SOC 690S Machine Learning in Causal Inference
Methodological Foundations and Sociological Applications

Time and Location

Fall 2025
Tuesdays 3:05–5:35PM
Erwin Square Plaza Room 753

Instructor Information

Professor Wenhao Jiang
Email wenhao.jiang@duke.edu
Office Hours Wednesdays 10AM - 12PM or by appointment

Course Overview and Prerequisites

This course introduces the foundations and frontiers of causal inference, machine learning (ML), and their integration in social science research. The course begins with linear regression and its statistical properties, especially in moderately high-dimensional settings, and motivate the use of ML when data are high-dimensional or exhibit nonlinear relationships. Core ML techniques, including regularized regression, tree-based models, and neural networks, are introduced, alongside the modern causal inference framework, covering potential outcomes, randomized experiments, and directed acyclic graphs (DAGs). The goal is to provide a unified foundation for understanding both predictive modeling and causal identification.

Building on these foundations, the course covers advanced topics in causal inference, many of which integrate machine learning methods for flexible estimation and robust inference, such as propensity score methods (PSM) and weighting, regression discontinuity design (RDD), difference-in-differences (DID), instrumental variables (IV), heterogeneous treatment effects, and synthetic control methods. We will also explore emerging methods for causal analysis using unstructured data, including feature engineering from text and image embeddings. The course closes with a forward-looking discussion on causal reasoning within ML.

The class will enable students to engage with current sociological literature that uses advanced methods and to apply these methods in their own work. The course assumes working knowledge of probability theory and statistical inference at the level of *Social Statistics I*. Basic understanding of matrix algebra and calculus is preferred but not required. `R` will be used for data simulation and some problem sets, and `Python` will be supplemented in some ML and embedding methods.

Expectations

Reading and participation (10%):
You are expected to read required readings prior to the class in which they are discussed. If you do not understand the readings, please come prepared to discuss what you did and did not follow.

Problem sets (30%):
During the semester, there will be 5 problem sets. The problem sets are relatively short and will be assigned bi-weekly. They are used to enhance your understanding of the materials. Sometimes `R` coding is required. Problem sets will be released on Tuesdays immediately after class and are due two weeks later by Monday, 11:59PM.

In-Class Midterm Exam (30%):
There will be an in-class midterm on October 21. The exam is open book and open note, but will be timed. You will not be required to write any code for the midterm.

Take-Home Final (30%):
There will be a take-home exam administered on December 13, in lieu of the official exam schedule of the class. This exam is open note and open book, but you may not communicate with class members. There may be questions that require you to use `R` on the final. You are welcome to use your own data or project for certain coding assignments or replication projects. Further details will be announced later in the semester.

Software

This class will use `R`, and prior experience at the level of *Social Statistics I* is assumed. `Python` will be used as a supplementary tool in certain cases, such as for machine learning and embedding-based models.

Use of Artificial Intelligence (AI)

While the use of AI is not permitted during the midterm exam, I encourage students to use AI tools to support their statistical learning, particularly for understanding the steps in statistical derivations. AI assistance is also permitted on problem sets and the take-home final, but students are expected to ensure that they fully understand each step in their solutions.

Course Materials

We will use three textbooks frequently throughout the course:

- [CML] Chernozhukov et al. 2025. *Applied Causal Inference Powered by ML and AI*
- [ISL] James et al. 2013. *An Introduction to Statistical Learning: with Applications in R*.

- [MHE] Angrist and Pischke 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*.

I recommend that you purchase these books if you like to have physical copies. You can digitally access the full text of CML, MHE, and ISL through the embedded hyperlinks. The following textbooks may also be useful for reference, particularly if you choose to delve further into topics covered in this course.

