FALL 2016
IR 312 Introduction to Data Analysis
School of International Relations
University of Southern California
Schedule: Tues, Thu 3:30 to 4.50pm, VKC 152

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

Course Description and Objectives:

Are democratic countries less likely to engage in interstate disputes? Do cash transfer programs reduce poverty in developing countries? What factors predict bilateral trade flows? Is it possible to detect electoral fraud just by looking at the distribution of vote counts across districts? Has income inequality increase across and within countries over the past few decades? Academic researchers and policy-makers increasingly rely on quantitative methods to answer these questions. As the sheer volume of data available grows, the ability to analyze data, interpret the results, and effectively communicate key findings has become an essential skill to conduct empirical research in the social sciences. The ability to extract valuable insights from quantitative data – often referred to as "data science" – is also a common demand by employers in the private sector.

This course aims at filling this gap by providing students with the fundamental statistical and computing skills necessary to apply data analysis methods to the study of International Relations. After the course, students will be able to:

1. Read and process data in multiple formats and conduct basic descriptive analysis
2. Create effective visual depictions of statistical patterns in data using visualization techniques
3. Formulate hypotheses about relationships between variables and test them with multivariate regression analysis in order to answer research questions
4. Measure uncertainty using confidence intervals and interpret these results building upon basic concepts in probability theory
5. Identify potential challenges to causal inference and design empirical strategies to address them
6. Conduct basic analyses of new types of data, such as text and networks

The course will follow a "learning-by-doing" approach and will place emphasis on gaining experience in analyzing data. Students are expected to do the required readings for each week and run the 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 data science. Most of the applications will be related to International Relations questions, but the course should be of interest to social science students more generally.

Students will learn how to use the open-source software R, a popular statistical programming environment, described by The New York Times as "the lingua franca" of data analysis in corporations and academia for its flexibility, scalability, and ease of use. R is freely available for download and runs on Macintosh, Windows, and Linux computers. It is much more powerful than other statistical software, such as Excel, SPSS, and STATA, but it is a bit more difficult to learn. No previous programming experience is required, and a variety of resources will be made available to the students in order to learn R as efficiently as possible.

Course Requirements and Grading:

Class participation: 15%
Problem sets: 50%
Midterm exam: 15%
Final exam: 20%

Class participation

Students are expected to attend every session, and do the assigned readings before each session. Doing the assigned readings implies not only reading the required pages of the book, but also running the code on their own, including the additional *swirl* exercises, and come to class with questions and ready to engage in a discussion about that week's topic. Every session will begin with an in-class quiz about the content of the readings. Participation in these quizzes is required, but they will not be graded.

Problem sets

There will be a total of seven 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 on a 0 to 10 scale.

Midterm and final exams

The midterm exam will be a 24-hour take-home exam to be posted on Blackboard at 11.59pm PST on October 12 and due at 11.59pm PST on October 13 (week 8 of the class). The final exam will be administered in class during the assigned time by the University after the end of the semester. Both tests will be cumulative and open-book, and will assess how well students understand the key statistical and programming concepts in the class up to that point in the semester. They will include both short conceptual questions and longer questions where students will be asked to interpret the output of a statistical analysis.

Online Forum:

We will have a dedicated Discussion Board on the course Blackboard site for discussion and student questions. The details will be announced at the beginning of the semester. All course-related questions (especially those related to coding in R) should be asked here. This way all students will be able to benefit from seeing the questions and answers. You will have the option to post a question anonymously.

Required Readings:

Imai, Kosuke (2016) *A First Course in Quantitative Social Science*. Princeton University Press. (QSS in the course outline below)

Monogan, Jamie (2015) *Political Analysis Using R*. Springer. (PAUR 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 all the code provided by the author.

The e-book version of PAUR is freely available through the USC library website. You can also purchase a discounted softcover edition of the book (at only $25) through the Springer "MyCopy" system. This book will complement the main textbook and offer you additional opportunities to improve your programming skills.

