Statistical Analysis Exercise (v22)

ECONOMIC GROWTH CENTER – YALE UNIVERSITY

Instructions

You have **48** hours from when you receive this exam to complete it and submit your solutions. Late submissions will be penalized. You may consult any resources you like, except for other people. List all resources that you consulted (except built-in software documentation) in your write-up. Consulting resources does not result in penalties. All necessary files for completing this exam are included in this email. This test is designed to be completed in any version of Stata, but you may complete it in another statistical program (such as R) if you convert the provided .dta datasets.

Be careful and methodical with your work. Ensure your code file is well formatted and annotated. Correctness counts, but so does clarity and brevity. Organize your work as though you are going to pass off the project to be maintained and extended by another RA. If you choose to include purely interactive commands like `summarize`, comment them out in your final code file.

The dataset you are working with simulates a dataset that has come from fieldwork and may contain errors. If you notice any outliers or other clearly incorrect values, deal with them appropriately.

If you become stuck, explain (in comments in your code file) what you would have done if you had more time or knew the correct commands. Complete as much of the exam as you can. If your answer depends on previous steps that you were unable to complete, you can still earn points for demonstrating that you would have arrived at the correct answer if you had successfully completed all previous steps.

Follow all instructions precisely. If you choose to deviate from the instructions, explain how and why you did so.

The instructions are not intended to be confusing. If any instructions are unclear, do your best, and tell us what was unclear.

Return the following deliverables in a zipped archive:

1. Your do-file(s), or other type of code file
2. A log file generated by the do-file itself, or other type of log file
3. Your final (merged) dataset
4. Any outputs (e.g. datasets, graphs, tables)
5. Either as comments in your code file or as a separate document:
    a. Answers to discussion/analysis questions
    b. A list of consulted resources
    c. A brief description of your experience with Stata or other programming languages[1]

Include the following certification in the body of your email:

I certify that the attached materials represent my own work and that I did not receive any aid on this exam. If found to be in contravention of this statement, I am aware that I will be disqualified from this and any other positions with Inclusion Economics at Yale University or the Economic Growth Center.

IMPORTANT: In an effort to make evidence-based changes to our recruiting process in order to reduce bias, all candidates are instructed to use a Candidate Number rather than their name in their submissions. Your Candidate Number was sent to you in the email that contained this attachment. Please use only this number in your submissions, including in your file paths in your .do files and your saved file names.

We will assess your answers in four categories:

1. Basic coding skills
2. Advanced coding skills
3. Code clarity and replicability
4. Economic intuition and interpretation of statistical tests

Dataset

From 2009 to 2016, a cluster randomized controlled trial was implemented by several researchers to study the impact of local banks' expansion on households' loans, savings, and insurance taking behavior. The research project sought to evaluate the impact of the intervention on a wide range of outcomes that reflect household well-being (e.g. household income, consumption), as well as individual well-being (e.g. women's empowerment, health).

The research implementation partner was a large financial institution in rural South India that randomly expanded bank infrastructure across villages (called service areas) in three districts of Tamil Nadu, India.

[1] Did you learn Stata from a class or on your own? Were you an RA or have you done independent research? How many projects have you completed? This helps us understand your background and potential. Please be honest if you have scant experience; many successful applicants have never used Stata prior to this exam.

In 2009, 101 service areas across three districts were identified, which formed 50 service area pairs. One service area "pair" is a triplet, containing one treatment area and two control areas. A bank branch was assigned to each service area, and branches in treatment areas were opened at the time of assignment, while branches in control areas were opened 18 -24 months later. More than 4,000 households were randomly selected across all service areas to be included in the study. The opening of bank branches happened in three rounds. Thus, there are three rounds of baseline surveys, and three rounds of endline surveys.

The data and description of this study have been altered for the purposes of this Stata exam. Correct answers will not conform to any published results.

Codebook

treatment_status.csv

- pair_id: uniquely identifies a service area control and treatment pair
- group_id: uniquely identifies one service area
- treated: indicator ==1 denoting a member of the treatment group

endline.dta

- hhid: unique identifier for each household
- totformalborrow_24: total formal borrowed amount (loans) in Indian Rupees in the past 24 months
- totinformalborrow_24: total informal borrowed amount (loans) in Indian Rupees in the past 24 months
- hhinc: self-reported total household income in the last 30 days (in rupees)
- survey_round: the round of the survey, either endline 1, 2 or 3
- hhnomembers: number of household members in each household

baseline_controls.dta

- This dataset contains baseline household demographics including gender, age, education of head of household, household religion, and household caste.

Exercises

1. DATA PREPARATION

a) Load the endline data.
b) Recode household debt and income variables as numeric values instead of strings, and replace "None" with 0.
c) Browse the variables in this dataset, and write a few sentences about the financial status of households in this sample, supported by this data. Feel free to use a table or figure to support your argument.
d) Top code household debt and income variables, replacing all values greater than three standard deviations above the mean with a value that is equal to three standard deviations above the mean.
e) Label the new top-coded variables.
f) Write a few sentences about why we might want to top code these types of survey responses from households, and give an example of another data quality or cleaning check you might want to implement in this type of data.
g) Create a total borrowed amount variable that equals the sum of formal and informal borrowed amounts.

h) Merge the endline data with the *treatment_status* dataset to assign a treatment status for each household.
i) Create a dummy variable for households that are below the poverty line using a daily per capita poverty line of 26.995 INR (which was equivalent to 1.90 USD at the time of data collection). Use the endline top coded "hhinc" variable, which contains self-reported household income over the past 30 days, in order to do this.
j) Write a few sentences about the strengths and limitations of using the dummy you created to assess a household's poverty status. If you were able to collect more data from these households, what types of additional questions might you ask?
k) Merge your working data with the baseline controls dataset, and save the merged data. If you need to make decisions about dropping mismatched values, please justify them in notes.

2. ANALYSIS

a) In a sentence or two, state a testable hypothesis about one of the possible impacts of this program, either on a particular outcome of interest, or for a particular sub-group of participants. Justify your prior (or prediction) for this particular treatment effect.
b) Choose a few baseline household variables, and perform t-tests or produce a balance table to test for the significance of differences between the treatment and control groups.
    i. Why did you choose these particular variables to test?
    ii. What are the results of the test, and what can they tell us about the validity of the experiment?
    iii. Please present the t-tests or balance checks in a table.
c) Regress (with OLS) the household income on the treatment dummy. Include pair fixed effects, and correct standard errors if necessary.
    i. Explain why you think it might be appropriate to use a fixed effects specification in this case, and how you would interpret the effect of the treatment on household income in this case. Interpret your results.
    ii. Briefly justify your choice of standard errors.
d) Generate a log income variable, and re-run the previous specification with log household income as the dependent variable.
    a. What are the key differences between the results of this regression and the results of your previous specification?
e) Re-run the previous regression including a set of household-level controls.
    a. Explain why you chose these controls, and if there are key differences in your results as compared to previous specifications.
    b. Export and save a regression table **suitable for publication** from these results.
f) Your research team needs to present data from this study before a policy audience. Using Stata, create a bar chart **suitable for publication** that summarizes the average borrowed amount for each income quartile, by treatment group. Save the chart as a PNG file.