730	-2.940656	1.146482e-03	True
2	dist_sec	2.453122	2.507662	-0.054540	4.266282e-02	True
5	fam_n	7.281327	7.302469	-0.021142	4.289667e-01	False
4	grc	3.531599	3.543050	-0.011450	6.895228e-01	False
14	grc97	3.531599	3.543050	-0.011450	6.895228e-01	False
12	hohage	43.648828	44.276918	-0.628090	2.259461e-06	True
8	hohedu	2.663139	2.590348	0.072791	1.038219e-02	True
11	hohsex	0.924656	0.922947	0.001709	5.721253e-01	False
9	hohwag	544.339544	573.163558	-28.824015	3.287285e-04	True
1	indig	0.325986	0.332207	-0.006222	2.459021e-01	False
6	min_dist	107.152915	103.237854	3.915060	7.055795e-16	True
3	sc	0.822697	0.815186	0.007511	9.646120e-02	False
15	sc97	0.822697	0.815186	0.007511	9.646120e-02	False
0	sex	0.519317	0.505052	0.014265	1.220744e-02	True
10	welfare_index	655.428377	659.579100	-4.150723	1.531678e-03	True

1.2.3 1.3 Interpretation

- A: Are there statistically significant differences between treatment and control villages at baseline?
- B: Why does it matter if there are differences at baseline?
- C: What does this imply about how to measure the impact of the treatment?

Discuss your results here

A. **There are statistically significant differences between treatment and control villages at baseline.**

B. **It matters if there are differences at the baseline because these differences are confounders with future differences. If there are differences observed in the future, we cannot be as certain that it was due to the treatment or due to the impacts of**

existing differences pre-treatment. Essentially there is a lack of true randomization pre-treatment, which limits the effectiveness of measuring the treatment effect.

C. **This weakens causal claims about the treatment as there are confounders that might lead to differences in spite of the treatment. In response to this limitation, a model that controls for the confounding variables with statistically significant differences shown above is preferred.**

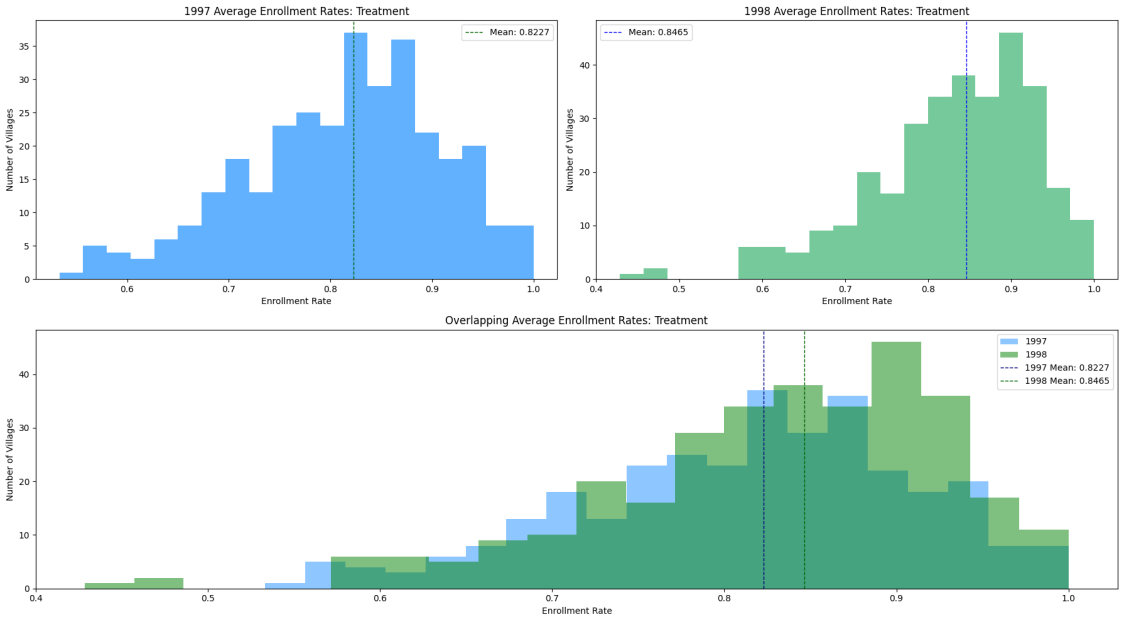
1.2.4 1.4 Graphical exploration, part 1

Visualize the distribution of village enrollment rates **among poor households in treated villages**, before and after treatment. Specifically, for each village, calculate the average rate of enrollment of poor households in treated villages in 1997, then compute the average rate of enrollment of poor households in treated villages in 1998. Create two separate histograms showing the distribution of these average enrollments rates, one histogram for 1997 and one histogram for 1998. On each histogram, draw a vertical line that intersects the x-axis at the average value (across all households). Does there appear to be a difference? Is this difference statistically significant?

