Final Team Project Guidelines

Causal Inference in Practice

Overview

This team project is an opportunity for you to apply what you've learned in this course to a real-world observational dataset of your choice. In small groups (3-4 students), you will identify a causal question, choose an appropriate dataset, apply suitable causal inference methods, and clearly communicate your findings. The goal is to develop your ability to formulate causal questions, assess identification strategies, and carry out analysis and interpretation in an applied setting.
This project is worth 25% of your final grade.

Deliverables

Each team must submit:

- A 10-minute recorded presentation
- A final written report (maximum 5 pages, not including tables/figures/references)

Project Components

1. **Choose a dataset**
   - Select a publicly available dataset suitable for your causal question
   - See recommended sources below

2. **Formulate a causal question**
   Your questions should be well-defined and feasible with the data
   - Example: *What is the effect of attending private school on college attendance?*
   - Make sure your treatment and outcome variables are clearly defined
   - Aim for a realistic but tractable setting – it's okay to simplify
   - Draw a DAG to represent your assumed data-generating process
   - Explicitly describe the hypothetical intervention (e.g., if we could manipulate X, would Y change?)

3. **Select and justify causal inference methods**
   Apply at least two different causal inference approaches:
   - Choose **one** from:
     - Matching
     - Outcome modeling (or, regression adjustment)
     - Propensity score methods
     - Inverse probability weighting
     - Or, any combination of the above
   - Choose **a second** from:
     - Instrumental variables
     - Regression discontinuity designs
     - Difference-in-differences
     - Other methods (if self-studied, include a reference)

   Explain why your chosen methods are appropriate for the causal question and data

4. **Conduct the analysis**
   Your analysis should include:
   - Data cleaning and preprocessing
   - Descriptive statistics
   - Estimation of the causal effect
   - Covariates balance check

5. **Interpret and discuss the results**
   Your report and presentation should:
   - Highlight the identification strategy
   - Discuss key assumptions
   - Acknowledge limitations of your analysis
   - Offer meaningful interpretation of your findings

Recommended Data Sources

- Add Health: Adolescent health and social networks
  (Our TA has provided guidelines for accessing the data and is available to assist)

Other potential places to find datasets include:

- ICPSR - Social science data archive
- IPUMS - Census, health, and international survey data
- Harvard Dataverse – Public datasets for academic research
- Kaggle Datasets - Curated public datasets
- OpenICPSR – Open-access research datasets

Important Dates

- **Final Presentations**: Upload your video by **June 11 at 11:59 PM**
- **Presentation Peer Review**: Watch and score **10 group presentations** between **June 12 - 17** (participation in peer review counts toward attendance, minimum 80% participation required)
- **Final Report Due**: Submit by **June 19 at 11:59 PM**

Evaluation

- **Presentation (30%)**: Judged on clarity, organization, and delivery
- **Peer review (20%)**: Based on ratings from your classmates
- **Final report (50%)**: Evaluated on structure, justification of methods, clarity of causal reasoning, implementation, and interpretation of results (must submit R program along with your report)