POLS0012 – 2019-2020
Causal Analysis in Data Science

This module introduces the rapidly-growing field of causal inference. Increasingly, social scientists are no longer willing to merely establish correlations and assert that these patterns are causal. Instead, there is a new focus on *design-based inference*, designing research studies in advance so that they yield causal effects. We will begin by asking what it means for *X* to cause *Y*, using the framework of potential outcomes. We will then look at the most popular research designs in causal analysis, including experiments (also known as randomised control trials), natural experiments that we can analyse with instrumental variables and regression discontinuity techniques, and causal inference over time with difference-in-differences and synthetic control. We will also evaluate 'observational' methods -- regression and the related technique of matching -- from the standpoint of causal inference. This course has a hands-on, practical emphasis. Students will learn to design effective studies and implement these methods in R, and will become critical consumers and evaluators of cutting-edge research. Examples will be drawn from economics, geography, political science, public health and public policy.

By the end of the module, students will be able to:

- Understand the concept of causation in the social sciences
- Distinguish between observational and causal analysis
- Design research studies that can yield causal effects
- Implement a range of techniques of causal analysis including experiments, matching, instrumental variables, regression discontinuity, difference-in-differences and synthetic control
- Evaluate the advantages and disadvantages of different research designs and methods
- Critically read and evaluate quantitative journal articles in the social sciences

LECTURES, TUTORIALS AND LEARNING MATERIALS

Each week there will be an introductory lecture followed by a tutorial. The lecture will last two hours and the tutorial will last one hour. The lectures will introduce students to the course material. The tutorials will be largely computer-based, learning to implement the techniques in R.

In contrast to other QStep courses, the module contains material that is not very mathematically challenging, but can be conceptually difficult. For that reason, the lectures rely far less on slides and much of the teaching will take place on the whiteboard. A few slides are used and will be provided in advance so that students can prepare for the class. To make sure that students have a good record of what is discussed in class, full typed lecture notes will be provided after each session. To do well in the course, you should complete the recommended readings before class, and then review the lecture notes and readings again after class to make sure you have understood everything.

ASSESSMENT

The module is assessed through the completion of two essays. Essay 1 is worth 33% of the final grade and Essay 2 is worth 67%.

- Essay 1 will be 1500 words and requires students to write a review of a published journal article, critically assessing it from the standpoint of causal analysis. This will be available at the start of term, but I recommend waiting until reading week before beginning work on it.
- Essay 2 will be 3000 words and comes in two parts. Part A will be a set of quantitative questions that require students to implement techniques from the course in R and write up the results, similar to the weekly tutorial assignments. Part B requires students to design an original research study using one of the techniques taught in the module. Part B will be available at the start of term so that students can work on it at their own pace.

Please remember that plagiarism is taken extremely seriously and can disqualify you from the module (for details of what constitutes plagiarism see http://www.ucl.ac.uk/current-students/guidelines/plagiarism). If you are in doubt about any of this, ask the tutor.

OTHER NON-ASSESSED WORK

The tutorials will allow students to apply and test their knowledge of the material covered on the module. You will be assigned exercises to complete in R, which may take longer to complete than the one-hour slot. If you do not finish during class time, you must finish them in your own time. Full solutions will be posted on the course Moodle page. From week 3 onwards, each week there will also be a short optional reading comprehension exercise that asks questions about the journal article assigned for the week's readings. It is designed to assist you in understanding original journal articles. Again, solutions will be posted on Moodle. It is optional, but very strongly recommended as it will provide essential practice for the first essay (the article review assignment).

READING MATERIALS

In order to fully understand of the concepts and techniques taught in this module, students will need to do background reading. Causal analysis is a relatively new and rapidly-evolving field. As such, there is no single textbook that covers the whole course, although we will read much of Gerber and Green's book on experiments and Angrist and Pischke's textbook on causal inference, both listed below. Other required readings on the techniques that we cover are drawn from a variety of other textbooks and journal articles. In addition, students must also read applied journal articles that implement the methods we learn about. Required articles are listed for each week. It is essential to read these each week, given that the first essay requires students to read and critically evaluate a journal article. It is not always necessary to read and understand every detail of each article; focus on how and why they apply the methods we learn about, and whether or not they do a good job.

