SOCIOL 212B: Quantitative Data Analysis. Winter 2025

Ian Lundberg, ianlundberg@ucla.edu, ianlundberg.org
Soonhong Cho, tnsehdtm@gmail.com, soonhong-cho.github.io

Course meeting
W 9–11:50am
Powell 320B

Office Hours
Th 2–3pm or by appointment:
calendly.com/ianlundberg/meeting
Haines 241C or Zoom: ucla.zoom.us/my/ianlundberg
Questions can also be posted on Piazza

Course description. This course is about answering social science questions using quantitative data. We will especially focus on how computational power is transforming the ways we can carry out quantitative research, covering both statistical and machine learning tools from the perspective of social science applications. The course especially emphasizes how to translate social science theories into quantities that can be estimated by algorithms designed for prediction. We will consider prediction in the service of both description and causal inference, building on ideas from SOCIOL 212A. The end product of the course is an extended abstract containing data analysis using the ideas from the course. For students continuing to 212C, the abstract can serve as the basis for the research project in that course. Students will leave the course prepared to connect social science theories to empirical evidence that can be produced by algorithms designed for prediction.

Learning goals. Students will learn to

- define a precise quantitative research question
- connect that question to predictions that can be made by statistical or machine learning algorithms
- make a principled argument for the choice of a particular learning approach

Who should take this course? The course is designed to support the development of quantitative social science research projects. The course is a good fit for PhD students in sociology, statistics, political science, economics, and other social sciences. PhD students from disciplines other than sociology should request a code from the instructor to enroll.

Prerequisite. Familiarity with basic probability and statistics (e.g., random variables, expectation, confidence intervals). Soc 212A is formally a prerequisite, but students who did not take Soc 212A are welcome to talk with me about whether Soc 212B would be a good fit for them.

Instructional format. Lecture with in-class exercises. Bring computers to class.

Course readings. Readings will be available online for free. See the course website for an updated schedule of readings and topics.

Many readings from books with free PDFs available online:

- Efron, B., & T Hastie. 2016. *Computer Age Statistical Inference: Algorithms, Evidence and Data Science.* Cambridge: Cambridge University Press.
- Hastie, T., R. Tibshirani, & J. Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* New York: Springer.
- Hernán, M.A., & J.M. Robins. 2024. *Causal Inference: What If?* Boca Raton: Chapman & Hall / CRC.

Statistical software. You can use any statistical software you prefer. I use R and will best be able to support you in R. In addition to R, we will attempt to provide Stata support where possible. Not all algorithms are available in Stata. If you are fluent in another software, you are welcome to use that. The focus of this course is on conceptual ideas, not a programming language.

Typesetting. While typesetting in LaTeX is a useful skill, it is not required. You may handwrite any assignment and upload a scanned copy. Whether you typeset or handwrite will not affect your grade.

Grading. Letter grade or Satisfactory / Unsatisfactory. Grades will be determined by:

1) Problem sets	50%
2) Peer reviewing on problem sets	10%
3) Class participation	10%
4) Extended abstract	20%
5) Feedback to two peers on their extended abstracts	10%

For details, see Assignments.

Late work. There will be a 10% per day penalty for late work, applied automatically in BruinLearn. In exceptional circumstances, talk to me for an exception.

Academic integrity. Each student in this course is expected to abide by the UCLA Academic Integrity policies. Any work submitted by a student in this course for academic credit must be the student's own work.

Statement on accessible education.[1] Your access in this course is important to me. If you are already registered with the Center for Accessible Education (CAE), please request your Letter of Accommodation in the Student Portal. If you are seeking registration with the CAE, please submit your request for accommodations via the CAE website. Students with disabilities requiring academic accommodations should submit their request for accommodations as soon as possible, as it may take up to two weeks to review the request. For more information, please visit the CAE website (www.cae.ucla.edu), visit the CAE at A255 Murphy Hall, or contact them by phone at (310) 825-1501.

Assignments

Assignments are due on the course website in PDF form. If your code is not embedded within your PDF, you must also upload your code file. All assignments are due on Tuesdays at 5pm.

1) Problem sets.
Problem sets are intentionally brief and focus on key concepts. The answer key will be posted each week after the deadline, and I encourage you to review what you might have missed.

Part of each problem set will be open-ended: how might the ideas in this problem set be useful for your research proposal?

There will also be space to write questions you are having about your project, to get feedback from peers. Problem sets will be written so that it is in many cases possible to substitute your own research data for the data given in the problem set.

2) Peer reviewing on problem sets. A principle on which this course is built is that you will offer feedback to one another. Following this principle, after each problem set is submitted you will be assigned to review a peer's work in Canvas.

3) Class participation. This class will not involve a lot of lecture. Most class sessions will involve some lecture followed by hands-on data analysis carried out in groups. If you participate, you can expect to receive full participation credit. If you are absent from class, talk to me and we can find a way for you to carry out the activity for the day independently on your own time for credit.

4) Extended abstract of a research paper.

The final project of the course is an extended abstract of a research paper. This might build on the proposal you wrote in 212A, or you are free to pivot to a new topic.

[1] This statement is based on guidelines from the Center for Accessible Education.

