

Applied Causal Machine Learning, Project One

Nazanin Khazra

Your first project is the foundation of your final paper. The goal is to apply the econometric and causal inference tools we cover in class to a real dataset and to build toward your own contribution. This is **not a replication exercise**. You may use the same data or topic as the published paper, but your task is to take the analysis one step further and add your own hypothesis and subsequently ask your own question. The contribution can be small at this stage because we will deepen it in Project 2, but it must still be real and yours.

Important Notes

1. Your paper should have a title that can be your research question as well (talk to us if you need help choosing), make it short and meaningful, it should not be a paragraph!
2. Write your name under the title, *don't* use the university log.
3. Format your PDF as a journal-style paper; $\text{\LaTeX}$ is strongly recommended.
4. Push your Python code (not the data) separately to your Github and share the link to your repository. There will be a question on Crowdmark that allows you to paste your link to your repository.
5. Avoid text that reads as filler or AI-generated content. Every sentence should say some important information, AI writes vague. Re-read the text if you use GPT and edit so the words are yours.
6. Keep your language clear and direct. In Economics, we don't like complicated writing, write simple, direct, use "I" since your papers are single-authored.
7. There is no strict page limit. Try not to go beyond 10 pages of text. The graphs add to the page numbers but if one only counts the text pages, it's best to keep the below 10 so we can read all the students papers carefully.
8. Graphs and tables should be in the paper and NOT in a separate section.
9. This project builds the base for Project Two: a strong start means a strong finish.

1 Introduction

Purpose of the Introduction

Your introduction sets the stage and convinces readers that your question is meaningful. It should do the following:

1. **Motivation.** Explain why this question matters. Why should economists or policymakers care? What is at stake?
2. **State the research question.** Clearly define the economic question or hypothesis you address. Your question can resemble that of your base paper, but you must extend it or ask something new. At this point, I have talked to all of you so you should be relatively clear about this part.
3. I usually start the paragraph by writing "The main research question is ...," be very clear about your question. Dedicate a paragraph to it.
4. **Literature Review and Contribution.** Describe how your work contributes to the literature and differs from prior studies. What is your "small but real" addition? This part is blended in the literature review. It should read as "paper A did this, but I'm doing this other thing" or "paper B found this but my initial findings are different". Tell a story about the literature and where you are standing. You can keep the same theme through the paper. A good paper has citations everywhere!
5. NEVER summarize a paper in your paper, a literature review is not a list of summarized papers. We have discussed this before. Let me know if you have any questions.
6. You should also briefly mention the identification strategy and assumptions. You will go more in depth and test them in the body.
7. **Preview of findings.** Briefly summarize your main results or insights. Add a paragraph and smoothly transition to other sections.

Writing Tips

Use the "why do we care?" and "so what?" tests. Keep it concise but rich. A good introduction reads like a story that motivates curiosity.

2 Data and Summary Statistics

What to Include

- **Data Source.** Describe your dataset, origin, unit of observation (firm, individual, region, etc.), sample size, and time period. Don't just say "from replication package", trace it to the original source and explain that.
- **Variables.** Define key variables (outcome Y , treatment D , and controls X). Discuss them narratively, not as bullet lists. **No bullet lists in the paper whatsoever! We don't do that! :D**
- **Summary Statistics.** Present descriptive tables (means, standard deviations) and visualizations (histograms, scatter plots, maps).
- You want to explain your data as much as possible. The purpose of a sum stat visualization or table is to **shed more light on your story**, so it should not be a random subset of variables! You should answer this question after each output "What do I learn from this table or visualization that is relevant to the research questions?", explain it for the reader, tell me how to think about the table. What numbers do I pay attention to? How do you interpret those numbers?
- Life lesson: **a good summary statistics section almost fully answers the research question!**, you should explain the story and the data with plots and tables so good that I believe your story even before seeing your regression! See if you can do this! That's a top 5 journal paper.
- **Data Limitations.** Note issues like missing data, measurement error, or selection bias. Add a brief paragraph about how you cleaned the data and the limitations. You can have an appendix and in details explain how you cleaned your data. It will be handy when you want to replicate your results or publish the paper. Document each step, we probably won't read that part but it will be a lifesaver for your future self.

How to Present

Each figure or table must tell a story. Don't just show numbers—interpret them. For instance: *"As table blah blah shows, treated firms had higher baseline productivity than controls; we address this through our identification design and ML models."*

3 Regression Results and Other Models

IMPORTANT: not everything needs to be significant, not everything has to make sense, we are not done yet! It's OK if your regression coefficients are not significant or your Rsquared is small. Just tell me why this is the case.