[ ]:
```
# Your code here
import matplotlib.pyplot as plt

# Filter the DataFrame for poor households in treated villages for 1997
ve_97T = progresa_df[(progresa_df['year'] == 97) &
                                    (progresa_df['poor'] == 'pobre') &
                                    (progresa_df['progresa'] == 'basal')]

# Group by village and calculate the average enrollment rate for 1997
vavg_97T = ve_97T.groupby('village')['sc'].mean()
avg_tot_97T = ve_97T['sc'].mean()

# Do the same for 1998
ve_98T = progresa_df[(progresa_df['year'] == 98) &
                                    (progresa_df['poor'] == 'pobre') &
                                    (progresa_df['progresa'] == 'basal')]

vavg_98T = ve_98T.groupby('village')['sc'].mean()
avg_tot_98T = ve_98T['sc'].mean()

print(f"Average Enrollment Rate (Treatment 1997):{avg_tot_97T} ")
print(f"Average Enrollment Rate (Treatment 1998): {avg_tot_98T}")
print(f"Simple Difference in Means: {avg_tot_98T - avg_tot_97T}")
t_stat, p_value = stats.ttest_ind(ve_98T['sc'],ve_97T['sc'], nan_policy="omit",␣
 ↪equal_var = False)
print(f"t-value: {t_stat}\np-value: {p_value}")

# Plotting
fig = plt.figure(figsize=(18, 10))  # Wider figure to accommodate the 3rd␣
 ↪subplot
```

```
# Histogram for 1997 on top
ax1 = fig.add_subplot(2, 2, 1)  # This means 2 rows, 2 columns, position 1
ax1.hist(vavg_97T, bins=20, alpha=0.7, color='dodgerblue')
ax1.axvline(avg_tot_97T, color='darkgreen', linestyle='dashed', linewidth=1,␣
 ↪label=f"Mean: {avg_tot_97T:.4f}")
ax1.set_title('1997 Average Enrollment Rates: Treatment')
ax1.set_xlabel('Enrollment Rate')
ax1.set_ylabel('Number of Villages')
ax1.legend()

# Histogram for 1998 on top
ax2 = fig.add_subplot(2, 2, 2)  # This means 2 rows, 2 columns, position 2
ax2.hist(vavg_98T, bins=20, alpha=0.7, color='mediumseagreen')
ax2.axvline(avg_tot_98T, color='blue', linestyle='dashed', linewidth=1,␣
 ↪label=f"Mean: {avg_tot_98T:.4f}")
ax2.set_title('1998 Average Enrollment Rates: Treatment')
ax2.set_xlabel('Enrollment Rate')
ax2.set_ylabel('Number of Villages')
ax2.legend()

# Overlapping histograms on the bottom
ax3 = fig.add_subplot(2, 1, 2)  # This means 2 rows, 1 column, position 2␣
 ↪(spanning both columns)
ax3.hist(vavg_97T, bins=20, alpha=0.5, color='dodgerblue', label='1997')
ax3.hist(vavg_98T, bins=20, alpha=0.5, color='green', label='1998')
ax3.axvline(avg_tot_97T, color='navy', linestyle='dashed', linewidth=1,␣
 ↪label=f"1997 Mean: {avg_tot_97T:.4f}")
ax3.axvline(avg_tot_98T, color='darkgreen', linestyle='dashed', linewidth=1,␣
 ↪label=f"1998 Mean: {avg_tot_98T:.4f}")
ax3.set_title('Overlapping Average Enrollment Rates: Treatment')
ax3.set_xlabel('Enrollment Rate')
ax3.set_ylabel('Number of Villages')
ax3.legend()

plt.tight_layout()
plt.show()
```

```
Average Enrollment Rate (Treatment 1997):0.8226968874033842
Average Enrollment Rate (Treatment 1998): 0.8464791213954308
Simple Difference in Means: 0.023782233992046597
t-value: 6.089840496639891
p-value: 1.1416386466097403e-09
```



Discuss your results here

There is a difference in means which is statistically significant at the 5% level based on a t-test conducted above. The calculated difference in means is around 0.024 (or 2.4 percentage points (not percent)) . The difference (for year 1998 minus 1997) is positive, indicating an increase in the average enrollment rate. The two histograms are pretty similar in shape with the 1998 histogram shifted slightly more to the right. Of course, the acutal magnitude of the change is not extreme, so the two historgams still overlap quite a bit. But since these is this level of change in such a short amount of time, it is not unreasonable to find it a practically significant change.

1.2.5 1.5 Graphical exploration, part 2

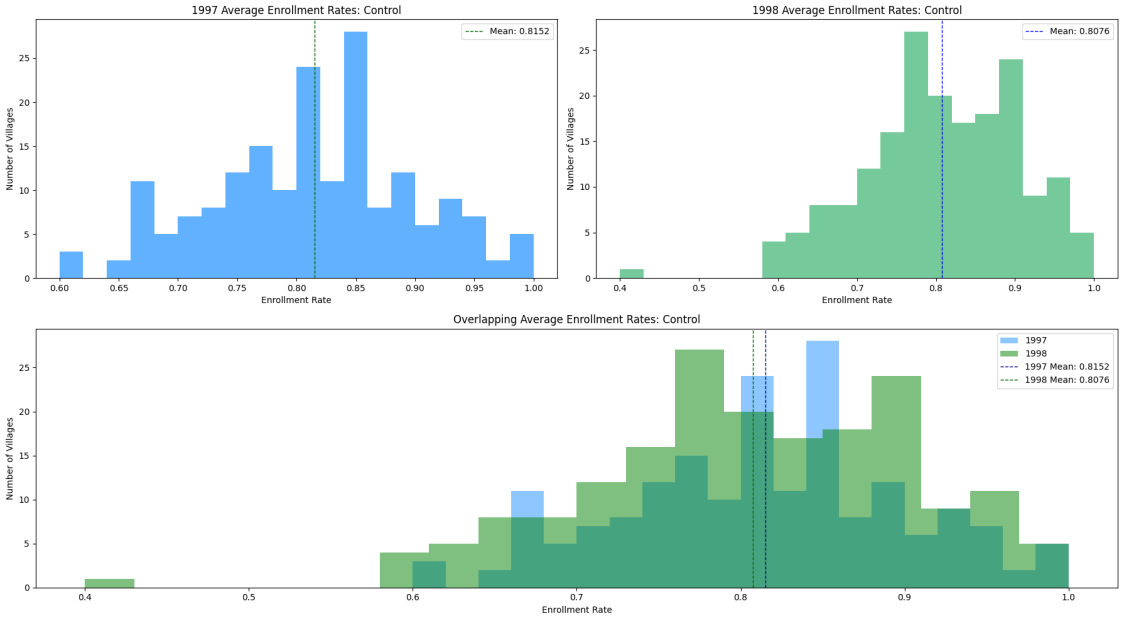
Repeat the above exercise for poor households in **control villages**, before and after treatment. Do you observe a difference in enrollment in control villages between 1997 and 1998? How does what you observe here affect how you might approach measuring the impact of PROGRESA?