- [CCI] Winship and Morgan 2007. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*
- [CIS] Rubin and Imbens 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*

Schedule

Week	Date	Topic	Problem sets	
			Assign	Due
1	Aug 26	Introduction: Motivation and Linear Regression		
2	Sep 2	Foundation: Machine Learning Basics		
3	Sep 9	Foundation: Machine Learning Advanced	1	
4	Sep 16	Foundation: Causal Inference Basics		
5	Sep 23	Foundation: Causal Inference Advanced	2	1
6	Sep 30	Core: PSM and Doubly Robust Estimation		
8	Oct 7	Core: Instrumental Variable Estimation	3	2
7	Oct 14	*Fall break*		
9	Oct 21	**In-class midterm**		
10	Oct 28	Core: Regression Discontinuity Design	4	3
11	Nov 4	Core: Panel Data and Difference-in-Difference		
12	Nov 11	Advanced: Heterogeneous Treatment Effect	5	4
13	Nov 18	Advanced: Unstructured Data Feature Engineering		
14	Nov 25	Advanced: Causal Reasoning in Machine Learning		5
	Dec 13	**Take-home final**		

The schedule is still tentative as of August 27, 2025. A detailed schedule of topics and readings, by class session, appears below. You are responsible for reading the *required readings* lists prior to the class section.

Aug 26: Introduction, Motivation, and Linear Regression

Required readings

- CML Preface and Chapter 1
- MHE Chapter 3

Additional readings

- ISL Chapter 2 and 3

Sep 2: Machine Learning Basics

Required readings

- ISL Chapter 5 and 6
- Molina Mario and Filiz Garip. 2019. "Machine Learning for Sociology." *Annual Review of Sociology, 45*(1), 27-45.

Additional readings

- ISL Chapter 3
- CML Chapter 3

Sep 9: Machine Learning Advanced

Required readings

- ISL Chapter 8 and 10
- Koch Bernard et al. 2025. "A Primer on Deep Learning for Causal Inference." *Sociological Methods & Research, 54*(2), 397-447.

Additional readings

- CML Chapter 4

Sep 16: Causal Inference Basics

Required readings

- MHE Chapter 2
- CML Chapter 2 and 5
- Lundberg Ian, Johnson Rebecca, and Stewart M. Brandon. 2021. "What is Your Estimand? Defining the Target Quantity Connects Statistical Evidence to Theory." *American Sociological Review*, *86*(3), 532-565.

Additional readings

- CCL Chapter 3 and 4

Sep 23: Causal Inference Advanced

Required readings

- CML Chapter 6 and 7
- Winship Christopher and David J. Harding. 2008. "A Mechanism-based Approach to the Identification of Age–Period–Cohort Models." *Sociological Methods & Research, 36*(3) 362-401.

Additional readings

- Blau M. Peter and Otis Dudley Duncan. 1967. *The American Occupational Structure*.

Free Press. *Chapter 5*: The Process of Stratification.
- Vaisey Stephen and Andrew Miles. 2017. "What You Can—and Can't—Do with Three-wave Panel Data." *Sociological Methods & Research, 46*(1), 44-67.

Sep 30: PSM and Doubly Robust Estimation

Required readings

- ISL Chapter 4
- Funk M. Jonsson et al. 2011. "Doubly Robust Estimation of Causal Effects." *American Journal of Epidemiology, 173*(7), 761-767.
- Sharkey Patrick and Felix Elwert. 2011. "The Legacy of Disadvantage: Multigenerational Neighborhood Effects on Cognitive Ability." *American Journal of Sociology, 116*(6), 1934-1981.

Additional readings

- Breen Richard and Ermisch John. 2024. "Using Inverse Probability Weighting to Address Post-Outcome Collider Bias. *Sociological Methods & Research, 53*(1), 5-27.
- Bang Heejung and James M. Robins. 2005. "Doubly Robust Estimation in Missing Data and Causal Inference Models." *Biometrics*, *61*(4), 962-973.

