GR5015 Data Analysis for the Social Sciences

Spring 2024
Thursdays 6:10PM–8:00PM
1102 International Affairs Building

Mike Z. He
Course Instructor
zh2263@columbia.edu
Office: 404B Altschul (Barnard)
Office Hours: Thursdays 4–6pm

Ellen Ren
Teaching Assistant
er2963@columbia.edu
Office: 404 Altschul (Barnard)
Office Hours: Mondays and Wednesdays 3–4pm

Course Overview

This course is meant to provide an introduction to probability and social statistics, tailored to the types of analyses and data issues encountered by QMSS students. The chief goal is to help students generate and interpret quantitative data in helpful and provocative ways. The hope is that by trying to measure the social world, students will see their thinking become clearer and their understandings of concepts grow more complex. They will also become competent at reading statistical results in social science publications and in other media.

Another important goal of the course is to teach students how to manipulate and analyze data themselves using statistical software. We will focus mainly on programming in both R and Python. There will be an R or Python write-up assignment every other week, although there will not be any physical lab sessions. These weekly assignments will be devoted to using these software programs to practice commands and to develop a paper using data of the student's choosing. The TA will hold additional weekly 1-hour lab sessions, and will lead students through labs, homework and/or lecture review.

Prerequisites: One semester of undergraduate statistics. Basic mathematics skills are assumed, and more advanced math will be introduced as needed.

Course Materials

We will be using as our textbook, Wooldridge, Jeffrey. 2008. *Introductory Econometrics: A Modern Approach*. South-Western College Pub; 4th Edition, ISBN=9780324581621. The book is on reserve at numerous libraries around the University. Feel free to use newer versions, if you prefer.

There are a number of helpful books for R:

- *Introduction to Econometrics with R*. Christoph Hanck, Martin Arnold, Alexander Gerber and Martin Schmelzer. At https://www.econometrics-with-r.org/index.html
- Using R for Introductory Econometrics. Florian Heiss. ISBN: 978-1-523-28513-6. Published in February 2016 at http://www.urfie.net/index.html. Consider the "Read Online" option
- *R for Data Science*. Garrett Grolemund & Hadley Wickham. at https://r4ds.had.co.nz

There are three very helpful Python books too:

- Downey, Allen B. *Think stats: exploratory data analysis*. O'Reilly Media, Inc., 2014, with the 2nd edition available for free download here.
- McKinney, Wes. *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython*. O'Reilly Media, Inc., 2017. 2nd edition. Information here.
- Sheppard, Kevin. "Introduction to Python for econometrics, statistics and data analysis." *University of Oxford, version* 3 (2017). Here for free download.

Other Readings. In some weeks, there will be additional readings from other sources.

Suggested Additional Readings. For more advanced students, additional readings can also be suggested, to see the concepts and methods in action in actual research articles and books – just let me know.

Course Requirements

The final grade of the course will be based of your fulfillment of each of the following requirements:

Attendance and Class Participation (10%): Students are expected to have read all the required readings before class, participate in class exercises and team assignments, and actively contribute to class discussion.

R/Python Lab Reports (30%): We will have R and/or Python data analysis assignments to write up and hand in. Students will be graded in terms of their ability to operate the program, select the most appropriate statistics for each type of analysis, interpret the statistics generated, and write brief summaries about what they have learned. In short, you will develop your own "social theory" using some data.

Midterm Exam (20%): We will have a midterm exam. This will include short answers and longer answer questions. More details will be announced later in the semester.

Independent Project (20%): There will be a semester-long independent project. This exercise will require students to integrate many of the skills and lessons learned throughout the semester into

a final research report, but more information will be given about this assignment as the semester progresses. The R and/or Python data analysis write-up assignments are designed to move the project along.

Group Final Presentation (15%): Towards the beginning of the semester (around Week 3), students will be assigned to teams composed of individuals with different skill sets, and will work with the group on various activities throughout the semester. This will culminate into a group presentation at the end of the semester on an analytical method of choice.

Group Peer Review (5%): To hold all group members accountable, you will complete a survey assessing each other's efforts at the end of the semester.

Late Submission Policy: All assignments are expected to be submitted on the due date. For every day after the submission date, 10% of the maximum grade will be deducted from the score.

All written work must be original and produced exclusively for this class. You are expected to follow the University's guidelines for the submission of written work.

