SPRING 2017
POIR 611 Introduction to Regression Analysis
School of International Relations
University of Southern California
Schedule: Mondays 5 to 7.50pm, VKC 105

Instructor: Prof. Barberá email: pbarbera@usc.edu
Office Hours: Tuesdays, 2pm-4pm Office: VKC 359A

Course Description and Objectives:

This course introduces PhD students to quantitative analysis in the social sciences. At the end of the semester, you will be able to:

1. Read and evaluate quantitative research in Political Science and IR
2. Test hypotheses about relationships between variables using quantitative methods, including regression analysis
3. Read and manipulate data in multiple formats for large-n research projects
4. Understand what additional training and skills you will need to conduct research, and the thorough grounding necessary for self-teaching
5. Provide you with a working knowledge of R to facilitate 2-4.

The course is roughly divided in three parts. Weeks 1-5 focus on learning description and inference for a single variable. We will cover the basics of probability theory and hypothesis testing. Weeks 6-11 introduce the workhorse of quantitative analysis – linear regression. This part will focus on the derivation, estimation, and interpretation of the linear model, and then solutions to violations of the linear regression assumptions. The final weeks of the semester will discuss more advanced topics, including techniques for causal inference, matrix algebra, time series analysis, and data visualization.

The course will follow a "learning-by-doing" approach and will place emphasis on gaining experience in analyzing data with the R programming language. Students are expected to do the required readings for each week and practice running code before each session. The lectures will build upon the content of the readings with a series of short in-class assignments that will introduce new statistical and programming concepts, which will then be applied to the analysis of data from published research papers or common tasks in quantitative social science. Most of the applications will be related to Political Science and International Relations questions, but the course should be of interest to social science students more generally.

Course Requirements and Grading:

Class participation: 10%
Problem sets: 50%
Replication paper: 30%
Student presentation: 10%

Class participation

Students are expected to attend every session, and do the assigned readings before each session. You should come to class with questions and ready to engage in a discussion about that week's topic.

Problem sets

There will be a total of eight problem sets, each focusing on applying different statistical and programming concepts covered in the course to new datasets in R. Students are encouraged to collaborate but in the submission they must write up the code and answers on their own, and report the name of other students that they worked with in a footnote on the first page. Each problem set will be graded as check plus, check or check minus. Problem sets with a check minus can be resubmitted to increase their grade to check. Late problem sets will be penalized with a 10% grade penalty for each 24 hours after the deadline.

Replication paper

Students are required to submit a research paper that replicates and extends an existing piece of scholarship, either a published article or book chapter, by May 10. The goal of this exercise is to demonstrate that you have the ability to conduct quantitative research. Note that it will not be sufficient with just reproducing the tables and figures in the original paper, you also need to extend the analysis in some way. You can either conduct additional robustness checks, add new variables or observations, test a new a hypothesis with the author's dataset, etc. The submission should include a written report of around 8,000 words in length, as well as the R code and data required to run the analyses in the report. The report should contain a summary of the published research, a description of the methodology employed, the datasets used, a comparison of the published results and the student's replication, and a discussion of any issues that arose during the replication exercise. The paper or chapter you replicate must be chosen in consultation with the instructor – please send me a PDF copy of the research you will be replicate and a one-paragraph summary explaining your choice by February 15.

Student presentation

Students will present (and give feedback on) the results of their replication papers during the last session of the semester. Presentations will be less than 10 minutes in length (5-10 slides) and should include a summary of: theory and hypotheses, research design, comparison of original and replicated results, and discussion of potential issues. The final research report should incorporate the feedback received during this session.

Required Readings and Software:

The primary software for the course is R. Make sure you install R and RStudio before the first day of class. Previous programming experience at the level of POIR 610 is required.

The required textbooks for the course are:

- Imai, Kosuke (2016) *A First Course in Quantitative Social Science*. Princeton University Press. (QSS in the course outline below)
- Gujarati, Damodar and Porter, Dawn (2008) *Basic Econometrics*, 5th Edition. McGraw-Hill. (GP in the course outline below)
- Gelman, Andrew and Hill, Jennifer (2006) *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Cambridge University Press. (GH in the course outline below)

QSS is forthcoming at Princeton University Press. The author has kindly given us permission to use the textbook free of charge in advance of its official release. Please do not share the PDF outside the class or post it online. This will be the main textbook of the class, and we will read most of it, and work through the code provided by the author.

We will rely on GP as required reading for the second half of the course and will read around half of it. GP is an excellent introduction to econometrics and I recommend you acquire it. Old editions (and the international edition) should be cheaper to buy and will also suffice for the course.

Two of the required readings will correspond to chapters in GH. It is a more applied book and a great self-teaching tool to learn about multilevel models, but most of its content goes beyond the scope of this course.

