SOSC 13200: Social Science Inquiry II. Section 3.

University of Chicago, Winter 2026.

Location: Cobb Hall 110
Course time: Mon/Wed 15:00–16:20
Timeframe: 01/05/2026–03/14/2026
Course github: https://github.com/UChicago-pol-methods/SOSC13200-W26

Instructor: Molly Offer-Westort; mollyow@uchicago.edu
Office hours: Mon 13:20–14:40 and Tue 14:00–15:20; book at https://calendar.app.google/wmSJFvfyxnDHFK5C9
Office: Pick Hall 328

Course description and objectives. In this course, you will learn to approach data thinking like a social scientist. That means thinking about where your data comes from and how it is measured, and assessing and describing relationships among variables. Throughout the course, you will be exposed to statistical tests and other quantitative methods used in social science research. You will also work on statistical software programming and data visualization.

This session of the course focuses on the theme of the "credibility revolution" in the social sciences: a move toward empirical studies built with quality data, rigorous research design, and appropriate strategies for analysis. The impact of this work was recognized by the 2021 Nobel prize in economics, awarded to David Card, Joshua Angrist, and Guido Imbens. We will also consider the contributions of carefully designed field experiments, and the work of 2019 economics Nobel prize winners, Abhijit Banerjee, Esther Duflo, and Michael Kremer.

Textbooks.

- Probability & Statistics:
  - Wackerly, Dennis, William Mendenhall, and Richard L. Scheaffer (2014). *Mathematical Statistics with Applications*. Cengage Learning.
  - For reference, not assigned
    https://www.stat.berkeley.edu/~stark/SticiGui/index.htm
- Coding in R:
  - Verzani, John. *simpleR – Using R for Introductory Statistics*. Online book: https://cran.r-project.org/doc/contrib/Verzani-SimpleR.pdf
  - Wickham, Hadley, Danielle Navarro, and Thomas Lin Pedersen (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag. Online book: https://ggplot2-book.org/

Mathematical proficiency. It is not assumed that students have had prior exposure to probability and statistics, and this course will not use calculus or linear algebra. Familiarity with these topics may be useful for some of the extra reference readings, but this is not required.

Statistical software. The statistical software used in this iteration of this class will be R, which we will use along with the RStudio interface. R is a statistical language and environment for data manipulation, calculation, and visualization; the software for R and RStudio are free to download online. R can be a very flexible and powerful tool, and it is widely used in statistical and social science research, as well as in some industry settings. Instructors in other sessions of this course may use Stata, which is also widely used and relevant in particular in the fields of economics and public policy. In general, basic familiarity with R, Stata, and python can be useful for research in the social sciences. You are not assumed to have prior experience with R for this class.

Computing. You will need access to a computer throughout the quarter with R and RStudio installed. If you do not have access to a private computer to install this software, University library computers are equipped with RStudio, and you should be able to save your assignments on Box while using these computers. If you have any issues with access to a computer, please reach out to me and we will find a solution.

Assignments. Homework consists of weekly assignments, submitted in R with compiled reports. Homeworks are graded on a 0–3 scale: 0 (not submitted), 1 (check–), 2 (check), 3 (check+). Homeworks are worth 24 points total. Participation is worth 13 points. The in-class midterm is worth 18 points, the in-person final is worth 30 points, and the take-home final is worth 15 points.

Component	Points	Due date
Participation	13	–
Homework 1	3	Fri, January 9 (11:59pm)
Homework 2	3	Fri, January 16 (11:59pm)
Homework 3	3	Fri, January 23 (11:59pm)
Homework 4	3	Fri, January 30 (11:59pm)
Homework 5	3	Fri, February 6 (11:59pm)
Homework 6	3	Fri, February 13 (11:59pm)
Homework 7	3	Fri, February 20 (11:59pm)
Homework 8	3	Fri, February 27 (11:59pm)
Midterm (in class)	18	Wed, February 4
Final (in person)	30	Exam week (TBA)
Final (take home)	15	Fri, March 6 (11:59pm)

Accommodations. Please reach out to me directly if you would like to request accommodations for the course to better facilitate your learning. Student Disability Services (https://disabilities.uchicago.edu/) is also available to provide you resources and support, and may provide approval for specific academic accommodations. Informing me in a timely manner will help me to ensure accommodations are met and I am able to implement an appropriate assessment of your learning.