Course Expectations

This course will be held fully in-person, and attendance will be taken for all sessions. Students should attend class each week and participate actively in class discussions and activities. Part of your course grade will be based on your attendance and participation in class (10% of the overall grade).

The University encourages individuals to stay home if sick, and as such, we will do all we can to help everyone stay connected to class and up-to-date on materials, if they must miss class due to illness or suspected illness.

Statement on Academic Integrity

Columbia's intellectual community relies on academic integrity and responsibility as the cornerstone of its work. Graduate students are expected to exhibit the highest level of personal and academic honesty as they engage in scholarly discourse and research. In practical terms, you must be responsible for the full and accurate attribution of the ideas of others in all of your research papers and projects; you must be honest when taking your examinations; you must always submit your own work and not that of another student, scholar, or internet source, **including generative AI software** (such as ChatGPT). Graduate students are responsible for knowing and correctly utilizing referencing and bibliographical guidelines. When in doubt, consult your professor. Citation and plagiarism-prevention resources can be found at the GSAS page on Academic Integrity and Responsible Conduct of Research.

Failure to observe these rules of conduct will have serious academic consequences, **up to and including dismissal from the university**. If a faculty member suspects a breach of academic honesty, appropriate investigative and disciplinary action will be taken following the Dean's Discipline procedures.

Statement on Disability Accommodations

If you have been certified by Disability Services (DS) to receive accommodations, please either bring your accommodation letter from DS to your professor's office hours to confirm your accommodation needs, or ask your liaison in GSAS to consult with your professor. If you believe that you may have a disability that requires accommodation, please contact **Disability Services** at 212-854-2388 or disability@columbia.edu.

Important: To request and receive an accommodation you must be certified by DS.

Note: The information in this syllabus is subject to change.

Course Schedule

Unless otherwise noted, the references below are to Wooldridge, 4e.

Class 1: Introduction to the Class (January 18)

Required Readings:

- Ch.19 – "Carrying Out an Empirical Project"

Class 2: Statistics, Data Structures, and Descriptive Statistics (January 25)

Required Readings:

- Ch. 1 – "The Nature of Econometrics and Economic Data"
- Appendices A.1 – "The Summation Operator and Descriptive Statistics"
- Appendices A.3 – "Proportions and Percentages"
- Appendices B.1 – "Random Variables and Their Probability Distributions"
- Appendices B.3 – "Features of Probability Distributions"

[Lab #1 due on January 31]

Class 3: Regression I – Simple Regression Model, Correlation, Goodness-of-Fit, and Hypothesis Testing (February 1)

Required Readings:

- Simple Regression Model
  - Ch. 2.1 – "Definition of the Simple Regression Model"
  - Ch. 2.2 – "Deriving the Ordinary Least Squares Estimates"
  - Ch. 2.4 – "Units of Measurement and Functional Form" (through "The Effects of Changing Units of Measurement on OLS Statistics")
- Correlation
  - Appendices B.2 – "Joint Distributions, Conditional Distributions and Independence"
  - Appendices B.4 – "Features for Joint and Conditional Distributions"
- Goodness-of-Fit and R^2
  - Ch. 2.3 – "Properties of OLS on Any Sample of Data"
- Hypothesis Testing
  - Ch. 4.1 – "Sampling Distributions of the OLS Estimators"
  - Ch. 4.2 – "Testing Hypotheses about a Single Population Parameter"
  - Ch. 4.3 – "Confidence Intervals"
  - Appendices B.5 – "The Normal and Related Distributions" (through "Additional Properties of the Normal Distribution")
  - Appendices C.5 – "Interval Estimation and Confidence Intervals"
  - Appendices C.6 – "Hypothesis Testing"

Recommended Other Readings:

- Abbott, Andrew. "Transcending general linear reality." *Sociological theory* (1988): 169-186.
- Berk, Richard A. "Chapter 1: Statistical Learning as a Regression Problem." *Statistical Learning from a Regression Perspective*. Springer International Publishing, 2016. 1-53.