Oct 7: Instrumental Variable Estimation

Required readings

- MHE Chapter 4
- CML Chapter 12 and 13

Additional readings

- Harding J. David et al. 2018. "Imprisonment and Labor Market Outcomes: Evidence from a Natural Experiment." *American Journal of Sociology, 124*(1), 49-110.
- Chyn Eric, Brigham Frandsen, and Emily Leslie. 2025. "Examiner and Judge Designs in Economics: A Practitioner's Guide." *Journal of Economic Literature, 63*(2), 401-439.

Oct 14: *Fall Break*

Oct 21: In-class Midterm

Oct 28: Regression Discontinuity Design

Required readings

- MHE Chapter 6
- CML Chapter 17
- Cattaneo Matias and Rocio Titiunik. 2022. "Regression Discontinuity Designs." *Annual Review of Economics, 14*(1), 821-851.

Additional readings

- Dell Melissa and Pablo Querubin. "Nation Building Through Foreign Intervention: Evidence from Discontinuities in Military Strategies." *The Quarterly Journal of Economics, 133*(2), 701-764.
- Card David et al. 2017. "Regression Kink Design: Theory and Practice." In *Regression Discontinuity Designs: Theory and Applications*, 341-382.

Nov 4: Panel Data and Difference-in-Difference

Required readings

- MHE Chapter 5
- CML Chapter 16
- Roth Jonathan et al. 2023. "What's Trending in Difference-in-Differences? A Synthesis of the Recent Econometrics Literature." *Journal of Econometrics, 235*(2), 2218-2244.

Additional readings

- Dafoe Allan. 2018. "Nonparametric Identification of Causal Effects Under Temporal Dependence." *Sociological Methods & Research, 47*(2), 136-168.
- Elwert Felix and Fabian T. Pfeffer. 2022. "The Future Strikes Back: Using Future Treatments to Detect and Reduce Hidden Bias." *Sociological Methods & Research, 51*(3), 1014-1051.

Nov 11: Heterogeneous Treatment Effect

Required readings

- CML Chapter 14
- Brand Jennie et al. 2021. "Uncovering Sociological Effect Heterogeneity Using Tree-based Machine Learning." *Sociological Methodology, 51*(2), 189-223.
- Künzel Sören et al. 2019. "Metalearners for Estimating Heterogeneous Treatment Effects Using Machine Learning." *Proceedings of the National Academy of Sciences, 116*(10), 4156-4165.

Additional readings

- Daoud Adel and Fredrik D. Johansson. 2024. "The Impact of Austerity on Children: Uncovering Effect Heterogeneity by Political, Economic, and Family Factors in Low-and Middle-Income Countries." *Social Science Research, 118*, 102973.
- Wager Stefan and Susan Athey. 2018. "Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests." *Journal of the American Statistical Association, 113*(523), 1228-1242.

Nov 18: Feature Engineering Using Unstructured Data

Required readings

- CML Chapter 10
- Vaswani Ashish et al. 2017. "Attention is All You Need." *Advances in Neural In-*

formation Processing Systems, 30.

Additional readings

- Athey Susan et al. 2024. "Labor-LLM: Language-based Occupational Representations with Large Language Models." *arXiv preprint arXiv:2406.17972*.
- Li He et al. 2025. "Transformer-Based Spatial-Temporal Counterfactual Outcomes Estimation." *arXiv preprint arXiv:2506.21154*.

Nov 25: Causal Reasoning in Machine Learning

Required readings

- Feder Amir et al. 2022. "Causal Inference in Natural Language Processing: Estimation, Prediction, Interpretation and Beyond." *Transactions of the Association for Computational Linguistics, 10*, 1138-1158.
- Kosuke Imai and Kentaro Nakamura. 2024. "Causal Representation Learning with Generative Artificial Intelligence: Application to Texts as Treatments." *arXiv preprint arXiv:2410.00903*.

Additional readings

- Kim Junsol and Byungkyu Lee. 2023. "AI-augmented Surveys: Leveraging Large Language Models and Surveys for Opinion Prediction." *arXiv preprint arXiv:2305.09620*.