There are many other excellent textbooks and articles that address the topics from the course in greater detail. The following is a list of recommended readings that you can rely on to supplement the required books:

- Diez, David M., Christopher D. Barr, and Mine Çetinkaya-Rundel. 2015. *Open-Intro Statistics*. 3rd edition. https://www.openintro.org/
- Wackerly, D., Mendenhall, W., & Scheaffer, R. L. (2007). *Mathematical statistics with applications*. Nelson Education.
- Wooldridge, J. M. (2015). *Introductory econometrics: A modern approach*. Nelson Education.
- Freedman, D., Pisani, R., & Purves, R. (2007). Statistics (4th edition). WW Norton and Company.
- Angrist, J. D., & Pischke, J. S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.
- Imbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

COURSE OUTLINE AND READINGS (subject to change)

Week 1. January 9. Course overview. Introduction to probability.

Probability theory: frequentist vs Bayesian frameworks. A review of set notation. Permutations and combinations. Calculating the probability of an event: the sample-point method and the event-composition method. Conditional, marginal, and joint probability. Bayes' rule.

R programming: review of for loops and functions.

Readings:
[required] QSS Chapters 6.1 and 6.2 (after class)
[application] Kosuke Imai, and Kabir Khanna. (2016). "Improving Ecological Inference by Predicting Individual Ethnicity from Voter Registration Record." *Political Analysis*, Vol. 24, No. 2 (Spring), pp. 263–272.

Week 2. January 16. No class (Martin Luther King's Day)

Problem set 1 due Tuesday January 17, 11.59pm

Week 3. January 23. Discrete and continuous probability distributions

The probability distribution for a discrete random variable. Bernoulli and binomial distributions. The probability distribution for a continuous random variable. Uniform and normal probability distributions. Expected value and variance of a random variable.

R programming: density, distribution function, quantile function and random generation from a random distribution.

Readings:
[required] QSS Chapters 6.3 and 6.4
[application] Arturas Rozenas. (forthcoming) "Detecting Election Fraud from Irregularities in Vote-Share Distributions," *Political Analysis.*

Week 4. January 30. Uncertainty

The law of large numbers. Central limit theorem. Bias and mean squared error of point estimators. Confidence intervals. Margin of error and sample size calculation in polls.

R programming: Monte Carlo simulations.

Readings:
[required] QSS Chapters 6.5 and 7.1
[application] Mosteller, Frederick (1997) "The Tennessee Study of Class Size in the Early School Grades", *Bulletin of the American Academy of Arts and Sciences* 50(7): 14-25.

Problem set 2 due Tuesday January 31, 11.59pm

Week 5. February 6. Hypothesis testing

Elements of a statistical test. Fisher exact tests. One-sample tests. Two-sample tests. Interpreting p-values. Power of tests.

R programming: computing t-tests.

Readings:
[required] QSS Chapter 7.2
[application] Beber, Bernd and Alexandra Scacco (2012) "What the Numbers Say: A Digit-Based Test for Election Fraud" *Political Analysis*, 20 (2): 211- 234.

Week 6. February 13. Linear regression.

Linear statistical models. The method of least squares. Assumptions of linear regression. The Gauss-Markov theorem. Multiple regression analysis: estimation and interpretation. Measuring model fit: R-squared and adjusted R-squared.

R programming: estimating linear regression models.

Readings:
[required] GP Chapters 3 and 7
[application] Fernández-Vázquez, Pablo, Barberá, Pablo, and Rivero, Gonzalo (2016) "Rooting Out Corruption or Rooting for Corruption? The Heterogeneous Electoral Consequences of Scandals", *Political Science Research and Methods* 4(2): 379-397.

Problem set 3 due Tuesday February 14, 11.59pm

Week 7. February 20. No class (President's Day)

Problem set 4 due Tuesday February 21, 11.59pm

Week 8. February 27. Linear regression with uncertainty.

Linear regression as a generative model. Standard errors of estimated coefficients. Confidence intervals for regression coefficients. Testing the overall significance of a multiple regression: the F test.

R programming: inference about predictions.

Readings:
[required] QSS Chapter 7.3
[recommended] GP Chapters 5 and 8.
[application] King, G., O. Rosen, M. Tanner, A. F. Wagner. (2008). "Ordinary Economic Voting Behavior in the Extraordinary Election of Adolf Hitler." *Journal of Economic History*, Vol. 68, pp. 951–996.

Week 9. March 6. Marginal effects. Simulation.

Scaling and units of measurement. Regression on standardized variables. Polynomial regression. Interactive models. The log-linear model. Consequences of model specification errors. Tests of specification errors. Overview of regression with binary dependent variables.