[ ]:
```
ve_97C = progresa_df[(progresa_df['year'] == 97) &
                                    (progresa_df['poor'] == 'pobre') &
                                    (progresa_df['progresa'] == '0')]

# Group by village and calculate the average enrollment rate for 1997
vavg_97C = ve_97C.groupby('village')['sc'].mean()
avg_tot_97C = ve_97C['sc'].mean()

# Do the same for 1998
ve_98C = progresa_df[(progresa_df['year'] == 98) &
                                    (progresa_df['poor'] == 'pobre') &
```

```
                                        (progresa_df['progresa'] == '0')]

vavg_98C = ve_98C.groupby('village')['sc'].mean()
avg_tot_98C = ve_98C['sc'].mean()

print(f"Average Enrollment Rate (Control 1997):{avg_tot_97C} ")
print(f"Average Enrollment Rate (Control 1998): {avg_tot_98C}")
print(f"Simple Difference in Means: {avg_tot_98C - avg_tot_97C}")
t_stat, p_value = stats.ttest_ind(ve_98C['sc'],ve_97C['sc'], nan_policy="omit",␣
 ↪equal_var = False)
print(f"t-value: {t_stat}\np-value: {p_value}")


# Plotting
fig = plt.figure(figsize=(18, 10))  # Wider figure to accommodate the 3rd␣
 ↪subplot

# Histogram for 1997 on top
ax1 = fig.add_subplot(2, 2, 1)  # This means 2 rows, 2 columns, position 1
ax1.hist(vavg_97C, bins=20, alpha=0.7, color='dodgerblue')
ax1.axvline(avg_tot_97C, color='darkgreen', linestyle='dashed', linewidth=1,␣
 ↪label=f"Mean: {avg_tot_97C:.4f}")
ax1.set_title('1997 Average Enrollment Rates: Control')
ax1.set_xlabel('Enrollment Rate')
ax1.set_ylabel('Number of Villages')
ax1.legend()

# Histogram for 1998 on top
ax2 = fig.add_subplot(2, 2, 2)  # This means 2 rows, 2 columns, position 2
ax2.hist(vavg_98C, bins=20, alpha=0.7, color='mediumseagreen')
ax2.axvline(avg_tot_98C, color='blue', linestyle='dashed', linewidth=1,␣
 ↪label=f"Mean: {avg_tot_98C:.4f}")
ax2.set_title('1998 Average Enrollment Rates: Control')
ax2.set_xlabel('Enrollment Rate')
ax2.set_ylabel('Number of Villages')
ax2.legend()

# Overlapping histograms on the bottom
ax3 = fig.add_subplot(2, 1, 2)  # This means 2 rows, 1 column, position 2␣
 ↪(spanning both columns)
ax3.hist(vavg_97C, bins=20, alpha=0.5, color='dodgerblue', label='1997')
ax3.hist(vavg_98C, bins=20, alpha=0.5, color='green', label='1998')
ax3.axvline(avg_tot_97C, color='navy', linestyle='dashed', linewidth=1,␣
 ↪label=f"1997 Mean: {avg_tot_97C:.4f}")
ax3.axvline(avg_tot_98C, color='darkgreen', linestyle='dashed', linewidth=1,␣
 ↪label=f"1998 Mean: {avg_tot_98C:.4f}")
```

```
ax3.set_title('Overlapping Average Enrollment Rates: Control')
ax3.set_xlabel('Enrollment Rate')
ax3.set_ylabel('Number of Villages')
ax3.legend()

plt.tight_layout()
plt.show()
```

```
Average Enrollment Rate (Control 1997):0.8151860030575845
Average Enrollment Rate (Control 1998): 0.807636956730308
Simple Difference in Means: -0.007549046327276487
t-value: -1.4342427265153408
p-value: 0.15151736204981645
```



Discuss your results here

There is a difference in means, but the magnitude is extremely close to 0. Consequently, it is NOT statistically significant at the 5% level based on a t-test conducted above. The calculated difference in means is around -0.0007 (or -0.7 percentage points (note: not a percent change but about the difference between 81.5% to 80.7%)). The difference (for year 1998 minus 1997) is negative, indicating, if anything, a decrease in the average enrollment rate. However, considering the high p-value, this difference is probably stochastic and has little bearing or meaning. The two histograms are pretty similar in shape and overlap much more than in previous question. This makes sense considering there is very little difference between the pre- and post- distributions.

1.3 Part 2: Measuring Impact

Our goal is to estimate the causal impact of the PROGRESA program on the social and economic outcomes of individuals in Mexico. We will focus on the impact of the program on school enrollment rates among the poor (those with poor==‘pobre’), since only the poor were eligible to receive PROGRESA assistance, and since a primary objective of the program was to increase school enrollment.

1.3.1 2.1 Simple differences: T-test

Begin by estimating the impact of Progresa using “simple differences.” Restricting yourself to data from 1998 (after treatment), calculate the average enrollment rate among **poor** households in the Treatment villages and the average enrollment rate among **poor** households in the control villages. Use a t-test to determine if this difference is statistically significant. What do you conclude?

[ ]:
```
warnings.filterwarnings('ignore')
filtered_21 = progresa_df[(progresa_df['poor'] == 'pobre') &
                   (progresa_df['year'] == 98)].drop(['year',
                                                      'folnum'],
                                                     axis = 1)
df_21 = filtered_21.groupby(['progresa'])['sc'].mean(numeric_only = True)
pooravg_C98 = df_21[0]
pooravg_T98 = df_21[1]

print(f'''Avg. Enrollment Rate-Poor Households (Control 1998): {pooravg_C98}
Avg. Enrollment Rate-Poor Households (Treatment 1998): {pooravg_T98}''')

#Performing t-test
t_value = stats.ttest_ind(filtered_21[filtered_21.progresa == 'basal']['sc']
                          ,filtered_21[filtered_21.progresa == '0']['sc']
                          , nan_policy='omit', equal_var = False).statistic
p_value = stats.ttest_ind(filtered_21[filtered_21.progresa == 'basal']['sc']
                          ,filtered_21[filtered_21.progresa == '0']['sc']
                          , nan_policy='omit', equal_var = False).pvalue
print('Simple Difference in Means:', pooravg_T98 - pooravg_C98 )
print('t-statistic : ', t_value)
print('p-value : ', p_value)
```

```
Avg. Enrollment Rate-Poor Households (Control 1998): 0.807636956730308
Avg. Enrollment Rate-Poor Households (Treatment 1998): 0.8464791213954308
Simple Difference in Means: 0.0388421646651228
t-statistic :  8.181477157107308
p-value :  2.9655072988948406e-16
```

Discuss your results here

The t-test results in a t-statistic that is statistically significant at the 5% level. Given this statistical significance, it implies that the difference between poor households in the control and treatment group for the same year is likely not 0. Bringing into consideration that the actual difference in means is positive (Treatment minus Control),

this also implies a higher average enrollment rate for those in the treatment than control.

