Applied Quantitative Methods for the Social Sciences I

Program:	Master in Social Sciences, Fall 2024
Room:	18.1.A04
Time:	Mondays, 10:00–13:00

Contact Information

Instructor:	Patrick Kraft, PhD
Office:	18.2.A19
Email:	patrickwilli.kraft@uc3m.es
Office Hours:	Mondays, 15:00–17:00, or by appointment.

I Overview

Course Description

This is a first course on statistical inference and modeling for use in social science research. It covers the theory of statistical inference, essential concepts in statistical modeling, justifications for and problems with common statistical procedures, and how to apply procedures to empirical social science data to draw conclusions relevant to positive social theory. We will pay particular attention to the motivation for statistical inference and modeling from the standpoint of social science. Lectures and reading will primarily cover theory and simple examples. Problem sets will cover both simple theoretical extensions and applications of tools we develop to real data. The topics we will discuss here include:

1. Programming with R
2. R Markdown & LaTeX
3. Probability theory
4. Statistical inference
5. Introduction to OLS
6. Data visualization.

Prerequisites

Students should have completed our intro course "Mathematics for Social Sciences and Basic Statistics" or its equivalent. Students should have a working knowledge of arithmetic, algebra, and elementary calculus. The course is suitable for students with a large range of prior exposure to statistics and mathematics. All students capable of gaining admission to our MA program can fully succeed in this class regardless of prior technical preparation other than the required skills listed above.

II Textbook and Required Material

There is one required textbooks for the course:

Imai, Kosuke, and Nora Webb Williams. 2022. *Quantitative Social Science: An Introduction in Tidyverse*. Princeton University Press

Additional required or optional readings will be available on Aula Global. You should complete the readings assigned for each week before the respective class session.

We will be using RStudio for the programming portion of the course. You can get started by installing R and RStudio on your computer. Next, you can work through RStudio's Posit Recipes, a set of tutorials that will help you familiarize yourself with basic programming concepts and R.

III Schedule

1: Introduction (September 9)

Our first week provides a broad overview of the content we are going to cover throughout the semester. We will discuss course requirements, administrative questions, and grading policies. Furthermore, we are going to take a first look at R and RStudio.

2: Causality (September 16 & September 23)

In this section, we consider causality, one of the most central concepts of quantitative social science. Much of social science research is concerned with the causal effects of various policies and other societal factors. Do small class sizes raise students' standardized test scores? Would universal health care improve the health and finances of the poor? What makes voters turn out in elections and determines their choice of candidates? To answer these causal questions, one must infer a counterfactual outcome and compare it with what actually happens (i.e., a factual outcome). We show how careful research design and data analysis can shed light on these causal questions that shape important academic and policy debates. We introduce various research designs useful for causal inference and apply them to additional studies concerning social pressure and voter turnout, as well as the impact of minimum-wage increases on employment. We also learn how to subset data in different ways and compute basic descriptive statistics in R

Required:

- ☐ Imai, Kosuke, and Nora Webb Williams. 2022. *Quantitative Social Science: An Introduction in Tidyverse*. Princeton University Press, chs. 1 & 2.

Recommended:

- ☐ Angrist, Joshua D, and Jörn-Steffen Pischke. 2008. *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press, chs. 1 & 2.

3: Measurement (September 30 & October 7)

Measurement plays a central role in social science research. In this section, we first discuss survey methodology, which is perhaps the most common mode of data collection. Surveys are effective tools for making inferences about a large target population of interest from a relatively small sample of randomly selected units. In addition to surveys, we also discuss the use of latent concepts, such as ideology, that are essential for social science research. These concepts are fundamentally unobservable and must be measured using a theoretical model. Thus, issues of measurement often occupy the intersection of theoretical and empirical analyses in the study of human behavior. Finally, we introduce a basic clustering method, which enables researchers to conduct an exploratory analysis of data by discovering interesting patterns. We also learn how to plot data in various ways and compute relevant descriptive statistics in R.

Required:

- ☐ *Problem Set 1 due on September 30.*
- ☐ Imai, Kosuke, and Nora Webb Williams. 2022. *Quantitative Social Science: An Introduction in Tidyverse*. Princeton University Press, ch. 3.

Recommended:

- ☐ Gailmard, Sean. 2014. *Statistical modeling and inference for social science*. Cambridge University Press, chs. 1 & 2.

4: Prediction (October 14 & October 21)

In this section, we discuss prediction. Prediction is another important goal of data analysis in quantitative social science research. Our first example concerns the prediction of election outcomes using public opinion polls. We also show how to predict outcomes of interest using a linear regression model, which is one of the most basic statistical models. While many social scientists see causal inference as the ultimate goal of scholarly inquiry, prediction is often the first step towards understanding complex causal relationships that underlie human behavior. Indeed, valid causal inference requires the accurate prediction of counterfactual outcomes.

