CS&SS / STAT 566 CAUSAL MODELING

WINTER 2019

Instructor: Thomas Richardson

E-mail: `thomasr@uw.edu`

Lecture: TTh 1:00-2:20 MUE 224

Office hour: T 3.00-4.00 in Padelford (PDL) B-313 (or by appointment)

Teaching Assistant: Rahul Biswas

E-mail: `rbiswas1@uw.edu`

Lab Time: M 3:30-4:20 SAV 117

Office hour: T 10:00-11:00 and F 1:00-2:00 Padelford (PDL) C-302

Web-site: `www.stat.washington.edu/tsr/s566`

Overview

This course provides an introduction to formal approaches to causality, based on potential outcomes (aka counterfactuals) and graphs.

These provide precise answers to questions such as:

- In what way does information coming from a randomized experiment differ from that resulting from a purely observational study?
- If subjects in a randomized experiment don't 'comply' with the treatment to which they were assigned, how (if at all) should this influence the conclusions that we draw?
- What is 'endogeneity' and how should it be 'adjusted' for?
- What is 'mediation'? How do we define a 'direct' effect?
- Can the measurement of 'background variables' or 'covariates' help to 'strengthen' inferences drawn from observational data? If so, which extra variables should be measured, and how should they be taken into account?
- What is 'confounding' ? When should variables be adjusted for?
- How can matching be used to adjust for confounding?
- What is the relationship between causal models or hypotheses and statistical models or hypotheses?

A central theme of this course will be that in order to properly answer these and other questions it is necessary to use a formal theory of causation; intuition alone is neither a safe nor a reliable guide. A causal theory aims to provide answers to causal questions in the same way that probability theory provides a formal method of answering questions about chance.

Methods and software will be presented that are based on these theoretical approaches.

Causal methods in the Social Sciences

Among other models, we will look at the method of 'instrumental variables', originating in Economics, and 'structural equation models' (SEM) which are widely used in the Social Sciences. We will consider some of the following questions:

- What is an 'instrumental variable'?
- What makes a SEM different from a regression model?
- What is being assumed when a structural equation model is constructed? Can all of these assumptions be tested?
- Given enough data will we ever/always be able to reject a SEM model which is false?
- Are there principled ways of searching for SEM models?

Texts:

- *Causal Inference in Statistics: A Primer.* J. Pearl, M. Glymour, N. Jewell, Wiley, 2016.
- *Mostly Harmless Econometrics.* J.D. Angrist, J.-S. Pischke, Princeton University Press, 2008.
- *Counterfactuals and Causal Inference.* S.L. Morgan, C. Winship, Cambridge University Press, 2007.

Supplementary Texts:

- *Single World Intervention Graphs (SWIGs): A Unification of the Counterfactual and Graphical Approaches to Causality: A primer.* T. S. Richardson and J. M. Robins. Second UAI Workshop on Causal Structure Learning, Bellevue, Washington, 07/15/2013.

  `http://bit.ly/2qjDF6T`
- *Causality: Models, Reasoning and Inference.* J. Pearl, Cambridge University Press, Cambridge, 2nd Edition, 2009.
- *Causal Inference for Statistics, Social, and Biomedical Sciences. An Introduction.* G. W. Imbens and D. B. Rubin, Cambridge University Press, 2015.

Credit

Credit will be assigned on the basis of Homework (80%) and Project (20%).

Students taking the course for credit will be required to do homework exercises, complete a project and make a poster presentation. (For more details on the project see below.)

Students will be asked to come up with topics and datasets for projects. If your dataset is used for a project I will add 0.1 to your grade (assuming it is $<$ 4.0!). Poster presentations will be 2.30-4.30 on Tuesday March 19. *If insufficient datasets are available, I will give a final exam.*

Pre-requisites

The pre-requisite for this class is one prior class in Statistics. I will try to briefly review most concepts before using them.

