Project Description

Mauricio Fernández-Duque

October 2025

Task Overview

In this project, you will conduct an original analysis using one of the identification strategies we have seen in class and then write up your findings in the style of a (very short) economics research paper. This document describes the project in detail; see also the slides posted on Canvas.

On the last day of class (Tuesday, November 18), you need to submit your group project. Each of you will then be quizzed on your group project individually.

Learning Objectives

1. Face a task econometricians often face: figuring out what causally-identified questions you can answer from a non-experimental dataset. You are given a dataset, will know the data-generating process, and have to use that information to figure out a good non-experimental analysis.
2. Analyze data using STATA, generating regression tables and graphs.
3. Collaborate with a team to write a small ecoomics paper.
4. Learn how to use AI productively.

Values Guiding AI Use

I would like you to follow the project in several steps I outline below. For each step, there will be better and worse approaches to using AI to assist you. It is important for you to learn how to use AI in ways that are helping make your work even better, instead of doing the work for you. The line between one and the other can be blurry, but I encourage you to reflect about what you want to get out of AI and how you should use it. Here are the underlying values I encourage you to follow.

- *The reason we're here is for you to learn, and AI can get in the way of your learning.* AI is an incredibly powerful tool that can do a lot of the work for you. If you let the AI do the work for you, you are not learning. If you do the work yourself and have

the AI complement your work, this can really help your learning. However, there's a wide gray area here—what exactly is the line between the AI doing the work and it helping your own work is by no means clear. Here you will have to use your judgment. Be aware of the "illusion of competence"—the idea that if the AI is producing the output and you're reporting it, you may get the sense that you've understood way more than you actually have. The same happens when a teacher effortlessly gives a natural explanation to a difficult topic—in class you will feel like you understood it, but it may just be that you're experiencing the teacher understanding it and mistaking her understanding for your own. There are other concerns with AI—about hallucinations, environmental impact, ethical implications. I believe these are more talked about so I won't elaborate on those, but you should think about how to navigate those concerns.

- *AI can be incredibly helpful if used correctly.* Given the objectives to learn, AI has a lot of tools at its disposal. However, you need to be mindful about seeking out those tools, and using AI in the right way. For example, you can ask AI to come up with an initial list of ideas and working off of those. But by doing so, you're not giving yourself the opportunity of acquiring the skill of coming up with the ideas, and you'll be anchored by whatever ideas the AI comes up with. Depending on the task (and figuring out your research strategy is one of those tasks), it may make more sense for you to do the effort of coming up with ideas and asking the AI for feedback on your ideas. As a general rule, asking the AI to act as a question-raising tutor or mentor instead of as a response-giver will help better serve the purpose of complementing instead of replacing your learning process.

- *We must face head-on the trust issues that AI raises between students and professors.* AI is a new, groundbreaking technology that has made it much easier to turn in well-performed work that does not correspond to your knowledge. As a professor, I dread the idea of grading a piece of writing that was not written by one of you, and this has gotten exponentially easier with AI. On the other hand, students deal with this as well. I believe students generally have qualms and concerns with the problematic use of AI, that students like learning, and that students are not looking for their work to be supplanted by AI slop. I don't want to deprive you of the opportunity to do something creative because the incentives to cheat suddenly shot up through no fault of your own. I think professors generally are dealing with how to resolve this tension. The project will evaluate you in ways that are difficult to fake through AI—such as through identifying a valid research strategy, meeting with me, and a hand-written portion. It will also give you ample opportunities to use AI in ways that can be good or bad for your learning. I will try to guide you towards the good ways, and trust that you'll follow along. On your end, I would like you to be open about how you're using AI with me, as that will help me guide you into better uses. Students are understandably worried that they'll be accused of cheating if they use AI, while being confronted with the reality of ubiquitous AI use (including by professors). So in the spirit of trust, I will not deduct points, and may award points, for open discussion about how you are using AI in order to have a constructive dialogue.

- *Use the Dartmouth chat!* DartmouthChat (Open WebUI) Admittedly this is not really

a value, but this is a natural place to put this. I have heard that some students avoid using the Dartmouth chat because they're afraid the chat can be used as evidence of their cheating. This is not only false (this is not a thing that Dartmouth does), it's actually backwards. Dartmouth does not retain your chat information, whereas the big AI companies do and use it for all sorts of purposes you probably wouldn't be thrilled about.

Step 1: Identify your research question and identification strategy

Due: 11:59pm on Thursday, November 6, by email (one submission per group)

Your grade depends in part on how strong your identification strategy is. Remember: the weaker your identification assumptions, the stronger your identification strategy.

I am setting the deadline at November 6 because by then you will have seen all the identification strategies, and in class on Thursday I will give feedback before submission (I encourage you to ask for feedback sooner).