Academic honesty. All students are expected to be familiar with and follow the University's academic honesty policies (https://studentmanual.uchicago.edu/academic-policies/academic-honesty-plagiarism/). No citation is required for code generated by AI, but you are responsible for understanding the code and ensuring it works correctly. You will be asked to provide input on our class AI policy, and then will be expected to follow it: https://github.com/UChicago-pol-methods/SOSC13200-W26/blob/main/ai-policy.md.

Course outline.

Week 1: Course introduction.

Class 1.1 (Monday 1/5).

- Overview of course objective: how to formalize and test a theory with data and statistical tools.
- Some motivating examples.
- Introduction to coding: what the software will be, how you will download it, basics of setting yourself up with good directory hygiene.

Readings.

- Tabarrok, A. (2021). A nobel prize for the credibility revolution. Marginal Revolution

Learning R.

- On your own, set up the swirl course on "R Programming with Email Notification" https://github.com/UChicago-pol-methods/R_Programming_E
- Run the following code: `install.packages(c("swirl")); library(swirl); install_course_github("R_Programming_E");swirl()Completesections1(BasicBuildingBlocks)and2(WorkspaceandFiles).`

Class 1.2 (Wednesday 1/7).

- What kinds of questions can we want to ask in social science research? How do we start formalizing a question that is tractable and falsifiable? What population is the question in reference to?

Readings.

- Holland, P. W. (1986). Statistics and causal inference. *Journal of the American*

statistical Association 81(396), 945–960 Skip sections 5, 8.2.
- King, G., R. O. Keohane, and S. Verba (1994). *Designing social inquiry: scientific inference in qualitative research*. Princeton University Press Chapter 1.

Homework 1 due Friday 1/9 at 11:59pm: Install and set up statistical software and class working directory on your computer. Submit a compiled R document in pdf with your name and date, showing the file working directory.

Week 2: Summarizing data numerically and visually.

Class 2.1 (Monday 1/12).

- Working with the Card and Krueger dataset, summarizing data numerically. Summaries of univariate data (sample mean, sample median, quantiles, sample variance, sample standard deviation).

Readings.

- Card, D. and A. B. Krueger (1994). Minimum wages and employment: A case study of the fast food industry in new jersey and pennsylvania. *American Economic Review 84*(4), 772–793
- We will come back to this paper again later this quarter to discuss the content and results; for this first pass, we will just focus on the data.

Learning R.

- In swirl, complete sections 3 (Sequences of Numbers), 4 (Vectors), 5 (Missing Values).

Statistical reference (not required).

- Wackerly et al. Chapter 1: What is Statistics?

Class 2.2 (Wednesday 1/14).

- Working with the Card and Krueger dataset, summarizing data visually.
- Graphical representation of different types of univariate data: histograms, density plots, boxplot, bar charts.
- Challenges with measurement.

Readings.

- Adcock, R. and D. Collier (2001). Measurement validity: A shared standard for qualitative and quantitative research. *American Political Science Review 95*(3), 529–546

Learning R.

- ggplot2 book Chapter 2: First steps.

Homework 2 due Friday 1/16 at 11:59pm: Submit a compiled R document in pdf with your name and date. Load in Card and Krueger dataset, and produce specified summary statistics and a plot.

Week 3: Probability as a model of the world.

No class Monday 1/19 (MLK Day).

Class 3.1 (Wednesday 1/21).

- How do we think about what it means for something to be random? Why is this mathematical abstraction useful for social science research?
- Overview of discrete probability.
- Conditional probability, Bayes Rule.
- Why understanding conditional probability is essential to understanding social phenomena, and why relying on your intuition can be wrong.

Readings.

- Stark, P. B. and D. A. Freedman (2003). What is the chance of an earthquake? Earthquake science and seismic risk reduction
- Gelman, A. (2007). The prosecutor's fallacy. Statistical Modeling, Causal Inference, and Social Science https://statmodeling.stat.columbia.edu/2007/05/18/the_prosecutors/
- Hill, R. (2004). Multiple sudden infant deaths–coincidence or beyond coincidence? *Paediatric and Perinatal Epidemiology* https://onlinelibrary.wiley.com/doi/10.1111/j.1365-3016.2004.00560.x

Learning R.