R programming: regression diagnostics.

Readings:
[required] GP Chapters 6 and 13
[recommended] GH Chapter 4.
[recommended] King, G. (1986). How not to lie with statistics: Avoiding common mistakes in quantitative political science. *American Journal of Political Science*, 666-687.

Problem set 5 due Tuesday March 7, 11.59pm

Week 10. March 27. Regression diagnostics. Model specification.

Analytical computation of marginal effects in regression models. Simulation of probability models. Summarizing linear regressions using simulation. Simulation for non-linear predictions.

R programming: Monte Carlo simulations applied to regression analysis.

Readings:
[required] GH Chapter 7 (skip 7.4)

[application] Franco, A., N. Malhotra, and G. Simonovits. 2014. "Publication Bias in the Social Sciences: Unlocking the File Drawer." *Science* 345 (6203): 1502–5.

Problem set 6 due Tuesday March 28, 11.59pm

Week 11. April 3. Multicollinearity and Heteroscedasticity. The matrix approach to linear regression.

The nature of multicollinearity. Theoretical and practical consequences of multicollinearity. The nature of heteroscedasticity. Detecting heteroscedasticity. Robust standard errors. Rudiments of matrix algebra. The matrix approach to the linear regression model.

R programming: robust standard errors

Readings:
[required] GP Chapters 10, 11, and Appendices B and C.

Week 12. April 10. Overview of causal inference.

The potential outcomes framework. Techniques for causal inference: natural experiments, matching, regression discontinuity design, difference-in-differences estimator, fixed effects.

R programming: matching, regression discontinuity, fixed effects.

Readings:
[required] GH Chapters 9 and 10.

Problem set 7 due Tuesday April 11, 11.59pm

Week 13. April 17. Time series analysis

OLS estimation in the presence of autocorrelation. Detecting autocorrelation. Stochastic processes. Unit root. Stationarity. Cointegration.

R programming: time series models in R

Readings:
[required] GP Chapters 12 and 21

Week 14. April 24. Data visualization. Course recap.

Principles for effective data visualization. How to avoid making misleading plots. What have we learned? Review of the course.

R programming: introduction to ggplot2

Readings:
[required] Kastellec, J. and E. Leoni (2007) "Using graphs instead of tables in political science" *Perspectives on Politics*
[recommended] Hughes, A. (2015) "Visualizing inequality: How graphical emphasis shapes public opinion" *Research and Politics*.

Problem set 8 due Tuesday April 25, 11.59pm

Week 15. May 5th (FRIDAY, noon-2pm, VKC 300A). Student presentations.

Academic Conduct:

Plagiarism – presenting someone else's ideas as your own, either verbatim or recast in your own words – is a serious academic offense with serious consequences. Please familiarize yourself with the discussion of plagiarism in *SCampus* in Section 11, *Behavior Violating University Standards* https://scampus.usc.edu/1100-behavior-violating-university-standards-and-appropriate-sanctions. Other forms of academic dishonesty are equally unacceptable. See additional information in SCampus and university policies on scientific misconduct, http://policy.usc.edu/scientific-misconduct.

Discrimination, sexual assault, and harassment are not tolerated by the university. You are encouraged to report any incidents to the *Office of Equity and Diversity* http://equity.usc.edu or to the *Department of Public Safety* http://adminopsnet.usc.edu/department/department-public-safety. This is important for the safety of the whole USC community. Another member of the university community – such as a friend, classmate, advisor, or faculty member – can help initiate the report, or can initiate the report on behalf of another person. *The Center for Women and Men* http://www.usc.edu/student-affairs/cwm/ provides 24/7 confidential support, and the sexual assault resource center webpage http://sarc.usc.edu describes reporting options and other resources.

Support Systems:

A number of USC's schools provide support for students who need help with scholarly writing. Check with your advisor or program staff to find out more. Students whose primary language is not English should check with the *American Language Institute* http://dornsife.usc.edu/ali, which sponsors courses and workshops specifically for international graduate students. If an officially declared emergency makes travel to campus infeasible, *USC Emergency Information* http://emergency.usc.edu will provide safety and other updates, including ways in which instruction will be continued by means of blackboard, teleconferencing, and other technology.

The *Office of Disability Services and Programs* (https://dsp.usc.edu/, DSP) provides certification for students with disabilities and helps arrange the relevant accommodations. Any student requesting academic accommodations based on a disability is required to register with DSP each semester. A letter of verification for approved accommodations can be obtained from DSP. Please be sure the letter is delivered to me as early in the semester as possible. DSP is located in GFS 120 and is open 8:30 a.m. – 4:30 p.m., Monday through Friday. The phone number for DSP is (213) 740-0776.