COURSE OUTLINE AND READINGS

Week 1. Aug. 23 & 25. Introduction

Why data analysis? Overview of the course and logistics. Introduction to R and first steps with RStudio.

Readings:
[required] QSS Ch. 1 (pages 9-30) & swirl exercises INTRO1 and INTRO2
[recommended] PAUR Ch. 1 (pages 1-12)
Application: Understanding world population dynamics

Week 2. Aug. 30 & Sep. 1. Univariate analysis

What is a variable? Types of variables. Describing a variable with summary and descriptive statistics. Visualizing a variable using barplots, histograms, and density plots.

Readings:
[required] QSS Ch. 2.6 (pages 69-75) & Ch. 3.1 – 3.3 (pages 73-84) & swirl exercises MEASUREMENT1
[recommended] PAUR Ch. 3 (pages 33-52)

Application: Measuring public opinion. Political Efficacy in China and Mexico.
Reading related to application: King, Gary, Christopher J.L. Murray, Joshua A. Salomon and Ajay Tandon (2004) "Enhancing the Validity and Cross-Cultural Comparability of Measurement in Survey Research." *American Political Science Review*, 98 (1): 191-207.

Problem set 1 due on Sep 2, 11.59pm.

Week 3. Sep. 6 & 8. Data manipulation

Reading data in different formats into R. Logical statements. Subsetting data. Recoding and generating new variables.

Readings:
[required] QSS Ch. 2.1 – 2.2 (pages 35-52) & swirl exercises CAUSALITY 1
[recommended] PAUR Ch. 2 (pages 13-32)

Application: The Efficacy of Small-Class Size in Early Education
Reading related to 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.

Week 4. Sep. 13 & 15. Causality.

Understanding causality: counterfactuals. Experimental and observational studies: internal validity, external validity, and ethical concerns. Endogeneity. Omitted variable bias. Estimating treatment effects with R.

Readings:
[required] QSS Ch. 2.4 – 2.7 (pages 53-75) & swirl exercises CAUSALITY 2.
Application: Indiscriminate Violence and Insurgency
Reading related to application: Lyall, J. 2009. "Does Indiscriminate Violence Incite Insurgent Attacks?: Evidence from Chechnya." *Journal of Conflict Resolution* 53(3): 331–62.

Problem set 2 due on Sep 16, 11.59pm.

Week 5. Sep. 20 & 22. Bivariate analysis.

Correlation coefficients and scatterplots to understand the relationship between two continuous variables. Frequency tables for two categorical variables. Conditional means for a categorical variable and a continuous variable.

Readings:
[required] QSS Ch. 3.6 – 3.7 (pages 106-125) & swirl exercises MEASUREMENT 2.
[required] PAUR Ch. 5.2 – 5.3 (pages 71-75)
Application: Voting in the United Nations Assembly General.
Reading related to application: Bailey, Michael A., Anton Strezhnev, and Erik Voeten. Estimating Dynamic State Preferences from United Nations Voting Data." *Journal of Conflict Resolution*, forthcoming.

Week 6. Sep. 27 & 29. Measurement.

Working with survey data. Types of survey sampling. Challenges: sample selection bias, unit and item non-response bias, social desirability bias. Basics of modular programming: loops, if/else statements.

Readings:
[required] QSS Ch. 3.4 (pages 96-103) and 4.1 (pages 133-150) & swirl exercises PREDICTION 1.
Application: Predicting the outcome of the 2008 U.S. Presidential elections using public opinion polls.

Problem set 3 due on Sep 30, 11.59pm.

Week 7. Oct. 4 & 6. Regression (I)

Linear regression and least squares. Estimating predicted values.

Readings:
[required] QSS Ch. 4.2 (pages 150-174) & swirl exercises PREDICTION2
[recommended] PAUR Ch. 6 (pages 79-98)

Application: The Electoral Effects of Corruption Scandals
Reading related to application: Fernández-Vázquez, P.; Barberá, P., and Rivero, G. (2016) "Rooting Out Corruption or Rooting for Corruption? The Heterogeneous Electoral Consequences of Scandals", Political Science Research and Methods 4(2): 379-397.