1.3.2 2.2 Simple differences: Regression

Estimate the effects of Progresa on enrollment using a regression model, by regressing the 1998 enrollment rates **of the poor** on treatment assignment. For now, do not include any other variables in your regression. Discuss the following:

- Based on this model, how much did Progresa increase or decrease the likelihood of a child enrolling? Make sure you express you answer in a sentence that a person with no technical background could understand, using appropriate units.
- How does your regression estimate compare to your t-test estimate from part 2.1?
- Based on this regression model, can we reject the null hypothesis that the treatment effects are zero?
- What is the counterfactual assumption underlying this regression?

[ ]:
```
# Your code here
from statsmodels.formula.api import ols

model_22 = ols(formula='sc ~ progresa', data=filtered_21).fit()
#Summary of the fitted model
model_22.summary()
```

[ ]:

Dep. Variable:	sc	**R-squared:**	0.003
Model:	OLS	**Adj. R-squared:**	0.003
Method:	Least Squares	**F-statistic:**	69.87
Date:	Mon, 05 Feb 2024	**Prob (F-statistic):**	6.64e-17
Time:	11:42:59	**Log-Likelihood:**	-11926.
No. Observations:	27450	**AIC:**	2.386e+04
Df Residuals:	27448	**BIC:**	2.387e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
Intercept	0.8076	0.004	220.676	0.000	0.800	0.815
progresa[T.basal]	0.0388	0.005	8.359	0.000	0.030	0.048

Omnibus:	7638.939	**Durbin-Watson:**	1.734
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	15767.534
Skew:	-1.767	**Prob(JB):**	0.00
Kurtosis:	4.140	**Cond. No.**	3.01

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Discuss your results here

Based on this model, when a household is enrolled in the progresa program on average, the likelihood a child enrolling in school is associated with a chance of enrollment higher by 0.0338 (i.e.a 3.88 percentage point higher change (not percent)). The re-

gression result mirrors very closely the difference in means used for the t-test from part 2.1. They both imply a 3.88 percentage point difference between enrollment likelihoods on average between poor individuals in 1998 who are in and not in the program. That means that on average, poor individuals in the treatment group in 1998 are associated with about a 4.8% (0.0388/0.8076 = 0.048) higher likelihood of enrollment than poor individual in the control group in 1998.

Based on this regression, it seems plausible and reasonable to reject the null hypothesis that the treatment effects are 0. The p-value for the coefficient on the `progresa` is extremely close to 0 and thus less that 0.05. In this case, I do still want to note that there were some possible confounds such as the significant baseline differences found in question 1.2. This model also doesn't account for time-varying confounders and pre-existing trends. Thus, while we can say that the treatment effects are likely non-zero, it's still good to be cautious about quantifying that effect precisely using this model

The counterfactual assumption underlying this regression is that without the treatment, in the year 1998, there would be no difference in the enrollment rates between poor individuals in villages assigned to received the program (Treatment) and villages who were not assigned to receive it (Control).

1.3.3 2.3 Multiple Regression

Estimate the above regression, but this time include a set of control variables. Include, for instance, age, distance to a secondary school, gender, education of household head, welfare index, indigenous, etc.

- How do the controls affect the point estimate of treatment effect?
- How do the controls affect the standard error on the treatment effect?
- How do you interpret the differences (or similarities) between your estimates of 2.2 and 2.3?
- Interpet the coefficient associated with the `dist_sec` variable. Is this evidence that the household's distance from a secondary school has a *causal* impact on educational attainment?

[ ]:
```
# Your code here

from statsmodels.formula.api import ols

model_23 = ols(formula=('''sc ~ progresa + age + dist_sec +
                        sex + hohedu + hohwag + min_dist+ dist_cap +
                        welfare_index + indig + hohage + hohwag + fam_n'''),␣
 ↪data=filtered_21).fit()
#Summary of the fitted model
model_23.summary()
```

[ ]:

Dep. Variable:	sc	**R-squared:**	0.272
Model:	OLS	**Adj. R-squared:**	0.272
Method:	Least Squares	**F-statistic:**	849.5
Date:	Mon, 05 Feb 2024	**Prob (F-statistic):**	0.00
Time:	11:42:59	**Log-Likelihood:**	-7541.1
No. Observations:	27263	**AIC:**	1.511e+04
Df Residuals:	27250	**BIC:**	1.521e+04
Df Model:	12		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
Intercept	1.4431	0.018	80.855	0.000	1.408	1.478
progresa[T.basal]	0.0337	0.004	8.418	0.000	0.026	0.042
age	-0.0657	0.001	-94.750	0.000	-0.067	-0.064
dist_sec	-0.0106	0.001	-12.012	0.000	-0.012	-0.009
sex	0.0304	0.004	7.847	0.000	0.023	0.038
hohedu	0.0079	0.001	9.224	0.000	0.006	0.010
hohwag	-7.23e-07	2.81e-06	-0.257	0.797	-6.24e-06	4.79e-06
min_dist	0.0004	6.34e-05	6.326	0.000	0.000	0.001
dist_cap	0.0002	3.7e-05	5.512	0.000	0.000	0.000
welfare_index	2.283e-05	1.82e-05	1.253	0.210	-1.29e-05	5.85e-05
indig	0.0188	0.005	3.961	0.000	0.009	0.028
hohage	-2.299e-05	0.000	-0.118	0.906	-0.000	0.000
fam_n	0.0006	0.001	0.727	0.467	-0.001	0.002

Omnibus:	2940.375	**Durbin-Watson:**	1.729
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	3992.254
Skew:	-0.930	**Prob(JB):**	0.00
Kurtosis:	3.237	**Cond. No.**	9.54e+03

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 9.54e+03. This might indicate that there are strong multicollinearity or other numerical problems.

Discuss your results here

- How do the controls affect the point estimate of treatment effect?
- How do the controls affect the standard error on the treatment effect?
- How do you interpret the differences (or similarities) between your estimates of 2.2 and 2.3?
- Interpet the coefficient associated with the `dist_sec` variable. Is this evidence that the household's distance from a secondary school has a *causal* impact on educational attainment?

The controls decrease the point estimate of the treatment effect by a magnitude of about 0.0051 to 0.0337. The controls decrease the standard error on the treatment by a magnitude of about 0.001. This differences are honestly quite minimal. Even after including controls for variables that had statistically significant differences at baseline year (1997), we see the treatment effect is not much different.

