SOSC 13200: Social Science Inquiry II

DETAILS
Winter 2021, Social Sciences Research Building 106
TuTh 12:30PM – 1:50PM (section 4) / 2:00 PM – 3:20 PM (section 6)
Course website on Canvas

INSTRUCTOR
Dr. William Schultz, wschultz@uchicago.edu
Office hours: Wednesday (1-3PM), contact me for other appointments
Office: Remote (Zoom links in Canvas announcements)

COURSE DESCRIPTION
In this quarter, 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 learn about statistical programming and data visualization.

This quarter in the sequence focuses on the theme of the "credibility revolution" in the social sciences: a move toward empirical studies built with quality data and rigorous research design. 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.

COURSE MATERIALS
As in the Fall Quarter, some assigned readings will be posted to Canvas. Others are linked in this syllabus. Finally, some assigned readings will come from textbooks that are available online:

- **Probability & Statistics:**
    - [IMS in the schedule below] Çetinkaya-Rundel, Mine, and Johanna Hardin (2021). *Introduction to Modern Statistics, First Edition.* OpenIntro. Free online book.
    - [For reference if you'd like more technical detail; not assigned] Wackerly, Dennis, William Mendenhall, and Richard L. Scheaffer (2014). *Mathematical statistics with applications*. Cengage Learning.
    - [For reference if you'd like an alternative approach to this material; not assigned] Gujarati, Damodar, and Dawn Porter. *Basic Econometrics.* McGraw-Hill.
- **Coding in R:**
    - Verzani, John. *simpleR – Using R for Introductory Statistics.* Free online book.
    - Wickham, Hadley, Danielle Navarro, and Thomas Lin Pedersen (2016). *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag. Free online book.

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 is necessary for a more advanced understanding of statistics (e.g.: the Wackerly et al. reference text suggested above), but it is not necessary for this class.

STATISTICAL SOFTWARE

We will use a statistical language called R for this class, in conjunction with the RStudio IDE (the program we'll use to run R code). R is a statistical language and environment for data manipulation, calculation, and visualization. Both R and RStudio are *free to download online*. R is a flexible and powerful tool. It is widely used across the social and natural sciences, as well as in many industry settings. Some instructors in other sections this quarter may use Stata, which is also widely used in social science, particularly in the fields of economics and public policy. In general, basic familiarity with both R and Stata is useful for social research. However, we will only use R.

EARNING YOUR GRADE

Homework consists of weekly assignments, compiled into a PDF using R Markdown (covered in class). R Markdown is a document formatting language you can use within R to create documents that weave together text, figures, and data analysis all at once. You can also use it create slides, like mine!

I've included brief descriptions of your homework assignments in the schedule below and will provide more detailed descriptions on Canvas. Homework is due at 11:59PM each Sunday (at the end of each week). During the last three weeks of the quarter, you will work towards a cumulative final project instead of submitting separate assignments. In this project you will re-analyze a dataset of your choice from published 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.

Your overall grade will be out of 100 points:
- **Participation (attendance and in-class participation)**: 10%
- **Final project**: 30%
- **Weekly homework**: 60%

MISSED/LATE ASSIGNMENTS

I accept missed assignments without penalty for documented medical reasons, family crises, call to active military duty or jury duty, religious holy days, and official University activities. I also give deference to parents with dependent children who are sick. Otherwise, I reserve the right to deduct 10% of an assignment's full worth for each day it is late.

EMAIL POLICY

Please include your first and last name and course information in the subject line of your email. When I receive your email, I will make effort to respond in a timely manner, usually within 48 hours. You may receive a reply sooner than that, but you should not expect an immediate response. Please treat all email correspondences with your instructor as you would treat any other professional exchange. I expect emails to be respectful and polite, to use correct grammar and complete sentences.

Schedule

I. Course Introduction

Homework 1 (due Sunday Jan 16th): *Install and set up R and R Studio on your computer. Write an R script that sets a working directory and loads the HW1 example dataset you can find on Canvas.*

Jan 11   Introduction [REMOTE]
- Tabarrok (2021). A Nobel Prize for the Credibility Revolution. *Marginal Revolution*.
- [Online tutorials, requires making a free account] RStudio primers 1.1 and 1.2 (Visualization basics and Programming basics), found here

Jan 13   Using data to learn about social phenomena [REMOTE]
- Holland, Paul W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*. [Read: Sections 1-4; Section 7]
- King, Keohane, and Verba (1994). Designing Social Inquiry: *Scientific Inference in Qualitative Research.* Princeton University Press. Chapter 1. [Skim pages 1-7, pages Read 7-28]