- In swirl, complete sections 6 (Subsetting Vectors), 7 (Matrices and Data Frames), 8 (Logic), and 9 (Functions).

Statistical reference (not required).

- Wackerly et al. Chapter 2: Probability; 2.1–2.4, 2.7, 2.10–2.13.

Homework 3 due Friday 1/23 at 11:59pm: Exercises in defining sample space, mapping events to probabilities. Using `sample()` to simulate random processes in R.

Week 4: Joint relationships.

Class 4.1 (Monday 1/26).

- Using Angrist & Krueger data to discuss identifying patterns in data, covariance, correlation.
- Education and earnings are correlated; preview correlation vs. causation with respect to the effect of education on earnings.

Readings.

- Angrist, J. D. and A. B. Keueger (1991). Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics 106*(4), 979–1014

Statistical reference (not required).

- Wackerly et al. Chapter 3: Discrete Random Variables and Their Probability Distributions; 3.1–3.3.
- Wackerly et al. Chapter 5: Multivariate probability distributions; 5.1–5.4, 5.7 (ignore continuous distributions).

Class 4.2 (Wednesday 1/28).

- Using Miguel and Kremer data, conditional means, difference in means.
- Data visualization: scatterplots, plotting conditional expectation function, faceting plots by category.

Readings.

- Ba, B., D. Knox, J. Mummolo, and R. Rivera (2021). The role of officer race and gender in police-civilian interactions in chicago. *Science 371*(6530), 696–702
- Miguel, E. and M. Kremer (2004). Worms: identifying impacts on education and health in the presence of treatment externalities. *Econometrica 72*(1), 159–217
- Additional content on the "Worm Wars" regarding replication issues with Miguel & Kremer; you do not need to read or listen to all of this:
  - Belluz, J. (2015). The "worm wars" explained. Vox
  - https://www.vox.com/2015/7/24/9031909/worm-wars-explained
  - Gordon, A. and M. Hobbes (2021). The "worm wars". Maintenance Phase podcast episode
  - https://open.spotify.com/episode/7ujpdCPK5hKPgUYDGV87aV?si=nmXeJSqnR0aRlF5S3

Learning R.

- Verzani, Chapter 4: Bivariate Data, pp. 32–40.

Statistical reference (not required).

- Huntington-Klein, Nick (2021). *The Effect*. Chapter 5.

Homework 4 due Friday 1/30 at 11:59pm: Data exploration, visualizing joint relationships.

Week 5: Uncertainty and inference.

Class 5.1 (Monday 2/2).

- Using Butler and Broockman data, randomization inference, introduction to hypothesis testing, confidence intervals, p-values and statistical significance.
- Sampling from a larger population; central limit theorem, one- and two-sample t-tests.

Readings.

- Butler, D. M. and D. E. Broockman (2011). Do politicians racially discriminate against constituents? a field experiment on state legislators. *American Journal of Political Science 55*(3), 463–477
- Gerber, A. S. and D. P. Green (2012). *Field experiments: Design, analysis, and interpretation*. WW Norton Chapter 3: Sampling Distributions, Statistical Inference, and Hypothesis Testing, Sections 3.1–3.5, pp. 51–70.
- Pager, D. (2003). The mark of a criminal record. *American journal of sociology 108*(5), 937–975

Statistical reference (not required).

- Athey, S., & Imbens, G. (2017). "The Econometrics of Randomized Experiments." In A. V. Banerjee and E. Duflo, editors, *Handbook of Field Experiments*, 1, pp. 73–140. North-Holland.
- Wackerly et al. Chapter 7: Sampling Distributions and the Central Limit Theorem; 7.1–7.3.
- Wackerly et al. Chapter 8: Estimation; 8.5–8.6.

Class 5.2 (Wednesday 2/4).

- In-class midterm.

Readings.

- No readings.

Homework 5 due Friday 2/6 at 11:59pm: Re-create the Pager data based on the description in the text. Formalize in words what hypotheses are being tested. Conduct the statistical analyses described. What is the reference population? Are these results generalizable? Why or why not?

Homework 6 due Friday 2/13 at 11:59pm: Replication of Butler and Broockman. What statistical tests are being used? How do you interpret the regression coefficients?

Week 6: Bivariate regression: best linear predictor.

Class 6.1 (Monday 2/9).

- With Butler and Broockman data, regression as the best linear predictor, least squares visually demonstrated.
- Interpreting regression coefficients.