Required:

- □ *Problem Set 2 due on October 14.*
- □ Imai, Kosuke, and Nora Webb Williams. 2022. *Quantitative Social Science: An Introduction in Tidyverse*. Princeton University Press, ch. 4.

Recommended:

- □ Fox, John. 2015. *Applied regression analysis and generalized linear models*. 3 ed. Sage Publications, chs. 5 & 7
- □ Wooldridge, Jeffrey M. 2013. *Introductory econometrics: a modern approach*. Cengage Learning, Appendix E.

5: Probability (October 28 & November 4)

Until now, we have studied how to identify patterns in data. While some patterns are indisputably clear, in many cases we must figure out ways to distinguish systematic patterns from noise. Noise, also known as random error, is the irrelevant variation that occurs in every real-world data set. Quantifying the degree of statistical uncertainty of empirical findings is the topic for the next section, but this requires an understanding of probability. Probability is a set of mathematical tools that measure and model randomness in the world. As such, this section introduces the derivation of the fundamental rules of probability, with the use of mathematical notation. In the social sciences, we use probability to model the randomly determined nature of various real- world events, and even human behavior and beliefs. Randomness does not necessarily imply complete unpredictability. Rather, our task is to identify systematic patterns from noisy data.

Required:

- □ *Problem Set 3 due on October 28.*
- □ *Research Project Proposals due on November 4.*
- □ Imai, Kosuke, and Nora Webb Williams. 2022. *Quantitative Social Science: An Introduction in Tidyverse*. Princeton University Press, ch. 6.

Recommended:

- □ Gailmard, Sean. 2014. *Statistical modeling and inference for social science*. Cambridge University Press, chs. 3-6.
- □ King, Gary. 2006. "Publication, publication." *PS: Political Science & Politics* 39 (01): 119–125

6: Uncertainty (November 11 & November 18)

Thus far, we have studied various data analysis techniques that can extract useful information from data. We have used these methods to draw causal inferences, measure quantities of interest, make predictions, and discover patterns in data. One important remaining question, however, is how certain we can be of our empirical findings. For example, if in a randomized controlled trial the average outcome differs between the treatment and control groups, when is this difference large enough for us to conclude that the treatment of interest affects the outcome, on average? Did the observed difference result from chance? In this section, we consider how to separate signals from noise in data by quantifying the degree of uncertainty. We do so by applying the laws of probability introduced in the previous chapter. We cover several concepts and methodologies to formally quantify the level of uncertainty. These include bias, standard errors, confidence intervals, and hypothesis testing. Finally, we describe ways to make inferences from linear regression models with measures of uncertainty.

Required:

- □ *Problem Set 4 due on November 11.*

- ☐ *Peer review due on November 18.*
- ☐ Imai, Kosuke, and Nora Webb Williams. 2022. *Quantitative Social Science: An Introduction in Tidyverse*. Princeton University Press, ch. 7.

Recommended:

- ☐ Gailmard, Sean. 2014. *Statistical modeling and inference for social science*. Cambridge University Press, chs. 7 & 8.
- ☐ Fox, John. 2015. *Applied regression analysis and generalized linear models*. 3 ed. Sage Publications, chs. 6 & 9

7: Final Project Presentations (November 25)

Required:

- ☐ *Problem Set 5 due on November 25.*
- ☐ *Final research project due on December 16.*
- ☐ Imai, Kosuke, and Nora Webb Williams. 2022. *Quantitative Social Science: An Introduction in Tidyverse*. Princeton University Press, ch. 8.

Recommended:

- ☐ Cinelli, Carlos, Andrew Forney, and Judea Pearl. 2020. "A Crash Course in Good and Bad Controls." *Available at SSRN 3689437*

Extra days in case one of the sessions listed above has to be canceled:

- December 2
- December 9
- December 16

Note: The schedule and readings may be subject to change depending on our progress during the semester.

IV Evaluation

Your final grade will be determined based on the following three components:

1. **Problem sets ($20\% = 4 * 5\%$):** You will work on bi-weekly problem sets. I strongly encourage you to work in groups and discuss each question with your peers. However, each student must write up and submit their own original solution. Problem sets have to be submitted via Aula Global before our lecture on the specified due date (i.e., by 10:00). *Of the 5 problem sets, I will take the average of the 4 highest grades*, meaning that you may skip one problem set without affecting your grade.
2. **Research project proposal (10%):** About halfway through the semester, you have to submit a proposal for a research project that utilizes the methods and techniques covered throughout the course. It should consist of 2 to 3 pages outlining your research question, dataset, and hypotheses.
3. **Proposal peer review (5%):** After submitting the proposal, you will be paired with one of your peers to give and receive constructive feedback for your projects.
4. **Research project (35%):** At the end of the semester, you are expected to submit your final research project. While you have to incorporate an original data analysis using R, you are free to choose any topic and/or data source you find interesting (and it may overlap with your other substantive coursework). Further details will be discussed in class.
5. **Final exam (30%):** The final exam will test you on all the material covered throughout the semester. It will focus on the theoretical questions related to statistical modeling and inference.

