**LPO 7870 Research Design and Data Analysis II**

## Peabody College, Department of Leadership Policy and Organizations

**Final Project**

**Spring 2024**

**Instructions**

1. **Please choose Option 1 or Option 2 below.**
2. **You may work alone or in groups of two. You may pair with any person you choose, and we can facilitate matches if needed (please email us ASAP if you are looking for a partner).**
3. **Due date: April 29th at 8:00 pm.**
4. **Late submissions will be penalized in proportion to the time of submission.**

**Final Project Option 1**

For this option, please choose one of the topics below. Under each topic, we have identified three published studies that address a similar research question. You are to write a five-page (double-spaced, standard margins) review of the three studies that addresses: (a) their research questions, data and sample studied, and main findings, (b) the research designs used, and (c) the advantages and disadvantages of each research design. The paper should conclude with a brief set of key take-away points for policymakers. This paper should not be a simple summary of each study in succession. Rather, it should engage with the three studies as a group: how have these different authors approached the topic? Do they come to similar or different findings? If different, what might account for their different findings? Do not use AI products on this assignment. Some additional advice for this project option is provided below (following the topic list).

**School Vouchers – U.S.**

Erickson, H. H., Mills, J. N., & Wolf, P. J. (2021). The Effects of the Louisiana Scholarship Program on Student Achievement and College Entrance. Journal Of Research on Educational Effectiveness, 14(4), 861-899. <https://doi.org/10.1080/19345747.2021.1938311>

Chingos, M. M., & Peterson, P. E. (2015). Experimentally estimated impacts of school vouchers on college enrollment and degree attainment. *Journal of Public Economics*, 122(0), 1--12. <http://dx.doi.org/10.1016/j.jpubeco.2014.11.013>

Waddington, R. J., & Berends, M. (2018). Impact of the Indiana Choice Scholarship Program: Achievement Effects for Students in Upper Elementary and Middle School. *Journal of Policy Analysis and Management*, 37(4), 783--808. <https://doi.org/10.1002/pam.22086>

**School Vouchers – International**

Angrist, J. (2002). Vouchers for Private Schooling in Colombia: Evidence from a Randomized Natural Experiment. *American Economic Review, 92*(5), 1535--1558. <http://dx.doi.org/10.1257/000282802762024629>

McEwan, P. J., & Carnoy, M. (2000). The Effectiveness and Efficiency of Private Schools in Chile's Voucher System. *Educational Evaluation and Policy Analysis, 22*(3), 213--239. <https://doi.org/10.3102/01623737022003213>

\*\* Note: for this paper, focus on the “effectiveness” results, not the cost function sections.

Muralidharan, K., & Sundararaman, V. (2015). The Aggregate Effect of School Choice: Evidence from a Two-Stage Experiment in India. *The Quarterly Journal of Economics, 130*(3), 1011--1066. <https://doi.org/10.1093/qje/qjv013>

**Financial Aid and the FAFSA – U.S.**

Bettinger, E. P., Long, B. T., Oreopoulos, P., & Sanbonmatsu, L. (2012). The Role of Application Assistance and Information in College Decisions: Results from the H&R Block FAFSA Experiment. *The Quarterly Journal of Economics*, 127(3), 1205--1242. <http://qje.oxfordjournals.org/content/127/3/1205.abstract>

Page, L. C., Sacerdote, B. I., Goldrick-Rab, S., & Castleman, B. L. (2022). Financial Aid Nudges: A National Experiment with Informational Interventions. *Educational Evaluation and Policy Analysis*. <https://doi.org/10.3102/01623737221111403>

Narayan, A. (2020). Does Simplifying the College Financial Aid Process Matter? *Economics of Education Review*, 75, 101959. <https://doi.org/10.1016/j.econedurev.2020.101959>

**Conditional Cash Transfers – Latin America**

Maluccio, J. A., & Flores, R. (2005). Impact Evaluation of a Conditional Cash Transfer Program: The Nicaraguan Red de Protección Social. Research Report 141 International Food Policy Research Institute, Washington, DC. <https://www.foodsecurityportal.org/sites/default/files/2020-10/Impact_Evaluation_of_Conditional_Cash_Transfer_Nicaragua_2005_0.pdf>

\*\* Note: this is a long report. Focus on the short-run effects on enrollment.

