097400

Introduction to Causal Inference

Project Guidelines

Goals:

1. Students will interact with the material throughout the semester, allowing them to develop a deeper understanding of Causal Inference
2. Students will learn from one another by presenting different stages of the project and hearing feedback from peers
3. Students will receive ongoing feedback from the course's staff throughout the semester, improving the quality of the final projects which comprise 85% of the course's grade.

Stage 1: Defining the Project's Research Question

Students will address the following questions:

1. What is the *causal* question you seek to answer?
2. What available material already exists around the question? (Try Google Scholar)
3. What data do they intend to use/gather to answer their question?
4. **The data must be observational, not a random trial. The treatment and outcome must be confounded.**

Positive Example:

Do vocational training programs improve participants' income?
Previous studies have found improved income for participants treated who received vocational training programs; however, the effect size differs depending on gender.[1]
Even if vocational programs do improve income, the effect may be small and diminish over time.[2]
We intend to look at the effects of vocational training programs in developing countries and intend to use data like the Mongolia Compact – Vocational education data found on the US data.gov website.

Negative Examples:

A. Misspecified causal question –
   What is the effect of gender (or belonging to a minority) on income?
   This is a problematic causal questions because we cannot estimate the counterfactual of being born a different gender. The "overlap/positivity" assumption is not met. In contrast, what is the effect of the gender submitted in a CV on getting invited to a job interview can be estimated. Because the gender written in the CV can be changed, while keeping everything else the same.
B. Use reliable sources for available material –
   Do not quote Wikipedia or online forums
C. Realistic data –

[1] LaLonde, Robert J. "Evaluating the econometric evaluations of training programs with experimental data." *The American economic review* (1986): 604-620.
[2] Hirshleifer, Sarojini, et al. "The impact of vocational training for the unemployed: experimental evidence from Turkey." *The Economic Journal* 126.597 (2016): 2115-2146.

Even if you do not have an exact source of data that you wish to use, it should be realistic to find the necessary data.

We want this stage to help you correctly specify causal questions and to make sure that your question is relevant and interesting.

You will meet with course personnel at least 2 times during the semester to help you correctly define the above details.

Grading (5% of project grade):
You do not need to submit your answers to the questions above. However, you are expected to meet at least once with Galit and at least once with Dovid to make sure that you have clear answers to these questions.
The grade is binary – if you have met with both staff members and have clear answers to the above questions, you get 100%. If you do not meet with the staff members, you will receive 0% for this part of the project.
Come prepared to the meetings with the course staff. Try telling your partner the answers to the three questions above. If you do not understand one of the questions, have clear questions.
During the second half of the semester, additional times will be offered for meetings with the staff about the project.

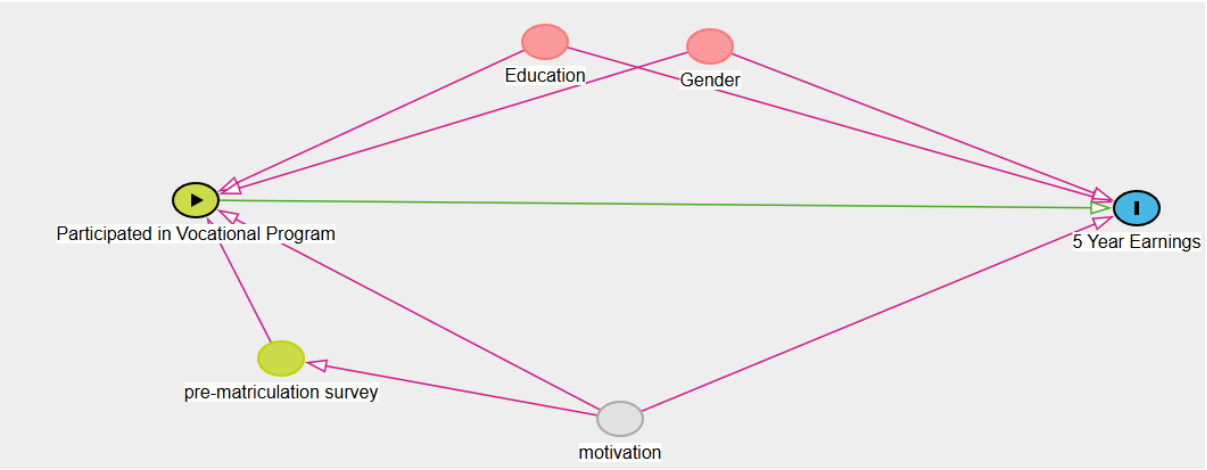
Stage 2: Project Research Proposal (Due date: 13/8/2025)

Students will address the following questions and submit a 5-page report:

1. What is the *causal* question you seek to answer? (see stage 1)
2. What knowledge already exists about the question? (cite relevant sources)
3. What data do you intend to use?
4. What are the causal assumptions made? (These may depend on the estimation methods you plan to use. We suggest using Pearl's causal graph to concisely present causal relationships if relevant.)
5. What challenges may affect the analysis?
6. What are the estimation methods you are planning to use? Explain.
7. What robustness checks are you planning to do?

