SOSC 13200: Social Science Inquiry II

Section 2

University of Chicago
Winter Quarter 2023
(online version of this syllabus here)

Course details:
Location: Cobb Hall 116
Course time: Tue Thu : 11:00 AM-12:20 PM
Course github: https://github.com/UChicago-pol-methods/SOSC13200-W23

Instructor:
Dr. Molly Offer-Westort, mollyow@uchicago.edu
Office hours: https://calendly.com/mollyow/office-hours
Office: Pick Hall 526 (please email to schedule in-person meetings)

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's 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:**
  - Huntington-Klein, Nick. (2021). *The effect: An introduction to research design and causality*. Chapman and Hall/CRC. Online book: https://theeffectbook.net/
  - [For reference, not assigned] Wackerly, Dennis, William Mendenhall, and Richard L. Scheaffer (2014). *Mathematical statistics with applications*. Cengage Learning.
- **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 both R and Stata is useful for research in the social sciences. However, in this version of the class, we will only use R. 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. During the last three weeks of the course, you will work on a final project, in which you will re-analyze a dataset of your choice, from published empirical academic work in the social sciences. In the final report, you will describe the data, hypotheses, what statistical tests are being implemented, and give your interpretation of results. Grade composition is:

- Homework: 60%
- Final project: 30%
- Participation: 10%

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. If you or your household is affected by the ongoing pandemic in a way that affects your ability to participate in or attend class, please reach out to me as well. 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.

Week 1: Course introduction

Class 1.1 Tuesday 1/3

- 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:

- A. Tabarrok, A Nobel Prize for the Credibility Revolution. https://marginalrevolution.com/marginalrevolution/2021/10/the-credibility-revolution-1.html

Class 1.2 Thursday 1/5

- What kinds of questions can we/might 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, Paul W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*.
- King, Keohane, and Verba (1994). Designing Social Inquiry: *Scientific Inference in Qualitative Research.* Princeton University Press. Chapter 1.

Learning R:

- Read explainer on writing code and setting your working directory
- On your own, do the Rstudio primers 1.1 and 1.2 https://rstudio.cloud/learn/primers

Homework 1 due Friday 1/6 at 5pm: 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 Tuesday 1/10

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

Readings:

- Huntington-Klein, Nick (2021). The Effect. Chapter 3.

- Card, David and Krueger, Alan B. (1994). Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania. *American Economic Review.*
    - *We'll come back to this paper again later this quarter to discuss the content and results; for this first pass, we'll just focus on the data.*

Learning R:
- Verzani, 1. Introduction, 2. Data, 3. Univariate Data pp. 1-19

Statistical Reference (not required):
- Wackerly et al. Chapter 1: What is Statistics?

Class 2.2 Thursday 1/12
- What goes into a data set?: Units of observation and variables
- Working with the Card and Krueger dataset, summarizing data visually; recreate Figure 1. Graphical representation of different types of univariate data: histograms, density plots, boxplot, bar charts.
- Challenges with measurement.

Readings:
- Adcock, Robert, and David Collier (2001). "Measurement validity: A shared standard for qualitative and quantitative research." *The American Political Science Review*.
- Deaton, A. (2016). Measuring and understanding behavior, welfare, and poverty. *American Economic Review*, 106(6), 1221-43.

Learning R:
- Ggplot2 book Chapter 2: First steps

Homework 2 due Friday 1/13 at 5pm: 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

Class 3.1 Tuesday 1/17
- 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

Readings:
- Mlodinow, Leonard (2008). *The Drunkard's Walk*. Random House Digital Inc. Chapter 2.
- Stark, P. B., Freedman, D. A., Mulargia, F., & Geller, R. J. (2003). What is the chance of an earthquake? *Earthquake science and seismic risk reduction.*

Statistical Reference (not required):

- Wackerly et al., Chapter 2: Probability; 2.1-2.4

Class 3.2 Thursday 1/19

- Conditional probability, Bayes Rule
- Why understanding conditional probability is essential to understanding social phenomena--and why relying on your intuition can be wrong

Readings:

- Buchanan, M., (2007). The Prosecutor's Fallacy, reproduced on Andrew Gelman's blog.
- Hill, R. (2004). Multiple sudden infant deaths–coincidence or beyond coincidence?. *Paediatric and perinatal epidemiology*.

Statistical Reference (not required):

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

Homework 3 due Monday 1/23 at 5pm: Exercises in defining sample space, mapping events to probabilities. Using sample() to simulate random processes in R.

