SOSC 13200/3: Social Science Inquiry II

Winter, 2020

- Instructor: Evgenia Olimpieva (*she, her, hers)*
- Time: Tuesdays and Thursdays, 11:00am-12:20pm
- Location: Harper Memorial Library 148
- E-mail: eolimpieva@uchicago.edu
- Office Hours: Fridays 3:00-5:00pm and by appointment, Pick 504. Sign up here.

1. Course Description

This course will empower you to understand and use data to answer exciting questions about the social world, as well as to evaluate the data-driven scholarly work of others. Second in the Social Science Inquiry sequence, the purpose of the course is to furnish students with the computing and statistical skills necessary to conduct quantitative research in social sciences and to complete their independent research projects in the spring quarter. We will begin by getting comfortable with the fundamentals of programming in R – one of the most in-demand programming languages in the world today. We will use R to learn the nuts and bolts of a workflow with datasets and in the process will wrestle with tough questions surrounding the challenges of measurement in social sciences. The second half of the course is devoted to learning the fundamentals of statistical inference, with a focus on linear models and some of the challenges and limitations associated with this commonly used statistical technique. Throughout the course, students will be applying the statistical and programming skills they have learned to produce a guided replication of Koch and Nicholson's 2016 study on the relationship between war casualties and voter turnout.

2. Objectives

By the end of the course, students will be able to:

- Produce reproducible work using R Markdown
- Clean and transform data using the *tidyverse* packages.
- Interpret, evaluate, and conduct basic data analysis.
- Conduct hypothesis tests.
- Conduct linear regression analyses, including dummies and interactions.
- Read and interpret the results of a linear regression table.
- Understand the difference between experimental and observational studies.
- Know why correlation does not equal causation.

3. Requirements and Evaluation

i. Grade Composition

Homework Assignments (60%): Throughout the quarter, students will complete five homework assignments. Homework assignments are due at 5:00pm on the assigned date, unless otherwise specified. Each assignment makes up 12% of the grade for the course.

Exercises (20%): On alternating weeks from the homework assignments, students will complete four short exercises. These are due at 5:00pm on the assigned date, unless otherwise specified. Each of these exercises makes up 5% of the grade for the course. These will be graded for completeness only, and the material will be reviewed during the next class.

Participation (20%): It is important that students attend class, having read the assigned readings, and prepared to be engaged, active members of the class. The first, minimal requirement for participation is attendance, *including arriving on time*. Students are always responsible for any information they miss. Beyond attendance, participation will be based on students' preparation for class and involvement in class discussions, smaller-group activities, helping one another during R workshop sessions, and asking questions and providing help for others on Piazza (which is the online discussion platform we will be using for this class, accessible through Canvas).

ii. Grade cutoffs:

Percentage receives at least	$\underbrace{< 60}_{F}$	$\underbrace{\leq 60}_{D}$	$\underbrace{\leq 70}_{C-}$	$\underbrace{\leq 74}_{C}$	$\underbrace{\leq 81}_{B-}$	$\underbrace{\leq 84}_{B}$	$\underbrace{\leq 91}_{A-}$	$\underbrace{\leq 95}_{A}$

4. Course Policies

i. Late Assignments:

Homework assignments will lose half of a letter grade for each full day they are late, until solutions are distributed to the class. After that point, they will receive no credit. Students may have one 48-hour extension (or two 24-hour extensions) with no questions asked during the course of the quarter**. Extensions only apply to assignments (not exercises).** When using the extension, students should email the instructor before the original deadline and make a note of that on their assignments. All work should be submitted through Canvas. If you are having trouble getting your PDF file to "knit," you may submit it as a .html file as long as you submit the corrected PDF within 24 hours. The .html file must still be complete and the answers should not change between the two.

ii. Academic Integrity:

