Poli 170A: Applied Data Analysis for Political Science

Summer Session I 2019

Instructor: Brandon Merrell, bmerrell@ucsd.edu

Lectures: Tuesdays and Thursdays, 8:00–10:50 in Center 201

Office Hours: Tuesdays 11:45am–1:00pm (and by appointment) in SSB 448

Online Content: http://TritonEd.ucsd.edu

Description: The growth of data analytics is changing the way policymakers, businesses, and individuals operate. Across all these categories, people are increasingly able to pose and answer complex questions using new evidence.

This course introduces students to various statistical techniques used in social science research. Part of this task is conceptual: helping students to think sensibly and systematically about research design. To this end, students will learn how theory, measurement, and data fit together to help us understand the concepts we care about. But part of our task is practical as well: students will interact with data, analyze it using the various techniques we cover, and present their findings in a paper and presentation for class. As part of this process, students will utilize "R," which is a powerful, flexible, and commonly known statistical tool in social science research. Our overall 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 policy reports, the news media, and academic research.

As a purely practical matter, this class is highly recommended for students who plan to write senior theses or pursue graduate work in the social sciences: the statistical and computing material you learn will be helpful for those undertaking such projects.

Prerequisites: This course does not have any formal prerequisites. Although the class is relatively heavy on applied statistical techniques, the mathematical demands are relatively light; our focus will very much be on helping students understand the intuition of the methods we use and helping them interpret the outputs of these techniques rather than their internal workings.

Because some students may be newcomers to statistical analysis, we will begin with an overview of foundational material that is necessary for all social science research. However, because this is an upper-division course and many students will already have completed other coursework in data science, our pace will be much quicker than POLI 30.

Software: We use the open-source statistical software R (http://www.r-project.org). R can be more powerful and flexible than other statistical software, such as SPSS and Stata, but it can also be more difficult to learn. We will also use RStudio (http://www.rstudio.com), a user interface that simplifies common operations. You should install both programs before the first day of class. If you have trouble, contact the TA. For help with R, I recommend Kosuke Imai's *Quantitative Social Science* textbook, which is used in several courses at UCSD, as well as Peter Dalgaard's *Introductory Statistics with R* and Andrew Gelman and Jennifer Hill's *Data Analysis Using Regression and Multilevel/Hierarchical Models.* I also recommend the following **free resources**:

- https://r4ds.had.co.nz
- https://www.jaredknowles.com/r-bootcamp/
- https://cran.r-project.org/web/packages/IPSUR/vignettes/IPSUR.pdf
- https://www.youtube.com/user/TheLearnR/videos
- https://www.rstudio.com/online-learning
- https://www.coursera.org/collections/learn-r
- https://stackoverflow.com/

Rules and Requirements: The course requirements consist of a closed-note exam (40%, July 25th), participation (15%), and a research project (45%). You must earn a passing grade on both the final and the research paper to pass the overall course. I use the following grading scale: "A-" = [90-93.$\bar{3}$), "A" = [93.$\bar{3}$-96.$\bar{6}$), "A+" = [96.$\bar{6}$-100], with other letter grades following analogous intervals.

Research Project: Your assignment is to write a 8-12 page paper in which you develop a theory and either conduct or propose a research design for evaluating that theory using empirical data. The project includes a 1-2 page research proposal (5%), a 15 minute presentation (15%), and the submission of a final research paper (25%).

Papers must be submitted to TurnItIn.com via TritonEd on Saturday, August 3rd before 6:00pm local time. Papers should be typed, 1.5-spaced, with size 12 font in a standard typeface. I am happy to read drafts of the paper throughout the term.

Academic Dishonesty: All work must be completed by the individual to whom it is assigned. Students are not permitted to use unauthorized assistance of any kind. Any student who is caught cheating or plagiarizing will receive a failing grade for the course and will be reported to the Academic Integrity Office for administrative sanction.