Class 4: Regression II – Multiple Regression Analysis, Hypothesis Testing in Multiple Regression (February 8)

Required Readings:

- Multiple Regression Analysis
  - Ch. 3.1 – "Motivation for Multiple Regression"
  - Ch. 3.2 – "Mechanics and Interpretation of Ordinary Least Squares"
  - Ch. 3.3 – "The Expected Value of the OLS" (from "Assumption MLR.4" and on)
  - Ch. 6.1 – "Effects of Scaling on OLS"
  - Ch. 6.3 – "More on Goodness of Fit and Selection of Regressors"
- Hypothesis Testing in Multiple Regression
  - Ch. 4.4 – "Testing Hypotheses about a Single Linear Combination of Parameters"
  - Ch. 4.6 – "Reporting Regression Results"

Recommended Other Readings:

- Let's Put Garbage Can Regressions and Garbage Can Probits Where They Belong by Christopher H. Achen.
- Westfall, Jacob, and Tal Yarkoni. "Statistically controlling for confounding constructs is harder than you think." *PLOS One* 11.3 (2016): e0152719.

[Lab #2 due on February 14]

Class 5: Regression III – Log Transformations, Categorical-by-Continuous Interactions (February 15)

Required Readings:

- Log Transformations
  - Ch. 2.4 – "Units of Measurement and Functional Form" (only "Incorporating Nonlinearities in Simple Regression")
  - Ch. 6.2 – "More on Functional Form" (only "More on Using Logarithmic Functional Forms")
- Categorical-by-Continuous Interactions
  - Ch. 7.1 – "Describing Qualitative Information"
  - Ch. 7.2 – "A Single Dummy Independent Variable"
  - Ch. 7.3 – "Using Dummy Variables for Multiple Categories"
  - Ch. 7.4 – "Interactions Involving Dummy Variables"

Recommended Other Readings:

- Friedrich, Robert J. "In defense of multiplicative terms in multiple regression equations." *American Journal of Political Science* (1982): 797-833.

- Gustavsson, Sara, et al. "Regression models for log-normal data: comparing different methods for quantifying the association between abdominal adiposity and biomarkers of inflammation and insulin resistance." *International Journal of Environmental Research and Public Health* 11.4 (2014): 3521-3539.

[Lab #3 due on February 21]

Class 6: Regression IV – Continuous-by-Continuous Interactions, Quadratics, and F-Tests (February 22)

Required Readings:

- Ch. 6.2 – "More on Functional Form" (only "Models with Interaction Terms" and "Models with Quadratics")
- Ch. 4.5 – "Testing Multiple Linear Restrictions: The *F* Test"

Class 7: Regression V – The Gauss-Markov Assumption and Asymptotics, Specification and Data Issues (February 29)

Required Readings:

- The Gauss-Markov Assumption and Asympototics
    - Ch. 2.5 – "Expected Values and Variances of the OLS Estimators"
    - Ch. 3.3 – "The Expected Value of the OLS"
    - Ch. 3.4 – "The Variance of the OLS Estimators"
    - Ch. 3.5 – "Efficiency of OLS: The Gauss-Markov Theorem"
    - Ch. 5.1 – "Consistency"
    - Ch. 5.2 – "Asymptotic Normality and Large Sample Inference"
    - Ch. 5.3 – "Asymptotic Efficiency of OLS"
- More Specification and Data Issues
    - Ch. 8 – "Heteroskedasticity"
    - Ch. 9.1 – "Functional Form Misspecification"
    - Ch. 9.5 – "Missing Data, Nonrandom Samples and Outlying Observations"
    - Ch. 9.6 – "Least Absolute Deviations Estimation"

Recommended Other Readings:

Normality Assumption

- Lumley, Thomas, et al. "The importance of the normality assumption in large public health data sets." *Annual review of public health* 23.1 (2002): 151-169.

Heteroskedasticity

- King, Gary, and Margaret E. Roberts. "How robust standard errors expose methodological problems they do not fix, and what to do about it." *Political Analysis* 23.2 (2014): 159-179.
- Rigobon, Roberto, and Dani Rodrik. "Rule of law, democracy, openness, and income." *Economics of transition* 13.3 (2005): 533-564.

