Introduction to Causal Inference

Spring 2016

Keio University

Instructor: Teppei Yamamoto
Email: teppei@mit.edu
Class Time: July 14–16 and 18, Periods 2–4; July 19, Periods 3–4
Class Room: South Building, Floor 7, Room 471
Office Hours: July 14–16, 18 and 19, 4:30–5:30PM

Purpose and Goals

This intensive short course provides a survey of empirical methods used for causal inference in political science research. We cover a variety of research designs and statistical methods for causal inference, including randomized experiments, matching, regression, sensitivity analysis, nonparametric bounds, instrumental variables, regression discontinuity, difference-in-differences, and causal mediation analysis. We analyze the strengths and weaknesses of these methods. Applications are drawn from various fields in social science, including political science, public policy, economics and sociology.

Prerequisites

There is no prerequisite for this class, but some prior exposure to basic probability theory and statistics will help you understand the materials better. Experience with real social science data, using regression models and preferably in R, will make you an ideal target of this course. However, any student with interest and motivation is welcome to participate in the course.

Requirements

The final grades will be determined based on the following requirements:

- **Class attendance, participation and engagement (10%)**: This course is taught within a short period and we will cover a broad range of materials at a fast pace. Therefore, your strong commitment and participation in each class is essential. Don’t miss a class or you will be totally lost! Your active participation during class is also important and counts towards your grade. Here are some notes regarding what to expect:
– You may interrupt me at any point for clarification or a question during my lecture. If any questions or ideas occur to your mind, don’t wait until the end of the class and come to me; I would rather you speak up during the class and share what you have to say with everyone else in the class.

– During my lecture I will occasionally ask questions to the class. Try and be the first one who responds, even if you are not 100% sure you know the correct answer. Saying something incorrect is far better than saying nothing at all.

– I understand that the class format is not ideal for a technical course like this one and that some people may find it difficult to stay perfectly concentrated throughout the four and half hours each day, even with breaks. Still, please try your best not to fall asleep during the class. If you do, I may ask you to leave the room. I also ask that you refrain from checking emails, Facebook, etc. during the class.

• **Problem sets (30%)**: You can only learn statistics by doing statistics. Therefore, the homework for this course is extensive, including daily assignments. The assignments will consist of both analytical and data analysis questions, and they will be posted on the course website immediately after the last class on each day. There will be two types of problems: required and optional. Only the required problems will be due at the beginning of class next day, and performance on those will count towards your final grade. The optional questions are not directly included in grade calculation, although you are encouraged to complete them as your time permits. Due to the intensive nature of this course, I expect many of you to be unable to work on the optional problems. However, I strongly encourage you to come back to them during the summer vacation as review materials. Solutions for the problems, including the optional ones, will be posted immediately after they are due.

  Additional notes on problem sets:

  – Late submission will not be accepted unless you ask for special permission from me in advance of the deadline. (Permission may be granted or not granted, with or without penalty, depending on specific circumstances.)

  – Working in groups is encouraged, but you must submit your own write-up of the solutions. In particular, you must not copy and paste someone else’s answers or computer code. We ask you to write down the names of the other students with whom you solved the problems together on the first sheet of your solutions.

  – For analytical questions, you should include your intermediate steps, as well as comments on those steps when appropriate. For computing and data analysis questions, include annotated code as part of your answers. All results should be presented so that they can be easily understood.

• **In-class quizzes (30%)**: There will be a 20-minute closed-book quiz at the beginning of the first class each day (except on the first day). The quizzes will test your understanding of the materials covered up until the previous day, but naturally focus will be given to more recent materials.

• **Research proposal (30%)**: For most of you the ultimate goal of taking this course should be to use the methods covered in the class for your own applied empirical research in political science or another field of social science. Therefore, the final assignment for this class will be a short research proposal (in place of a final exam or a term paper). Your proposal is due in one week after the class (i.e. **July 26**) and should meet the following formatting requirements:

  – Your proposal should be approximately 10 pages (12-point font, A4-sized paper) and written in English or Japanese.

  – Your proposal should look like a typical empirical journal article minus an actual empirical analysis. That is, the proposal should start with an abstract and consist of an introduction and sections on theoretical motivations, research design, data description, and statistical methodology.