The coefficeint on `dist_sec` is -0.0117 implying that all else constant, on average the enrollment rate for poor students in 1998 across treatment and control is 1.17 per-

centage points lower for every one kilometer increase in a household's/individual's distance to the nearest secondary school. The standard error and p-value on this coefficient are quite low, implying that the estimate might be precise. However, there is no way to make a causal claim about this particular variable. There is no experimental mechanism guaranteeing that the effect of distance from school on school enrollment is not impacted by endogeneity. Thus, this relationship may simply be associative or could be due to other reasons like self-selection bias. Maybe individuals who value education less choose to move further away. Regardless, of what the true story is, there isn't enough of a support for a causal claim.

1.3.4 2.4 Multiple Regression Revisited

For the same set of control variables that you used in 2.3, carry out the following alternative estimation procedure.

- First, regress the 1998 enrollment of the poor on the control variables, **without including the treatment assignment.**
- Second, use this model to obtain predicted values of the 1998 enrollment for each child in the sample used to estimate the model in step 1.
- Third, compute a new value for each child, which is the difference between the actual 1998 enrollment and the predicted enrollment value from step 2.
- Finally, regress the difference from step 3 on treatment assignment.

Compare the point estimate and the standard error on treatment assignment that you obtained in step 4 to their analogues in 2.3. Explain the results.

[ ]:
```
# Your code here

from statsmodels.formula.api import ols
import statsmodels.api as sm

model_24_first = ols(formula=('''sc ~ +  age + dist_sec +
                        sex + hohedu +hohwag + min_dist+ dist_cap +
                        welfare_index + indig + hohage + hohwag + fam_n'''),␣
 ↪data=filtered_21).fit()
filtered_21['fitted_sc'] = model_24_first.fittedvalues
filtered_21['residuals'] = filtered_21['sc'] - filtered_21['fitted_sc']

model_24_second = ols(formula=('''residuals ~ progresa'''), data=filtered_21).
 ↪fit()
model_24_second.summary()
```

[ ]:

Dep. Variable:	residuals	**R-squared:**	0.003
Model:	OLS	**Adj. R-squared:**	0.003
Method:	Least Squares	**F-statistic:**	70.19
Date:	Mon, 05 Feb 2024	**Prob (F-statistic):**	5.63e-17
Time:	11:42:59	**Log-Likelihood:**	-7541.5
No. Observations:	27263	**AIC:**	1.509e+04
Df Residuals:	27261	**BIC:**	1.510e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
Intercept	-0.0207	0.003	-6.605	0.000	-0.027	-0.015
progresa[T.basal]	0.0334	0.004	8.378	0.000	0.026	0.041

Omnibus:	2932.774	**Durbin-Watson:**	1.729
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	3979.059
Skew:	-0.928	**Prob(JB):**	0.00
Kurtosis:	3.235	**Cond. No.**	3.01

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Discuss your results here

Compare the point estimate and the standard error on treatment assignment that you obtained in step 4 to their analogues in 2.3. Explain the results.

The results are almost identical. The standard errors are identical while the coefficients differ slightly. Considering how small the difference is, it might be due to precision loss or rounding with such small numbers. Both results are similar because they are essentially accomplishing the same task. In the two step process, the residuals produced in the first-step regression are uncorrelated with the controls variables. The residuals represent represent the portion of the average enrollment rate not explained by the control variables. Thus, this step allows us to subsequently isolate and measure the effect of the treatment variable. In the second-step regression, one can simply isolate the treatment effect now that the controls have already been accounted for. The difference is simply simultaneously or sequentially controlling for confounders.

1.3.5 2.5 Difference-in-Difference, version 1 (tabular)

Thus far, we have computed the effects of Progresa by estimating the difference in 1998 enrollment rates across villages. An alternative approach would be to compute the treatment effect using a difference-in-differences framework.

Begin by estimating the average treatment effects of the program for poor households using data from 1997 and 1998. Specifically, calculate the difference (between 1997 and 1998) in enrollment rates among poor households in treated villages; then compute the difference (between 1997 and 1998) in enrollment rates among poor households in control villages.

Display your results in a 2x2 table where the rows are Control/Treatment and the columns are 1997/1998.

- What is your difference-in-difference estimate of the impact, and how does it compare to your earlier (simple difference) results?
- What is the counterfactual assumption underlying this estimate?

[ ]:
```
# Your code here
# Create a pivot table
pivot_table = progresa_df[progresa_df.poor == 'pobre'].pivot_table(values='sc',␣
 ↪index='progresa', columns='year', aggfunc='mean')

# Calculate the differences
pivot_table['first_diff'] = pivot_table[98] - pivot_table[97]

control_diff = pivot_table['first_diff'].iloc[:2][0]
treat_diff = pivot_table['first_diff'].iloc[:2][1]
diff_in_diff = (treat_diff - control_diff)
print(f"Difference-in-Differences: {diff_in_diff}")
# Display the table
dfi.export(pivot_table, 'pivot_table.jpeg')
Image('pivot_table.jpeg')
```

```
Difference-in-Differences: 0.031331280319323085
```

[ ]:

year	97	98	first_diff
progresa			
0	0.815186	0.807637	-0.007549
basal	0.822697	0.846479	0.023782

Discuss your results here

The difference in difference estimate is about 0.03133 compared to an earlier single difference estimate of about 0.0388. The single difference is larger but only slightly. The counterfactual assumption underlying the difference-in-difference estimate is that in the absence of the treatment, the change over time in school enrollment from 1997 to 1998 for poor housholds would have been same for the treatment and the control group. Another way to phrase that is: The difference between the control and treatment groups would have been the same in 1997 and 1998 in the absence of the treatment.

1.3.6 2.6 Difference-in-Difference, version 2 (regression)

Now use a regression specification to estimate the average treatment effects of the program in a difference-in-differences, for the poor households. Do this (i) first without including any control

variables; and then (ii) do it a second time including at least 5 control variables.

- What is your estimate (i) of the impact of Progresa? Be very specific in interpreting your coefficients and standard errors, and make sure to specify exactly what units you are measuring and estimating.
- Does your estimate of the impact of Progresa from (i) change when you add control variables as in (ii)? How do you explain these changes, or the lack of changes on the `progresa` coefficient between (i) and (ii)?
- How do the estimates from (i) and (ii) compare to the difference-in-difference estimates from 2.4 above? What accounts for these differences, if any exist?
- What is the counterfactual assumption underlying regression (ii)?

