SOCI 103M Quantitative Research in Sociology with R

Prof. Ke Nie
Department of Sociology
knie@ucsd.edu

12-12:50p, MWF
Meetings in-person in HSS 2150
Office Hours: By email request on Zoom

1 Course Overview

Sociological research requires rigorous investigations of empirical data. There are diverse types of empirical data we use for conducting research, from historical archives, participatory observations, and surveys, to numbers and statistics. There is no predetermined hierarchy among these data sources; different types of data are good at answering different kinds of questions. With that said, in recent years, the development of computational tools that enable the collection, analysis, and presentation of massive digital data invites us to take advantage of them for revisiting the old wisdom or breaking new grounds in sociology.

This course is designed for students who are interested in leveraging computational tools to manage and analyze quantitative data for answering questions relevant to sociology and social sciences more broadly. In this course, I aim to provide the necessary training that will equip you with the fundamental skills to pursue a career in social science research with a concentration on quantitative analysis. For those who wish to build a career in data analytics or data science, this course may also be helpful since it requires a similar set of skills. However, the course is primarily tilted toward grappling with sociological inquiries. This means that, instead of merely teaching coding and programming, the course will focus on how to study sociological questions through them.

The course integrates training for three important sets of skills: research design, data manipulation, and research communication. They are all intertwined so it is practically impossible to separate them. With regard to the technical end, we will primarily use `R`, which is widely used in both academia and industry for data tasks including data wrangling, data analysis, and data visualization. By the end of the quarter, you will achieve the following objectives:

- Design a research project of your interest that is also relevant to sociology and social sciences;
- Master the fundamentals of `R` programming with `R Studio`;
- Use `R Markdown` to communicate your research;
- Write a research paper and present it in class.

2 Prerequisites

This course does not require previous coding experience. We will start from scratch, and you will be able to nail it as long as you follow the course closely. Yet, since we will be using our computer a lot, possessing a personal computer will be much easier for taking the course, although you can always go to a computer lab

on campus to finish the assignments. `R` supports multiple operating systems, including Windows, macOS, and Ubuntu, so most personal computers will work just fine.

As we will be dealing with numbers, a preliminary knowledge of mathematical statistics (e.g., probability theory, parameter estimations, hypothesis testing, linear regressions) will be helpful. We will review them as we go into the course, but the reviews will only be shallow due to the alternative focus of this course. I will encourage you to use outside resources, as indicated in the section of reading materials below, for your reference.

3 Course Requirements

Attendance (10% of the final grade; the following percentages mean the same unless otherwise specified). You are expected to attend all classes with active participation. With that said, I understand that absences for personal emergencies or health-related issues are sometimes necessary. Therefore, you are allowed to have 3 absences without penalty over the quarter (except for Week 10, where absences are evaluated on a case-by-case basis), and you don't need any excuses for them. However, you need to **inform me via email** regarding your absence **no later than the second day of the class** you miss. For example, if you miss a class on Wednesday, you need to email me about your absence by Thursday even though you don't need to explain to me why. If there is no email, there will be an automatic penalty of 1% of the final grade until the exhaustion of the 10%, and the penalty cannot be appealed. If you know you will have to miss more than 3 classes, please inform me ahead of time so that we can make arrangements.

Interactive programming training (10%). There will be interactive programming training all along the way for you to practice your programming skills. The training is set up on an interactive coding platform called Datacamp. The good thing about Datacamp is that it provides you with a user-friendly interface where you can simply focus on coding by following the instructions. You will need to set up your own account on Datacamp using the link I send to you after our first meeting to join our Datacamp Group for free to complete the exercises. For most weeks, there will be 2 training sets per week. Each training set typically comprises some video instructions and a series of coding prompts. You can easily finish them if you follow the instructions. The training sets of the week are typically **due by the beginning of the Wednesday's and Friday's class**, so please complete them promptly. Late submissions will result in an automatic 0.5% deduction from the final grade.

Problem sets (20%). There will be 2 take-home problem sets, one in Week 3 and the other in Week 7. The problem sets will be posted by Monday and due by Friday of the week. Generally, the problem set will ask you to code for analyzing some given datasets to answer sociological questions. You are supposed to use `R Markdown` to solve the problem sets, where the code, the results, and your response will be integrated into one single pdf/html file for submission. Late submissions will be deducted 1% from the final grade every 24 hours.

