TOULOUSE SCHOOL OF ECONOMICS
L3-Spring 2022
INTRODUCTORY ECONOMETRICS

In this homework, you put yourself in the shoes of a team of economists that is asked to assess the potential effects of awareness-raising campaigns on female voting participation in a recent election in Pakistan.

In Pakistan, women are 15 percentage points less likely to vote in an election compared to men. Many potential reasons have been given to explain this behavior: it may be that the cost of participation is too high because of cultural stereotypes or mobility constraints, or it may be that men want to control if and how women vote, and so women choose not to vote at all.

To check what kind of campaign might reduce this gap in voting behavior, your colleagues conducted a door-to-door experiment in 12 villages in the rural region of Sindh: They educated some women in some neighborhoods about the importance of voting (Treatment 1: *treated1*). Some other women in some other neighborhoods received education about the importance of voting but also information regarding the balloting process (Treatment 2: *treated2*).

Which neighborhoods were "targeted" with which information was chosen as follows:

1. Each village was divided up into neighborhoods.
2. Neighborhoods were *randomly* assigned to be targeted by the campaign or not.
3. Each targeted neighborhood was *randomly* assigned to receive either treatment 1 or treatment 2.
4. However, within each of these neighborhoods, some women were *randomly* chosen to not receive the treatment. In other words, there are some untreated women in the same neighborhood as treated women.

So in some targeted neighborhoods women were educated about the importance of voting or not treated at all. In other targeted neighborhoods they were either educated about the importance of voting *and* got information regarding the balloting process, or not treated at all.

Before the campaign, your colleagues collected data on characteristics of the women in the 12 villages. As the treatment was randomly assigned, there should be no systematic differences between the women in the different groups at this stage. During the campaign, your colleagues noted which neighborhoods were targeted and which women treated by which of the two treatments. After the campaign, they collected data on the women's voting participation. Your job is now to evaluate the data to understand the effects of the campaign.

The dataset "gw2018.sas7bdat" is available on Moodle. The variables in the dataset are the following:

Treatment dummies		=1 if the woman ...
targeted_neighborhood	Dummy	... resided in a neighborhood with treated individuals
*treated*1	Dummy	... was treated by treatment 1 (If treated1=1, then treated2=0))
*treated*2	Dummy	...was treated by treatment 2 (If treated2=1, then treated1=0)
Characteristics of the woman/household before the start of the campaign:		
village_code	Categorical	Village code
neighborhood_code	Categorical	Neighborhood code
age	Numeric	Age in Completed years
land	Numeric	Size of owned land
hhsize	Numeric	Total number of household members
num_women	Numeric	Total number of women in the household
house_quality	Numeric	House quality index from pca
any_school	Dummy	1 if the woman has any schooling
d_married	Dummy	1 if the woman is married
tv_access	Dummy	1 if the house has access to a TV
cable_access	Dummy	1 if the house had cable access
advice_pir	Dummy	1 if the household follows the advice of a spiritual guide in important decisions
d_married	Dummy	1 if the woman is married
Outcome variables measured after the campaign:		
voted	Dummy	1 if the woman voted in the election

Questions

Problem 1. (3.5 points) A first overview of the data and some manipulation.

1. (0.5 point) Initialize a library called "project" in which you save the dataset. What is the total number of observations in the dataset?

2. (0.5 points) Distinguishing between qualitative variable and quantitative variables, compute simple statistics for all variables (excluding school id).

3. (1 points) What percentage of women 30 years old or younger were taught both about the importance of voting and the balloting process?

4. (1.5 points) For each village, what was the percentage of neighborhoods that was targeted?

Problem 2. (4.5 points) Consider the model M1

$$x_i = \beta_0 + \beta_1 treated1_i + \beta_2 treated2_i + u_i$$
$$E[u_i | treated1_i, treated2_i] = 0$$

where x_i is a woman characteristic or some information about the household she belongs to. Consider $x_i = any\_school_i$, a dummy equal 1 if the woman has any schooling (but the answer would be similar if we considered the other characteristics too e.g. *age*, *d_married*, *land*, *hhsize* etc.). [*In this exercises and the exercises that follow, regresssion will only require you too run the simple linear regressions found in the TPs. **Do not** attempt to correct for heteroskedasticity or use alternative estimation methods when asked to estimate a model.]

1. (1 point) What sign and magnitude do you expect from β_1 and β_2? Explain your intuition.

2. (3.5 points) Test that β_2 is statistically different from zero at 10% significance level.

   (a) State the null and the alternative hypothesis.

(b) Write the formula for the test statistic and its distribution under the null hypothesis (clearly state the degrees of freedom).
Then, provide the value of the test statistic for this test.

(c) Provide the decision rule. Then, find the critical value of the test.