[ ]:
```
# Your code here

model_26first = ols('''sc ~ progresa + C(year) + progresa * C(year)''',
                    data = progresa_df[progresa_df.poor == 'pobre']).fit()
print(model_26first.summary())
model_26second = ols('''sc ~ progresa + C(year) + progresa * C(year) + age +␣
 ↪dist_sec +
                         sex + hohedu +hohwag + min_dist+ dist_cap +
                         welfare_index + indig + hohage + hohwag + fam_n''',
                    data = progresa_df[progresa_df.poor == 'pobre']).fit()
print(model_26second.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                     sc   R-squared:                       0.001
Model:                            OLS   Adj. R-squared:                  0.001
Method:                 Least Squares   F-statistic:                     28.31
Date:                Mon, 05 Feb 2024   Prob (F-statistic):           2.76e-18
Time:                        11:43:00   Log-Likelihood:                -26242.
No. Observations:               58372   AIC:                         5.249e+04
Df Residuals:                   58368   BIC:                         5.253e+04
Df Model:                           3
Covariance Type:            nonrobust
===============================================================================
===================
                                    coef    std err          t      P>|t|
[0.025      0.975]
-------------------------------------------------------------------------------
-------------------
Intercept                          0.8152      0.003    233.182      0.000
0.808       0.822
progresa[T.basal]                  0.0075      0.004      1.691      0.091
-0.001       0.016
C(year)[T.98]                     -0.0075      0.005     -1.480      0.139
-0.018       0.002
progresa[T.basal]:C(year)[T.98]    0.0313      0.006      4.835      0.000
```

```
0.019       0.044
==============================================================================
Omnibus:                    15346.988   Durbin-Watson:                   1.397
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            30608.651
Skew:                          -1.711   Prob(JB):                         0.00
Kurtosis:                       3.937   Cond. No.                         7.67
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
                            OLS Regression Results
==============================================================================
Dep. Variable:                     sc   R-squared:                       0.283
Model:                            OLS   Adj. R-squared:                  0.283
Method:                 Least Squares   F-statistic:                     1634.
Date:                Mon, 05 Feb 2024   Prob (F-statistic):               0.00
Time:                        11:43:01   Log-Likelihood:                -16476.
No. Observations:               58005   AIC:                         3.298e+04
Df Residuals:                   57990   BIC:                         3.312e+04
Df Model:                          14
Covariance Type:            nonrobust
==============================================================================
==================
                                     coef    std err          t      P>|t|
[0.025      0.975]
------------------------------------------------------------------------------
------------------
Intercept                            1.4111      0.012    114.216      0.000
1.387       1.435
progresa[T.basal]                    0.0033      0.004      0.859      0.390
-0.004       0.011
C(year)[T.98]                        0.0280      0.004      6.429      0.000
0.019       0.037
progresa[T.basal]:C(year)[T.98]      0.0308      0.006      5.583      0.000
0.020       0.042
age                                 -0.0658      0.000   -143.065      0.000
-0.067      -0.065
dist_sec                            -0.0096      0.001    -15.459      0.000
-0.011      -0.008
sex                                  0.0332      0.003     12.420      0.000
0.028       0.038
hohedu                               0.0072      0.001     12.158      0.000
0.006       0.008
hohwag                           1.055e-06   1.93e-06      0.547      0.585
-2.73e-06    4.84e-06
min_dist                             0.0004   4.39e-05      8.607      0.000
0.000       0.000
```

```
dist_cap                          0.0002   2.55e-05      7.235      0.000
0.000       0.000
welfare_index                  2.013e-05   1.26e-05      1.599      0.110
-4.54e-06    4.48e-05
indig                             0.0242      0.003      7.407      0.000
0.018       0.031
hohage                            0.0002      0.000      1.711      0.087
-3.36e-05       0.000
fam_n                             0.0001      0.001      0.208      0.835
-0.001       0.001
==============================================================================
Omnibus:                     5460.754   Durbin-Watson:                   1.492
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             7170.195
Skew:                          -0.859   Prob(JB):                         0.00
Kurtosis:                       3.110   Cond. No.                     9.65e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 9.65e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Discuss your results here

In the first regression, the impact of the the program or DiD estimator is 0.0313 with a standard error of 0.006 and p-value close to 0. This means the estimate is precise and statistically significant at the 5% level. The coefficient indicates that on average, school enrollment rates increased by 3.13 percentage points more in the treatment group than in the control group from 1997 to 1998.

After adding controls, this coefficient drops slightly, down to 0.0308 with the same standard error and p-values. Even after controlling for other variables, the DiD estimator remains statistically significant and relatively unchanged. This means that after accounting for both the temporal changes and potential failures in the randomness of program assignment (seen in question 1.2), there is still on average a positive association between being in the program and school enrollment. If considering the program as impacting school enrollment, this suggests that Progresa has a positive and statistically significant effect on the treated population compared to the control group over the studied period, after controlling for other observable factors and introducting a experimental design that attempts to isolate the treatment effects.

Compared to the tabulations in 2.5, we can see that the tabulation results mirror the DiD estimator produced in the regressoin without any controls. This is likely due to the fact that the tabulation does not control for any of the confounders controlled for in (ii). By controlling for these confounders in (ii), the idea is to estimate a more accurate treatment effect even if both results still indicate a positive treatment effect overall.

The counterfactual assumption underlying regression (ii) is the the parallel trends

assumption. That in the absence of the treatment, the difference for the treatment group between year 1998 and 1997 would have been the same as the difference for the control group between year 1998 and 1997.

1.3.7 2.7 Spillover effects

Thus far, we have focused on the impact of PROGRESA on the school enrollment of poor households. Repeat your analysis in 2.5, instead focusing on (a) the impact of PROGRESA on the school enrollment of non-poor households, and (b) the impact ofo PROGRESA on *other outcomes* of poor households that might plausibly have been affected by the PROGRESA program. * Do you observe any impacts of PROGRESA on the school enrollment of the non-poor? * Regardless of whether you find evidence of spillovers to non-poor, describe one or two reasons why PROGRESA *might* have impacted non-poor households. Give concrete examples based on the context in which PROGRESA was implemented. * Do you observe any impacts of PROGRESA on other aspects of the welfare of poor households?

[ ]:
```
# Your code here

#non-poor household analysis
pivot_table = progresa_df[progresa_df.poor == 'no pobre'].
 ↪pivot_table(values='sc',
                                                                  ␣
 ↪index='progresa',
                                                                  ␣
 ↪columns='year',
                                                                  ␣
 ↪aggfunc='mean')