Quiz (10%). There will be an in-class quiz, which will be similar to the problem sets in structure. The only difference is that you are supposed to complete it in class within 50 minutes. While taking the quiz, you should **NOT** consult with any other students with any communication tools, but you will be allowed to use any other resources, including online resources, to answer the quiz. Late submissions will result in a 1% deduction from the final grade every 24 hours.

Collaboration on a research paper (30%). The whole class will form 12 research groups to work on a research

project. Given the size of the class (25 students), most of you will form a group of 2 and only one group of 3. Your goal is to investigate a research question of your interest, which needs to be relevant to sociology or social science more broadly and eventually produce a research paper at the end of the quarter.

To form a research group, you can find your own partner and register on Canvas. For those who do not wish to find a partner themselves, you will be randomly assigned to a group. Grading will be based on group outcomes rather than individual performance, so each group member will be given the same grade for the research paper. It also implies that you should be responsible for finding the way of collaboration that works best for your group.

The paper does **NOT** need to cover original research, although it will be highly recommended. This means that you can either use a dataset that has already been published or collect and work on your own data. However, the paper needs to be written in a similar structure as a published paper, which means that it should include a literature review (why is your question important and how did others manage to answer this question), elaborations on data & methods, discussions on findings, a conclusive remark on the implications of the findings, and a bibliography.

You need to write the paper in `R Markdown` and knit it (I will explain it as we go) as an **html** file; however, you do **NOT** need to show your code in the final paper. I will collect all the papers and integrate them into a website dedicated to this course.

I will not set a word limit for the paper, although it does not need to be too long: a 2500- to 3000-word paper (excluding the bibliography) is totally acceptable. I will grade the paper primarily based on its completeness, i.e., whether the paper has a clear structure and answer the proposed question clearly, while the other aspects will be of secondary relevance.

I will dedicate roughly one class each week for you to work on the research project, and I will give out assignments that ask for each piece of the paper as outlined in the course schedule below. Late submissions will result in a 1% deduction from the final grade every 24 hours.

Research presentation and commentation (20%). In Week 10, you will need to present your research in front of the whole class. The way how you do it is completely at your discretion. You can use slides or any other audiovisual aids for the presentation as long as you think it can expressly address your research to the audience. Each presentation should not exceed 7 minutes, and it will be followed by a 3-minute Q&A. The presentation will be graded based on the clarity of the delivery.

Assuming there will be 12 groups doing 12 presentations, we will randomly spread the groups into the 3 meetings of Week 10 so there will be 4 presentations per day. In addition to that, for each group that does the presentation on the day, I will also randomly select one group from each of the other two days to give written comments on the presentation and send the comments to the presenting group by the next day on Canvas. For example, if a group presents on Monday, a randomly selected group that presents on Wednesday and a randomly selected group that presents on Friday will write a comment on the presentation and send the comment to the presenting group by Tuesday. Each commenting group will only need to write one comment for the presenting group, so 2 comments for each presenting group. The comment does not need to be long; it only needs to address the strength and the weakness of the presentation and raise 2 questions that the presenters did not touch on in their presentation or Q&A. The presenting group needs to respond to the comments and answer the questions on Canvas by the end of Finals Week. Particularly, in the response, the presenting group should address how they deal with the questions in their final paper.

4 R Programming

We teach the course in R, which is an open-source computing language that is very widely used in statistics. You can download it for free from www.r-project.org. The website itself provides many great resources, in particular the manuals, to learn R. Some good qualities of R include: it contains various canned functions for most commonly used algorithms, which makes our job here quite easy; it has strong, production-quality graphical capabilities, which is cool in itself; and there are many jobs out there that require R proficiency. Some sources that might be helpful for you to get started and get deep:

- A nice way to start you off are some video tutorials on YouTube like this one by the channel R Programming 101 and this video series by the John Hopkins University.
- A more comprehensive guide can be checked out in Wickham, H., & Grolemund, G. (2016). R for data science: import, tidy, transform, visualize, and model data. " O'Reilly Media, Inc.". They basically put the whole book online for free.
- Irizarry, R. A. (2019). Introduction to data science: Data analysis and prediction algorithms with R. CRC Press. Also, a free online book, focusing slightly more on modeling and more advanced techniques compared to the one above.
- R-bloggers, a useful blog that posts news and tutorials on R.

5 Datacamp

Please check out the "Sign up for Datacamp group" section on Canvas to sign up for a Datacamp account and join our Datacamp group as soon as you decide to come on board.

