ICPSR Summer Program

Causal Inference  
Course Syllabus

Instructors:  
Dominik Hangartner  
Marco R. Steenbergen

Summer 2011
Short Course Description

This course provides an introduction to statistical methods used in causal inference and is geared specifically toward students in the social sciences. Using the potential outcomes framework of causality, topics covered include research designs such as randomized experiments and observational studies. We explore the impact of noncompliance in randomized experiments, as well as nonignorable treatment assignment in observational studies. To analyze these research designs, the methods covered include matching, instrumental variables, difference-in-difference, and regression discontinuity. Examples are drawn from economics, political science, public health and sociology. Prerequisites for the course are knowledge of multiple regression using linear algebra and some familiarity with generalized linear models. The course will rely on R for computation.

Overview and Objectives

Causal inference from observational studies or "broken" randomized experiments (due to e.g. noncompliance) historically has been viewed as problematic, or even illegitimate, by most statisticians. This has led to the paradoxical situation that statisticians provided countless sophisticated methods to model conditional expectations for association and prediction, while being almost completely silent about how to validly estimate the effect of a cause---although that is exactly what most of the social and behavioral sciences is all about.

While randomized experiments are rightfully considered to be the gold standard for causal inference, they are quite often not a realistic option because of ethical and / or practical limitations. Then, one typically has two choices: to give up and miss the opportunity to learn anything about the problem at hand or to turn to observational studies. By taking this class, you opt for the second option.

However, causal inference from observational studies or broken randomized experiments is notoriously difficult, since (for point identification) it necessarily involves unidentifiable and hence untestable assumptions about the magnitude and direction of confounding. Hence, the plausibility of causal inferences from observational data always depends on the amount and strength of substance-matter knowledge and the quality of the research
This course aims at providing you with the tools to engage in proper causal inference. The specific goals are twofold: (1) to sensitize you to the many mines that one could step on while making causal claims and (2) to provide a set of research designs and statistical techniques that can help you avoid some of these mines. We pursue these goals within the broad framework of the so called Rubin Causal Model (RCM) which is based on Neyman's potential outcomes notation.

The hope is that, by the end of the class, you will be sensitive to the assumptions that causal inference entails and that you will be able to work through research designs and methodologies that, as much as possible, avoid implausible assumptions.

**Organization**

Many problems of causal inference from observational studies revolve around the concept of confounders, i.e. true causes of an effect that may render the impact of a putative cause spurious. There are different ways of handling confounders, depending on whether they are observed or not. After proving a general introduction to causation and causal inference (Module I), we begin by considering research designs in which the confounders are unobserved but rendered impotent through randomization (Module II). Since randomization is not always feasible, we may have to rely on other methods of controlling for confounders. In Module III, we consider designs in which the confounders are observed and can be controlled statistically. In Module IV, we focus on cases in which at least some of the confounders are unobserved.

**Prerequisites**

Prerequisites for the course are knowledge of multiple regression using linear algebra and some familiarity with generalized linear models. The course will rely on R for computation. If you need to review material, please consult this excellent textbook:


If you need to review some R basics, you may want to have a look at


**Materials**

The following textbooks are required:


In addition to the required books, you may wish to obtain a copy of the textbook:


Evaluation

Your course grade is determined on the basis of your homeworks. Most of these exercises will require that you use the statistical package R, which may be downloaded free of charge from http://www.r-project.org/. There will be no final exam.

Schedule

Module I  Causality and Causal Inference

Session 1  Causality I: Causal Inference Using Potential Outcomes
Today we will introduce the topic of causal inference. We will define causal effects based on the potential outcomes framework of Neyman and Rubin, encounter the fundamental problem of causal inference, and discuss confounding as what separates association from causation, and observational studies from randomized experiments.

Readings:


Further Readings:


Session 2
Causality II: Different Models of Causality
We discuss the logic of the Neyman-Rubin Causal Model. We relate the model to alternative and complementary conceptions of causation in philosophy of social science, with special emphasis on counterfactual theories based on the "closest possible worlds" (Lewis 1973) and Woodward's (2003) "manipulability theory" of causation and causal explanation.

Going further back in time, we review Mill's method of difference and contrast it with Fisher's randomized experiment as the "reasoned basis" of inference.

In addition, we revisit the controversy about "No Causation Without Manipulation" (Holland 1986) and discuss the importance of well-defined interventions for meaningful causal inference.

