Treatment Effects, Causal Inference, and Machine Learning (XXX-XXX)
Fall 2019
Duke University

Instructor	Justin Kirkpatrick
Email	justin.kirkpatrick@duke.edu
Day & Time	Tuesdays 10:05-11:20am
Office	4120A Environment Hall
Office Hours	Wednesdays 11-12
Credits	3

We've all heard the quote "correlation is not causation." But when can we say something is causal? In policy analysis, we often need causality to make claims about policy effectiveness. The purpose of this course is to equip you with the tools and the intuition necessary to read, digest, criticize, and design policy research that makes causal claims. One fundamental part of this is to be able to identify and analyze sources of endogeneity and bias - for instance, cities with strict gun control tend to have high rates of gun violence; a naive person would conclude that gun control causes violence, but in reality, cities with gun crime problems tend to respond by introducing stricter gun control.

Course Overview

This course will focus on preparing you to design and execute your own research projects, and introduce you to the econometric tools of causal inference. We will examine traditional regression-based estimators including fixed effect models, difference-in-differences, instrumental variables, and regression discontinuity, and will cover the cutting edge in the use of machine learning in causal inference. Evaluation will come from weekly reading responses, a revisable midterm paper, and a final peer-reviewed group project, along with three problem sets where you will re-create analysis from canonical papers. The final project will use the tools and concepts from the class in a manner well-suited for an independent, publishable research project, or your Master's Project.

We will leverage in-class discussion rather than relying on 75-minute lectures each day. Whenever possible, I will limit lecture to 30-45 minutes, using the additional class time to further explore the topic at hand with paired discussion, examples, and conceptual brainstorming.

Course Objectives

By the end of this course, you will be able to:

1. Explain the difference between correlational studies and causal inference.
2. Define endogeneity and list the three most common types of endogeneity bias.
3. Compute basic regression estimates directly from data
4. Summarize identification strategies used in published research
5. Compare and Contrast the main methods for causal inference and state the assumptions in each of them.

6. Apply the concepts to a new data set or research question by developing and writing a research proposal.
7. Appraise proposals and justify the methods and assumptions.
8. Integrate machine learning techniques with causal frameworks.

Prerequisites
ENV710 and high school calculus is required. Familiarity with **R** statistical software is highly advised.

Required Texts
Angrist, Joshua D., and Jörn-Steffen Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2008. (any edition, available in paperback on Amazon)

Cook, Thomas D., Donald Thomas Campbell, and William Shadish. *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin, 2002.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*. Vol. 1. New York: Springer series in statistics, 2001. Available free in pdf from Trevor Hastie's website: https://web.stanford.edu/~hastie/Papers/ESLII.pdf

Attendance and Participation
This course requires active participation in class and completion of all assigned readings. Therefore, attendance is an important part of your grade. More than two unexcused absences will result in a decrease in your total grade by 3% (e.g. from a 92 to an 89). Because we will leverage in-class activities and discussions rather than rely on daily lectures, it is of the utmost necessity that you arrive on time and prepared for discussion, especially having read the assigned papers.

Assignments and Assessments
Please note: **no late work is accepted**. Sakai deadlines will be strictly set.

Weekly reading responses
A reading response of 2-3 paragraphs will be required each week, due **Sunday night at 11:59pm** before each Tuesday class. Responses are to be posted to the Sakai board and set to "visible" for the class. The response will address *one* of the assigned papers or chapters, and will include (i) any strengths or weaknesses the paper displays, and (ii) potential applications or expansions of the concept or model the paper presents (other questions the model may be able to answer, or other ways the model could be adapted). The purpose of this assignment is to get you thinking critically about a paper. Evaluation is pass/fail. (10%)

Response Responses
After reading responses are posted, you will comment on the reading response posted by each member of your group. Your response should expand on the original response, and should add some substance. Response responses are due **By 11:04am the day of class.** Evaluation is pass/fail. (5%)

Problem Sets

Three problem sets where you will use **R** to recreate analyses from canonical papers will be assigned throughout the course. Each assignment will include the data to be used, instructions that you will need to follow, and questions you will need to answer. Evaluation will be based on proper execution of the instructions in **R** and on your answers to each question. (25%)

Final Project (Masters students only)
In groups of 4 to 5, you will draft and present a research proposal to address a policy question of your group's choosing. The draft research proposal will include the following topics:

- Clearly present the policy question and answer the question "why should we care?"
- Identify previous published research on the topic to give context to the question.
- Propose a method of answering the question using methods covered in class or from outside sources.
    - Identify the appropriate model(s) and discuss the reasons for their use in answering your question.
    - Identify potential areas of weakness, including potential threats to identification.
    - Propose and discuss robustness checks.
- Describe the data necessary to execute the analysis (*note: you will not be required to obtain the data or execute the analysis*).
- Devise a research proposal budget

The written portion of the final project will be evaluated by the instructor. (25%)

In addition to the draft proposal, your group will present your proposal to the class in a short (5-8 minute) presentation near the end of the course. You will be evaluated only on your presentation clarity; un-graded feedback on the content of the proposal will be given to improve the final written draft. (10%)