The main textbooks for this course are:

- Alan S. Gerber and Donald P. Green. *Field Experiments: Design, Analysis and Interpretation*. WW Norton and Co., 2012
- Joshua D. Angrist and Jorn-Steffen Pischke. *Mastering Metrics: The Path from Cause to Effect*. Princeton University Press, 2015

Many other textbooks cover parts of the course, often in a more advanced fashion. Here's a list of works to consult for more information on certain topics; we'll also read individual chapters from some of them:

- Joshua D. Angrist and Jorn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press, 2009
- James Druckman, Donald Green, James Kuklinski and Arthur Lupia (eds.). *Cambridge Handbook of Experimental Political Science*, Cambridge University Press, 2011
- Thad Dunning. *Natural Experiments in the Social Sciences: A Design-Based Approach*. Cambridge University Press, 2012
- Guido Imbens and Donald Rubin. *Causal Inference for Statistics, Social and Biomedical Sciences: An Introduction.* Cambridge University Press, 2015
- Stephen L. Morgan and Christopher Winship. *Counterfactuals and Causal Inference: Methods and Principles for Social Research 2nd ed.* Cambridge University Press, 2014
- Rebecca Morton and Kenneth Williams. *Experimental Political Science and the Study of Causality: from Nature to the Lab*. Cambridge University Press, 2010
- Paul R. Rosenbaum. *Observation and Experiment: An Introduction to Causal Inference*. Harvard University Press, 2017

WEEKLY OUTLINE

Note: for journal articles marked "[C]", a reading comprehension quiz will be available on Moodle to help prepare you Essay 1

Week 1. Statistical Preliminaries

We'll start the module with a recap of some core skills and concepts from statistics, as well as R code. Some of the module uses techniques that you haven't seen since first year, so we'll use this first class to bring everyone up to speed in preparation for the rest of the module.

Week 2. Causation and Randomised Experiments

We'll develop a counterfactual model of causation that explains the distinction between correlation and causation, illustrated by epidemiological debates about diets and health outcomes. We'll use the model to examine why randomised experiments offer a solution to the "fundamental problem of causal inference", and we'll learn how to analyse experiments using average treatment effects.

Required Reading:

- Gerber and Green, Chapters 1 and 2.1-2.6
- Gary Taubes, "Do We Really Know What Makes Us Healthy?", *New York Times Magazine,*16th September 2007. Available at: http://www.nytimes.com/2007/09/16/magazine/16epidemiology-t.html

Supplementary Reading:

- Angrist and Pischke, Chapter 1
- Imbens and Rubin, Chapters 1-2

Week 3. Randomised Experiments: Internal Validity

Experiments are statistically simple, but complex to administer in practice. We'll cover the concept of internal validity: does an experiment truly uncover a causal effect? We'll learn how to use balance tests to detect failures of randomisation, as well as how to cope with attrition. A famous experiment on class size reduction in primary schools provides a key example of the challenges of achieving internal validity in practice.

Required Reading:

- Gerber and Green, Chapters 2.7, 4.2-4.4 and 7
- [C] Alan Krueger (1999). "Experimental estimates of education production functions." *Quarterly Journal of Economics* 114 (2): 497-532[focus on pp. 497-517]

Supplementary Reading:

- Gerber and Green, Chapter 8
- Morton and Williams, Chapter 7.1-7.2

Week 4. Randomised Experiments: Inference and External Validity

This week we'll finish learning to analyse experiments by looking at a new inference technique (Fisher's Exact Test, aka randomisation inference). Then we'll look briefly at external validity. The aim of experiments is to learn about causal effects in the real world, but they may take place in artificial settings or on samples that differ from the populations that we care about. We'll ask how much we can hope to learn from experiments and how policy-makers can use experimental results in practice.

Required Reading:

- Gerber and Green, Chapter 3
- Dani Rodrik (2008). "The new development economics: we shall experiment, but how shall we learn?" *Harvard Kennedy School Research Paper*
- [C] David Broockman and Josh Kalla (2015). "Campaign Contributions Facilitate Access to Congressional Officials: A Randomized Field Experiment." *American Journal of Political Science* 60 (3): 545-558.

Supplementary Reading:

- Gerber and Green, Chapter 11
- Morton and Williams, Chapters 7.3-9

Week 5. Observational Studies and Causal Inference: Matching, Propensity Scores and Regression

In many cases it is impossible to carry out experiments. Matching, often using propensity scores, offers a close analogy to experiments in an observational setting and involves a similar set of assumptions to regression. We'll learn how to do matching, asking how closely observational

methods can approximate experiments. Examples are drawn from literature on smoking and health, and violence in civil wars.

