DS-UA 0201: Causal Inference

Credits: 4 credits

Course description and impact

We often want to know the relationship between cause and effect. Almost every domain has significant cause and effect questions that can drive decisions with significant consequences. Here are some examples:

- Would providing formula to mothers in Africa reduce the rate of infant mortality?
- Did the reduction of US taxes on corporations increase employment?
- Will stopping advertising for a month greatly reduce sales?
- Does completing a data science major increase happiness?

You've probably read that "correlation does not imply causation." But that raises the question: What exactly is causation and how can it be measured? This course answers that question and provides tools for understanding causation.

We will start with the situation in which you are able to design and implement the data gathering process, called the experiment. We'll define causation in the context of various liberal arts disciplines, explore underlying theory, identify preconditions required for A to cause B, show how to design robust experiments, and discuss how to understand threats to the validity of less- than-robust experiments. We'll cover randomized controlled experiments, inclusion of a control group, blind and double-blind experiments, and pitfalls.

Often one does not have the ability to design and experiment. Instead one is handed observational data, data gathered for some purpose other than to establish cause and effect. Starting about 20 years ago, significant progress has been made in deducing cause and effect from such data. Some of these techniques are complex and require both advanced probability and significant programming skills.

Techniques that depend on the investigators' ability to design and execute an experiment are well known. Often, however, the investigators cannot design an appropriate experiment and must work with existing data. Techniques for dealing with "observation data" have arisen comparatively recently. All of them require working with the data carefully.

In this course, we will cover experimental design and then turn to those careful approaches, where we will consider such approaches as quasi-experiments, regression discontinuities, differences in differences, and contemporary advanced approaches.

Learning objectives

After successfully completing the course, students are able to

- Understand randomized control experiments in the humanities, social sciences, and sciences.
- Articulate what might have gone wrong in such experiments.
- Design and implement computationally-enhanced approaches to deduce causality from observational data.

Office hours

The course instructor is available two hours per week for one on one meetings with students.
[Specifics added here for finalized syllabus]

Topics covered

The course is structured into a sequence of lectures and accompanying assignments.

The assignment consists of labs and homeworks.

- Labs are short exercises done in class and submitted in class.
- Homeworks are longer exercises designed to take a week.

This course uses the R language because there are robust libraries for causal inference that are aligned with one of the textbooks. The first several weeks are used to teach R programming skills.

There are no stand-alone essay-type assignments. Instead writing is embedded in the labs and homeworks. During the labs, students write computer programs. The computer programs are complex to write, as often each part must be cohesive with the others. During the homeworks, students write computer programs and provide short essays on the interpretation of the data and the implications for the data around some decision or problem that the data inform. The writing length would typically be two to three pages for these essays.

The textbooks for the course are:

- (AP) Joshua D. Angrist and Jorn-Steffen Pischke, *Mastering Metrics, The Path from Cause to Effect*, Princeton University Press, 2015.
- Kosuke Imai, *Quantitative Social Science: An Introduction*, Princeton University Press, 2017.
- (PM) Judea Pearl and Dana MacKenzie, *The Book of Why: The new Science of Cause and Effect*, Basic Books, 2018.

- (RLD) cran.r-project.org, R Language Definition. Available on-line at https://cran.r-project.org/doc/manuals/r-release/R-lang.html
- (WG) y Hadley Wickham and Garrett Grolemund, *R for Data Science*, O'Reilly, 2017.

A weekly schedule is below. Each week has 2 lectures and 1 lab. Some weeks have a homework that is due the following week.