Final Paper and Presentation (PhD students only)
PhD students enrolled in the course will not participate in the Final Project. Instead, PhD students will select an estimation method not covered in class (e.g. Split-sample IV; quantile regression) and will make a presentation to the class. The presentation will introduce the estimator, discuss the estimator's statistical properties including identifying assumptions and asymptotic variance, and provide 1-2 examples of applications. In addition to the presentation, PhD students will submit a short paper summarizing these points and detailing the history of the estimator in the literature. The purpose of the history portion of the assignment is to gain an understanding into the process of establishing estimators, and understand how statistical advances are made in response to analytical needs. (35%)

Tests

There will be one midterm for the class on **November XX, 2019**. The midterm will be in-class, and will be multiple-choice, short-answer, and short-essay format. The questions will pertain to the concepts covered in lecture, reading, and assignments. *The short-answer and short-essay sections of the midterm will be revisable for students scoring below a B*. (25%)

There is no final exam for this course. Your group project and presentation will take its place.

Grading
All grades are considered final unless otherwise noted (see *Midterm*, above). Any request for a re-grade beyond simple point adding mistakes will require that the entire assignment be re-graded by the instructor. Any points previously awarded by the grader may be changed in either direction in the re-grade.

Resources
The University has a wide range of resources available to you. These include the Writing Center (link), ESL resources (link), the Academic Resource Center (link), and the Perkins Library Data Librarian (link) who is well-versed in publicly available (and internal-to-Duke) dataset useful for researchers. If you don't already use Google Scholar, it is an infinitely useful tool for finding academic research papers of the sort you will be reading in this class.

Students with Disabilities
Students with disabilities may contact the Student Disabilities Access Office at disabilities@aas.duke.edu to arrange for accommodations in the class. Only arrangements made in conjunction with the SDA office will be honored.

Academic Integrity
Duke University is a community dedicated to scholarship, leadership, and service and to the principles of honesty, fairness, respect, and accountability. Citizens of this community commit to reflect upon and uphold these principles in all academic and non-academic endeavors, and to protect and promote a culture of integrity. To uphold the Duke Community Standard:

- I will not lie, cheat, or steal in my academic endeavors;
- I will conduct myself honorably in all my endeavors; and
- I will act if the Standard is compromised.

Course Schedule

Week	Topic	Reading	Assignments Due
1-1 (Sep 6)	Stats Review	Skim SCC (Shadish, Cook, and Campbell) Ch 1 MHE (Mostly Harmless…) Ch 1	
1-2	Regression & T-tests	SCC Ch 2	
2-1	Causality, validity, and the Rubin Model	Lecture Notes #2	
2-2	Bias and endogeneity	SCC Ch 3	
3-1	Conditional means, Omitted Variables, and fixed effects	MHE Ch (Some paper that is particularly bad for OVB – a wage paper?)	

3-2	Random Assignment	MHE Ch 2	
4-1	Experiments in Economics (RCT's)		
4-2	Selection on Observables - Matching	MHE	
5-1	Applications of Matching	MHE, Andam et al (2008 PNAS)	
5-2	Difference-in-differences I	Lecture Notes #3	
6-1	Difference-in-differences II	Lecture Notes, MHE	**Problem Set 1 due 9:00am before class**
6-2	Applications and Examples of Diff-in-diff	Kirkpatrick & Bennear (2014), Kotchen (2012)	
7-1	Instrumental Variables (IV) I	Lecture Notes, MHE	
7-2	Instrumental Variables (IV) II	Lecture Notes, MHE	
8-1	Applications and Examples of IV	Chay and Greenstone	
8-2	Regression Discontinuity	Lecture Notes, MHE	
9-1	Regression Discontinuity – time trends, diagnostics, and applications	van der Klaauw (Econometrica 2002)	
9-2	Recap: Selection on Unobservables		**Problem Set 2 due 9:00am before class**
10-1	***Midterm***		
10-2	Asymptotic variance		
11-1	Controlling for spatial effects	Anselin?	
11-2	Synthetic Counterfactual	Abadie, Diamond and Haimueller (2010 JASA)	
12-1	Applications of Synthetic Counterfactual	Sills et al (2015 PLoS One)	
12-2	Machine Learning Overview – Prediction vs. Causation	Hastie, Athey and Imbens (2017 JEP)	
13-1	Machine Learning I – Random Forests	Hastie	
13-2	Machine Learning II – LASSO	Tibshirani (1996)? Something lighter?	
14-1	Machine Learning III – Neural Nets, Deep learning	Augmented SCM (Ben-Michael et al. 2018)	
14-2	Asymptotic properties of Random Forests, LASSO	Wager and Athey (2017 JASA)	
15-1	Using machine learning in causality	Athey (2018 working paper *Impact of…economics*), Belloni, Chernozhukov, and Hansen (2012, 2014	

		series), Manresa (2016)	
15-2	Applications of machine learning in causal frameworks	Burlig et al (2016), Cicala (2017)	
16-1	Group Presentations I		
16-2	Group Presentations II		
Dec 10, 2019			**Problem Set 3 due 9:00 am (last day of instruction)**
December 20, 2019	No final exam. Group project due at Final exam time.		Final Group Project due on Sakai @11:59pm

Acknowledgements
This syllabus and course structure draws heavily from Dr. Lori Bennear's ENV850 course at Duke University.