It is very important that you sign up promptly, as we will be running interactive programming exercises on this platform. Failure to register will be detrimental to your grade. Please let me know as soon as possible if you run into problems.

After signing up, you will have access to all the exercises I assign to you. You can check out the assignments either directly on Datacamp or on Canvas. Please remember that these exercises are typically **due by the start of Wednesday and Friday's class, respectively, unless otherwise noticed**. These deadlines will **NOT** be addressed in our syllabus, so please always be aware of the deadlines, which I will specify in the title of the assignment on Canvas.

You will also have free access to all other courses on Datacamp until the end of the quarter. Feel free to explore them if you are interested.

6 Reading Materials

No books are required to purchase for this class, as I will post all the readings on Canvas, our online course portal. This is to mitigate the financial burden on you. Please do not disseminate the readings; just use them for the sole purpose of this course.

However, as we will cover some technical aspects, especially those regarding statistical inferences, that are relevant yet not central to the course, I will recommend the following books for your reference along the way:

- Imai, K. (2018). Quantitative social science: an introduction. Princeton University Press.
- Healy, K. (2018). Data visualization: a practical introduction. Princeton University Press.
- Agresti, A. (2018). Statistical methods for the social sciences (5th edition). Pearson.
- Lewis-Beck, C., & Lewis-Beck, M. (2015). Applied regression: An introduction (Vol. 22). Sage publications.
- Angrist, J. D., & Pischke, J. S. (2009). Mostly harmless econometrics: An empiricist's companion. Princeton university press.
- Morgan, S. L., & Winship, C. (2015). Counterfactuals and causal inference. Cambridge University Press.

I will also recommend online resources if that is what you prefer. In fact, you will notice that a lot of learning will happen by searching things online, and that is not a problem as long as you make judgments about the validity of the information carefully and cautiouslly. If you feel more comfortable following more established resources, you can first of all take advantage of the free access I give you to Datacamp to explore the large repertoire of courses they have, including more advanced use of `R`, as well as other programming languages. You can of course also explore other learning platforms, such as Coursera (find here the free UCSD courses for UCSD students) or Khan Academy for most of the technical training. You can even try Crash Course Statistics if you will, although one caveat is that it will not dive deep into the topics and some videos may contain errors.

7 Academic Integrity

Academic integrity is a big deal. You must complete all academic work assigned to you yourself, without any unauthorized aid.

Violations must be reported to the UCSD Academic Integrity Review Board. If you are unsure what constitutes a violation, please refer to the website of the Academic Integrity Office at UCSD.

Know that unintentional plagiarism is not distinguishable from intentional plagiarism – so please be sure to cite properly anytime you quote or paraphrase.

8 Tentative Course Schedule (Subject to Alteration)

We will meet on Mondays, Wednesdays, and Fridays for most weeks without a university-recognized holiday. Generally speaking, we will discuss substantive topics on Mondays, combat coding on Wednesdays, and work on your research project on Fridays, but that is not always the case. We will make adjustments according to the needs.

Please see below a detailed breakdown of what each day's meeting will entail. Please be aware of all the deadlines and checkpoints.

Please also note that:

- There will be 2 Datacamp exercises every week except for Week 10 and **their deadlines are always by the start of the Wednesday's and Friday's class (12p), respectively**. These deadlines are **NOT** presented

below for the sake of clarity.

- Please read all the required readings before the first day of the week.

Week 01

Jan 9 (M)		Introduction to the course
Jan 11 (W)	[Discussion]	Quantitative research in sociology and social sciences
Jan 13 (F)	[Programming]	Set up the environment for `R`

Readings:

- Luker, K. (2015). What's it all about? In *Salsa dancing into the social sciences: Research in an age of info-glut*. Harvard University Press.
- Ledford, H. (2020). How Facebook, Twitter and other data troves are revolutionizing social science. *Nature*, 582(7812), 328-331.

Week 02

Jan 16 (M)		**NO CLASS** - Martin Luther King Jr. Holiday
Jan 18 (W)	[Discussion]	Propose a research question
Jan 20 (F)	[Programming]	`R` foundations - operators, packages, and import data

Readings:

- Mears, A. (2017). Puzzling in sociology: On doing and undoing theoretical puzzles. *Sociological Theory*, 35(2), 138-146.
- Pepinsky, T. T. (2019, February 7). *On puzzles and political science*. Personal blog.