Please familiarize yourself with the University's Academic Integrity & Student Conduct policies. You are responsible for following these policies. Since getting help from others is an intrinsic part of learning programming, you may work in groups on exercises and on the coding portions of the assignments (but not on other, theoretical or mathematical questions). If you do work in a pair or a group, write the names of people you have worked with on the top of your assignment. While you are allowed to get help from others, *you cannot copy and paste* other students' code. All code you write must be your own, even if you received help from me or your classmates or based it on code you find online. If asked, ***you must be able to explain each line of code in your submission***. Plagiarism and other infractions of academic integrity will result in an automatic zero for the assignment and could result in failing the class. If you have any questions about what counts as independent work, please raise them before turning in an assignment.

iii. Technology:

Laptops should be used only for class purposes. Cell phones must be off and put away during class. Any infractions to either of these policies will result in reduction in your attendance and participation grades.

iv. Contact:

Email is the easiest and most reliable way to contact me. I am very invested in your learning and want to help you as much as I can along the way. However, being formal and respectful is a pre-requisite for you to receive a reply from me. I will try to reply to all of the e-mails within 24 hours – please do not expect an immediate response. I highly encourage you to come talk to me during office hours about any topics that you find confusing. I also highly encourage you to **utilize Piazza**. Before e-mailing me with a question regarding the exercises and assignment, check if your question has already been addressed on a Piazza discussion thread and, if not, share your question with others there. Remember that this is one way in which you can earn your participation points!

v. Disabilities:

If you have a disability that entitles you to a specific course accommodation, please contact me after speaking to the Coordinator for Student Disabilities Services, as early as possible in the quarter. If you are unable or do not wish to file your disability with the university and have an access need that could be met by means of a reasonable accommodation, let me know. You do not need to discuss with me the nature of your disability. I am invested in your ability to fully participate in this course and will try to accommodate your needs as much as I can.

vi. Name and Gender Pronouns:

Professional courtesy and sensitivity are especially important with respect to individuals and topics dealing with differences of race, culture, religion, politics, sexual orientation, gender, gender variance, and nationalities. Class rosters are provided to the instructor with the student's legal name. I will gladly honor your request to address you by an alternate name. I will also gladly honor your request to use your correct gender pronouns.

5. Readings

There are two required books for this course:

1) The first is a free online textbook *Introduction to Statistics*, developed primarily by David M. Lane. It is in the public domain and is available through the "Online Statistics Education: A Multimedia Course of Study" project. You can find it here.

2) The second book is *OpenIntro Statistics* by Diez, Barr and Çetinkaya-Rundel, the Third Edition. This book is available for free through UChicago Library. You can also find it here. I will refer to this book as "OIS" in the schedule.

3) We will also be using the in-class R Notes that I have written for this class. These will be available through Canvas.

You might also find the following resources helpful. These readings are is not required, but are rather additional resources for further learning:

1) *Introductory Statistics with R*, by Peter Dalgaard, is available electronically through the library. You can find it here when on campus or by using the library's proxy server.
2) *R for Data Science* by Hadley Wickham and Garrett Grolemund, which can be freely accessed here.
3) *A Beginner's Guide to R*, by Alain F. Zuur, Elena N. Ieno, and Erik H.W.G. Meesters, available here (on campus or using the University's proxy server)

6. Statistical Software

We will be using R as the main statistical software for this course, which is free and publicly available. Students are expected to download both R and R Studio to their personal computers before the first class. For homework assignments, students will also need to download LaTeX. I will provide you with detailed instructions on how to install these programs. Students should bring computers, with the aforementioned programs to each class. If there is a problem for any

reason, please come speak to me about other possible solutions. **On January 6th from 1-2:30pm, we will have a lab devoted to help you resolve installation issues, should you have any. Regenstein Library, Group Study Room 403.**

7. Getting Help

In addition to the textbooks and the class resources I have provided above, the other (in fact, the first) place you should look for help is Google. We will talk about the best practices for that during the first lecture. There is a broad community of R users who have posted helpful explanations and who answer one another's questions online. Figuring out how to troubleshoot when coding with the help of the internet is part of what I expect you to learn in this course as even experienced coders often need to do this. For some of the questions you ask me, my response will be to go to Google!