# Calculate the differences
pivot_table['first_diff'] = pivot_table[98] - pivot_table[97]

control_diff = pivot_table['first_diff'].iloc[:2][0]
treat_diff = pivot_table['first_diff'].iloc[:2][1]
diff_in_diff = (treat_diff - control_diff)
print(f"Difference-in-Differences: {diff_in_diff}")
# Display the table
print(pivot_table)

model_27first = ols('''sc ~ progresa + C(year) + progresa * C(year)''',
                 data = progresa_df[progresa_df.poor == 'no pobre']).fit()
print(model_27first.summary())
model_27second = ols('''sc ~ progresa + C(year) + progresa * C(year) + age +␣
 ↪dist_sec +
                         sex + hohedu +hohwag + min_dist+ dist_cap +
                         welfare_index + indig + hohage + hohwag + fam_n''',
                 data = progresa_df[progresa_df.poor == 'no pobre']).fit()
print(model_27second.summary())
```

```
Difference-in-Differences: 3.4275264633842895e-05
year            97        98  first_diff
progresa
0         0.762587  0.776337    0.013750
basal     0.795264  0.809049    0.013785
                            OLS Regression Results
==============================================================================
Dep. Variable:                     sc   R-squared:                       0.002
Model:                            OLS   Adj. R-squared:                  0.002
Method:                 Least Squares   F-statistic:                     6.332
Date:                Mon, 05 Feb 2024   Prob (F-statistic):           0.000276
Time:                        11:43:01   Log-Likelihood:                -5448.5
No. Observations:               10425   AIC:                         1.090e+04
Df Residuals:                   10421   BIC:                         1.093e+04
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
==================
                                 coef    std err          t      P>|t|
[0.025      0.975]
------------------------------------------------------------------------------
------------------
Intercept                        0.7626      0.009     89.682      0.000
0.746       0.779
progresa[T.basal]                0.0327      0.011      2.978      0.003
0.011       0.054
C(year)[T.98]                    0.0138      0.013      1.079      0.280
-0.011       0.039
progresa[T.basal]:C(year)[T.98]  3.428e-05      0.016      0.002      0.998
-0.032       0.032
==============================================================================
Omnibus:                     2002.335   Durbin-Watson:                   1.443
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             3446.137
Skew:                          -1.408   Prob(JB):                         0.00
Kurtosis:                       2.994   Cond. No.                         7.36
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
                            OLS Regression Results
==============================================================================
Dep. Variable:                     sc   R-squared:                       0.292
Model:                            OLS   Adj. R-squared:                  0.291
Method:                 Least Squares   F-statistic:                     303.7
Date:                Mon, 05 Feb 2024   Prob (F-statistic):               0.00
Time:                        11:43:01   Log-Likelihood:                -3614.1
No. Observations:               10334   AIC:                             7258.
```

```
Df Residuals:                     10319   BIC:                                7367.
Df Model:                            14
Covariance Type:            nonrobust
==============================================================================
==================
                                   coef    std err          t      P>|t|
[0.025      0.975]
------------------------------------------------------------------------------
------------------
Intercept                          1.4366      0.035     40.955      0.000
1.368       1.505
progresa[T.basal]                  0.0243      0.009      2.612      0.009
0.006       0.043
C(year)[T.98]                      0.0404      0.011      3.741      0.000
0.019       0.062
progresa[T.basal]:C(year)[T.98]   -0.0032      0.014     -0.231      0.817
-0.030       0.024
age                               -0.0685      0.001    -59.812      0.000
-0.071      -0.066
dist_sec                          -0.0151      0.002     -8.532      0.000
-0.019      -0.012
sex                                0.0286      0.007      4.227      0.000
0.015       0.042
hohedu                             0.0081      0.001      6.609      0.000
0.006       0.010
hohwag                          8.274e-07   3.13e-06      0.265      0.791
-5.3e-06    6.96e-06
min_dist                        -9.24e-05      0.000     -0.794      0.427
-0.000       0.000
dist_cap                           0.0006   7.63e-05      7.972      0.000
0.000       0.001
welfare_index                   8.306e-05   2.85e-05      2.917      0.004
2.72e-05       0.000
indig                              0.0146      0.010      1.399      0.162
-0.006       0.035
hohage                             0.0002      0.000      0.469      0.639
-0.001       0.001
fam_n                             -0.0046      0.001     -3.160      0.002
-0.007      -0.002
==============================================================================
Omnibus:                      841.465   Durbin-Watson:                   1.509
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              804.007
Skew:                          -0.623   Prob(JB):                    2.58e-175
Kurtosis:                       2.440   Cond. No.                     1.54e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
```

specified.
[2] The condition number is large, 1.54e+04. This might indicate that there are strong multicollinearity or other numerical problems.

[ ]:
```
# Your code here


#spillover for poor household analysis
pivot_table = progresa_df[progresa_df.poor == 'pobre'].
 ↪pivot_table(values='hohwag',
                                                                    ␣
 ↪index='progresa',
                                                                    ␣
 ↪columns='year',
                                                                    ␣
 ↪aggfunc='mean')
# Calculate the differences
pivot_table['first_diff'] = pivot_table[98] - pivot_table[97]

control_diff = pivot_table['first_diff'].iloc[:2][0]
treat_diff = pivot_table['first_diff'].iloc[:2][1]
diff_in_diff = (treat_diff - control_diff)
print(f"Difference-in-Differences: {diff_in_diff}")
# Display the table
print(pivot_table)

model_27first1 = ols('''hohwag ~ progresa + C(year) + progresa * C(year)''',
                  data = progresa_df[progresa_df.poor == 'pobre']).fit()
print(model_27first1.summary())

model_27second1 = ols('''hohwag ~ progresa + C(year) + progresa * C(year)
                      + age + dist_sec +
                        sex + hohedu +hohwag + min_dist+ dist_cap +
                        welfare_index + indig + hohage + fam_n
''',
                  data = progresa_df[progresa_df.poor == 'pobre']).fit()
print(model_27second1.summary())
```

```
Difference-in-Differences: 0.0
year              97          98  first_diff
progresa
0         573.163558  573.163558         0.0
basal     544.339544  544.339544         0.0
                            OLS Regression Results
==============================================================================
Dep. Variable:                 hohwag   R-squared:                       0.000
Model:                            OLS   Adj. R-squared:                  0.000
```

