CoRE-IPD Summer School in Ghana
Problem Set on RCT and DID

Leonard Krapf

July 2025

Setting

The file tanzania.dta contains a fictional dataset of 8400 individuals from 80 Tanzanian villages. Study participants were repeatedly interviewed in three survey waves (0,1,2). After wave 1, a subset of villages was randomly selected with 50% probability to participate in a financial inclusion campaign aimed at increasing the usage of saving accounts. Therefore, the effect of the treatment, if any, would show in wave 2. In this treatment, each individual received their own savings account linked to their mobile phone. Usage of the account was measured by asking participants whether they had used the account within the last week before the survey. You, as a researcher, have been asked to perform an evaluation of this policy intervention. Solve the exercises listed below. When asked to interpret/compare your results, please pay special attention to causality, effect sizes, and statistical significance.

Question 1

a) Given that the villages were randomly selected into treatment, you decide to simply compare the means of villages participating in the financial inclusion campaign (treatment) to those which did not (control). Interpret the difference in means. Hint: You should do this using only the rows from the second wave.

b) Assuming the RCT was truly successful in randomization, perform an OLS regression analysis (without any extra controls) in order to estimate the effect of the intervention on the probability of using the savings account in the week before the interview. Interpret your results. Are the conclusions different from a comparison in means?

c) Assuming randomization was successful, will this regression yield an average treatment effect (ATE) or an average treatment of the treated (ATT)? Explain your answer.

Question 2

a) You suspect that the implementing official of your study has rigged the treatment assignment process in a way that favored villages closer to her home village. Would this bias your results in any way and if so, under which conditions?

b) Still focusing on wave 2, check whether the distribution of distance to the official's home village is significantly correlated with the probability of treatment. Then, check whether the education of the household head is significantly correlated with the probability of treatment. You can use a simple OLS regression for each.

c) Assuming these two variables did affect the selection process, can you adapt your RCT calculations accordingly? If yes, do so and interpret your results. Does it seem that the RCT can be trusted to have been completely random?

Question 3

Consider these assumptions to hold. Wave 0 is the prebaseline, 1 is the baseline, and wave 2 is the endline.

a) Estimate the DiD regression, without any controls. How do your results compare to those in 1b?

b) You would like to rerun the D-in-D with controls. Control for both distance of the household to the official's village and the number of years of household head's education. How do your results about the effect of the treatment change from what you obtained in 3a?

c) What does that tell you about using controls at the individual level (when the treatment is at a larger group/geographic level such as the village) in a D-in-D framework? Does the causal estimate of a Diff-in-Diff rely on the control and treatment groups having the same "baselines" of covariates (such as individuals having on average the same education, or the same distance to the official's village)?

Question 4

Now you want to try to assess the plausibility of the key assumption on which your Diff in Diff in based, discussed in 3b. Plot the outcome of interest (savings account usage) from wave 0 to wave 1 to wave 2, for both control and treatment. Create a vertical intercept at wave 1. Comment on what you see.

Question 5

Now you want to run some placebo regressions to further help increase your confidence in your D-in-D results.

a) Create a regression with a fake treatment time - let wave 1 be the start of the post-treatment period, instead of wave 2, and estimate the Diff in Diff regression. Comment on your results and whether they are good news (inspire confidence in your original Diff in Diff) or not.

b) Go back to the "true" treatment times, but now let the outcome of interest be something else - monthly income, instead of savings account usage. Run the Diff in Diff regression with the fake treatment outcome. Comment on your results and whether they are good news (inspire confidence in your original Diff in Diff) or not.

BONUS

a) Discuss which method (OLS based on RCT, or D-in-D) seems more plausible for getting a causal estimate in this dataset, and therefore which results you want to rely on. What do these results say about the effect of the treatment intervention in these villages, i.e. about the success of the inclusion campaign in affecting the use of a savings account?

b) If the intervention was successful, what more information would you need to know to recommend to policymakers whether it is "worth it" to pursue such an intervention?