Week 8. Oct. 11 & 13. Regression (II). MIDTERM EXAM.

Regression and causality. Understanding model fit: residuals, R squared.

No readings: course review.

Midterm exam due on Oct 15, 11.59pm.

Week 9. Oct. 18 & 20. Regression (III)

Multivariate regression. Understanding marginal effects.

Readings:
[required] QSS Ch. 4.3 (pages 174-196) & swirl exercises PREDICTION3
Application: The gravity model of international trade.
Reading related to application: Bergstrand, Jeffrey H. (1985) "The gravity equation in international trade: some microeconomic foundations and empirical evidence." *The Review of Economics and Statistics*, 67 (3): 474-481.

Week 10. Oct. 25 & 27. Probability

Probability: frequentist vs Bayesian, permutations, sampling with and without replacement. Random variables and probability distributions. Law of large numbers and the central limit theorem.

Readings:
[required] QSS Ch. 6.1, 6.3 and 6.4 (excerpts) & swirl exercises PROBABILITY1
Application: Electoral forensics. Detecting election fraud in Russia and Nigeria.
Reading related to application: Beber, Bernd and Alexandra Scacco (2012) "What the Numbers Say: A Digit-Based Test for Election Fraud" *Political Analysis*, 20 (2): 211-234.

Problem set 4 due on Oct 29, 11.59pm.

Week 11. Nov. 1 & 3. Uncertainty

Standard error and confidence intervals. Hypothesis testing: one-sample and two-sample tests. Linear regression with uncertainty.

Readings:
[required] QSS Ch. 7 (pages 339-422) & swirl exercises UNCERTAINTY1, 2, 3
Application: Sex ratios and the price of agricultural crops in China.
Reading related to application: Qian, Nancy (2008) "Missing Women and the Price of Tea in China: The Effect of Sex-Specific Earnings on Sex Imbalance." *Quarterly Journal of Economics*, 123 (3): 1251-1285.

Problem set 5 due on Nov 5, 11.59pm.

Week 12. Nov. 8 & 10. Networks.

Introduction to network analysis: network components (nodes/edges), types of networks (directed/undirected), centrality measures, community detection, network visualization with R.

Readings:
[required] QSS Ch. 5.2 (pages 220-235) & swirl exercises DISCOVERY2
Application: International trade networks.
Reading related to application: De Benedictis, Luca, and Lucia Tajoli (2011) "The World Trade Network." *The World Economy*, 34 (8): 1417-1454.

Week 13. Nov. 15 & 17. Text

Importing text data into R. Term frequencies and word clouds. Basics of regular expressions. Topic discovery.

Readings:
[required] QSS Ch. 5.1 (pages 203-219) & swirl exercises DISCOVERY1
Application: How world leaders communicate on social media.
Reading related to application: Barberá, Pablo and Thomas Zeitzoff (2016) "The New Public Address System: Why Do World Leaders Adopt Social Media?" *Working paper.*

Problem set 6 due on Nov 19, 11.59pm.

Week 14. Nov. 22. Guest lecture.

Data analysis with R at Facebook, by Drew Dimmery (Facebook data scientist)

Readings: TBA

Week 15. Nov. 29 & Dec. 1. Data visualization. Course recap.

Principles for effective data visualization. Customizing plots in R: adding a legend, changing the size and color of figure elements, generating figure facets. How to avoid making misleading plots. What have we learned? Review of the course.

Readings:
[required] Kastellec, J. and E. Leoni (2007) "Using graphs instead of tables in political science" *Perspectives on Politics*
[recommended] QSS Ch. 8 (pages 373-376)
Application: Visualizing income inequality within and across different countries.

Reading related to application: Hughes, A. (2015) "Visualizing inequality: How graphical emphasis shapes public opinion" *Research and Politics*.

Problem set 7 due on Dec 3, 11.59pm.

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. The *Office of Disability Services and Programs* http://sait.usc.edu/academicsupport/centerprograms/dsp/home_index.html provides certification for students with disabilities and helps arrange the relevant accommodations. 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.