Stata Test: Pre-Doctoral Position with Prof. Nishith Prakash and Prof. Priya Mukherjee

Instructions:

The test should ideally take about 3 hours to complete, but you will be given a window of **Maximum 4 hours** to submit the test. You will be notified via email as soon as the test is sent and your time starts 15 minutes after you confirm the receipt of the test. In case of a technical error, please reach out to the sender immediately, otherwise we would not be able to compensate you for the lost time.

All necessary files for completing the test are included in the Stata test folder. When you submit the test, we want to see a complete record of your work. Please attach a **do-file** that completes all problems, a **log file** that shows the output of your code, and the relevant **tables and figures**. upload here.

.

Please Note: All the documents - Log file, do-file, pdf, tables, Latex source codes etc. should be **compressed in a single zip folder** and saved in the following format - *"FirstName_LastName"* (For eg: Abhijit_Banerjee)

Remember that you will be graded not only on the answers you give, but also on the process that you take to arrive at your answers (i.e. your do-file and comments). This exam is meant to judge your general reasoning and problem-solving ability on top of your Stata skills, so if there are any questions that you cannot answer using Stata, please explain what steps you would have taken had you known the appropriate commands, and you will be given the appropriate partial credit. Keep in mind that some questions are somewhat open-ended/ambiguous and that there is no "correct" answer (although that does not mean that there are no incorrect answers!). Please make sure to explain your reasoning and give as full an answer as you can.

As a Research Associate (RA), you will be working in a fast-paced, team-oriented environment. As such, RAs should write their do-files and comments clearly and efficiently so that it is easy for others on the team to understand and collaborate on. In addition, as the project develops, the dataset will change over time. As much you can, write code that will produce correct answers even as your dataset changes over time. For example, if a question were asking for a

maximum value for income, good code would accurately display the maximum, even if another 100 observations were collected.

Using outside resources is allowed, but please mention what resources you have used if you do so, and do not consult with other people about the test. Stata's internal help files, though, are often the most helpful resource out there.

Data quality assurance is key to the success of our project. With each module, in addition to the questions, please check for data quality and any potential problems with the data that should be flagged, and include this in the do-file you submit.

Make your code reusable so that anybody with access to your Stata .do file and dataset can replicate the results by simply executing the .do file. Ensure that you follow a consistent naming pattern across all your files and folders.

1 PART-A (60 Points)

1. **Importing:** We are running a randomized evaluation in India, with the enumerators sending the data collected daily to you. You have been assigned to do quality checks, known as high-frequency checks.

   (a) Our team has obtained data from a partner organization with extra variables that were not asked in our survey, and so we are interested in adding this to our dataset. However, the file that the organization has sent to you, *New Variables.csv*, is not in **.dta** format. Import this dataset into Stata.

   (b) Our main dataset is called *Main Dataset.dta* – open it and then merge in the *New Variables* dataset to it.

   (c) New people have been surveyed today, and their survey responses have been sent to you in the dataset - *New Observations.dta*. Add these observations to our main dataset.

2. **Descriptive Statistics:** - Now that we have our data in Stata, we would like you to play around with the data a bit and provide us with some descriptive statistics.

   (a) Create a summary statistics table which has at least the mean, median and standard deviation of the following variables: *income, age, surveytime.*

   (b) Now create a tabular representation of the above statistics in a publishable format - preferably in Latex or rtf and save the table.

   (c) Checking the uniqueness of our ID variables is a critical task. Is the *hhid* variable unique (i.e. different values for every observation) in the dataset? If it isn't, list out these duplicated IDs. What might we want to do to try and resolve an issue of duplicate IDs?

(d) With the information that you possess so far, create a missing values plot of any 4 variables of your choice. Provide appropriate labels, title, and subtitle to the plot

(e) Create histograms and overlay them with k-density plots for any 2 variables of your choice. Provide appropriate labels, title, and subtitles to your plots.

3. **Cleaning:** - After collecting all of our data, we want to clean it up so that it is ready to be used for analysis.

   (a) A key component of data cleaning is removing personally identifiable information (PII). When working with this dataset, we wouldn't want surveyor names to be known. Replace surveyor names with corresponding numbers. For example, each value of the surveyor "Benjamin" should be replaced with "1", and likewise a unique id should be assigned to all the other surveyors. This variable should be in numeric format.

   (b) If you browse the data, you'll see a number of negative values such as -999. These are missing values but, as of now, provide us with little information. Referring to the associated table "Missing Value Codes" and only for variables *burglaryyn, vandalismyn, trespassingyn,* recode these negative values with their corresponding extending missing value

2 PART-B (40 Points)

Research Question: Effect of increases in tenure security through land ownership rights in the household labor supply.

Some essential details about the study: It is a Survey of 2750 urban households of Peru conducted in March 2000. The Sample was drawn randomly based on the universe of households that in the 1993 Census declared to be living informally. The Stratification is on the city level, with clusters of 10 households randomly sampled by neighborhood.

Identification Strategy: Difference-in-difference (DID) strategy : Comparison of the difference in labor supply of program beneficiary and non-beneficiary households in early neighborhoods to the difference between beneficiaries and non-beneficiaries in late neighborhoods (future program - after the survey).

1. Import the data set *analysis_sample.dta*

2. Report the *Survey Design.* (Primary Sampling Unit (PSU), Stratification, Standard Error Clustering, Finite Population Correction (FPC).)

3. **Balance Table:** Prepare a balance table (or Difference-in-Means Table) for Households with Land Titles before the intervention (variable *hastitle*=1) vs those without Titles (variable *hastitle*=0). Use the following variables to check the balancedness: *agehd, avgage, elemhd, members, pctmale*. On the basis of the balance table, comment on whether the randomization was successful or not.

4. **kdensity Plot:** Households with Land Titles before the intervention have the variable *hastitle*=1 and likewise, those without Titles have been assigned a value of 0. Similarly, those who were program beneficiaries have a variable *enter*=1 and 0 likewise for the rest. Lastly, those households who did not have land rights before the program and were living in squatters have *squat*=1 and 0 for the others. Generate four kdensity plots for the variable **totwkhrs** (HH Total Weekly Work Hours) for the various categories and use the following labels:

   - "No program, Not Titled:" if *hastitle*=0 *enter*=0
   - "Program, Titled:" if *hastitle*=1 *enter*=1
   - "No program, Squatter" if *enter*=0 *squat*=1
   - "Program, Squatter" if *enter*=1 *squat*=1