A-164/API-211 Program Evaluation (Spring 2025)

Rucinski

**Problem Set 3 – Assignment**

**Problem Set Policies:**

* This assignment is due by **5pm on Friday, April 4**, uploaded to Canvas a Word doc or PDF. A 24-hour extension is granted automatically. To request an extension longer than 24 hours, email [melanie\_rucinski@gse.harvard.edu](mailto:melanie_rucinski@gse.harvard.edu). We encourage you to submit by the deadline to give yourself a break over the weekend.
* Please type your solutions. You can use the equation editor/insert equation tool in Word to help you typeset any equations or math. **Include your Stata or R code at the end of your submission**. You can copy and paste the code directly into Word.
* You are welcome to discuss the problem set in small groups, and to submit questions to the teaching team and other students via email or Slack. Please write the names of any students you worked closely with at the top of your submission. **You must submit individually, and** **all work you submit should be your own**. To avoid misunderstandings, we ask that you **do not share any *written* answers to these questions** in shared documents, emails, chats, etc.
* Please familiarize yourself with [HGSE policy on the use of generative AI](https://registrar.gse.harvard.edu/AI-policy). You may use ChatGPT, Claude, etc. to refine or translate your ideas and to help with writing code. You may **not** use AI to *generate* ideas for this or any other assignment. **You should never, under any circumstances, upload any portion of this assignment or the data to a generative AI tool.** If you do use AI to assist you in completing this assignment, please credit the tool you used as you would a human collaborator, and explain how you used it.
* If you have questions about appropriate use of AI or find yourself turning to AI to write answers for this assignment, or if you are considering intentionally copying an answer written by another student, please be in touch with a member of the teaching team. We would far prefer to help you in office hours or grant you an extension than to have a conversation about academic misconduct.
* If you encounter any issues or questions regarding the coding for this problem set that you are unable to answer with the help of the internet or your peers, please post to the #coding-helpdesk channel on Slack. **Please do not email or DM members of the teaching team with coding questions.**
* Once your assignment is graded, please pay careful attention to the written feedback, and seek to address this feedback in future assignments.

**Part 1: Pathways to Education**

**Background**

*Pathways to Education*

Pathways to Education (<http://www.pathwaystoeducation.ca/>) is a comprehensive youth support program developed to improve academic outcomes among those entering high school from very poor socioeconomic backgrounds in Toronto, Canada. The program includes proactive advocates assigned to each student, free weekly tutoring, group activities, career counseling, and college transition assistance, combined with immediate and long-term incentives to reinforce a minimum degree of mandatory participation.

The program began in **2001** for students entering grade 9 and living in *Regent Park*, a large public housing project in downtown Toronto. It continued for each grade 9 cohort after that. The program was also expanded in **2007** to entering grade 9 students in two other public housing locations: *Thistletown* and *Lawrence Heights*.

For more details on this program, see Oreopoulos, Brown, and Lavecchia (2014) (<http://www.nber.org/papers/w20430>) from which the data for this assignment comes.

*Data*

The data are adapted for the needs of our class. Here are variables relevant to the analysis:

* cohort: year that a student entered Grade 9 (2000 - 2007)
* project: the name of the public housing project the student is from, coded as a numeric variable with labels. (The numeric codes corresponding with Regent Park, Thistletown and Lawrence Heights are 51, 63 and 36 respectively.)
* regent\_park: an indicator variable equal to 1 if the student is from Regent Park, 0 otherwise
* lawrence\_heights: an indicator variable equal to 1 if the student is from Lawrence Heights, 0 otherwise
* thistletown: an indicator variable equal to 1 if the student is from Thistletown, 0 otherwise
* never\_treated: an indicator variable equal to 1 if the public housing project a student is from never participated in the program, 0 otherwise
* female: an indicator variable equal to 1 if the person if female, 0 otherwise
* immigrant: an indicator variable equal to 1 if the person is an immigrant, 0 otherwise
* ageingrade9: age when entered grade 9
* graduated5: an indicator equal to 1 if the person graduated from high school within 5 years, 0 otherwise

*A note on coding*

No code is directly provided on this problem set; however, the code *solutions* from problem set 2 contain most of the commands you need to complete this assignment. Coding assistance is an acceptable use of generative AI.

1. Use a linear regression to estimate the difference in graduation rates between students living in Regent Park who were in 9th grade in 2001-2007 (treated students) and students living in Regent Park who were in 9th grade in 2000 (untreated students). (Hint: we are looking for a simple bivariate regression here. You do not need to include any controls.)
   * 1. Write a sentence interpreting your estimated difference and its statistical significance.
     2. Explain in a single sentence why this difference might not measure the causal impact of the program on graduation rates.
   1. Next, run a regression to estimate a difference-in-differences effect of the program on 5-year high school graduation rates (you do not need to include any controls beyond the required elements of a DiD regression). In this regression, you should restrict your analysis to students who were in 9th grade between 2000-2006 (exclude the 2007 cohort) and cluster your standard errors by housing project. Regent Park is the treatment group, and all other projects are in the comparison group.
      1. Write a sentence interpreting your difference-in-differences estimate and its statistical significance.
      2. Is this estimate larger or smaller than your estimate in (1)? Explain why in a single sentence.

In the following questions, we will estimate an effect of Pathways to Education on graduation rates using two-way fixed effects (setting aside [methodological concerns with this approach](https://www-sciencedirect-com.ezp-prod1.hul.harvard.edu/science/article/pii/S0304407623001318)).

* 1. Use a two-way fixed effects (TWFE) regression to estimate a single treatment effect of the program across all three treated locations. Your regression should include fixed effects for cohort, fixed effects for housing project, and a treatment indicator equal to 1 for treated programs in treated years (Regent Park in 2001-2007, Thistletown and Lawrence Heights in 2007 only). It need not contain any additional controls. You should continue to cluster your standard errors by housing project. Write a sentence interpreting your treatment effect estimate and its statistical significance.
  2. Refer to the *cohort* fixed effects from your regression. Are graduation rates generally increasing, decreasing, or staying the same over time? Explain in one sentence how you know. (If the code you used in part 3 did not automatically output the fixed effects estimates, you can run the same regression again, but include indicator variables for each cohort, which are algebraically identical to fixed effects. If you do this correctly, you should retrieve the same treatment effect estimate.)
  3. Name a *specific* variable that is controlled for by the *cohort* fixed effects in your TWFE regression (this should be something that is not in the provided data). Explain in 1-2 sentences *why* this variable is controlled for by the *cohort* fixed effects, including any assumptions required for this variable to be contained in the cohort fixed effects.
  4. Name a *specific* variable that is controlled for by the *project* fixed effects in your TWFE regression (this should be something that is not in the provided data). Explain in 1-2 sentences *why* this variable is controlled for by the *project* fixed effects, including any assumptions required for this variable to be contained in the project fixed effects.
  5. Name a *specific* variable, not contained in the data, that is **not** controlled for by *either* the cohort *or* the project fixed effects.
     1. Explain (in not more than 3 sentences) why this variable is not controlled for by either set of fixed effects.
     2. In 3-5 sentences, state whether you think that omitting this variable is likely to bias our treatment effect estimate, and explain why or why not.