  – The key section of your proposal will be the research design section, where you should focus on the discussion of your **identification strategy**. That is, you should explain, clearly and rigorously, what
assumptions have to be satisfied in order for your proposed research to produce a reliable estimate of a causal effect, and how and why your research design makes those assumptions likely to hold. Making a good argument for your identification strategy is an essential component of a good research proposal for this course.

– The theoretical motivation section should be concise and should not take up more than two pages of your proposal. In particular, you should avoid lengthy literature review. You should instead focus on the discussion of why your proposed research is interesting and important (not only to yourself but also to general audience in political science) clearly and succinctly.

– The data description section is optional and can be omitted if you do not already have a concrete dataset, data source, or specific mode of data collection in mind. However, even if you do not have any of those yet, you are encouraged to include a section or paragraphs discussing possible options for data collection to show that you are proposing a realistic empirical project that could actually be conducted with sufficient time and resource.

Course Website

Course materials (lecture slides, problem sets, quiz questions, solutions, readings, etc.) will be posted on the course website, which can be accessed at:


Notes on Computing

In this course, we will use R for the purpose of computation and data analysis. I expect that you will solve problem set questions using R. For those of you who are not familiar with the language, the first problem set will include questions that are designed to get you started with R.

Books

• **Recommended books:** There is no single textbook for the course. However, many of the core materials are covered in the following book. The book is very popular among political scientists and applied economists and written in a relatively accessible fashion. I recommend that you purchase a copy. (Japanese translation is also available, but I recommend the original version, which is actually less expensive.)


If you find the above book too technical, the same authors have also written a more accessible version of essentially the same book. (I use this one for my undergraduate causal inference class which assumes no prior knowledge on statistics or probability theory.)


The following books cover materials in this course that are not discussed in the above two books. They are roughly at the same level of technicality as the first book but written from somewhat different perspectives which some methodologists find more appealing. These are not as essential as the first book, but still highly recommended for those who are interested in causal inference.
Optional books: The following books are more advanced textbooks and monographs on causal inference. They go beyond the level of this course and not required for successfully completing the course. However, they are recommended for those of you who have stronger technical background and want to understand the materials more deeply.


Course Outline

Note that recommended readings are in **bold**.

1 Statistical Models for Causal Analysis

Causality as counterfactuals, potential outcomes, the Fundamental Problem of Causal Inference, identification and estimation, causal estimands, interference, causal graphs and other causal models, sufficient component causes

Readings:

- **Morgan and Winship:** Chapters 1, 2 and 3.
- **Angrist and Pischke:** Chapter 1.
2 Randomized Experiments

2.1 Identification and Estimation

Identification of causal effects under randomization, covariate adjustment, blocking, practical considerations

Readings:

- Angrist and Pischke: Chapter 2.
- Gerber and Green: Chapters 2, 3 and 4.
2.2 Inference

Neyman variance, clustered designs, randomization inference, bootstrap, power analysis

Readings:


3 Observational Studies

3.1 Identification

Selection on observables, post-treatment bias, subclassification

Readings:

- Angrist and Pischke: Chapter 8.1

3.2 Matching and Weighting

Covariate matching, balance checking, propensity scores

Readings:

- Morgan and Winship: Chapter 5.


### 3.3 Regression

OLS as an estimator of causal effects

Readings:

• Angrist and Pischke: Chapter 3.

• Morgan and Winship: Chapters 6 and 7.
3.4 Partial Identification and Sensitivity Analysis

Nonparametric bounds, sensitivity analysis

Readings:

- Morgan and Winship: Chapter 12

4 Instrumental Variables

Treatment noncompliance, principal stratification, local average treatment effects, Wald estimator and two-stage least squares

Readings:

- Angrist and Pischke: Chapter 4
5 Regression Discontinuity

Sharp and fuzzy designs, identification, estimation, falsification checks

Readings:

- Angrist and Pischke: Chapter 6

6 Difference in Differences and Methods for Panel Data

Selection on time-invariant unobservables

Readings:

• *Angrist and Pischke: Chapter 5*
7 Causal Mechanisms

Direct and indirect effects, sequential ignorability, sensitivity analysis and research designs

Readings:


