Economics 172: Problem Set #2

Due on October 20, 2025 at 11:00pm

Zachary Brandt
zbrandt@berkeley.edu

Problem 1: RCT

This question uses an adapted dataset based on Muralidharan, Singh, and Ganimian's (2019) paper *Disrupting Education? Experimental Evidence on Technology-Aided Instruction in India*. The paper is available online and the replication data is available on ICPSR. Download the adapted dataset from bCourses.

This project evaluated the impact of a center-based and technology-aided after-school educational program on math and Hindi performance among middle schoolers living in low-income neighborhoods in urban India. The technology-based curriculum was designed to be high-quality, adaptive, and engaging. Approximately 600 middle schoolers were recruited to participate in the study. Half of these recruited students were randomly allocated by lottery to receive a voucher to participate in the program (treatment group), and half were not (control group). For the purposes of this question, you can assume that 100% of those assigned to treatment participated in the program and that 0% of those assigned to control participated in the program.

Below is a list of the variables included in the dataset, with a brief description of each. Note that BL refers to versions of variables collected at baseline (collected before the program began) while EL refers to variables collected at endline (collected at the conclusion of the program).

Variable	**Description**
student_id	Identification numbers that uniquely identify students
student_age	Age of the student (collected at baseline)
student_female	Indicator (0/1) for whether the student is female
student_grade	Grade of the student (collected at baseline)
treatment	Indicator (0/1) for treatment status of the student
BL_math_percent, EL_math_percent	Math score in percent correct
BL_hindi_percent, EL_hindi_percent	Hindi score in percent correct
BL_ses_index, EL_ses_index	Household wealth index

1. Import the data and generate a table of summary statistics. What is the range of ages and grade levels within this sample? What are average scores for math and Hindi at baseline? At endline?
2. Next you'll check whether the treatment and control groups are balanced in terms of each of the following variables: (1) age, (2) sex, (3) household wealth index at baseline, (4) math scores at baseline, and (5) Hindi scores at baseline. (Hint: To do this, you'll run a separate regression for each of these five variables. Use `stargazer` or other preferable command to output the tables.) Before you code up any regressions, write down the regression you plan to run for at least one of the variables. Which parameter represents the coefficient of interest? What do you expect the estimate of this parameter to be (positive, negative, zero) and why?
3. Use the `lm` command to run the regressions in R. Are there significant differences between the treatment and control groups for any of these five variables? What is the purpose of this exercise, and what are you able to conclude?
4. Next, you'll estimate the impact of the treatment on math and Hindi scores at endline.
   (a) First, write the two regressions you will run to estimate these treatment effects. In words, what will each of these parameters capture?
   (b) Run the two regressions using the `lm` command and use the `stargazer` command to produce tables containing the results of these two regressions. Interpret your results. What is the effect of the treatment on each of math and Hindi scores? Are these estimated treatment effects statistically significant? Explain.

Problem 1 continued on next page...

Solution

Statistic	N	Mean	St. Dev.	Min	Max
student_id	533	310.246	177.595	1	619
treatment	533	0.493	0.500	0	1
student_age	533	12.413	1.357	10	15
student_female	533	0.771	0.421	0	1
student_grade	533	7.182	1.101	4	9
BL_math_percent	533	0.316	0.124	0.010	0.758
BL_hindi_percent	533	0.435	0.167	0.041	0.923
BL_ses_index	533	−0.053	1.657	−5.548	4.117
EL_math_percent	533	0.504	0.179	−0.009	1.007
EL_hindi_percent	533	0.555	0.193	0.072	1.005
EL_ses_index	533	−0.059	1.661	−5.681	4.128

i) Above is a table of summary statistics I have generated for the data.

ii) To check whether the treatment and control groups are balanced in terms of each of the variables age, sex, household wealth at baseline, math scores at baseline, and Hindi scores at baseline, I plan to run the following regression.

$$\text{student\_age}_i = \alpha + \beta \cdot \text{treatment}_i + \epsilon_i$$

where β represents the coefficient of interest, the average difference in student_age between the treatment and control groups. I expect the estimate of this paramter to be zero since the treatment and control groups should be balanced in terms of this variable and the four others mentioned.

	Dependent variable:				
	student_age	student_female	ses_index	math_percent	hindi_percent
	(1)	(2)	(3)	(4)	(5)
treatment	0.153	0.001	−0.191	−0.014	0.010
	(0.117)	(0.036)	(0.143)	(0.011)	(0.014)
Constant	12.337***	0.770***	0.041	0.323***	0.430***
	(0.083)	(0.026)	(0.101)	(0.008)	(0.010)
Observations	533	533	533	533	533
R^2	0.003	0.00000	0.003	0.003	0.001
Adjusted R^2	0.001	−0.002	0.001	0.001	−0.001
Std. Error (df = 531)	1.356	0.421	1.656	0.124	0.167
F Statistic (df = 1; 531)	1.707	0.002	1.766	1.572	0.522

iii) Above is the regression table for the five variables. There are no statistically significant coefficients on the treatment variable for any of the regressions, suggesting that there are no significant differences between treatment and control groups across these five variables.

Problem 1 continued on next page...

iv) To estimate the impact of the treatment on math and Hindi scores, I will run the following two regressions.

$$\text{EL\_math\_percent}_i = \alpha + \beta \cdot \text{treatment}_i + \epsilon_i$$
$$\text{EL\_hindi\_percent}_i = \alpha + \beta \cdot \text{treatment}_i + \epsilon_i$$

The coefficient of interest in both of these regressions is the paramter β, which captures the effect of the treatment on math and Hindi scores.

	Dependent variable:	
	EL_math_percent	EL_hindi_percent
	(1)	(2)
treatment	0.077***	0.065***
	(0.015)	(0.016)
Constant	0.466***	0.523***
	(0.011)	(0.012)
Observations	533	533
R^2	0.047	0.029
Adjusted R^2	0.045	0.027
Residual Std. Error (df = 531)	0.175	0.190
F Statistic (df = 1; 531)	26.040***	15.689***

Problem 2: Diff-in-Diff

We often study the effect of interventions on those who actually received treatment. However, it is possible that economic interventions can create positive externalities for non-recipient households, for instance, through market interactions or financial transfers between households. In this part of the problem set, we will estimate the spillover effects of a fictitious cash transfer program in a rural Tanzanian setting.

Starting in 2015, imagine that thousands of households in rural northern Tanzania received large cash transfers of roughly US $1,000 from the international non-governmental organization GiveDirectly (GD). These cash transfers were targeted to relatively poor households within certain villages but not others. The cash grants were unconditional, and the amount is equivalent to more than 50% of total annual income for many recipient households. GD expanded cash transfers in some villages and not others in a non-randomized fashion. Then in 2018, approximately three years after the distribution of the cash transfers, the households were surveyed and measures were collected on many economic, social and life outcomes, including household per capita income.

The goal of this question is to estimate the impact of these transfers on the households deemed to be too wealthy to be eligible for GD transfers. Comparisons will be made across these ineligible households in the "cash" villages (locations where poor households received cash transfers from GD) versus in control villages (where poor household did not receive GD transfers). Any effects on these ineligible households can be considered spillovers since none of them received cash transfers through the program.

1. Discuss the econometric assumptions needed to make a difference-in-differences (DD) approach appropriate in the case of understanding the spillover impacts of cash transfers on household incomes.

2. Why is it important that we have access to data from both the baseline (pre-intervention) and endline (post-intervention) survey rounds?

3. How does the lack of randomization in the allocation of cash transfers across villages affect the estimation of treatment effects, and how might it lead to omitted variable bias?

4. Please download the data to be used in the analysis. The dataset `Ec172_Fall25_PS2Q2_data.csv` is a partial extract of project data (although the actual data has been modified in various ways). Each observation (row) in the dataset represents one household in one time period.

   All households in this problem set dataset are **ineligible** for the GD cash transfers (i.e., they are too wealthy to receive cash transfers). There are two observations for each household, one from the baseline survey (time=0) and one from the endline survey (time=1). Using the `lm.cluster` command, determine the average difference at baseline (time=0) between the households in villages where eligible households received cash transfers (`cash`=1) versus the households in control villages that did not receive transfers (`cash`=0) for each of the following two characteristics: household income per capita (`income_pc`) and if the household has an elderly member (`elder`=1). Importantly, these characteristics were collected by the survey team in the baseline survey before the cash transfers were sent out. (Consider these two characteristics one by one, that is, you should run two separate regressions here.) Make sure to account for the correlation among households in the same village by "clustering" your standard errors by village (`village_code`).