Schady, N., & Araujo, M. C. (2008). Cash transfers, conditions, and school enrollment in Ecuador. *Economía* 8 (2), 43–70. <https://www.jstor.org/stable/20065524>

Schultz, T. P. (2004). School Subsidies for the Poor: Evaluating the Mexican *Progresa* Poverty Program. *Journal of Development Economics* 74 (1): 199–250. <https://www.sciencedirect.com/science/article/pii/S0304387803001858>

Additional advice for Project Option 1

* Some of these papers cover a lot of territory. *Focus on the big picture* and on how the three papers connect. What is the central research question(s) addressed by each paper? What is the primary empirical approach taken by the papers?
* Some of these papers are technical and/or long and in places will feel over your head. *This is OK*. We do not expect you to fully understand *everything* that the authors are doing. The goal is to uncover the key ideas, research question, design, and results. It is a valuable skill to be able to do this with technical papers.
* Some additional prompts that may help you get started:
  + Is the data non-experimental in nature, or the result of a randomized experiment?
  + What are the key outcomes (dependent variables) used in the paper? What is the key explanatory variable?
  + How does the research design fare in terms of its *internal validity* (ability to support causal inferences)? How does it fare in terms of *external validity* (ability to generalized to other populations)?
  + Is there anything you find unconvincing about the paper? Are there possible alternative explanations for its findings?
  + What makes the research question an interesting or important one, from a policy perspective? Why might a policymaker like to know a convincing answer to this question? How might a policymaker use these results?

**Final Project Option 2**

This option includes both conceptual questions and analysis of data from a Conditional Cash Transfer (CCT) program. There is no page minimum or maximum associated with this option, but responses to all of the below questions must be included. Sumit a Word document with formatted tables and the relevant parts of your Stata output. Part of your grade will be based on the clarity of submission, including text, tables, and graphs. BEFORE YOU ATTEMPT ANYTHING: read over the questions carefully and think. Think about the comparisons you are making in each case. Some of these items are challenging. You may consult with your classmates and Sara, Kathryn, and the professors for advice.

**\*\*\***

This project uses data from a CCT program. The program gave cash transfers to low-income households in exchange for student attendance in school. **The objective of the exercise is to determine if the policy resulted in *lower* educational expenditures at the household level.** This is important to determine since part of the effect of the CCT may come from a reduced burden on households to pay for education.

A pilot of the program was implemented in several communities. Here are a few important details about the program:

* Authorities identified communities for the pilot that had desirable characteristics: a majority of low-income households, accessibility to the government, etc. The total list of potential communities was longer than could be served by the program. As such, the government used randomization to determine treatment.
* Information on assets and human capital (e.g., educational attainment) of all households in ALL communities were collected. Based on that information, an index was created, and households were deemed eligible or not for the program (see the variable “eligible”). Poor, eligible households have scores lower than 750 points.
* Communities were selected at random for the pilot program. (The variable “treatcom” equals 1 for communities selected for treatment and equals 0 otherwise). Within the treated communities, only eligible households were entitled to the program (the “eligible” variable).
* Some eligible families in the treated communities did not take up the treatment (see the variable “takeup”.)

We are providing relevant parts of the dataset that address each of the points above. All datasets include the same variables. The data include baseline measures (variable “round”=0) and one follow-up data collection, one year after the start of the program (“round”=1). Here is a description of the variables:



**The main outcome variable is “oee”, out of pocket education expenditure per person**; the values are in monthly per-capita dollars. The data has the following variables measuring characteristics of the household, head of the household, and spouse/partner: head’s age (age\_hh); age of spouse (age\_sp); education of the head (educ\_hh) and of spouse (educ\_sp); ethnicity (ethnicity\_hh); whether the head is female (female\_hh); the size of the household at baseline (hhsize\_basal); if the household had a dirt floor (dirtfloor\_basal); if it has a bathroom (bathroom\_basal); and the number of hectares of land (landhectars\_basal).

Other variables are:

* local: id of the locality (local)
* hhid: id of the household
* round: 0 if baseline, 1 follow-up
* takeup: if the family participated in the program (or not)
* eligible: eligibility of the hh; e.g. the family has an score<=750, the family is eligible.
* treatcom: if the community was part of the pilot
* score: poverty index (family is poor if score<=750).