```
Method:                 Least Squares   F-statistic:                     8.614
Date:                Mon, 05 Feb 2024   Prob (F-statistic):           1.03e-05
Time:                        11:43:01   Log-Likelihood:            -5.2158e+05
No. Observations:               65392   AIC:                         1.043e+06
Df Residuals:                   65388   BIC:                         1.043e+06
Df Model:                           3
Covariance Type:            nonrobust
================================================================================
                                   coef    std err          t      P>|t|
[0.025      0.975]
--------------------------------------------------------------------------------
Intercept                        573.1636      6.306     90.888      0.000
560.803     585.524
progresa[T.basal]                -28.8240      8.019     -3.595      0.000
-44.541     -13.107
C(year)[T.98]                 -3.437e-13      8.918  -3.85e-14      1.000
-17.480      17.480
progresa[T.basal]:C(year)[T.98] -7.448e-13     11.340  -6.57e-14      1.000
-22.227      22.227
==============================================================================
Omnibus:                    80974.047   Durbin-Watson:                   0.328
Prob(Omnibus):                  0.000   Jarque-Bera (JB):         15508664.221
Skew:                           6.706   Prob(JB):                         0.00
Kurtosis:                      77.243   Cond. No.                         7.85
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
                            OLS Regression Results
==============================================================================
Dep. Variable:                  hohwag   R-squared:                       1.000
Model:                             OLS   Adj. R-squared:                  1.000
Method:                 Least Squares   F-statistic:                 2.472e+33
Date:                Mon, 05 Feb 2024   Prob (F-statistic):               0.00
Time:                        11:43:02   Log-Likelihood:             1.7050e+06
No. Observations:               64962   AIC:                        -3.410e+06
Df Residuals:                   64947   BIC:                        -3.410e+06
Df Model:                          14
Covariance Type:            nonrobust
================================================================================
                                   coef    std err          t      P>|t|
[0.025      0.975]
--------------------------------------------------------------------------------
```

```
Intercept                          3.594e-13    3.5e-14     10.282      0.000
2.91e-13    4.28e-13
progresa[T.basal]                  2.021e-12   1.11e-14    182.348      0.000
2e-12    2.04e-12
C(year)[T.98]                      2.141e-12   1.24e-14    173.058      0.000
2.12e-12    2.16e-12
progresa[T.basal]:C(year)[T.98] -3.301e-12   1.56e-14   -211.168      0.000
-3.33e-12   -3.27e-12
age                               -4.316e-15   1.25e-15     -3.447      0.001
-6.77e-15   -1.86e-15
dist_sec                           4.942e-14   1.71e-15     28.877      0.000
4.61e-14    5.28e-14
sex                               -2.952e-14    7.6e-15     -3.886      0.000
-4.44e-14   -1.46e-14
hohedu                             1.035e-14   1.69e-15      6.128      0.000
7.04e-15    1.37e-14
hohwag                                1.0000   5.51e-18   1.81e+17      0.000
1.000       1.000
min_dist                          -1.839e-15   1.24e-16    -14.820      0.000
-2.08e-15    -1.6e-15
dist_cap                           2.366e-16   7.27e-17      3.256      0.001
9.42e-17    3.79e-16
welfare_index                     -2.527e-15   3.59e-17    -70.382      0.000
-2.6e-15   -2.46e-15
indig                              1.464e-14   9.23e-15      1.586      0.113
-3.45e-15    3.27e-14
hohage                             1.516e-15   3.77e-16      4.017      0.000
7.76e-16    2.26e-15
fam_n                              2.828e-14    1.7e-15     16.672      0.000
2.5e-14    3.16e-14
==============================================================================
Omnibus:                     4143.570   Durbin-Watson:                   2.789
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             5415.010
Skew:                           0.593   Prob(JB):                         0.00
Kurtosis:                       3.771   Cond. No.                     9.57e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 9.57e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Discuss your results here

- Do you observe any impacts of PROGRESA on the school enrollment of the non-poor?
- Regardless of whether you find evidence of spillovers to non-poor, describe one or two reasons why PROGRESA *might* have impacted non-poor households. Give concrete examples based on the context in which PROGRESA was implemented.

- Do you observe any impacts of PROGRESA on other aspects of the welfare of poor households?

Running the same model with the non-poor data set shows a DiD estimator close to 0 with an extremeley large p-value. This implies the coefficient is not significant at the 5% level, and a causal impact is very unlikely.

There are many reasons why the progresa program may have impacted non-poor households even if non-poor school enrollment wasn't effected. For example, it may be that many non-poor people work with and for the educational institutions in a village or other insitutions that offers services that people pay for with money from the CCT. Thus the CCTs could increase the wages of those who already qualify as non-poor. More students paying to go to school also means that there is more money to go into the fixed costs of running the school like infrastructure and teacher salaries. While some of this will have to go to accomodating new students, shared resources will also be improved. This can improve the quality of the education for all students, not just those in the program. If that is the case, more non-poor people may be willing to enroll in school as well, considering the quality of education will go up. This is possible but needs to be accompanied by a reasonable increase in the number of teachers, textbooks, etc. Alternatively, if the opposite is true and schools become overcrowded and can't manage to provide appropriate resources for new students, there might be non-poor individuals withdrawing from the schools due to the decrease in education quality.

I tried to see if the progresa program had impacts on other aspects of the welfare of poor households for a number of difference response variables in the data. For the most part, the actual DiD estimator was extremely close to if not actually 0 with a p-value of almost 0, which implies the true relationship is likely no-relationship. I tested this for multiple response variables and got pretty consistent zero-valued DiD estimators even after including controls. This seems to imply that the impacts of the progresa on other aspects of household welfare were minimal if not non-existent, at least in terms of the data use for this analysis. Maybe there are some spillover effect that were not recorded in the data.

1.3.8 2.8 Summary

- Based on all the analysis you have undertaken to date, do you believe that Progresa had a causal impact on the enrollment rates of poor households in Mexico?

Discuss your results here

I would say there is prety strong evidence of a causal impact. The randomized assignment, while not perfect, was pretty good. Even then, controlling for time-variant confounders as well as baseline differences was accomplished by using the DiD regression model with control variables. While the impact seemed diminished after these controls, it was still existent and positive even after accounting for these confounders. This leads me to believe that there is a statistically significant causal impact of the program that increases student enrollment, specifically from poor households. Overall. the experimental design allowed for relatively effective isolation of the treatment effect.