Applied Quantitative Methods for the Social Sciences II

Universidad Carlos III de Madrid

Program: Master in Social Sciences, Spring 2025
Room: 18.1.A04
Time: Thursdays, 10:00–13:00

Contact Information

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

I Course Description

This class is the second graduate course for quantitative methods in our MA program that builds explicitly on the foundation laid in AQMSS-I. In that class, you learned the nuts and bolts behind statistical inference as well as the statistical software R. In this course, we extend those tools to cover various linear and non-linear models. Therefore, the main goal for this semester is not only to become a proficient *consumer* of quantitative research, but also to set the groundwork to pursue your own projects using advanced statistical modeling techniques. The topics that we will intensively discuss include:

1. Properties of Ordinary Least Squares (OLS)
2. Model Specification and Diagnostics
3. Panel Regression and Multilevel Modeling
4. Maximum Likelihood Estimation (MLE)
5. Generalized Linear Models
6. Model Visualization and Interpretation.

The goal of this course is three-fold: (1) to prepare students to conduct research using appropriate statistical models and to communicate their results to a nontechnical audience; (2) provide a foundation in the theory of maximum likelihood so students can investigate and implement a wide range of statistical models; and (3) provide students with the tools necessary to learn more advanced statistical methods in the future.

Please work through the assigned readings ahead of time. We expect everyone to come to class fully prepared. Expect that this will take considerably longer than in a substantive seminar. Do not skip equations! Instead, take notes, prepare questions and team-up with others to answer them, or as last resort, bring them up in class. After every class we expect you to go over the lecture notes and your notes once again.

The course will be conducted as mix of lecture and practical computer lab sessions. During the lecture sessions we will highlight the central concepts and ideas of the readings (and at that time they should not be new to you!). In the more practical sessions, students will learn to program the statistical models introduced in the lecture. We will also work through examples collaboratively. Our pace will adapt according to your progress. Nevertheless, understand that the bulk of learning in this course will take place outside of the classroom, by reading, practicing using statistical software, and doing problem sets.

Prerequisites

Students should have completed our intro course "Mathematics for Social Sciences and Basic Statistics" as well as "Applied Quantitative Methods for the Social Sciences I". Based on these previous courses, students are expected to have a working knowledge of basic calculus, matrix algebra, probability theory, linear regression, and statistical computing using R.

II Textbooks

Required Readings

The following textbook is required for this course. Please purchase it at the beginning of the semester:
Urdinez, Francisco, and Andres Cruz. 2020. *R for Political Data Science: A Practical Guide*. CRC Press.

Note that this is an *applied* R textbook; it introduces key concepts in statistical modeling without relying on too much mathematical notation or technical proofs. It also provides useful coding examples using the `tidyverse`. The content of this textbook (in combination with the other required readings listed below) covers the baseline of knowledge and skills that will be sufficient to successfully pass this course.

Recommended Readings

In addition to the required readings that help you pass this course, I strongly encourage you to work through the recommended readings to delve deeper into specific topics and understand them at a more technical level. The following textbook covers the nuts and bolts of linear and generalized linear models:

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

The author also provides an R companion that includes great coding examples (the only disadvantage being that they don't utilize the `tidyverse`):

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

In case you are looking for an even more rigorous treatment of linear and generalized linear models, I recommend the following textbook:

- Hansen, Bruce. 2022. *Econometrics*. Princeton University Press.

Additional recommended readings will be provided on Aula Global.

III Schedule

Note: The schedule and readings may be subject to change depending on our progress during the semester. Required readings (excluding the textbook itself) will be available on Aula Global.

1: Introduction (January 30)

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 review the material from AQMSS-I including a brief discussion of the final exam.

Recommended:

- ☐ Skim Urdinez and Cruz 2020, chs. 1-4.
- ☐ King, Gary. 2006. "Publication, publication." *PS: Political Science & Politics* 39 (01): 119–125

2: Regression and Causality (February 6)

This week starts with a big picture perspective by connecting this course to the remaining methods curriculum in MA program focused on research design and causal inference. We are also going to talk about some useful properties of OLS and how it can help us identify causal effects.

Required:

- ☐ Angrist, Joshua D, and Jörn-Steffen Pischke. 2008. *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press, chs. 1-3 (you can skip sections 3.3-3.5).

Recommended:

- ☐ Gailmard, Sean. 2014. *Statistical modeling and inference for social science*. Cambridge University Press, ch. 10

- ☐ Fox 2015, chs. 1, 5, 6
- ☐ Fox and Weisberg 2018, ch. 4
- ☐ Hansen 2022, chs. 1-3

3: Applied Regression Modeling and Interpretation (February 13)

In this week we continue our review of OLS properties and key assumptions in linear models. We are also going to revisit questions about model specification, coefficient interpretation, and hypothesis testing.