You are going to produce several sets of estimates of the causal effect of this program. Please proceed systematically. Again, please think about the relevant comparisons (and the fact that there are two points at which data were collected: at baseline, before anything happens (round=0), and at follow up, after implementation of the program (round=1).

1. Explain Take up. (20 points)

For this question, use **Question1\_Dataset\_Final**. It contains only information at baseline (round==0) for eligible families.

Before estimating the effect of the program, you are asked to analyze take up (variable takeup). Your hypotheses is that takeup (which takes a value of 0 or 1) depends on characteristics of the household at baseline. You should include the following variables: age\_hh; educ\_hh; female\_hh; hhsize\_basal; dirtfloor\_basal; bathroom\_basal; and landhectars\_basal.

1. One way to do this is to estimate a linear probability model of takeup as a function of the characteristics of the household. Please write down the main equation of the model, explaining each variable and the potential problems associated with this model.
2. Estimate the model. Does gender of the head of the household (female\_hh) predict take-up?
3. Another approach would be to estimate a logit model. Please describe the main equation of the model, explaining each variable. What is the advantage of this model over a linear probability model?
4. Estimate the logit model, report odds-ratio coefficients, and the marginal effects (after the logit regression use the “margins” command. Note you will need to choose values of the predictor variables at which to calculate the marginal effects. We recommend doing this at the means.
5. Compare your estimates from the linear probability model and the logit model.
6. Randomized control trial (RCT) (40 points)

For this question, use **Question2\_Dataset\_Final**. (It contains only eligible households, for the two rounds of data)

Now, you are going to estimate the experimental effect of the program. You would like to estimate the effect of the program on the treated people (treatcom=1 if treated, 0 otherwise). The proposed model is the following (applicable to time/round 1):

1. In this case, please explain the assumption . Do you think that the assumption holds in this case? Why or why not?
2. Please present a test that shows that the baseline characteristics (age\_hh, age\_sp, educ\_hh, educ\_sp, female\_hh, dirtfloor\_basal, bathroom\_basal, landhectars\_basal) for eligible households are balanced between treatment and control communities ().
3. Now estimate the following regression using observations from time/round 1:
4. Is the program effective in reducing expenditure in the household?
5. Is this effect significant?
6. Provide a 90% confidence interval for the effect.
7. Now re-estimate the regression in part (c), adding controls for baseline characteristics of the household (age\_hh, age\_sp, educ\_hh, educ\_sp, female\_hh, dirtfloor\_basal, bathroom\_basal, landhectars\_basal) Interpret the coefficient on “treatcom”? Is this estimator unbiased? Please explain.
8. Which test would you do to see if the coefficients on age\_hh and educ\_hh are jointly equal to zero? Please conduct the test and report the results.
9. Now you want to estimate if the program has any differential effect based on the gender of the head of the household (variable “female\_hh”). Please explain the following regression, applied to observations in time/round 1:

What does the coefficient capture?

1. Now estimate the regression in part (f). Interpret the coefficients and in this context? Are they statistically significant?
2. Regression Discontinuity Design (RDD) (40 points).

For this question, use the data **Question3\_Dataset\_Final**. It contains only information on treatment communities, for both eligible and non-eligible households.

In this question, you will estimate the program using the targeting index (score variable). Recall that in treated communities there are two types of households: poor (eligible) and non-poor (non-eligible).

* 1. Please describe the general idea of the Regression Discontinuity design, including the main assumption of the method necessary to estimate causal effects.
  2. One important test for an RDD is that households “just above” and “just below” the cutoff point have similar characteristics at baseline. Use a bandwidth ranging from 700 to 800 points on the score variable to propose a test, perform that test, and interpret the substantive results of your output. Provide the Stata command you ran and copy-paste your results output.
  3. Another important property for a valid RDD is that the forcing variable (“score”, in this case) does not have manipulation at the cutoff point. Please provide a histogram of the score variable. Do you think that there is manipulation in this case? Provide the Stata command you ran and copy-paste your figure.
  4. Now, use households between 700 and 800 points on the score variable to estimate the effect of the program (using a regression discontinuity design). What is the substantive interpretation of your results? Provide the Stata command you ran and copy-paste your results output.
  5. Is your answer in part (d) similar to the effect you found using the RCT design in Question 2c? Would we expect to see the same results as the RCT? Why or why not?
  6. What is the effect of decreasing the bandwidth at both ends by 25 points (i.e., ranging from 725 to 775)? Show your results.