For a *multivariate regression*, you need the treatment variable, your control variables, and an argument as to why the selection on observables assumption is plausible.

For a *DiD analysis*, we need a treatment group, control group, and a treatment date (the time at which the treatment effects the treatment group but not the control group). If you decide to run a DiD, your first step is to decide how to setup your DiD analysis in this way. Specifically, you need to identify the following:

- Treatment group that has an increase in some treatment.
- Control group that does not have a large increase in that treatment.

You need to argue why the parallel trends assumption is reasonable.

For an *instrumental variables* approach, you need your instrument, treatment and outcome variables. You need to argue why the relevance assumption, exclusion restriction, and monotonicity are likely to hold.

For a *regression discontinuity* design, you need to have a running variable, a threshold, a treatment and an outcome variable. You need to justify why it is plausible that the potential outcomes curves are smooth at the threshold.

You can limit yourself to the data you have, or look for additional data from outside sources. Additional data is more work and hard to pull off, so I suggest sticking to the data at hand. Talk to me if you are considering adding more data.

After discussing in your groups for a while, try to agree on an angle and then send me the following information: (i) your identification strategy; (ii) the information described above per identification strategy; (iii) a justification for your identification assumptions. Think about this and write it up in conjunction with Step 2, below.

Step 2: Identify the outcome of interest and the relevant subgroups for analysis.

Due: 11:59pm on Thursday, November 6, by email (one submission per group)

At the same time as you are deciding on your treatment and control groups, you should decide what outcome variable you will analyze and what subgroups, if any, your analysis will focus on.

In conjunction with thinking about what outcome you want to focus on, you should also think about what subgroup(s) your analysis should focus on. You should think about groups whose outcomes you may be particularly interested in, or who may be more likely to be affected by the treatment. I will try to give some guidance on this, but there may be some trial-and-error involved, as well.

AI can be helpful in Steps 1 and 2 to give you feedback on your proposed identification strategy, but I would not use it to come up with the idea itself. The most important reason to not use it is that this is a key part of the learning objective of this task. On a practical level, you should not use it because the AI will lack important knowledge of the setting that may not fully be described by the data handout, so you should think about what information may be missing for you to strengthen your confidence in your proposed research strategy and seek feedback form me. I will either have the answers or make up acceptable answers you can assume are true.

I recommend you think of as many identification strategies as you can, get feedback from me about which makes most sense, and use that information to land on what you will do.

Step 3: Download the datasets for your analysis and produce a new, clean dataset.

"Due": Tuesday, November 11 5pm

We will look at the data together during class time on Thursday, November 6. By the end of class on Tuesday, November 11, you need to show me that you have downloaded the data that you need and constructed the final dataset that you will use in your analysis.

Once you have downloaded your data, you should write a STATA program (.do file) that reads in the raw files and performs any manipulations that are necessary to produce your analysis dataset: you may need to make sample limitations and/or construct variables. This do-file can either stand alone and save a Stata dataset for analysis, or it can be the same do-file that ultimately does the analysis.

Step 4: Preliminary analyses.

Due: 11:59pm on Monday, Friday, November 14, by email (one submission per group)

Summary statistics: You could produce a table that reports the mean of each of variable in your analysis (outcomes and key control variables) for the treatment group before and after treatment and control group before and after treatment (e.g., a table with 4 columns). This should not be raw Stata output pasted into the paper, but instead you should make

a table (e.g. in Excel) to present the information. STATA has a command that helps you build table output called `outreg2`, and ChatGPT can be helpful for figuring out how to turn your regression output into tables.

You should include plots, depending on your identification strategy.

- *Selection on observables.*— Plot the regression residuals and the fitted line if you are following a selection on observables approach.

  Make a table with your regression estimates (similar to those discussed in class and in conceptual exercises). Your table should have a few columns corresponding to, for example, specifications with different sets of control variables.

- *Difference in difference.*—DiD graph for your main outcome: On the same graph, plot the means of your key outcome variable for both your treatment and control groups over time. You should use this graph to judge whether it looks like "parallel trends" is a valid assumption in your case or not. Do the best you can to make sure it is clear which group is which in your graph and also when the policy actually changes (i.e., with a vertical line). Note: There will be no penalty for not having parallel trends in your assignment. You will be graded based on doing things correctly and then interpreting them correctly.

  DiD table: Make a table with your DiD regression estimates (similar to those discussed in class and in conceptual exercises). Your table should have a few columns corresponding to, for example, specifications with different sets of control variables.

- *Instrumental variables.*— Graph the residuals of the first stage, reduced form, and two stage least squares. Include a table with these specifications, possibly adding (justified) controls.