II. Summarizing data numerically and visually

Homework 2 (due Sunday Jan 23rd): *Submit a compiled R markdown pdf with your name and date. Load in the Lyall et al. dataset and produce specified summary statistics and a plot.*

Jan 18   Numerical summaries of data [REMOTE]
- Card, David and Alan B. Kreuger. (2004). "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania," American Economic Review, 84(4): 772-93.
- Verzani, 1. Introduction, 2. Data, 3. Univariate Data. pp. 1-19

Jan 20   Visual summaries of data [REMOTE]
- Adcock, Robert, and David Collier (2001). "Measurement validity: A shared standard for qualitative and quantitative research." *The American Political Science Review*. [529-536]
- Deaton, A. (2016). Measuring and understanding behavior, welfare, and poverty. *American Economic Review*. [Read: 1221-1230, 1238-1239]
- Wickham et al. (2016). Chapter 2: First steps

III. Probability as a model of the world

Homework 3 (due Sunday Jan 30th): *Defining a sample space and simulating random processes in R.*

Jan 25   Randomness and discrete probability
- 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.*

Jan 27   Conditional probability
- 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*.

IV. Joint relationships

Homework 4 (due Sunday Feb 6th): *Exploratory data analysis and visualizing joint relationships*

Feb 1   Covariation
- 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.

Feb 3   Change in the conditional mean
- Eggers, Andrew C. and Jens Hainmueller. (2009). "MPs for Sale? Returns to Office in Postwar British Politics," *American Political Science Review,* 103(4): 513-33.

V. Uncertainty and inference

Homework 5 (due Sunday Feb 13th): *Explain the hypotheses Pager is testing. Conduct the statistical analyses described. What is the reference population? Are these results generalizable? Why/why not?*

Feb 8   Intro to hypothesis testing, 1
- IMS Chapters 12 and 13, available here. Note that the preceding Ch. 11 covers randomization inference. It may be helpful to skim through if you found that topic difficult to follow. We likely won't get to the bootstrap (Ch. 12) this week, but we'll return to it.
- Butler, D. M., & Broockman, D. E. (2011). Do politicians racially discriminate against constituents? A field experiment on state legislators. *American Journal of Political Science*. [Abstract, Introduction, Conclusion; we'll look at their data in class this week]

Feb 10   Intro to hypothesis testing, 2
- Pager, D. (2003). The mark of a criminal record. *American Journal of Sociology.*

VI. Bivariate regression: best linear predictor

Homework 6 (due Sunday Feb 20th): *Replication of some of Butler and Broockman's analyses. What statistical tests are being used? How do you interpret the regression coefficients?*

Feb 15   Ordinary least squares (OLS) regression, 1
- 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.

Feb 17   Ordinary least squares (OLS) regression, 2
- IMS Chapter 7: Linear regression with a single predictor

- IMS [Chapter 24: Inference for linear regression with a single predictor](#)

VII. Multiple regression: model building

Final project part 1 (send by Sunday Feb 27th if you want ungraded feedback)*: data set selection and descriptive analysis.*

Feb 22 Estimating and interpreting multiple regression models
- Banerjee, A., Duflo, E., Glennerster, R., & Kinnan, C. (2015). The miracle of microfinance? Evidence from a randomized evaluation. American economic journal, 7(1), 22-53.

Feb 24 Causal inference with multiple regression
- 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.*
- 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).

VIII. Challenges to inference and external validity

Final project part 2 (send by Sunday March 6th if you want ungraded feedback)*: multiple linear regression and data visualizations.*

Mar 1   Pitfalls of p-values and hypothesis testing
- Amrhein et al. (2019). Retire Statistical Significance. *Nature*.
- Gerber and Malhotra (2008). Do Statistical Reporting Standards Affect What is Published? *Quarterly Journal of Political Science*.
- Play around with "Hack Your Way To Scientific Glory," https://projects.fivethirtyeight.com/p-hacking/

Mar 3   Research design and the trustworthiness of findings
- [Available by searching online] *Bohannon (2015).* I Fooled Millions into Thinking Chocolate Helps Weight Loss. Here's How*. Gizmodo.*

IX. Moving beyond regression, and beyond hypothesis testing

Final project (due Tuesday March 15th, 11:59PM)*: finalize your report*

Mar 8   Alternative empirical questions and alternative tools
- 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*

Mar 10   Wrap-up and review