The course will presume some familiarity with basic mathematical concepts, such as what constitutes a 'proof'.

If you are having trouble following material that is being presented please let me know as soon as possible (in class, by coming to office hours, or via e-mail). I am always happy to provide additional examples, or clarifications, if requested. (I am not a mind-reader, so please let me know if it would be helpful!)

Registration

Owing to limited spaces in the class, if you do not show up during the first week you may be dropped in order to allow others to register.

Outline coverage of material (by week)

1. Motivation; Counterfactuals & Causal Effects
2. Graphs and independence: d-separation;
3. Graphs and counterfactuals;
4. *Presentation of data-sets*

   Measures of dependence: Odds Ratios, Risk Ratios, Risk differences
5. Imperfect Experiments and Instrumental Variables
   Handling Non-Compliance
6. Direct Effects and Mediation
7. Structural Equation Models (SEMs)
   Understanding the constraints implied by SEMs
8. When does a regression coefficient give a causal effect?

9. Handling Confounding when structure is known: non-parametric identification and the *do*-calculus

10. When little is known about causal structure: exploratory search.

Assigned Readings

Week 1: Introduction to Counterfactuals

- Morgan and Winship: Section 1.1; Sections 2.1-2.5;
- Angrist and Pischke: Chapter 1, and Chapter 2.1-2.2;

Projects

The group project counts for 20% of the grade.

Groups will be assigned by the instructor so as to ensure the presence of someone with a proposed topic in each group and diversity of disciplines, among other criteria.

Typical projects include:

- Applying methods of causal inference to a data-set;
- Investigating methods of causal inference, either analytically, or via a simulation study;
- Critique/summary of a recent research paper dealing with causality.

Please start thinking now about possible datasets for projects. Projects from your own research, lab or department are highly appropriate. Any data should be available for analysis, i.e. the data should now be accessible, and analysis by a project group of two students from the class (including yourself) should be allowed by the owners of the data. If your dataset is used for a project I will add 0.1 to your final grade (assuming it is < 4.0!).

Please contact me to talk about possible projects. All proposed projects will be presented in two-minute talks in class on January 22 and 24. After that, I will ask you for your preferences and assign project groups. I will be happy to meet with you to discuss possible approaches and software that is available (both before and after projects are proposed.)

There are examples of projects from the past few years on the class web-site. I will also provide some links to data repositories.

The components of the project are as follows:

1. A one-page topic statement. Due Tuesday, February 5 in class. (5%).
2. A five minute presentation (no more!) of your chosen topic on February 14 and/or 19 in class. (10%).
3. Presentation of final results (40%).

   There will be a Poster Session from 2.30-4.20 on Tuesday March 19 (Finals Week).
4. A written report of no more than 15 pages (including all materials, and in 12 point font size with one-inch margins on all sides).

   Due Thursday, March 21 at 2.30pm by email. Counts for 45%.

Additional details for each component:

One-page statement: This should include your names, the title of the project, a brief description of the problem and a few key references. It should state clearly what the data are and what the causal questions of interest are. It should not exceed one page. Please send me this by email (pdf format).

Initial presentation (five minutes): Based on the one-page statement. I recommend that you prepare two to four slides, and/or provide a handout. You should also rehearse the presentation, with an emphasis on timing.

Final poster presentation: This should present and summarize the problem or subject that you analyzed. your main results, and some idea of what you did. This will take place

Written report: This should summarize the problem, the causal questions that you investigated, and your analysis. Do not include large amounts of raw computer output in this report (the 15-page limit includes all output, graphs and appendices). If computer results or plots are relevant to your argument, include them in the main body of the report, labeling them as Table 1, Figure 1, etc. A suggested outline for the report is:

(a) state the problem and the questions being asked, include enough detail to be self-contained;

(b) state your methods of analysis;

(c) present your results;

(d) describe/discuss your conclusions.

Though each group will produce one report, I will ask each project member to identify sections of the report on which they were 'first author'.