If you cannot solve the problem after reviewing class materials and looking online, you should post about it on **Piazza**. There will be a separate discussion for each exercise or assignment. You can also post your questions about R Notes. This will allow you to get help from myself and your classmates. It will also help your any classmates who are facing the same issue. Most importantly, it will help me to understand where you are struggling, to address it as early and efficiently as possible and to improve the course to better facilitate your learning. So, please, raise questions on Piazza if there is anything in the course that you are having difficulty with!

I also encourage you to respond to questions from your classmates if you know how to solve the problem. Remember, that participating on Piazza counts towards your class participation points! When posting, please include a quick statement about what you tried looking up online, as it may keep others from repeating the same steps and we may be able to suggest other things to try. Also, please remember to look through the questions your classmates have posted to ensure that it has not already been answered! When posting a chunk of code, make sure that we can reproduce it on our personal computers without your particular coding history (this might not make sense right now, but it will after Week 1).

8. Course Outline/Schedule

0. **Mon, Jan 6 –** Installation laboratory. Feel free to come whether or not you are experiencing difficulties downloading and installing the required software! But you should especially plan on coming if you are having issues. Location and time: **Regenstein Library, Group Study Room 403, 1-2:30pm.**

1. **Tue, Jan 7** – Class introduction and introduction to R: RStudio, R Markdown, packages, getting help, organizing your workflow, importing data.

   Before class:
   - Please come to class with a laptop that already has R, RStudio and LaTex installed and working.
   - Bring headphones to class.

   Reading:
   - Coding and naming in R handout

2. **Thu, Jan 9 --** Variables and measurement

   Before class:
   - Finish Chapter 1 of the R Notes.
   - Treisman "Causes of Corruption", Introduction (pp. 2-4) and pp. 16-18 available on Canvas.

   Reading:
   - Kellstedt and Whitten, *The Fundamentals of Political Science Research*, pp. 91-96
   - OIS, Section 1.2.1, pp. 9-13
   - Optional: Lane, Section H.1 of the Introduction ("Levels of Measurement" pp. 34-39 in PDF)

Monday, Jan 13 –Exercise 1 due by 5p.m.

3. **Tue, Jan 14** – Graphs and Visuals; ggplot2.

   Before class:
   - Complete the Introduction to the Part I of the ggplot 2 course on DataCamp. It should be free.
   - Finish Chapter 2 of R notes.

Reading:

- Kellstedt and Whitten, *The Fundamentals of Political Science Research*, pp. 104-109.
- Lane, Chapter 2 – "Graphing Distributions," pp. 65-122 (in PDF)

4. **Thu, Jan 16 –** Descriptive Statistics, Measures of Central Tendency and Spread.

   Before class:

   - Read Koch and Nicolson "Death and Turnout: The Human Costs of War and Voter Participation in Democracies" available on Canvas.

   Reading:

   - Lane, Section 1.3 – "Descriptive Statistics, pp. 15-19 (in PDF), and Chapter 3 – "Summarizing Distributions," except 3.7 – "Additional Measures of Central Tendency," pp. 123-135 and 140-163 (in PDF)

Monday, Jan 20 –Assignment 1 due by 5p.m.

5. **Tue, Jan 21 –** Inference 1: Population vs Sample, Normal distribution and its properties; Iteration with for loops.

   Reading**:**

   - OIS, Chapter 3.1-3.2 "Normal Distribution", pp. 127-141
   - Lane, Chapter 7 – "Normal Distribution," pp. 248-271 (in PDF)

6. **Thu, Jan 23 –**Inference 2: Variability in the estimates, Sampling distribution of the mean and the sums; Standard Errors and Confidence Intervals

   Before class:

   - Finish Chapter 5 of R Notes, especially the section on loops

   Reading**:**

   - OIS Chapter 4 "Foundations for Inference", sections 4.1-4.3, pp. 168-180
   - Lane, Chapter 9 – "Sampling Distributions," pp. 300-316 (in PDF)
   - Optional but fun: Watch two short Khan Academy videos on Central Limit Theorem and Sampling distribution of the sample mean.

Monday, Jan 27 – Exercise 2 due by 5p.m. (CLT Visualization)

7. **Tue, Jan 28** – Data transformation with dplyr.