Required Models for Project One

Include and discuss every model we covered in class. You cannot skip or select a subset of methods. This is an exam! :D You can take models out when you want to send your paper for publication, but for the course, keep everything in. Specifically:

1. Original regressions (e.g., OLS baseline) from the reference paper. **Important:** the regression tables (and all tables in the paper) should be formatted as a journal paper. Word doc tables that are not publishable, screenshot, or python tables are not acceptable.
2. Identification assumption tests (parallel trends, exogeneity) in the form of graphs or tables.
3. Summary-statistics-based visualizations of identification tests (event studies, balance plots).
4. The causal regressions that apply to your setting, e.g. DID. Again, these regressions should be in a journal format, don't give me ugly word doc tables :D
5. Regularization: LASSO and Ridge. You can present them as short tables (selected numbers) or figures.
6. Robustness checks: If applicable, show the results from alternative specifications (e.g., different models, inclusion of control variables) to demonstrate the robustness of your findings.
7. Machine learning: Decision Tree, Random Forest, Bagging, and Boosting. Show me a Tree, discuss the interactions you see and the threshold values (not all of them, only the interesting findings). Interpret your RF importance matrix.
8. A few DAGs with clear explanations about open/closed paths and how they link to regressions. See if you can make them look nice, it may be hard,... it's ok to include what Python gives you if you don't have time for this submission, but fix it for your P2. Though it's easy to do with latex, upload the image to GPT and ask for the latex code, it's two seconds.

Regression Specification, write the model specification for EACH model

For example, your base model might look like:

$$Y_i = \alpha + \beta D_i + X_i^\top \gamma + \varepsilon_i$$

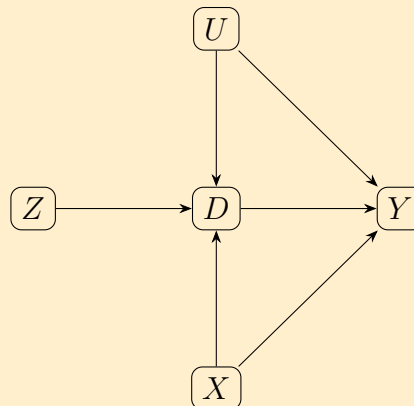
Interpret β carefully and motivate why causal inference requires assumptions or identification strategies.

Identification Strategy (this part may be similar to the original paper)

Before showing causal results, explain identification clearly. If you use:

- **DID:** show pre-trends and parallel trend tests.
- **Experiments :**
 1. Randomization Validity – show a balance table comparing baseline covariates between treatment and control groups.
 2. No Spillovers / SUTVA – show a conceptual diagram or network plot illustrating that one unit's treatment can't affect another's outcome.
 3. No Differential Attrition / Noncompliance – show a summary table or bar chart comparing attrition or compliance rates across treatment groups.

Example DAG



Explain how each path is closed by controls or instruments and how this maps onto your regression or ML model. You can simplify variables and write them in groups of variables to show a meaningful graph. If you have 100 variables and you want to draw each, the graph will not be a useful visualization.

Results Presentation

- Present results in clear tables (coefficients, SEs, N, fit statistics).
- For ML models, show metrics (accuracy, MSE, feature importance).
- Use graphs (partial dependence, variable importance).
- Include robustness checks: alternative controls, specifications, or samples. (if you have time, if not, leave it for p2)

4 Conclusion or Summary of Main Findings

What to Cover

- The main economic insights from your analysis.
- How your findings answer your research question.
- Unexpected or contrasting results and how they relate to existing literature.
- NEVER introduce new material, citations, or concepts in the conclusion. Just summarize, short, sweet.

Advice

Interpret your results economically, not just statistically. What mechanisms might explain your patterns? What does this imply for theory or policy? How big is the number you found in dollars?

5 Future Steps (Optional)

Purpose of This Section

This section bridges Project One and Project Two. Discuss:

- How you plan to extend or refine your existing analysis.
- New data, or variables you will add.
- How your small contribution will become a deeper causal investigation in P2.

6 References and Appendix

Formatting

List all cited works using APA or Chicago style. Include all datasets, papers, and software packages referenced. If you are using latex, use href to make the intext citations clickable.

Appendix Content (Optional)

- **Important:** big ugly tables that you have to keep, goes to appendix and are cited in the paper. But this is only if you have no other way of making the table small and keep it in the main text.
- Data cleaning process in details.
- Additional robustness tables or figures.
- Technical derivations or supplemental models.
- Code snippets that support your analysis (if makes sense).

Final Thoughts

Take a moment to appreciate how far you have come in such short time! The paper in front of you is fully yours and it your contribution to the literature even if tiny! You own it, your name is on it, you should be very proud of your work.

Well,... I'm proud of you!

Looking forward to reading your projects team! :)