Required Reading:

- Angrist and Pischke, Chapter 2 [including Appendix pp. 82-5]
- Elizabeth Stuart (2010). "Matching methods for causal inference: a review and look forward." *Statistical Science* 25 (1), pp. 1-21
- Donald Rubin (2007). "The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials." *Statistics in Medicine* 26 (1): 20-36

Supplementary Reading:

- Angrist and Pischke, *Mostly Harmless Econometrics* Chapter 3
- Imbens and Rubin, Chapters 12-18
- [C] Jason Lyall (2010). "Are co-ethnics more effective counterinsurgents? Evidence from the second Chechen war." *American Political Science Review* 104 (1): 1-20

Week 6. Compliance, Instrumental Variables and Natural Experiments

Instrumental variables is a powerful technique that has been used in two different settings. We'll first learn how to use instrumental variables to analyse randomised experiments where some units fail to comply with the experiment. Second, natural experiments - where an outcome occurs randomly without the intervention of the analyst - have become increasingly popular in the social sciences. We'll define natural experiments and learn how to analyse them using the method of instrumental variables.

Required Reading:

- Gerber and Green, Chapters 5 and 6
- [C] David Clingingsmith, Asim Ijaz Khwaja and Michael Kremer (2009). "Estimating the impact of the Hajj: religion and tolerance in Islam's global gathering." *Quarterly Journal of Economics* 124 (3): 1133-1170

Supplementary Reading:

- Angrist and Pischke, *Mostly Harmless Econometrics*, Chapters 3-4
- Dunning, Chapters 1 and 4

Week 7. Instrumental Variables and Natural Experiments in Practice

We'll use recent studies to illustrate how instrumental variables can work when applied to natural experiments and how they can go wrong. We'll discuss studies on the effect of western TV on support for communism in East Germany, the relationship between police numbers and crime, the political impact of the US 'Tea Party' protest movement, and how participation in the Hajj pilgrimage alters the beliefs of Muslims.

Required Reading:

- Angrist and Pischke, Chapter 3

- [C] Jens Hainmueller and Holger L. Kern (2009). "Opium for the masses: how foreign media can stabilize authoritarian regimes." *Political Analysis* 17 (4): 377-399
- Steven D. Levitt (1997). "Using electoral cycles in police hiring to estimate the effects of police on crime." *American Economic Review* 87 (3): 270-290
- Madestam *et al* (2014). "Do political protests matter? Evidence from the tea party movement." *Quarterly Journal of Economics* 128 (4): 1633-1685

Supplementary Reading:

- Dunning, Chapters 7-10

Week 8. Regression Discontinuity Designs

Regression discontinuity analysis involves a natural experiment where treatment is assigned based on an arbitrary rule, like exceeding a threshold. We'll learn how to do the analysis, looking at a paper on whether British MPs are able to use office to enrich themselves.

Required Reading:

- Angrist and Pischke, Chapter 4
- [C] Andy Eggers and Jens Hainmueller (2009). "MPs for Sale? Returns to Office in Postwar British Politics." *American Political Science Review* 103 (4): 513-533

Supplementary Reading:

- Angrist and Pischke, *Mostly Harmless Econometrics*, Chapter 6
- Dunning, Chapter 3

Week 9. Causal Inference over Time: Difference-in-Differences and Fixed Effects

Difference-in-differences or fixed effects can be used for causal inference with panel data, when a treatment varies over time in some units but not others. We'll look at a very famous example that over-turned economists' thinking on minimum wages.

Required Reading:

- Angrist and Pischke, Chapter 5
- [C] David Card and Alan Krueger (1994). "Minimum wages and employment: a case study of the fast food industry in New Jersey and Pennsylvania." *American Economic Review* 84 (4): 772-793

Supplementary Reading:

- Angrist and Pischke, *Mostly Harmless Econometrics*, Chapter 5

Week 10. Synthetic Control Analysis

The new method of synthetic control is useful for causal inference over time with a small number of units, particularly when the treatment occurs in only one unit. We'll learn how to create a synthetic control case to compare to the treated unit, based on an optimal combination of untreated units. We'll look at applications including the impact of tobacco control measures and German reunification.

Required Reading:

- Alberto Abadie, Alexis Diamond and Jens Hainmueller (2015). "Comparative politics and the synthetic control method." *American Journal of Political Science* 59 (2): 495-510
- Alberto Abadie, Alexis Diamond and Jens Hainmueller (2010). "Synthetic control methods for comparative case studies: estimating the effect of California's tobacco control program." *Journal of the American Statistical Association* 105 (490): 493-505