Before class:

- Complete the first module ("Data Wrangling) of the Introduction to Tidyverse course on DataCamp (Links to an external site.). It should be free.
- Optional: Sections 12.1-12.4 of Chapter 12 "R for Data Science" and Chapter 13.

8. **Tue, Jan 30 –** Inference 3: Univariate Hypothesis Testing

Reading:

- OIS Chapter 4 "Foundations for Inference", sections 4.3-4.5, pp. 180-202
- (Optional) Wooldridge "Introductory Econometrics", Appendix C-6b pp. 695-703.

Monday, Feb 3 -- Assignment 2 due by 5p.m

9. **Tue, Feb 4 –** Data types; tidy data. Merging data with dplyr and tidying data with tidyr.

Before class:

- Complete the first module of Cleaning Data (Links to an external site.) in R on DataCamp
- Complete the first module of Joining Data (Links to an external site.) with dplyr on DataCamp

Reading:

- Ch. 12-13 of R for Data Science (Links to an external site.)
- Wooldridge, Introductory Econometrics, Ch1

10. **Thu, Feb 6 --** Correlation, line fitting, residuals.

Reading:

- OIS, Chapter 7, pp 331-339.
- Lane, Chapter 4 - "Describing Bivariate Data," through the section "Computing Pearson's r," pp.164 – 177 (in PDF)

Monday, Feb 10 -- Exercise 3 due by 5p.m.

11. Tue, Feb 11 – Bivariate Regression, Part 1

Reading:

- OIS, Chapter 7, pp 331-339.
- Lane, Section 14.1 – "Introduction to Linear Regression," pp. 462-467 (in PDF)

12. Thu, Feb 13 – Bivariate Regression, Part 2

Reading:

- Lane, Section 14.2, "Partitioning the Sums of Squares" through 14.4, "Inferential Statistics for b and r," pp. 468-481 (in PDF)
- OIS, pp. 351-356, "Inference for linear regression"

Monday, Feb 17 -- Assignment 3 due by 5p.m.

13. Tue, Feb 18 -- Multivariate Regression, Part 1

Reading:

- Lane, Section 14.7 – Introduction to Multiple Regression, pp. 495-506 (in PDF)
- OIS, Chapter 8.1 "Introduction to Multiple Regression", pp. 372-377

14. Thu, Feb 20 -- Multivariate Regression, Part 2

Reading:

- Buttolph Johnson and Reynolds, Political Science Research Methods, pp. 527-550

Monday, Feb 24 -- Exercise 4 due by 5p.m.

15. Tue, Feb 25 -- OLS extensions: Dummy variables and Interactions.

Before Class:

- Complete Ch1 "Parallel Slopes (Links to an external site.)" of "Multiple and Logistic Regression" DataCamp course.

- Kellstedt and Whitten, The Fundamentals of Political Science Research, pp. 202-212

16. Thu, Feb 27 -- OLS Assumptions

Reading:

- OIS, Chapter 8.3 "Checking model assumptions using graphs", pp. 382-385
- Watch Ben Lambert's video

Monday, March 2 -- Assignment 4 due by 5p.m.

17. Tue, March 3 -- Challenges in OLS and Causality; A very quick into LPM, Logit and Probit.

Before class:

- Read Section 2 on "Causality and Experiments" from The Foundations of Data Science by Ani Adhikari and John DeNero.
- Kellstedt and Whitten, The Fundamentals of Political Science Research, pp. 212-220

18. Thu, March 5 – A quick introduction to OLS with panel and cross-sectional data; Clustered Standard Errors

Before class:

- Read Ch.10, "Regression with Panel Data" from *Introduction to Econometrics* by James H. Stock, Mark W. Watson (PDF uploaded on Canvas)

19. Tue, March 10 – Decisions in modeling and how things can go wrong

Before class:

- Read Anthony Fowler and Andre Hall's paper "Do Shark Attacks Influence Presidential Elections?
- Skim Achen and Bartels article (which Fowler and Hall are in conversation with)
- **Fri, March 13 -- Assignment 5 due by 5 p.m.**