Required:
- ☐ Urdinez and Cruz 2020, ch. 5
- ☐ Check out `https://modelsummary.com/`, read the "Get started" vignette and work through the remaining tutorials.

Recommended:
- ☐ Fox 2015, chs. 7, 9
- ☐ Fox and Weisberg 2018, ch. 5
- ☐ Hansen 2022, ch. 4

4: Modeling Interactions and Simulating Quantities of Interest (February 20)

This week is all about turning complex regression results into easily interpretable visualizations by simulating quantities of interest. Special emphasis will be put on understanding regression models with interaction terms.

Required:
- ☐ *Problem Set 1 due on February 20.*
- ☐ King, Gary, Michael Tomz, and Jason Wittenberg. 2000. "Making the most of statistical analyses: Improving interpretation and presentation." *American journal of political science* 44 (2): 347–361
- ☐ Brambor, Thomas, William Roberts Clark, and Matt Golder. 2006. "Understanding Interaction Models: Improving Empirical Analyses." *Political Analysis* 14 (1): 63-82
- ☐ Check out `https://marginaleffects.com/`, read the "Get started" vignette and work through the remaining tutorials.

Recommended:
- ☐ Berry, William D, Matt Golder, and Daniel Milton. 2012. "Improving tests of theories positing interaction." *The Journal of Politics* 74 (03): 653–671
- ☐ Rainey, Carlisle. 2017. "Transformation-induced bias: Unbiased coefficients do not imply unbiased quantities of interest." *Political Analysis* 25 (3): 402–409
- ☐ Rainey, Carlisle. 2023. "A careful consideration of CLARIFY: simulation-induced bias in point estimates of quantities of interest." *Political Science Research and Methods*: 1–10
- ☐ Rittmann, Oliver, Marcel Neunhoeffer, and Thomas Gschwend. 2024. "How to Improve the Substantive Interpretation of Regression Results when the Dependent Variable is Logged." *Political Science Research and Methods*

5: Regression Diagnostics and Model Fit (February 27)

In this week we discuss various techniques to assess model fit as well as potential violations of regression assumptions. We will also briefly talk about using regression for case selection in qualitative research.

Required:
- ☐ Urdinez and Cruz 2020, ch. 6

Recommended:
- ☐ Fox 2015, chs. 11-13
- ☐ Fox and Weisberg 2018, ch. 8

- ☐ Hansen 2022, chs. 7, 9

6: Missing Data in Regression Models (March 6)

Social science data is often incomplete. This week introduces different strategies to deal with missing data and explains when and how multiple imputation can outperform listwise deletion. We will also discuss censoring and selection models in this context.

Required:

- ☐ *Problem Set 2 due on March 6.*
- ☐ Honaker, James, and Gary King. 2010. "What to do about missing values in time-series cross-section data." *American Journal of Political Science* 54 (2): 561–581
- ☐ Honaker, James, Gary King, and Matthew Blackwell. 2011. "Amelia II: A Program for Missing Data." *Journal of Statistical Software* 45 (7): 1-47

Recommended:

- ☐ Check out the documentation for the `mice` package in R (`https://amices.org/mice/`) and work through the tutorials. The package author also wrote a textbook on missing data imputation that is available online: `https://stefvanbuuren.name/fimd/`.
- ☐ Fox 2015, ch. 20
- ☐ Hansen 2022, ch. 27
- ☐ King, G., J. Honaker, A. Joseph, and K. Scheve. 2001. "Analyzing incomplete political science data: An alternative algorithm for multiple imputation." *American Political Science Review* 95 (1): 49–70
- ☐ Lall, Ranjit. 2016. "How multiple imputation makes a difference." *Political Analysis* 24 (4): 414–433
- ☐ Pepinsky, Thomas B. 2018. "A note on listwise deletion versus multiple imputation." *Political Analysis* 26 (4): 480–488
- ☐ Arel-Bundock, Vincent, and Krzysztof J Pelc. 2018. "When can multiple imputation improve regression estimates?" *Political Analysis* 26 (2): 240–245

7: Regression Models for Panel Data (March 13)

In this week we extend the linear regression framework to deal with hierarchical data structures such as time-series cross-sectional data. Specifically, we are going to discuss the distinction between fixed and random effects models as well as their application.

Required:

- ☐ Urdinez and Cruz 2020, ch. 7

Recommended:

- ☐ Fox 2015, ch. 23
- ☐ Fox and Weisberg 2018, ch. 7
- ☐ Hansen 2022, ch. 17
- ☐ Gelman, Andrew, and Jennifer Hill. 2007. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, chs. 11-13

8: Maximum Likelihood Estimation (March 20)

This week completes our discussion of linear models by introducing the likelihood theory of inference. Based on this foundation, we will develop a general framework to model various types of discrete data generating processes.

Required:

- ☐ *Problem Set 3 due on March 20.*
- ☐ King, Gary. 1998. *Unifying political methodology: The likelihood theory of statistical inference*. University of Michigan Press, chs. 1-4

Recommended:

- ☐ Fox 2015, Appendix D
- ☐ Hansen 2022, ch. 5

9: Binary Outcomes (March 27)

In this week we introduce logit and probit models for dichotomous outcomes. We will also talk about how to apply previously covered techniques for simulating quantities of interest in the context of generalized linear models.

Required:

- ☐ Urdinez and Cruz 2020, ch. 8

Recommended:

- ☐ Fox 2015, ch. 14.1
- ☐ Fox and Weisberg 2018, ch. 6
- ☐ Hansen 2022, ch. 25

10: Nominal and Ordinal Outcomes (April 3)

This week further extends our statistical toolkit to model nominal and ordinal outcomes.

Required:

- ☐ *Problem Set 4 due on April 3.*
- ☐ King, Gary. 1998. *Unifying political methodology: The likelihood theory of statistical inference*. University of Michigan Press, ch. 5

Recommended:

- ☐ Fox 2015, ch. 14.2
- ☐ Fox and Weisberg 2018, ch. 6
- ☐ Hansen 2022, ch. 26

11: Other Outcomes: Counts, Durations, ... (April 10)

This week provides a brief overview of modeling techniques for other data generating processes such as counts and durations.

Required:

- ☐ *Peer review due on April 10.*
- ☐ Urdinez and Cruz 2020, ch. 9

Recommended:

- ☐ Fox 2015, ch. 15
- ☐ Hansen 2022, ch. 26

Easter Break (April 17)

No class.

Session 12 (April 24): Final Project Presentations

In the last class of the semester, you will present your final research projects. This is another great opportunity to receive feedback before submitting the final project on May 8.

Required:

- ☐ *Problem Set 5 due on April 24.*

- ☐ *Final research project due on May 8.*
- ☐ Urdinez and Cruz 2020, ch. 9

Labor Day (May 1)

No class.

Buffer day (May 8)

Extra day in case one of the sessions listed above has to be canceled.

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 the statistical models discussed in the lecture.

V 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.

VI Acknowledgements

I have adapted the ideas and language from the work of several educators for this syllabus and the course material. For instance, I have borrowed liberally from other courses on social science research methods and statistics, as taught by Matthew Blackwell, Andrew Gelman, Thomas Gschwend, Gary King, Michael Peress, 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.

Arel-Bundock, Vincent, and Krzysztof J Pelc. 2018. "When can multiple imputation improve regression estimates?" *Political Analysis* 26 (2): 240–245.

Berry, William D, Matt Golder, and Daniel Milton. 2012. "Improving tests of theories positing interaction." *The Journal of Politics* 74 (03): 653–671.

Brambor, Thomas, William Roberts Clark, and Matt Golder. 2006. "Understanding Interaction Models: Improving Empirical Analyses." *Political Analysis* 14 (1): 63-82.

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.

Gelman, Andrew, and Jennifer Hill. 2007. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Hansen, Bruce. 2022. *Econometrics*. Princeton University Press.

Honaker, James, and Gary King. 2010. "What to do about missing values in time-series cross-section data." *American Journal of Political Science* 54 (2): 561–581.

Honaker, James, Gary King, and Matthew Blackwell. 2011. "Amelia II: A Program for Missing Data." *Journal of Statistical Software* 45 (7): 1-47.

King, G., J. Honaker, A. Joseph, and K. Scheve. 2001. "Analyzing incomplete political science data: An alternative algorithm for multiple imputation." *American Political Science Review* 95 (1): 49–70.

King, Gary. 1998. *Unifying political methodology: The likelihood theory of statistical inference*. University of Michigan Press.

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

King, Gary, Michael Tomz, and Jason Wittenberg. 2000. "Making the most of statistical analyses: Improving interpretation and presentation." *American journal of political science* 44 (2): 347–361.

Lall, Ranjit. 2016. "How multiple imputation makes a difference." *Political Analysis* 24 (4): 414–433.

Pepinsky, Thomas B. 2018. "A note on listwise deletion versus multiple imputation." *Political Analysis* 26 (4): 480–488.

Rainey, Carlisle. 2017. "Transformation-induced bias: Unbiased coefficients do not imply unbiased quantities of interest." *Political Analysis* 25 (3): 402–409.

Rainey, Carlisle. 2023. "A careful consideration of CLARIFY: simulation-induced bias in point estimates of quantities of interest." *Political Science Research and Methods*: 1–10.

Rittmann, Oliver, Marcel Neunhoeffer, and Thomas Gschwend. 2024. "How to Improve the Substantive Interpretation of Regression Results when the Dependent Variable is Logged." *Political Science Research and Methods*.

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