Week 4: Joint relationships

Class 4.1 Tuesday 1/24

- 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:

- Huntington-Klein, Nick (2021). The Effect. Chapter 4.
- Angrist, Joshua D., and Alan B. Krueger. "Does compulsory school attendance affect schooling and earnings?" *The Quarterly Journal of Economics* 106.4 (1991): 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 Thursday 1/26

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

Readings:

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

- Miguel, Edward, & Kremer, Michael. (2004). Worms: identifying impacts on education and health in the presence of treatment externalities. *Econometrica*, 72(1), 159-217.

Learning R:
- Verzani, 4: Bivariate Data, pp. 32-40

Homework 4 due Monday 1/30 at 5pm: Data exploration, visualizing joint relationships.

Week 5: Uncertainty and inference

Class 5.1 Tuesday 1/31

- Using Butler and Broockman data, randomization inference, introduction to hypothesis testing, confidence intervals, p-values and statistical significance.

Readings:
- Butler, Daniel M., & Broockman, David E. (2011). Do politicians racially discriminate against constituents? A field experiment on state legislators. *American Journal of Political Science*.
- Gerber, A. S., & Green, D. P. (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.

Statistical reference (not required):
- S. Athey and G. Imbens. (2017). Chapter 3 - The Econometrics of Randomized experiments. In A. V. Banerjee and E. Duflo, editors, *Handbook of Field Experiment*s, 1, pp. 73-140. North-Holland.

Class 5.2 Thursday 2/2

- Sampling from a larger population; central limit theorem, one- and two-sample t-tests.

Readings:
- Pager, Devah. The mark of a criminal record. *American Journal of Sociology* 108, no. 5 (2003): 937-975.

Statistical reference (not required):
- 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

Homework 5 due Monday 2/6 at 5pm: 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/why not?

Week 6: Bivariate regression: best linear predictor

Class 6.1 Tuesday 2/7

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

Readings:

- Bueno de Mesquita, E., & Fowler, A. (2021). *Thinking clearly with data: A guide to quantitative reasoning and analysis.* Princeton University Press. Chapter 5: Regression for Describing and Forecasting.
- Huntington-Klein, Nick (2021). The Effect. Chapter 13.1 only.

Class 6.2 Thursday 2/9

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

Readings:

- Huntington-Klein, Nick (2021). The Effect. Chapter 13.2 through 13.2.5 inclusive only.

Statistical reference (not required):

- Wackerly et al. Chapter 11: Linear Models and Estimation by Least Squares; 11.1-11.4

Homework 6 due Monday 2/13 at 5pm: Replication of Butler and Broockman. What statistical tests are being used? How do you interpret the regression coefficients?
Get approval for your final project dataset.

Week 7: Multivariate regression: model building

Class 7.1 Tuesday 2/14

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

Readings:

- Banerjee, A., Duflo, E., Glennerster, R., & Kinnan, C. (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 Thursday 2/16

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

Readings:

- Card, D. and Krueger, A. B. (1994). Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania. *American Economic Review.*
    - *Now we'll come back to this paper, and look at their approach to analysis.*
- Gerber, A. S., Green, D. P., & Kaplan, E. H. (2014). 1. The Illusion of Learning from Observational Research. In *Field experiments and their critics* (pp. 9-32). Yale University Press.

Homework 7 due Monday 2/20 at 5pm: Final project part 1: data set selection and descriptive analysis.

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

Class 8.1 Tuesday 2/21

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

Readings:

- Amrhein et al. (2019). Retire Statistical Significance. *Nature*.
- Gerber and Malhotra (2008). Do Statistical Reporting Standards Affect What is Published? Publication Bias in Two Leading Political Science Journals. *Quarterly Journal of Political Science*.
- Play around with "Hack Your Way To Scientific Glory," https://projects.fivethirtyeight.com/p-hacking/

Class 8.2 Thursday 2/23

- How our research design and data informs our analyses and generalizability
- In-class time to discuss research projects

Readings:

- Bohannon (2015). I Fooled Millions into Thinking Chocolate Helps Weight Loss. Here's How. *Gizmodo*.

Homework 8 due Monday 2/27 at 5pm: Final project part 2: multiple linear regression and data visualizations.

Week 9: Moving beyond multiple linear regression: other approaches

Class 9.1 Tuesday 2/28

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

Readings:

- Lundberg et al. (2020). What Is Your Estimand? Defining the Target Quantity Connects Statistical Evidence to Theory. *American Sociological Review.*
- Athey, S., & Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*

Class 9.2 Thursday 3/2

- Course wrap-up and review.

No readings

Homework 9 due Friday 3/3 at 5pm: Final project part 3: finalize report.