1. Week 1: The R language; RLD 1, 2, 3, 4; WG 2 (3 pages), 4 (3 pages), 6 (4 pages)
    a. Lecture topics: RStudio, variables and data types, assignment and objects, control flow, functions, scripts, using modules in libraries, writing libraries yourself, evaluation of function arguments (call by value and evaluate in context)
    b. Lab on: install RStudio, writing programs.
2. Week 2: Important Libraries I; RLD 5, 6, 8, 9
    a. Lecture topics: data frames and tibbles (WG 7 (5 pages)), readr (read and write files, WG 8 (21 pages)), tidyr (observations-in-rows data, WG 9 (22 pages)) ggplot2 (graphics; WG 1 (32 pages))
    b. Lab on : practice with libraries
    c. Homework: 1 assigned
3. Week 3
    a. Lecture topics: Introduction to causal inference. Credibility revolution. Levels of causation (in the sense of Pearl). Potential outcomes (Rubin) framework. Subjectivity of causal inference; Review of conditional probability, especially Bayes' Rule and its role in causal inference.
    b. Reading: AP Ch 1 (47 pages), PM Ch1 (31 pages), PM Ch 3 (43 pages)
    c. Lab on: I, Ch1 (31 pages) and I Ch 6 (73 pages)
4. Week 4
    a. Lecture topics: Selection bias, randomized control trials and their problems. Internal and external validity.
    b. Reading: AP Ch1. PM Ch 1
    c. Lab on: I Ch 2 (44 pages)
5. Week 5
    a. Lecture topics: Review of regression I (ordinary least squares, regression coefficients, "controls," interaction effects).
    b. Reading: AP Ch2 (52 pages), PM Ch 4 (33 pages)
    c. Lab on: I Ch 4 (67 pages)
    d. Homework 3 assigned
6. Week 6
    a. Lecture topics: Review
    b. Midterm Exam (in class)
    c. Lab: Midterm Q&A
7. Week 7
    a. Lecture topics: Review of regression II (conditional ignorability and its problems, confounders/lurking variables.
    b. Reading: AP Ch2, PM Ch4
    c. Lab on: I Ch4

8. Week 8
    a. Lecture topics: quasi-experiments I (instrumental variables, independence, exclusion and other assumptions. Education/charter school case study.)
    b. Reading: AP Ch 3 (50 pages)
    c. Lab on: I Ch 4
    d. Homework 4 assigned
9. Week 9
    a. Lecture topics: quasi-experiments II (Regression discontinuity. Thresholds, running variables, “sorting.” Case study on incumbency bias in the United States politics.)
    b. Reading: AP Ch4 (32 pages)
    c. Lab on: I Ch 4
10. Week 10
    a. Lecture topics: Quasi-experiments III (differences-in-differences, parallel trends, case study on fast food wages problem)
    b. Readings: Ap Ch 5 (32 pages)
    c. Lab on: I Ch 4
    d. Homeworks 5 assigned
11. Week 11
    a. Lecture topics: directed acyclic graphs and the paradoxes they “resolve,” introduction to causal diagrams
    b. Reading: PM Ch 6 (31 pages)
    c. Lab on: to be determined
12. Week 12
    a. A: Lecture Topics: Big data, machine learning, and causal inference
    b. Reading on: PM Ch 10 (23 pages)
    c. Lab on: to be determined
    d. Homework 6 assigned
13. Week 13: Designed at discretion of the instructor
14. Week 14: Designed at the discretion of the instructor
15. Week 15
    a. Lecture topics: summary, final exam review
    b. Lab: active study time for the final

Course assessment

All assignments (labs and homeworks) must be entirely be the student’s own submissions. Any sharing or copying of assignments is considered cheating and will result in an F in the course. A second cheating incident will, by CAS rules, result in a one-semester suspension from the College.

Students accumulate up to 100 points during the course.

- Up to 10 points for completing labs
- Up to 12 points for completing homeworks on time
- Up to 32 points for the midterm
- Up to 46 points for the final.

Grades will be determined using this scale:

Grade in Course	**Points Earned**
A	94 - 100
A-	90 - 93
B+	87 - 89
B	84 - 86
B-	80 - 83
C+	76 - 79
C	72 - 75
C-	70 - 71
D+	66 - 69
D	62 - 65
D-	60 - 61
F	Less than 60

Moses statement

Disability Disclosure Statement: Academic accommodations are available for students with disabilities. The Moses Center website is www.nyu.edu.csd. Please contact the Moses Center for Students with Disabilities (212-998-4980 or mosescsd@nyu.edu) for further information. Students who are requesting academic accommodations are advised to reach out to the Moses Center as early as possible in the semester for assistance.