- *Regression discontinuity.*— Add the standard regression discontinuity graph for your outcome variable, perhaps for balance tests, and for your first stage if you are pursuing a fuzzy regression discontinuity.

  Add the standard regression discontinuity tables from class / Mastering Metrics.

AI can be helpful in Steps 3 and 4 to figure out how to do a lot of the things you need to do, such as giving you the right STATA codes for different tables and graphs. Here's what I think you should be concerned about, which is what I'm concerned about when I use AI to help me code. If I don't know how STATA works (and how econometrics works, for that matter), I will not be able to evaluate whether what the AI is giving me is what I need, or what I would need to change about the code for it to work, or how to guide the AI to give me the correct code. So I'm not going to put precise limits on what you can and cannot ask the AI to do. But I want you to go into it with the philosophy that you are using the AI to simultaneously help you complete the task and to actively learn why the task is completed in a certain way.

One way I think AIs can be very helpful as a teaching tool si the following. Go to Wharton Generative AI Labs Prompt Library | All Prompts, which is a "prompt library", a repository of useful prompts. You'd do well to play around with it for a bit. My specific suggestion is

for you to use the "Tutor" prompt. Literally paste the prompt into the Dartmouth chatbot. It will turn into a tutor that will ask you what you are interested in learning about. Ask it to teach you about the STATA commands you need to run your analyses.

Step 5: Write up your findings.

Papers due at 8pm on Monday, November 17 by email (one submission per group)

The final write-up will include the following elements:

1. Introduction: In the introduction, you should state your research question and provide motivation for examining your outcome and subgroup(s).

2. Data. Explain the sources of your data, which variables you are using, and exactly how you constructed those variables in cases where clarification is required. You should also explain how your sample was created, etc.

3. Your graphs and tables, numbered so that they can be easily referenced

4. Empirical Strategy. Here is where you explain what you are going to do – your identification strategy (differences in differences), the regression equation you will estimate, and what assumptions are required to answer your question of interest.

5. Results. Present and discuss your findings.

6. Conclusion. Briefly restate the motivation, and outline the approach you took and the results you found. If you have further thoughts (e.g. about future work) that would be natural to put in the conclusion, you can do so, but it is not required.

7. Statement on AI use. At most half-a-page explaining how you used AI in your work.

The first three components of the paper should probably be about 3-4 pages of text (not inclusive of the graphs and tables). The last three components of the paper should probably also be about 3-4 pages, although you may need more space than that to cover everything (and it will also depend on spacing, etc.).

You should also hand in the do and log files for your data construction and analysis.

It may be helpful to look at other papers that have followed a similar identification strategy for inspiration on how your project should be written up. I can help you find such papers.

Arguing against blind AI use in writing is a bit of a tougher sell than arguing against AI use in coding. You can see the AI output as a piece of writing which you can evaluate, so why would you be worried about it unknowingly producing something you don't want? (Perhaps you don't even read what it wrote, in which case I really can't help you). The reason I think blind AI use in writing is bad goes deeper than the reason against blind AI use in coding. Writing is thinking. To learn to write is to learn to reason correctly. Writing is one of the most powerful tools you can develop for any of the jobs you will do (and outside of those jobs as well), and economics generally gives few opportunities for you to develop

that critical skill. AI makes these opportunities even easier to dispense with, and that is at your own cost.

If you were seasoned econometricians, I would be fine with you asking the AI to fill out some paragraphs that are pretty standard throughout empirical work, such as describing a regression table. Have the AI take a first pass, and then just clean it up a bit. But you have never written such a paragraph. So it is very valuable for you to write it, which requires you to think about it and interpret it. By pawning off your thought process you are losing out on the knowledge you could be acquiring. In contrast, I think that asking the AI to give editing suggestions on text you have written can be very helpful and pedagogic

Step 6: Individual quiz on your project

On Tuesday, November 18 I will have an in-class and-written quiz about your project. The questions are about what the project did and how to interpret the findings. You will not be able to have the paper in front of you while you answer the quiz. The point of the quiz is not to ask you very specific details, but to test your understanding of what you did, why you did it, what you found, and what your results mean.

The reason for holding a quiz is twofold. First, although group projects are great and have many benefits, teasing apart individual from group contributions is often difficult. The quiz allows me to understand how much each person understands of the project and contributed to it. Second, in a world of ChatGPT it is now too easy to have a paper automatically written. This is not great because you are not learning. The way I think of ChatGPT and other LLMS is that they will be the tool you use for running regressions. But if you do not know which regressions to run and why, you will not be able to take proper advantage of these tools. Moreover, in a world of ChatGPT a person who knows how to mechanically run regressions but does not understand which to run and why is quickly becoming obsolete. I want to make sure you understand which regression to run and why because I think that is where the biggest value-added is going forward.