Readings.

- Bueno de Mesquita, E. and A. Fowler (2021). *Thinking clearly with data: A guide to quantitative reasoning and analysis*. Princeton University Press Chapter 5: Regression for Describing and Forecasting.

Class 6.2 (Wednesday 2/11).

- Inference for linear regression.
- Linear regression with experiments and dummy variables; relationship to t-tests.

Readings.

- None; spend this time working on researching data for your projects.

Statistical reference (not required).

- Huntington-Klein, Nick (2021). *The Effect*. Chapter 13.1; 13.2 through 13.2.5 inclusive only.
- Wackerly et al. Chapter 11: Linear Models and Estimation by Least Squares; 11.1–11.4.

Week 7: Multivariate regression: model building.

Class 7.1 (Monday 2/16).

- Regression as linearization of conditional expectation function.
- With Banerjee et al. data, interpreting regression coefficients (and standard errors) in multivariate regression.

Readings.

- Banerjee, A., E. Duflo, R. Glennerster, and C. Kinnan (2015). The miracle of microfinance? Evidence from a randomized evaluation. *American Economic Journal: Applied Economics 7*(1), 22–53

Statistical reference (not required).

- Wackerly et al. Chapter 11: Linear Models and Estimation by Least Squares; 11.11–11.12.

Class 7.2 (Wednesday 2/18).

- When can we apply causal interpretations to regression coefficients from observational data?

Readings.

- Card, D. and A. B. Krueger (1994). Minimum wages and employment: A case study of the fast food industry in new jersey and pennsylvania. *American Economic Review 84*(4), 772–793
- Now we will come back to this paper, and look at their approach to analysis.
- Gerber, A. S., D. P. Green, and E. H. Kaplan (2014). The illusion of learning from observational research. In *Field experiments and their critics*, pp. 9–32. Yale University Press

Homework 7 due Friday 2/20 at 11:59pm: Case File 1 (Data + design audit). Short memo + reproducible R output on data provenance, unit of analysis, population, missingness, and core descriptive tables/plots.

Week 8: Regression: inference, challenges to inference, and external validity.

Class 8.1 (Monday 2/23).

- Pitfalls of p-values and hypothesis testing; what is the alternative?

Readings.

- Amrhein, V., S. Greenland, and B. McShane (2019). Retire statistical significance. *Nature 567*, 305–307
- Gerber, A. S. and N. Malhotra (2008). Do statistical reporting standards affect what is published? publication bias in two leading political science journals. *Quarterly Journal of Political Science 3*(4), 313–351

Class 8.2 (Wednesday 2/25).

- Statistical power, effect sizes, sample size.

Readings.

- Coppock, A. 10 things to know about statistical power. EGAP
- Lakens, D. (2022). Sample size justification. *Collabra: Psychology 8*(1), 33267

Homework 8 due Friday 2/27 at 11:59pm: Case File 2 (Primary analysis + interpretation). Short memo + reproducible R output with bivariate and multivariate regression, an alternate specification, and interpretation with visuals.

Week 9: Moving beyond multiple linear regression: other approaches.

Class 9.1 (Monday 3/2).

- Different types of research questions: prediction, heterogeneity.
- Some alternative tools, and how machine learning is used in the social sciences.

Readings.

- Lundberg, I., R. Johnson, and B. M. Stewart (2020). What is your estimand? defining the target quantity connects statistical evidence to theory. *American Sociological Review 85*(1)
- Athey, S. and G. W. Imbens (2019). Machine learning methods that economists should know about. *Annual Review of Economics 11*, 685–725

Class 9.2 (Wednesday 3/4).

- Declaring hypotheses in advance / pre-analysis planning.

Readings.

- Chen, L. and C. Grady. 10 things to know about pre-analysis plans. EGAP
- Monogan, J. (2015). Research preregistration in political science: The case, counterarguments, and a response to critiques. *PS: Political Science & Politics 48*(2), 425–429

Take-home final (project components, exam-structured). The take-home component of the final assessment is a take-home exam that uses either the same capstone dataset with new prompts and a required extension, or a fresh but similar dataset (TBA). You will submit a PDF report (R/Rmd to PDF) plus source files; the prompt will include data/measurement, descriptives/visualization, modeling/inference, credibility reflection, and a short bridge-to-next-quarter experiment component.