V Additional Resources

As we work through the course material, some of you may want additional information on the underlying mathematical concepts, while others want to dig deeper into programming. Here is a list of additional textbooks that you might find helpful in either case:

	Mathematics / Statistics	*Programming / R*
Recommended	Gailmard (2014) Wooldridge (2013) Angrist and Pischke (2008) Fox (2015)	Wickham, Çetinkaya-Rundel, and Grolemund (2023) Urdinez and Cruz (2020) Verzani (2014) Fox and Weisberg (2018)
Optional	DeGroot and Schervish (2012) Casella and Berger (2021)	Teetor (2011) Matloff (2011)

I particularly recommend Wickham, Çetinkaya-Rundel, and Grolemund (2023), which is freely available online at the following link: https://r4ds.hadley.nz/

There are countless other free resources available, but I want to highlight two great sets of YouTube videos in case you want to learn more about specific topics covered in our course. You'll find links to these videos on Aula Global as well:

- Gary King's lecture videos on Quantitative Social Science Methods:
  https://www.youtube.com/watch?v=qs2uCuDL20Q&list=PL0n492lUg2sgSevEQ3bLilGbFph4l92gH
  https://projects.iq.harvard.edu/gov2001
- Andrew Heiss's lecture videos on Causal Inference and Data Visualization:
  https://www.youtube.com/c/AndrewHeiss/playlists.
- Lastly, I have taught courses similar to this one at the University of Wisconsin-Milwaukee in the past. You can find my old lecture videos—which cover a lot of the same topics—here:
  https://www.youtube.com/channel/UCmXfZJxXiwypm7f0vnKO_DA/playlists

VI AI policy

In this course, students should not use artificial intelligence tools to carry out the work or exercises proposed by the faculty. In the event that the use of AI by the student gives rise to academic fraud by falsifying the results of an exam or work required to accredit academic performance, the Regulation of the University Carlos III of Madrid of partial development of the Law 3/2022, of February 24th, of University Coexistence, will be applied.

VII Acknowledgements

I have adapted the ideas and language from the work of several educators for this syllabus and the course material. For example, I have borrowed liberally from other courses on social science research methods and statistics, as taught by Sean Gailmard, Kosuke Imai, Gary King, Michael Peress, Thomas Gschwend, and others. I appreciate their contributions to the discipline and thank all educators who make their teaching material available to others. To pay it forward, I will share my own material with anyone who is interested.

References

Angrist, Joshua D, and Jörn-Steffen Pischke. 2008. *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.

Casella, George, and Roger L Berger. 2021. *Statistical inference*. Cengage Learning.

Cinelli, Carlos, Andrew Forney, and Judea Pearl. 2020. "A Crash Course in Good and Bad Controls." *Available at SSRN 3689437*.

DeGroot, Morris H, and Mark J Schervish. 2012. *Probability and statistics*. Pearson Education.

Fox, John. 2015. *Applied regression analysis and generalized linear models*. 3 ed. Sage Publications.

Fox, John, and Sanford Weisberg. 2018. *An R companion to applied regression*. 3 ed. Sage Publications.

Gailmard, Sean. 2014. *Statistical modeling and inference for social science*. Cambridge University Press.

Imai, Kosuke, and Nora Webb Williams. 2022. *Quantitative Social Science: An Introduction in Tidyverse*. Princeton University Press.

King, Gary. 2006. "Publication, publication." *PS: Political Science & Politics* 39 (01): 119–125.

Matloff, Norman. 2011. *The art of R programming: a tour of statistical software design*. No Starch Press.

Teetor, Paul. 2011. *R cookbook*. O'Reilly Media, Inc.

Urdinez, Francisco, and Andres Cruz. 2020. *R for Political Data Science: A Practical Guide*. CRC Press.

Verzani, John. 2014. *Using R for introductory statistics*. CRC Press.

Wickham, Hadley, Mine Çetinkaya-Rundel, and Garrett Grolemund. 2023. *R for data science: import, tidy, transform, visualize, and model data*. 2nd ed. O'Reilly Media, Inc.

Wooldridge, Jeffrey M. 2013. *Introductory econometrics: a modern approach*. Cengage Learning.