ARE 202C – Problem Set #1
Due April 10th *(at the beginning of class)*

This problem set requires you to download the Nicaragua_RCT.csv file from Canvas (Files > Problem Sets > Data). You may use any software package to complete the problem set, however, I will only provide hints and suggestions for Stata. To save yourself time, and to ensure you can replicate the work you do for the problem sets, I highly recommend you write your commands in a Do-file. For those of you relatively new to Stata, you can read up on Do-files here (http://www.stata.com/manuals13/u16.pdf). To load the dataset into Stata, start your Do-file with the line **import delim using** *file_name* (for help with any command, such as the import command for example, you can always just type **help import** in the command window). If you are using an older version of Stata, you can instead type **insheet using** *file_name*.

Once the dataset is loaded, you should see the 9 variables listed in the table below. Before we get started, I want to make sure it is very clear what the differences are between the "assign", "complier", and "treat" indicator variables. Initially, households were randomly assigned to a treatment group (assign = 1) or a control group (assign = 0). Only treatment group households were eligible to receive the treatment in the first stage of the program.

Each household, regardless of whether they were assigned to the treatment or control group, was also asked whether they would like to receive treatment. Households that selected to receive treatment have a value of 1 for "complier". Households that didn't select to receive the treatment have a value of 0 for "complier". Finally, if a household was assigned to the treatment group (assign = 1) and wanted to receive treatment (complier = 1), then they actually received treatment (treat = 1) in the first stage of the program. All other households did not (treat = 0).

The outcome variable of interest is the household's "income" from the specific type of farming that they do on their land. The other household variables {job, age, education, capital, land} were all observed before assignment and treatment occurred.

Variable Name	Description
assign	Assignment to "treatment" group (1 = treatment; 0 = control)
complier	1 = HH's that wanted "treatment"; 0 = HH's that didn't want "treatment"
treat	1 = HH's that were treated; 0 = HH's that were not treated
income	Annual per capita income from main crop (in $'s, adjusted for PPP)
job	1 = cattle; 2 = grain; 3 = yuca (cassava root)
age	Age of head of household
education	Number of years of education obtained by head of household
capital	Total value of mobile capital used on the farm
land	Farm size (in manzanas; 1 manzana = 1.7 acres)

1. If the randomization was successful, what would be the value of the following expression: $E[complier_i | assign_i = 1] - E[complier_i | assign_i = 0]$? Compare the mean value of "*complier*" for treatment group households ("*assign*" = 1) and the mean value of "*complier*" for control group households ("*assign*" = 0). Are they significantly different? (The **ttest** command in Stata conducts difference-of-means tests.)

2. If the randomization was successful, what other variable(s) should be uncorrelated with "*assign*"? Again, comparing sample means among these variables, is there any evidence that the treatment and control households are unbalanced?

3. We can define the Avg. Effect of Treatment on the Treated (ATT), that is, the effect of "*treat*" on "*income*" among the treated households, using the following linear model:

$$income_i = \alpha + \beta^{ATT} \cdot treat_i + \varepsilon_i$$

   If we estimate the model above on the full set of households using OLS (regressing "*income*" on "*treat*"), will $\hat{\beta}$ be an unbiased estimate of β^{ATT}, the sample ATT? Why or why not?

4. Alternatively, we could estimate the following model:

$$income_i = \delta + \beta^{ITT} \cdot assign_i + \mu_i$$

   β^{ITT} is called the Average Intention-to-Treat (ITT) effect. Estimate the model above using OLS. In words, interpret what $\hat{\beta}^{ITT}$ means. Is the OLS estimate of $\hat{\beta}^{ITT}$ an unbiased estimate of the true sample average ITT effect (β^{ITT})? Why or why not?

5. Do you expect the true value of β^{ATT} to be smaller or larger than your estimate of $\hat{\beta}^{ITT}$ from above? Why?

6. Now let's rethink the model from Question #3 above. Instead of estimating the model using the full set of households, can we produce an unbiased estimate of β^{ATT} by estimating the model on a subset of the households? If so, what subset of the households should we focus on and what is your estimate of the Average Effect of Treatment-on-Treated? Why is it an unbiased estimate?

7. Suppose we wanted to know if the avg. effect of Treatment-on-Treated (β^{ATT}) differs across different types of farmers (cattle, grain, or yuca). First, define the following three indicator variables:

$$I_{1,i} = \{1 \text{ if } job_i = \text{cattle; 0 otherwise}\}$$
$$I_{2,i} = \{1 \text{ if } job_i = \text{grain; 0 otherwise}\}$$
$$I_{3,i} = \{1 \text{ if } job_i = \text{yuca; 0 otherwise}\}$$

This can be done in a variety of ways (one option is to use the **tabulate**, or **tab**, command and using the **gen(***var_name***)** option).

If we estimate the following model using OLS, again focusing on the same subset of homes used in Question #6,

$$income_i = \alpha + \left(\beta_1 \cdot I_{1,i} + \beta_2 \cdot I_{2,i} + \beta_3 \cdot I_{3,i}\right) \cdot treat_i + \epsilon_i$$

will the OLS estimates of $\{\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3\}$ be unbiased estimates of $\{\beta_1, \beta_2, \beta_3\}$? Why or why not?

8. If your answer to **#7** above was “No, the OLS estimates are not guaranteed to be unbiased”, then is there a simple way to amend the model to produce unbiased estimates of $\{\beta_1, \beta_2, \beta_3\}$? After estimating the “amended” model from above, do you find evidence that the treatment has heterogeneous effects among the different types of farmers?