Positive Example:
Does admission in a vocational training programs improve participants' income?
We intend to use the New Vocational Training Study of Uzbekistan.
We believe consistency holds in our study because *admission* to the vocational training program is either true or not for each participant. There are not multiple levels of treatment. Therefore, each participant's observed outcome accurately reflects the potential outcome of the observed treatment.
We have mapped our causal assumptions to this causal graph[3].

[3] Try using https://www.dagitty.net/



We anticipate two main problems with our data: 1) There is data imbalance between the treatment group (participated in vocational training) and the control group (did not participate in vocational training). This may affect the "positivity/overlap" assumption. 2) We also believe that "motivation" may be a hidden confounder that affects both whether the participant joins the vocational training program and his income over the following five years. Our data contains a survey given to participants before signing up. One of the questions relates to the participant's motivation and we believe that adjusting on this sufficiently reduces confounding.

Negative Examples:

A. Misspecified Causal Question – see above
B. Nonsensical Causal Assumptions –
   Causal inference always requires assumptions and there are almost always hidden confounders that impact the believability of our findings. Students should show awareness of the limitations of their causal design and seek ways to mitigate the problems. For example, the five-year earnings of someone with a bachelor's in data science will obviously differ from someone with a high school education. This will also impact the likelihood of signing up for vocational training. Therefore, education should appear as a confounder. Maybe gender is less impactful? Maybe living near a city critically affects both and should appear in the graph above. Try to explain some of your more basic assumptions.
C. Challenges – Although we may not have covered all the methodologies for causal effect estimation, by looking at the data we can get a feel for different challenges that may arise. Some of these challenges will appear in the causal graph (like the "motivation" example above). If you are defining your own intervention, that may also be a challenge. Some are regular statistical-ML data challenges, like data imbalance or scarcity. Challenges should be causal or data related.

Comment: We want this stage to help you **correctly specify the causal question** and to make sure that you have found relevant data that will help you. We also want to help you formalize the causal assumptions that you make and allow you to consider how you will deal with problems stemming from causal assumptions or data constraints. We want to see that you have carefully considered appropriate methodologies for estimating your causal question.

Grading (10% of project grade) will be based on:
Whether you correctly specified a good causal question and found appropriate data to answer that question. Assuming the answer for the first two questions is yes, we examine the analysis you made. Specifically, that the causal assumptions/graph you made is reasonable and correct. Whether you fully identified the challenges in answering the causal question you want to answer, and whether you draw a correct estimation plan.

Stage 3: Final Project Report and Presentation (Due dates: Report 4/9/25, Presentation 7-8/9/25)

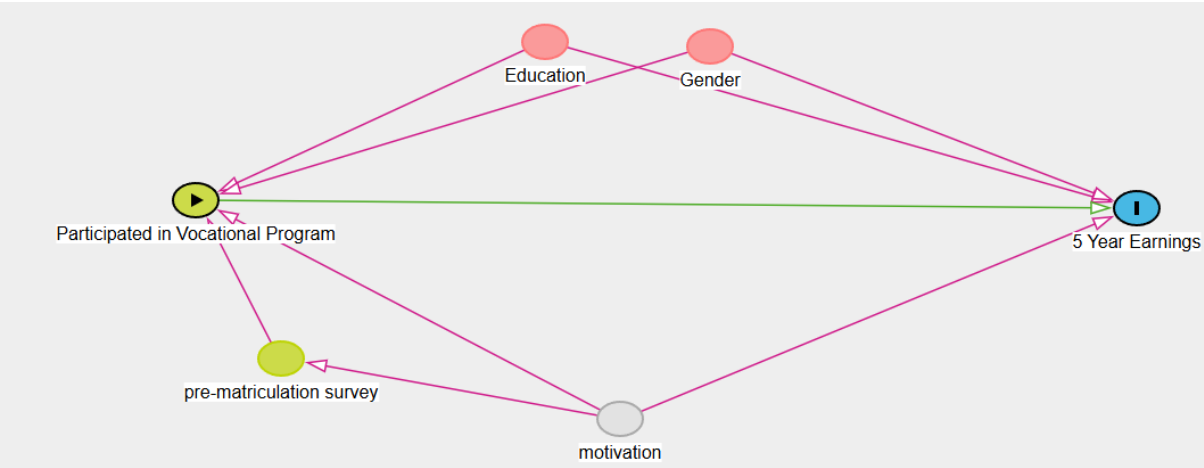
Students will submit 10–15-page report on their project, addressing the following questions:

1. What is the *causal* question they seek to answer?
2. What knowledge already exists about the question?
   If your causal question has previously been researched in the literature, what methods were used? Are the previously found results "causal" based on what we have learned in the course.
3. What data was used?
4. State the methods that you used for estimation and the causal assumptions necessary for these methods (e.g,, SUTVA, IVs, Target Trial, Causal Graph).
   a. Explain why these causal assumptions are reasonable.
   b. You must consider uncertainty (confidence intervals or statistical tests).
5. Present and analyze your results.
6. Present and analyze robustness checks
7. Discussion
   a. What conclusions can be drawn from your results? How confident are you in the results?
   b. What are the limitations of your analysis?
   c. *If your causal question has previously been researched in the literature, are your results similar to previously found results? If they are different, what could explain that difference?*

In addition, students must submit the following files: (a) the data used for this project (b) executable code for all the analysis performed in this project, (c) a statement of the ways in which you used ChatGPT in performing this project.

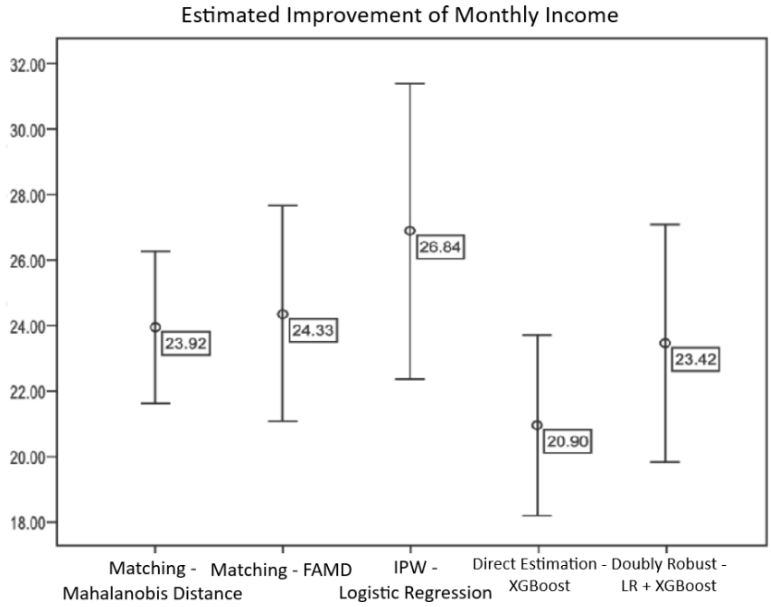
Positive Example:
Do vocational training programs improve participants' income?
We intend to use the New Vocational Training Study of Uzbekistan.
We have mapped our causal assumptions to this causal graph[4].

[4] Try using https://www.dagitty.net/



Show findings regarding propensity (like we saw in the Tutorials)

Show results of causal estimation that account for uncertainty.



A reasonable conclusion from the above data is that all methods indicate a positive monthly effect on income. One could consider how the challenges they considered in Stage 2 influence their results. Perhaps, data imbalance contributed to high variance in the IPW estimator, etc.

Examples of robustness checks are:
1. If you dealt with missing data using mean imputation, what happens if you impute using the median? Or regression?
2. If you used matching with Mahalanobis distance. What happens if you use a different distance metric?
3. If you trimmed the data using specific thresholds (e.g., for outliers). How sensitive are your results to small changes in those thresholds?
4. If you used a specific time-gap in a natural experiment. How sensitive are your results to small changes in those time specification?

Comment: This example is not comprehensive. It represents a minimal structure for some parts of Stage 3.

Prepare a 10-minute presentation on your project.

Grading:
Report (75% of project grading) will be based on:
Whether you correctly specified a good causal question and found appropriate data to answer that question. Assuming the answer for the first two questions is yes, we examine the analysis you made. Specifically, that the causal assumptions/graph you made is reasonable and correct. Whether you fully identified the challenges in answering the causal question you want to answer, and whether you applied a correct estimation plan. Have you done enough robustness checks and do they cover all the challenges you identified? The correctness and completeness of results (e.g., accounting for uncertainty using confidence intervals and/or statistical tests) and the depth of your discussion of these results (e.g., addressing modelling concerns/limitations, avoid erroneous conclusions).

Presentation (10% of project grading):
The presentation should be brief (5-10 minutes) and clear. It is not necessary to grab the audience's attention or invest time in an aesthetically pleasing design. It is important that you are fluent in explaining what you did and why.