Readings:

Further Readings:

Module II Randomized Experiments
Session 3
Randomized Experiments
We quickly review the logic of randomized experimentation, a research design that is widely believed to maximize internal validity and that is becoming ever more popular in the social sciences. We pay special attention to Fisher's randomization inference, in which randomization is the "sole and reasoned basis for inference". We discuss factorial and randomized-block designs, paying attention to how these designs function.

Readings:

Further Readings:
Session 4  
**Causality III: Different Approaches to Causal Inference**

We will discuss the graphical approach to causal inference as introduced by Pearl (e.g. 2000) and Spirtes, Glymour, and Scheines (1993). We discuss some distinctive advantages and drawbacks of directed acyclical graphs (DAGs) and show how they complement the potential outcomes framework as a different, but equivalent language (Lauritzen 2004). Based on DAGs, we revisit and old debate about the (im-) possibility to draw causal inference without substance-matter knowledge.

Readings:

Morgan and Winship. Chapter 1.6, pp. 24-30.

Further Readings:


**Homework 1: Assessing the effect of watching more Sesame Street on child cognitive development; program effect of watching TV vs. program effect of encouragement to watch.**

Module III  Causal Inference Under Exogeneity

Session 5  
**Subclassification and Matching I**

The advantage of randomized experiments is that potential confounders can be safely ignored since they will not bias estimates of treatment effects. But randomization is not always practical, nor is it always ethical. How can one ensure valid causal inference in a world without randomization? Today, we discuss designs which assume that selection into the treatment groups is based on observables. We start by considering two very intuitive methods, subclassification and exact
matching techniques.

Readings:


Further Readings:


**Session 6**

**Matching II**

Next, we discuss matching techniques that are based the propensity score and on Euclidean distance. We also consider some practical issues with matching such as severe sample size reduction, outliers/common support, and estimating standard errors. With respect to the latter topic, we consider the (somewhat surprising) failing of bootstrapping methods for valid standard errors.

Readings:


Further readings:


**Session 7**

**Matching III**

Today, we consider weighting as a nifty alternative to matching. In addition, we compare the OLS regression with matching estimators.

Readings:

Angrist and Pischke. Chapter 3.3.1-3.3.3, pp.69-91.

**Session 8**
Matching VI
Review / Lab Session:

Introduction to the R package `Matching" (Sekhon 2007).


Module IV  Causal Inference When the Confounders Are Not All Observed

Session 9
Difference-in-Differences
Confounders cannot always be observed and if they cannot, then
alternatives to matching have to be found. One such alternative arises in
the context of panel data or repeated cross-sections. Here one can take
the difference between pre- and post-tests and then compare those across
groups. To the extent that the differences in the confounders have
remained constant over time, then this estimator can
produce valid causal inferences.

Readings:

Angrist and Pischke. Chapter 5, pp. 221-246.

Further Readings:

Card, David and Alan B. Krueger. 1994. Minimum Wages and Employment:
The American Economic Review 84: 772-793.

Industrial and Labor Relations Review 43: 245-257.

Session 10
Synthetic Control
If you would like to use Difference-in-Difference but don't have
a good control unit: synthesize one (Abadie and Gardeazabal 2003, Abadie,
Diamond, Hainmueller 2007). This estimator is a simple but potentially
wildely applicable generalization of the Difference-in-Differences
estimator.

Readings:

Abadie, Alberto, Diamon, Alexis and Jens Hainmueller. 2007.
Synthetic Control Methods for Comparative Case Studies: Estimating the
Session 11
Instrumental Variables I
Instrumental variable (IV) methods can be used to address unobserved confounders in the context of cross-sectional data. Today, we discuss the basic logic of IV-techniques, focusing in particular on the LATE estimator.

Readings:
Angrist and Pischke. Chapter 4.1-4.4.4

Further readings:

Session 12
Review / Lab Session:
Illustration of Synthetic Control methods.

Session 13
Instrumental Variables II
In many settings, an instrument is only valid conditional on some covariates. As we will show, the LATE estimator is only able to identify average treatment effects conditional on covariates under very restrictive assumptions. Abadie's LARF estimator develops an ingenious weighting scheme to get unbiased treatment effects under heterogeneity for the subsample of compliers only.

Readings:
Angrist and Pischke. Chapter 4.5-4.7

Session 14
Lab Session:
Illustration of LATE and LARF estimators.

Homework 4: LATE and LARF.

Session 15
Regression Discontinuity Designs I
RDDs arise when selection into the treatment group depends on a covariate score that creates some discontinuity in the probability of receiving the treatment. Capturing the treatment effect in these designs can be thought of as a special case of IVs. Although RDDs are not yet widely used in the social sciences, they appear to have a great deal of potential for application.

Readings:


Further Readings:


