Regression Analysis for Political Science II

[84-702] – Spring 2024

Updated: January 16, 2024

Professor Jonathan Cervas
Email: cervas@cmu.edu
Location: POS Conference Room Posner Hall
Time: Tuesday 7:00p-9:50p Eastern
Office Hours Tuesday 6p-7p and by appointment (arrange via email)
CMU Academic Calendar

"*One can learn data analysis only by doing, not by reading*" - Kosuke Imai

How do we evaluate empirical claims about events we see around the world? Can we measure discrimination in job hiring? What is the best way to predict election outcomes? What factors drive the onset of civil wars? Is it possible to determine what members of Congress are more or less liberal given their voting record? These are just a few of the numerous question that social scientists are tackling with quantitative data. Beyond academia, companies and non-profits have invested heavily in data science techniques to learn about their users, platforms, and programs. Data scientists at these institutions are essentially applied social scientists and employ many of the same techniques you will learn in this course. The goal is to provide students with the foundation necessary to analyze data in their own research and to become critical consumers of statistical claims made in the news media, in policy reports, and in academic research.

Course Description

This course encompasses two primary objectives. Initially, it equips students with a comprehensive understanding of writing an good master's thesis. Students will gain insights into the essential components of the thesis and their significance, aiding them in developing their thesis proposals. By the conclusion of this course, students will have prepared a preliminary thesis draft. While this draft may not include the final data, learners will be expected to identify data sources and formulate a detailed strategy for completing their thesis.

The second objective focuses on providing students with quantitative methodologies necessary for investigating empirical queries pertinent to their research themes and hypotheses. The curriculum extends beyond the basics of Ordinary Least Squares regression, delving into models tailored for varied dependent variables. It addresses challenges like missing data, the utilization of textual and spatial data, and the intricacies of experimental and regression discontinuity designs. Additionally, the course explores simulations, the fundamental problem of social science research, i.e., causality, prediction models, and probability. While some topics may seem familiar, revisiting them is intended to enhance the understanding of these complex methods.

Learning Objectives

The primary aim of this course is to equip you with the necessary skills for successfully writing your thesis. As you progress through the semester, you will acquire a diverse set of analytical tools, essential for exploring the

empirical aspects of your research questions. Furthermore, the course is designed to hone abilities that will be invaluable in your subsequent roles as data specialists.

Upon successful completion of this course, you will have the proficiency to:

- Effectively engage in political science writing that meets professional standards.
- Navigate through the research process, recognizing various research methodologies and potential obstacles.
- Craft a robust research plan.
- Perform data cleaning tasks.
- Execute statistical evaluations.
- Formulate a comprehensive proposal for your master's thesis.

Prerequisite Knowledge

I assume that everyone enrolled in this course has satisfactorily completed RAPS I.

Material

Our primary text will be Data Analysis for Social Science by Kosuke Imai and Elena Llaudet. Problem Sets will be taken from material in ***DSS*** and from Imai's other book, Quantitative Social Science (*QSS*). They will be made available, but can also downloaded as a ZIP at https://press.princeton.edu/student-resources/data-analysis-for-social-science and https://press.princeton.edu/student-resources/quantitative-social-science

(***DSS***) Elena Llaudet and Kosuke Imai. 2022. Data Analysis for Social Science. Princeton University Press. https://press.princeton.edu/books/hardcover/9780691199429/data-analysis-for-social-science

(***Sage Research Methods***). *Project Planner*. 2017. https://methods.sagepub.com/project-planner/. DOI: 10.4135/9781526408495.

We may also use various material found on the internet. I offer some additional resources you may want to explore on your own if you want more advanced or specific methods.

Grading

I care the most about what you learn, not what numerical/letter summary of that learning you get at the end of the semester. So I would love to not have grades at all, but the university does not agree. Thus, we have grades to help encourage you to put effort into learning the course material, and to satisfy the university overlords.

Assessment

The course grade will be a weighted average of the following components:

Category	Percent of Final Grade
Problem Sets (~10)	25%
Thesis Reports	25%
Presentations	20%
Thesis Proposal	30%

R Programming

In this course, the R programming language will be utilized exclusively. You will be required to run all code within the R environment and employ RMarkdown for all Assignments. We will thoroughly explain these tools during the initial class session. It is essential that you have R, RStudio, and RMarkdown installed and operational on your computer by the second week of the course. To ensure your readiness, there is a brief Assignments with detailed instructions, which must be completed by the deadline specified on Canvas and in the schedule provided below.

Install R on your system. Follow these step-by-step instructions if you have issues: https://techvidvan.com/tutorials/install-r/

I have found this tutorial to be helpful. Please complete it before the second week of class: https://www.stephenpettigrew.com/r/

Here is another one, including an RMarkdown Demo: - https://www.miacosta.net/teaching/r-tutorials

Your ability to independently seek and access assistance is a vital aspect of your learning experience with a statistical programming language. If you encounter challenges with R or require support for statistical analysis during the course, please utilize the following resources in the specified order:

1. **In-App Help:** Within R, you can access help for any command by entering ? followed by the command in the R console. For example, ?summary provides details on the summary command. Please note that R's built-in help documentation can be complex.
2. **Online Search:** When facing specific questions, problems, or error messages in R, conduct a Google search followed by "in R" (e.g., "changing legend colors using barplot() in R"). Many individuals have likely encountered similar issues, and websites such as Stack Overflow often offer solutions to a wide range of queries. Additionally, you can seek assistance from ChatGPT at this stage, adhering to our AI usage policy. ***This should be your initial or secondary step before reaching out to the instructor or others for assistance.***
3. **rseek.org - rstats Search Engine:** Utilize rseek.org, a specialized search engine that filters Google results for R-related information.
4. **Course Resources:** Review the R code and tutorials shared within our course materials.
5. **Collaboration with Peers:** Embrace collaborative learning by engaging with your fellow students, as collaborative problem-solving is encouraged in this academic environment.
6. **Online Learning Materials:** Explore these online learning resources:
    - Cookbook for R
    - Quick-R: Home Page
    - R Koujue - Statistical Software - Research Guides

These resources have been curated and maintained by experts in the field, and you are encouraged to reach out to them directly for clarifications or additional guidance.

7. **Contacting Course Instructors:** In case you require further assistance, feel free to reach out to your course instructor via email or during their office hours. Ensure you provide a thorough description of your issue, along with your code and relevant screenshots. This information will assist them in understanding and addressing your problem. While you are always welcome to engage with your instructors for discussions on course content or any other academic matters, please make an earnest attempt to utilize the above resources for specific coding questions or R-related challenges. Instructors may inquire about your prior efforts when you reach out.

Swirl

Each week, students should utilize swirl lessons that are available at https://github.com/kosukeimai/qss-swirl and can be answered within R. The topics correspond to the book chapters. These are for your development and are ungraded.

To start the review questions, users must first install the swirl package and then the lessons for this book using the following three lines of commands. Note that this installation needs to be done only once at the beginning.

```
install.packages("swirl", repos='http://cran.us.r-project.org') # install the package
library(swirl) # load the package
install_course_github("kosukeimai", "qss-swirl") # install the course
swirl()
```

Assignments

Problem Sets

To only read about data science is about as useful as reading the entire DMV handbook and memorizing all state driving laws, and then to show up to a Formula 1 race expecting to win. You need to drive.

Thus, in this course, you will have problem sets to complete throughout the semester that will give you an opportunity to apply the statistical techniques you are learning. They will usually be focused on data analysis in general and will often involve a real dataset. I encourage students to rely on peer working groups as they work on these questions, but each student will submit their own work individually. We will have class time to work on these. You will turn in the problem set in the form of a ".Rmd" file on Canvas by Sunday at 11:59pm.

Presentations

Each student will be expected to give two presentations: One right after spring break evaluating the status of their proposal, and a final presentation of their proposal on the last day of class.

Final Thesis Proposal The main written product of the course is the Thesis Proposal, which will be signed off by your Thesis Advisor and uploaded to Canvas before May 13 to receive credit.

Tentative Course Schedule (Subject to Change as Semester Progresses):

Week 1. 01/16 - Course Introduction, Introduction to R and RStudio

- Readings:
  - ***DSS*** 1-1.6
- Key *R* Concepts:
  - +, -, *, /, <-, ", (), sqrt(), #
- Assignments:
  - Install *R*.
  - Practice basic arithmetic in *R*

Week 2. 01/23 - Observations and Variables, Computing and Interpreting Means

- Readings:
  - ***DSS*** 1.7-1.10
  - ***Sage Research Methods***: Overview, Philosophy of Research, Defining a Topic (including videos)
  - **CMIST Master's Thesis Guidelines**
- Key Concepts: dataframes, observations, variables, unit of observation, i, character vs. numeric variables, binary vs. non-binary variables, n, mean or average, $\sum$, unit of measurement
- Key *R* Concepts:
  - setwd(), read.csv(), View(), head(), dim(), $, mean()
- Assignments:
  - **Problem Set**: #1
  - Review the CMIST **"People"** page for potential Thesis Advisors. CMIST Core Faculty, Postdoctoral Fellows and Senior Lecturers can serve as primary thesis advisors; anyone (internal or external to CMIST) may serve as a secondary advisor. Arrange a meeting with them before Week 4 class.
  - *Canvas Submission*: Consider the important questions in the ***Sage Research Methods*** regarding access and constraints, as well as the final "Checklist." With these in mind, write a summary of your topic of research interest, why you think it is a "good" research topic, and any challenges you foresee. Also list two potential primary thesis advisors you plan to interview next week and why (review their profiles and CVs online), and any secondary advisors you have in mind.
- Class Activities:
  - We will discuss the syllabus, proposal expectations, and structure of a master's thesis.
  - Each student will introduce their possible research topic(s), associated challenges, and potential advisors.

Week 3. 01/30 - Estimating Causal Effects with Randomized Experiments, Does Social Pressure Increase the Probability of Turning Out to Vote?

- Readings:
  - ***DSS*** 2-2.7
- Key Concepts: causal relationships, treatment (X) vs. outcome variables (Y), potential outcomes, factual vs. counterfactual outcomes, fundamental problem of causal inference, individual vs. average causal effects, randomized experiments, random treatment Assignments, treatment and control groups, pre-treatment characteristics, the difference-in-means estimator
- Key *R* Concepts:
  - ==, ifelse(), []
- Assignments:
  - **Problem Set**: #2

Week 4. 02/06 - Survey Research and Exploring One Variable at a Time, Exploring the Relationship Between Two Variables

- Readings:

  - ***DSS*** 3-3.7
  - ***Sage Research Methods***: Developing a Researchable Question
- Key Concepts: sample, representative sample, random sampling, table of frequencies, table of proportions, histogram, descriptive statistics (mean, median, standard deviation, and variance), scatter plot, correlation
- Key *R* Concepts:
  - table(), prop.table(), hist(), median(), sd(), var(), ^, plot(), cor()
- Assignments:
  - **Problem Set**: #3
  - **Research Question & Hypothesis** (500-750 words):
    * Propose at least one research question and explain why it is both important and incompletely understood by existing research. Support your explanation with at least three sources, properly cited. Finally, propose at least one working hypothesis that you think should help to answer the research question.
    * Post your draft RQ on Canvas discussion board. Comment on each of your peer's RQs.

Week 5. 02/13 - Predicting Binary & Non-Binary Outcomes Using Linear Regression

- Readings:
  - ***DSS*** 4-4.4.9
- Key Concepts: prediction and correlation, predicted ($\widehat{Y}$) vs. actual outcome (Y), prediction errors ($\widehat{\epsilon}$), the least squares method, the linear regression model, $\widehat{Y} = \widehat{\alpha} + \widehat{\beta}X$, interpretation of coefficients, intercept ($\widehat{\alpha}$) and slope ($\widehat{\beta}$), $\triangle\widehat{Y} = \widehat{\beta}\triangle X$, interpretation of coefficients, intercept ($\hat{\alpha}$) and slope ($\hat{\beta}$), $\Delta Y = \hat{\beta}\Delta X$, R^2, relationship between R^2 and correlation
- Key *R* Concepts:
  - lm($Y \sim X$), abline()
- Assignments:
  - **Problem Set**: #4, #5
  - **Follow-up to Research Question & Hypothesis** (500-750 words)
    * Re-write your research question to be more specific, including units/measurement and scope (time, geography, etc).
    * Re-write your hypothesis(es), further defined and labeled H1, H2, etc.
    * Include variable definitions for the dependent and independent variables.
    * Summarize the data you intend to use in your research - it could be helpful to review the data used by authors in your literature review bibliography to see what units they used and how they measured them

Please make a 30min appointment with your thesis advisor to ensure your research question is "researchable" before proceeding with the next Assignments. Refer to the ***Sage Research Methods*** reading for what makes a question "researchable:" https://methods.sagepub.com/project-planner/developing-a-researchable-question

Week 6. 02/20 - Estimating Causal Effects with Observational Data and the Problem of Confounders, Controlling for Confounders Using Multiple Linear Regression

- Readings:
  - ***DSS*** 5-5.4.2
  - ***Sage Research Methods***: Literature Review
  - Knopf, Jeffrey W. 2006. "Doing a literature review." PS: Political Science and Politics 39(1): 127-132.
- Key Concepts: observational studies vs. randomized experiments, confounders (Z), interpretation of $\widehat{\alpha}$ and $\widehat{\beta}$ when X is binary and identifies treatment Assignments, multiple vs. simple linear regression models, new interpretation of coefficients
- Assignments:
  - **Problem Set**: #6
  - **Literature Review** (~1200 words, sources excluded):
    - Identify at least eight academic studies directly related to your research topic. Summarize each study, explaining its research question, central argument, and the main empirical findings (if applicable). Also, briefly mention how each paper contributes to our understanding of your topic and what questions they still leave unanswered. The template provided on Canvas can be a great resource when compiling and organizing your summaries of relevant literature, but the actual literature review should be a written report.

Week 7. 02/27 - Internal vs. External Validity

- Readings:
  - ***DSS*** 5.5-5.7
  - ***Sage Research Methods***: Research Design
  - McDermott, Rose. 2011. "Internal and External Validity." In Cambridge Handbook of Experimental Political Science, eds. James N. Druckman et al. Cambridge: Cambridge University Press, p. 27–40. https://www.cambridge.org/core/product/11F8092DCA6BECA1BBB922BFA08B4E23.
- Key Concepts: internal validity, external validity
- Assignments:
  - **Problem Set**: #7
  - **Theoretical Framework** (~750 words, excluding sources):
    - Theories are formulated to explain, predict, and understand phenomena and, in many cases, to challenge and extend existing knowledge, within the limits of the critical bounding assumptions. The theoretical framework is the structure that can hold or support a theory of a research study; it introduces and describes the theory which explains why the research problem under study exists. Develop a theoretical framework (including a conceptual framework; see reading) to explain a causal relationship between the variables you are researching. Make sure to support your argument on at least four sources

Week 8. 03/05 - SPRING BREAK

Week 9. 03/12 - Mid-semester Presentation (7-10min):

- Readings:

  - ***Sage Research Methods***: Research Ethics
  - Powner, Leanne. 2014. Empirical research and writing: A political science student's practical guide, Chapter 4.

Prepare a 5-slide presentation on your proposal progress thus far. Be sure to practice so it does not exceed 20min! This is a good chance to practice being succinct. Upload to Canvas. Include:

(1) Title slide (time to think of a draft title for your thesis!),
(2) RQ (& why it's important – hint: reference lit review),
(3) Hypotheses & H0,
(4) Threats to Validity & considerations, and
(5) Next steps.

Everyone will provide comments on how to move forward on each project.

Week 10. 03/19 - Probability

- Readings:
  - ***DSS*** 6-6.8
  - Grant, Cynthia & Osanloo, Azadeh. 2014. "Understanding, Selecting, and Integrating a Theoretical Framework in Dissertation Research." Administrative Issues Journal 4(2): 12-26.
- Key Concepts: probability, random variables, probability distributions, Bernoulli vs. normal distribution, the standard normal distribution, population parameters vs. sample statistics, the law of large numbers, the central limit theorem
- Assignments:
  - **Problem Set**: #8

Week 11. 03/26 - Hypothesis Testing with Coefficients

- Readings:
  - ***DSS*** 7-7.6
- Key Concepts: hypothesis testing, test statistic, standard error of $\hat{\beta}$
- Key *R* Concepts: summary()$coef
- Assignments:
  - **Problem Set**: #18

Week 12. 04/02 - Do Small Classes Increase Probability of Graduating from High School?, Do Women Promote Different Policies than Men?

- Readings:
  - ***DSS*** 7.7
  - ***Sage Research Methods***: Data Collection

- **Problem Set**: #9
- **Methodology** (~1000 words):
  - Describe two research designs, a qualitative and a quantitative one. You could follow only one in your thesis, but it is still a useful exercise that will help you to explore the potential empirical limits of your thesis. First, discuss the reasoning for your case selection and your data collection strategy. Make sure to justify the cases you will select, the data you will analyze, and the methods that you will use for testing your theory. If you have preliminary data, I encourage you to show it in tables and figures.
  - For the qualitative RD, discuss in detail whether you will conduct case studies, interviews, focus groups, ethnographic research, archival analysis, or process tracing. For example, if you plan to conduct interviews, make sure to describe who you will interview, why, and the type of questions you plan to ask. Also, whether the interviews will be structured, semi-structured, or unstructured, and closed-ended, open-ended, hypothetical, etc. For the quantitative portion, explain the units of analysis, how you will collect the data, and the research techniques that you will use to test your hypotheses. For example, explain whether you will conduct surveys, experiments, game theory, or statistical analyses. If you will use statistics, explain the techniques to be used, and justify them.

Week 13. 04/09 - Does Social Pressure Affect Turnout?

- Readings:
  - ***Sage Research Methods***: Planning and Practicalities
  - Assignments:
    - **Expected Findings** (500-750 words + visuals):
      - Building from the culmination of your work thus far describe what you expect to see in the outcomes of your research, based on your hypotheses. Include "dummy graphics" of data visualizations that show your expected findings using the resources above. Reiterate how you expect your findings to contribute to the broader literature on this topic, taking into account any remaining threats to validity that could not be controlled for.

Week 14. 04/16 - Is There Racial Discrimination in the Labor Market?

- Flex day
- Assignments:
  - **Problem Set**: #10
  - **Abstract** (500-750 words):
  - Now that you have completed the bulk of your proposal, it's time to write the introduction! The abstract is a brief but specific statement of the project's objectives, methods, and impact. The abstract should offer readers a glimpse of your intended work. Please address what you hope to accomplish, what methodological approaches you intend to utilize, what resources you intend to use, and why the project is important and relevant to you, to the field of political science, and to the global community

Week 15. 04/23 - Thesis Proposal Presentations

- Readings:
  - Refer to this online data visualization resource guide from our academic data librarian: https://guides.library.cmu.edu/data101/visualizingdata

- Assignments:
  - **Timeline**: Work closely with your advisor to develop a detailed timeline for your research and final thesis write-up including what you will do during summer break. This can be submitted in whatever format you choose (graphic, chart, list, calendar, etc.), but should include regular advisor meetings, hard deadlines and detailed descriptions of deliverables at each deadline.

Additional Resources:

- Gelman, Andrew, Jennifer Hill, and Aki Vehtari. 2020. Regression and Other Stories. Cambridge: Cambridge University Press. https://www.cambridge.org/core/books/regression-and-other-stories/DD20DD6C9057118581076E54E40C372C
- Alvarez, R. Michael, ed. 2016. Computational Social Science: Discovery and Prediction. Cambridge: Cambridge University Press. https://www.cambridge.org/core/books/computational-social-science/FB97BD1704D957183899DE120BEE2E4B
- Druckman, James N. 2022. Experimental Thinking: A Primer on Social Science Experiments. Cambridge: Cambridge University Press. https://www.cambridge.org/core/books/experimental-thinking/C43F73D2255BAD1CB47E39C05E51B399 Free PDF
- Grimmer, Justin, Margaret E. Roberts, and Brandon M. Stewart. 2022. Text as Data: A New Framework for Machine Learning and the Social Sciences. Princeton University Press. https://press.princeton.edu/books/hardcover/9780691207544/text-as-data
- Dunning, Thad. 2012. Natural Experiments in the Social Sciences: A Design-Based Approach. Cambridge: Cambridge University Press. https://www.cambridge.org/core/books/natural-experiments-in-the-social-sciences/96A64CBDC2A2952DC1C68AF77DE675AF
- Best, Henning, and Christof Wolf. 2023. The SAGE Handbook of Regression Analysis and Causal Inference. London. https://methods.sagepub.com/book/regression-analysis-and-causal-inference