[Deadline (Jan 20): Form your research group]

Week 03

Jan 23 (M)	[Discussion]	Data collection
Jan 25 (W)	[Programming]	Introduction to `R Markdown`
Jan 27 (F)	[Research]	Where do I collect my data, and how can it speak to my research question?

Readings:

- Jahnke, L. M., & Asher, A. (2012). The problem of data: Data management and curation practices among university researchers. *The Problem of Data*, 3-32.
- Heinrich, A., & Klein, E. (2021). Challenges for the management of qualitative and quantitative data: The example of social policy–related data collections. *Global Social Policy*, 21(1), 138-143.

[Deadline (Jan 27): Problem set #1]

Week 04

Jan 30 (M)	[Discussion]	Data quality and limitations
Feb 1 (W)	[Programming]	Data wrangling with `tidyverse`: numbers, strings, and data frames
Feb 3 (F)	[Research]	What is wrong with my data, and what can I do about it?

Readings:

- Lewis, K. (2015). Three fallacies of digital footprints. *Big Data & Society*, 2(2).
- Lokshin, M. M. (2018, July 5). *Data quality in research: what if we're watering the garden while the house is on fire?* World Bank Blogs.
- Dodds, L. (2020, January 31). *Do data scientists spend 80% of their time cleaning data? Turns out, no?*. Lost Boy Blog.

[Deadline (Feb 3): Submit research question]

Week 05

Feb 6 (M)	[Discussion]	Measurement
Feb 8 (W)	[Programming]	Data wrangling with `tidyverse`: play with pipes and aggregate summaries
Feb 10 (F)		**In-class midterm quiz**

Readings:

- Bail, C. A. (2014). The cultural environment: Measuring culture with big data. *Theory and Society*, 43(3), 465-482.
- Fink, E. L. (2017). Validity, Measurement of. In *The SAGE encyclopedia of communication research methods*. SAGE Publications.

Week 06

Feb 13 (M)	[Discussion]	Handle missing and abnormal data
Feb 15 (W)	[Programming]	Data wrangling with `tidyverse`: merge datasets
Feb 17 (F)	[Research]	What is missing from my data?

Readings:

- Porter, J. R., & Ecklund, E. H. (2012). Missing data in sociological research: An overview of recent trends and an illustration for controversial questions, active nonrespondents and targeted samples. *The American Sociologist*, 43(4), 448-468.
- Silaparasetty, V. (2020, Feb 5). *Guide to Handling Missing Values in Data Science*. Medium.

[Deadline (Feb 17): Submit literature review]

Week 07

Feb 20 (M)		**NO CLASS** - Presidents' Day Holiday
Feb 22 (W)	[Programming]	Data visualization with `tidyverse`
Feb 24 (F)	[Research]	How to tell a story from graphics?

Readings:

- Healy, K. (2019). Look at data. In *Data visualization: A practical introuduction*. Princeton University Press.
- Sarkar, D. (2018, September 12). *A Comprehensive Guide to the Grammar of Graphics for Effective Visualization of Multi-dimensional Data*. Medium.

[Deadline (Feb 24): Problem set #2]

Week 08

Feb 27 (M)	[Discussion]	Identify correlations
Mar 1 (W)	[Programming]	Correlation and Causation: statistical modeling in `R` (Part I)
Mar 3 (F)	[Research]	Can I find interesting correlations from my data?

Readings:

- Altman, N., & Krzywinski, M. (2015). Points of Significance: Association, correlation and causation. *Nature Methods*, 12(10).
- Ritter, D. (2014, March 19). *When to act on a correlation, and when not to*. Harvard Business Review.
- Vigen, T. *Spurious correlations*. Personal Blog.

[Deadline (Mar 3): Submit data & methods]

Week 09

Mar 6 (M)	[Discussion]	The conditions and pitfalls of correlation
Mar 8 (W)	[Programming]	Correlation and Causation: statistical modeling in `R` (Part II)
Mar 10 (F)	[Research]	Am I making the right claim about the data?

Readings:

- Wang, Y., & Kosinski, M. (2018). Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of personality and social psychology*, 114(2), 246.
- Arcas, B. A., et al. (2017, May 6). *Physiognomy's new clothes*. Medium.

Week 10 The Grand Presentations

Mar 13 (M)	[Presentation]	TBD
Mar 15 (W)	[Presentation]	TBD
Mar 17 (F)	[Presentation]	TBD

[Deadline: Submit comments by the second day of the presenting group]