   Hint: To cluster SEs we can use the `lm.cluster` command or the `feols` command. To use the `lm.cluster` command in R, first install the packages `miceadds` and `sandwich`, and then load them into your library. The `lm.cluster` command functions just like `lm`, but with the added option to specify a clustering variable.

   For the `feols` command, install and load the `fixest` package. Clustering is specified at the end of the command using `, cluster  village_code`. The syntax of `feols` is otherwise similar to `lm`, except

that you do not include `formula =` before the regression equation.

(a) Report the regression output for the two regressions mentioned above, and interpret the coefficients. Please also discuss the standard errors and t-statistics. Taken together, are ineligible households similar in cash villages versus control villages along these two dimensions? How does this finding inform the discussion of the validity of the difference-in-differences (DD) approach that you laid out above, if at all?

(b) Carry out a difference-in-differences (DD) analysis estimating the spillover impact of GD cash transfers among ineligible households, where the outcome of interest is per capita income (`income_pc`). You will need to use data from both time periods. Once again, make sure to account for the correlation among households in the same village by clustering by village (`village_code`).

(Hint: recall that a DD analysis requires the inclusion of separate explanatory variables for post (`time`), treatment (`cash`), and an interaction between the two variables. Note you will need to create a new variable in your dataset for this interaction term.)

What is the spillover impact among ineligible households when poor households in their village receive large cash transfers? Is this effect significantly different from zero at 95% confidence?

(c) Discuss at least two reasons why cash transfers received by other households in one's village might generate spillover effects over time even among non-recipient households. Given these likely mechanisms, does the magnitude of the estimated effect in part (b) seem plausible? If not, why not?

Solution

1. The econometric assumption needed to make a difference-in-differences approach appropriate in this case of estimating the spillover impact of cash transfers on household income is the parallel trends assumption. That is, in the absence of the treatment, the treatment and control groups experience the same trend in outcomes over time. In this case, that means that the ineligible households in "cash" villages and control villages would have followed the same trend in household incomes if there had been no cash transfers to eligible households in their respective villages.

2. Access to data from both the baseline and endline survey rounds is important because without baseline data, it is impossible to estimate the difference-in-differences effect. The treatment counterfactual trend is estimated using the baseline data, which is subtracted from the post-treatment, endline data difference to isolate the treatment effect.

3. While without randomization, a difference-in-differences approach removes time-invariant omitted variable bias if the trends are in parallel. However, if there are differences between the treatment and control group villages that determine different growth trajectories in household income over time, then the lack of randomization can lead to omitted variable bias.

4. (a) The regression output for household income per capita on cash village status at baseline shows a coefficient of 4.061 with a standard error of 11.437, resulting in a t-statistic of 0.355. This indicates that there is no statistically significant difference in household income per capita between ineligible households in cash villages and control villages at baseline.

   For the regression of having an elderly member on cash village status at baseline, the coefficient is -0.026 with a standard error of 0.024, yielding a t-statistic of -1.100. This suggests that there is a marginally significant difference in the proportion of households with elderly members between cash and control villages at baseline.

   Taken together, these findings suggest that ineligible households are relatively similar in cash

villages versus control villages along these two dimensions, which supports the validity of the difference-in-differences approach.

(b) The difference-in-differences analysis estimating the spillover impact of GD cash transfers on per capita income yields a coefficient of 4.061 on cash village status with a standard error of 11.438, resulting in a t-statistic of 0.355. This indicates that the spillover effect among ineligible households is positive and statistically significant at the 95% confidence level. On the post variable (time), the coefficient is 158.990 with a standard error of 20.143, yielding a t-statistic of 7.893, indicating a significant increase in income over time for all households. The interaction term between cash village status and time has a coefficient of 62.190 with a standard error of 33.512, resulting in a t-statistic of 1.856. This suggests that the spillover effect of cash transfers on ineligible households is positive but not statistically significant at the 95% confidence level.

(c) There are several reasons why cash transfers received by other households in one's village might generate spillover effects over time among non-recipient households. First, cash transfers can increase overall economic activity in the village, leading to higher demand for goods and services that benefits non-recipient households. Second, cash transfers may lead to increased social interactions and sharing of resources among households, which can improve the economic well-being of non-recipient households. Given these mechanisms, the magnitude of the estimated effect in part (b) seems plausible, as the positive spillover effect aligns with the expected economic dynamics in the village following cash transfers to eligible households.