Late Assignments and Missed Exams: Make-up assignments are only offered under valid and documented circumstances. If you know you will miss an exam for a legitimate reason, notify me at least a week in advance. Email is perfectly acceptable. If you cannot contact me in advance, you must do so as soon as possible. I will work with you to resolve reasonable problems, but it is your responsibility to arrange with me to take a makeup exam. All make-up work must be submitted 48 hours prior to the grade submission deadline.

Attendance: Class attendance is not mandatory but will probably improve your performance on assignments. Some material is also easier to learn when you hear someone explain it and/or when you have an opportunity to discuss it with others.

Grades and Appeals: You will be graded solely on your academic performance. This includes clarity of thought, knowledge of the material, composition, spelling, and grammar. Students can appeal grades that they believe are incorrect. Grade appeals will consist of a single typed page that identifies the problem and presents a reasoned argument that the grade fits the appeal criteria.

Disability: Students who will request accommodations should register with the Office for Students with Disabilities (University Center 202; 858.534.4382) and provide me with documentation outlining appropriate accommodations. I am happy to meet with you during my office hours to discuss your needs.

Course Schedule:

Part I: Fundamentals of Data Analytics

Meeting #1: Introduction, Theories, and Measurement (Tuesday, July 2nd)

- Course introduction and logistics
- Data analysis examples
- Requirements and assignments
- What are theories and why are they important?
- Characteristics of good theories
- Concepts and units of analysis
- Operational definitions
- Measurement error
- Reliability and validity
- Types of variables

No Class (July, 4th holiday)

Meeting #2: Causality, Experiments, Sampling, and Surveys (Tuesday, July 9th)

- Making comparisons
- Association, prediction, and causation
- Counterfactuals and types of causality
- Inference errors
- Causal mechanisms
- Optimal experiments
- Comparability and random assignment
- Problems and challenges in conducting experiments
- Natural experiments
- Samples and sample statistics
- Selection effects, representativeness, and randomization
- Observer effects and social desirability bias
- Other challenges with surveys

Meeting #3: Descriptive Statistics and Hypothesis Testing (Thursday, July 11th)

- Summarizing data
- Averages and central tendency
- Dispersion and standard deviation
- Describing distributions
- Normal distributions and z-scores
- Sampling distributions and standard errors
- The central limit theorem and t-scores
- Null and alternative hypotheses
- Statistical significance and type-I or type-II errors
- Chi-square tests, difference of means, difference of proportions
- Confidence intervals

Meeting #4: Linear Regression Assumptions and Diagnostics (Tuesday, July 16th)

- Graphing bivariate relationships
- Linear and non-linear relationships
- Scatterplots, correlation, and linear association
- Regression and prediction
- Measuring goodness of fit
- Multiple regression, interaction effects, and F-tests
- Outliers, residuals, heteroscedasticity, and multicollinearity

Meeting #5: Causal Inference Approaches (Thursday, July 18th)

- Matching
- Regression Discontinuity Design
- Difference-in-Difference
- Instrumental Variables
- Selection Models

Individual Meetings and Project Progress Check) (Tuesday, July 23rd)

Meeting #6: Exam (Thursday, July 25th)

Meeting #7: Binary, Ordinal, Nominal, and Count Dependent Variables (Tuesday, July 30th)

- Binary outcome models: logit and probit
- Odds ratios and predicted probabilities
- Ordered outcome models
- Nominal outcome models
- Count outcome variables
- Poisson distribution
- Overdispersion and contagion
- Negative binomial and zero-inflated poisson
- Presenting results and introduction to Zelig

Meeting #8: Duration Models and Panel Data (Thursday, August 1st)

- Duration dependent variables
- Exponential distribution, Weibull distribution, and hazard rates
- Censoring, proportional hazards, and time-varying covariates
- Multilevel or hierarchical data
- Panel and time-series cross-section data
- Unit similarity and exchangeability
- Fixed effects and random effects
- Modeling serial autocorrelation
- Course recap and next steps

Meeting #9: Student Research Presentations (Saturday, August 3rd)