By an "extended abstract," we mean a product of 4–6 pages or more that motivates your research question and presents some preliminary analyses of data. Figures and tables can be embedded in the text. The goal is to be helpful to you, so if you are moving ahead and wish to submit a longer draft that is also acceptable.

- 10 points. Motivate the question. Motivate why a general reader (outside of your own subfield) should care about the question.
- 10 points. Define the estimand. Tell us what you seek to estimate. This should involve
  - unit-specific quantity: some factual or counterfactual outcome defined for every unit
  - target population: a well-defined set of units over which you are aggregating
  - summary statistic: how you aggregate over the units, such as by a mean or median

  The estimand should not involve any model parameters (e.g., coefficients).
- 10 points. Data. Tell us what data you are using and how the data were collected. Knowing how the data came to be is critical for the next point.
- 10 points. Assumptions. What do you assume about the process that generated your data, in order to draw conclusions? Assumptions may involve tools such as Directed Acyclic Graphs or statements of conditional independence.
  - Note: It is ok to have heroic assumptions that are very doubtful. In 212C, you can refine the analysis to address some concerns about doubtful assumptions.
- 10 points. Estimation. What statistical learning procedures do you use to estimate your estimand from the data?
- 10 points. Results. Explain your results in terms accessible to a reader of the New York Times. You should report estimates that readers can understand even if they do not understand the parameters in your model.

5) Feedback on extended abstracts.
You will provide written feedback to two peers on their extended abstracts. The written feedback to each author should be between half a page and one page. You should start by summarizing the author's work. Then, consider the evaluation criteria above. Offer suggestions for improvement, and also emphasize what you see as the strengths in the proposal. How could the author use ideas from the course to make the paper even better?

Schedule of Topics (tentative)

1. Jan 8. Asking research questions without $\hat{\beta}$
   - Part 1: Asking research questions
     - Lundberg, Johnson, & Stewart 2021
   - Part 2: A $\hat{Y}$ approach to regression: Approximating a conditional expectation function
     - Berk 2020 Ch 1 p. 1–5, stopping at paragraph ending "...is nonlinear." Then p. 14–17 "Model misspecification..." through "... will always be in play."
     - Generalization to categorical outcomes
   - Part 3: Organizing your workflow
     - Reproducibility guidelines for the American Political Science Review. See also Stodden 2015.
     - For R users: Wickham 2023 on Quarto for reproducible workflows

2. Jan 15. Algorithms for prediction. (Note: After the morning class, students are encouraged to attend Brandon Stewart's afternoon tutorial in CCPR.)
    - Penalized regression (Efron & Hastie 7.3)
    - Multilevel models
3. Jan 22. Algorithms for prediction
    - Trees, forests, and boosting (Efron & Hastie 8.4, 17)
4. Jan 29. Data-driven selection of a prediction function
    - discussion paper to read before class: Salganik et al. 2020
    - topics for class
        - Define the task
        - Mimic the task
        - Cross validation
        - Ensembles: Super Learner
    - References for after class:
        - Hastie, Tibshirani, & Friedman Ch 7 on model selection
        - Efron & Hastie 12.2 on cross-validation
        - Hoffman tutorial on Super Learner. Historical reference: Van der Laan et al. 2007
5. Feb 5. Panel data (Note: Morning class aligned with Yiqing Xu's afternoon tutorial in CCPR)
    - Forecasting, interrupted time series, lagged dependent variables, difference in difference, fixed effects
        - Discussion paper to read before class: Athey et al. 2021 sections 1–3
6. Feb 12. Inference: Bootstrap and beyond
    - discussion paper to read before class: Efron & Tibshirani p. 1–6.
    - topics for class
        - the estimator function
        - nonparametric bootstrap standard error
        - quantile confidence intervals
        - normal approximation confidence intervals
        - invariance of CIs to monotone transformations
        - bootstrap analogues for complex survey samples. Example: Replicate weights for the Current Population Survey: IPUMS documentation
    - For future reference: Horowitz 2019, Efron & Hastie Ch 10–11.
7. Feb 19. Nonparametric identification for causal and population inference
    - Discussion paper to read before class: Hernán 2016
    - The target trial
        - Reference: Hernán & Robins 2016
    - The adjustment set via DAGs

  - Reference: Greenland et al. 1999

8. Feb 26. Prediction for causal and population inference
   - The parametric g-formula
     - Reference: Hernán & Robins Ch 13.
     - Reference: Hoffman tutorial on G-computation.
9. Mar 5. Weighting for causal and population inference
   - Discussion paper to read before class: Groves 2011
   - topics
     - weighting in design-based settings
       * population inference in probability samples
       * causal inference in randomized experiments with unequal assignment probabilities
     - weighting in non-design-based settings
       * population inference in non-probability samples
       * causal inference in observational studies
   - References: Hartman et al. 2015, Egami & Hartman 2023
10. Mar 12. Doubly robust estimation
    - Augmented Inverse Probability Weighting (AIPW)
      - Glynn & Quinn 2010
    - Targeted Mimimum Loss-Based Estimation (TMLE)
      - Katherine Hoffman's tutorial
11. Course appendix: Topics to be discussed briefly during the course to the degree they are relevant to student projects and which will be supported by an online appendix to the course website.
    - Missing data
    - Factor analysis and scale construction
    - Mediation and structural equation models