Outliers

- Ruiter, Stijn, and Nan Dirk De Graaf. "National context, religiosity, and volunteering: Results from 53 countries." *American Sociological Review* 71.2 (2006): 191-210.
- Van der Meer, Tom, Manfred Te Grotenhuis, and Ben Pelzer. "Influential cases in multilevel modeling: A methodological comment." *American Sociological Review* 75.1 (2010): 173-178.
- Ruiter, Stijn, and Nan Dirk De Graaf. "National religious context and volunteering: More rigorous tests supporting the association." *American Sociological Review* 75.1 (2010): 179-184.
- Jasso, G. (1985). Marital Coital Frequency and the Passage of Time: Estimating the Separate Effects of Spouses' Ages and Marital Duration, Birth and Marriage Cohorts, and Period Influences. *American Sociological Review*, 50(2):224-241.
- Kahn, J.R. and Udry J.R. (1986). Marital Coital Frequency: Unnoticed Outliers and Unspecified Interactions Lead to Erroneous Conclusions. *American Sociological Review*, 51(5):734-737.
- Jasso, G. (1986). Is It Outlier Deletion or Is It Sample Truncation? Notes on Science and Sexuality. *American Sociological Review*, 51(5):738-742.

Quantile Regression

- Budig, Michelle J., and Melissa J. Hodges. "Differences in disadvantage: Variation in the motherhood penalty across white women's earnings distribution." *American Sociological Review* 75.5 (2010): 705-728.
- Killewald, Alexandra, and Jonathan Bearak. "Is the motherhood penalty larger for low-wage women? A comment on quantile regression." *American Sociological Review* 79.2 (2014): 350-357.
- Budig, Michelle J., and Melissa J. Hodges. "Statistical models and empirical evidence for differences in the motherhood penalty across the earnings distribution." *American Sociological Review* 79.2 (2014): 358-364.

Missing Data Imputation

- Matthew Blackwell, James Honaker, and Gary King. 2017. "A Unified Approach to Measurement Error and Missing Data: Details and Extensions." *Sociological Methods and Research*, 46, 3, Pp. 342-369.

[Lab #4 due on March 6]

Class 8: Models for Binary Outcomes – Linear Probability Model and Binary Logistic Regression (March 7)

Required Readings:

- Ch. 7.5 – "A Binary Dependent Variable: The Linear Probability Model"
- Ch. 17.1 – "Logit and Probit Models for Binary Response"
- Appendix C.4 – "General Approaches to Parameter Estimation" (only "Maximum Likelihood")

Recommended Other Readings:

- Mood, Carina. "Logistic regression: Why we cannot do what we think we can do, and what we can do about it." European sociological review 26.1 (2010): 67-82.
- Allison, Paul. "In Defense of Logit – Part 1 and Part 2" MARCH 28, 2017. Statistical Horizons blog.
- Park, Hyeoun. "An introduction to logistic regression: from basic concepts to interpretation with particular attention to nursing domain." *Journal of Korean Academy of Nursing* 43.2 (2013): 154-164.
- Challenger accident example from Gary King

[Midterm due on March 10]

Class 9: Models for Count Outcomes and Survival Data – Poisson Regression and Cox Proportional Hazards Models (March 21)

Required Readings:

- TBD

[Lab #5 due on March 27]

Class 10: Bigger Issues with Hypothesis Testing, OLS and its Assumptions, with Possible Bayesian Improvements and Beyond (March 28)

Required Readings:

- Gelman, Andrew. "The failure of null hypothesis significance testing when studying incremental changes, and what to do about it." Apr 21 2017.
- Gill, Jeff. "The insignificance of null hypothesis significance testing." *Political Research Quarterly* 52.3 (1999): 647-674.
- Andrew Gelman, "Bayesian statistics: What's it all about?**"** on 13 December 2016, 8:47 pm. Statistical Modeling, Causal Inference, and Social Science
- Rasmus Bååth. "Bayesian First Aid: Two Sample t-test" February 24, 2014. R-Bloggers.

Class 11: First Differences Analysis (April 4)

Required Readings:

- Ch. 13.3 – "Two-Period Data Analysis"
- Ch. 13.4 – "Policy Analysis with Two-Period Panel"
- Ch. 13.5 – "Differencing with More Than Two Time Periods"

[Lab #6 due on April 10]

Class 12: Data Reduction Techniques – Scales, Factor Analysis, and Cluster Analysis (April 11)

Required Readings:

- Introduction to Cronbach's Alpha – Dr. Matt C. Howard

- Factor Analysis: Definition, Methods & Examples
- Principal Component Analysis (PCA) For Dummies
- K-means properties on six clustering benchmark datasets - Applied Intelligence

Class 13: Final Presentations (April 18)

Required Readings:

- None

Class 14: Final Presentations, Review, and Course Wrap-Up (April 25)

Required Readings:

- TBD

[Independent Project due on May 5]