(d) Do you reject the null hypothesis? Motivate your answer using:

- the test statistic
- the p-value

Problem 3. (6 points) Consider the model M2 (Main effect):

$$voted_i = \beta_0 + \beta_1 treated1_i + \beta_2 treated2_i + u_i$$
$$E[u_i | treated1_i, treated2_i] = 0$$

Where the outcome *voted* is an indicator if the woman voted or not and the independent variables *treated1* *treated2* are as described in the table above.

1. (0.5 points) Explain what we want to test in this regression.

2. (1 point) What sign and significance do you expect for β_1 when running M2? And for β_2? Which of them two should be higher? Consider carefully the definitions of those variables. Explain your intuition.

3. (1 point) First, regress the model (M2) above and get $\hat{\beta}_1$ and $\hat{\beta}_2$. Then, create the variable $age2 = age^2$. Include the following variables in the model: *age*, *age*2, *any_school*, *advice_pir* and *house_quality*. Estimate the new model and interpret the estimated coefficients for all the variables. In particular, get the values of $\tilde{\beta}_1$ and $\tilde{\beta}_2$, the coefficients in this new model on *treated*1 and *treated*2 respectively. Finally *compare* the values of $\hat{\beta}_1$ and $\hat{\beta}_2$ with $\tilde{\beta}_1$ and $\tilde{\beta}_2$.

4. (2 points) So far, we have considered each treatment separately. *For this and the next question*, we will treat them as one, using the variable *treated*; that takes the value of 1 if either *treated*1 or *treated*2 are equal to 1. Recall, the neighborhoods within villages were randomly assigned to either a control group or *treated1* or *treated2*. Consider now the following model (M3),

$$voted_i = \beta_0 + \beta_1 treated_i + + u_i$$
$$E[u_i | treated_i] = 0$$

   (a) (0.5 points) Construct the variable *treated*.

   (b) (1.5) Now, consider the population of individuals that live in a treated neighborhood. Estimate the effect of receiving any treatment on the non-treated households in targeted neighborhoods. In other words, we are estimating the effect of living in a targeted neighborhood but not actually receiving the treatment. To do this, use the variables that you consider relevant (see the list of variables above). Interpret your results and be careful in your explanation. [Hint: evaluate the previous model (M3) when the individual receives and does not receive the treatment].

5. (1.5 points) Officers from the World Bank are interested in promoting woman voting in other regions of Pakistan. They ask you for statistical advice to convince the government of Pakistan to allow them to carry the project: some people in the government fear that the program could actually have negative average effect on voting. Others think the average effects could be too large. They ask you to:

   (a) (1 point) Give an estimate of the lower bound and upper bound of the average effect of the program (M3) with 95% confidence. State clearly all the steps you use. Interpret the bounds.

   (b) (0.5 point) Give a recommendation based on your previous results and argue why you think this is a good policy.

Problem 4. (5.5+1 points) In much of the literature on interventions in developing countries, researchers have realized that effectiveness of these voting interventions can depend on local culture and dynamics within the family. In the following question, we will explore family characteristics and the gender ratio in Pakistan and test a hypothesis that these may matter to interventions that try to change voting turnout.

1. (0.5 + 1 point) In order to study the potential effect of family dynamics on voting interventions, we want to construct variables that capture the gender and age ratio within each sample household. The World Bank's data on Pakistan in 2011 suggests that the female proportion of total population is 48.5%, that the proportion of adult females of the total female population is 49%.

   (a) (0.5 points) With this, construct the variable: The number of adult females that each household should have, given the statistics above and the size of each household. Name this variable: *expected_women*.

   (b) (1 bonus point) Plot a single histogram with both the distribution of the expected number of women and the actual number of women.

2. (0.25 points) The number of expected women in each household based on the Pakistani population is quite different from that in our sample. Show this by constructing $more\_women = num\_women - expected\_women$, and provide the summary statistics for *more_women*. Comment on the distributional differences using the mean of this constructed variable.

3. (2.75 point) Does this difference in within household gender ratio matter to voting and the effect of the treatment? Construct the gender ratio variable $interact\_treated\_mw = treated * more\_women$. Regress M4:

   $$voted_i = \beta_0 + \beta_1 treated_i + \beta_2 more\_women + \beta_3 interact\_treated\_mw + u_i$$

   Test jointly at the 5% significance level if both β_2 and β_3 are different from zero. Give an interpretation for the meaning of $\beta_2 = \beta_3 = 0$ and $\beta_3 = 0$ separately.

4. (1 point) Plot a scatterplot for the relationship between *more_women* and household size, with the linear regression plotted in the plot. For what household size does the surplus of women in a household reverse sign?

5. (1 point) Study the plot of the predicted residuals in the regression for M4. Is the predicted residuals what you expected to find? What is different from the typical setting? Check the mean of the predicted residuals. Is it still zero? Why?