Advanced Methods for Social Media and Text Data

Christopher Fariss (cjfariss@umich.edu)

Introduction to the Course

Course Content

This course focuses on the research design and data analysis tools used to explore and understand social media and text data. The fundamentals of research design are the same throughout the social sciences, however the topical focus of this class is on computationally intensive data generating processes and the research designs used to understand and manipulate such data at scale. By massive or large scale, I mean that there are lots of subjects/connections/units/rows in the data (e.g., social network data like the kind available from twitter), or there are lots of variables/items/columns in the data (e.g., text data with many thousands of columns that represent the words in the document corpus), or the selected analytical tool is a computationally complex algorithm (e.g., a Bayesian simulation for modeling a latent variable, a random forest model for exploratory data analysis, or a neural network for automatically classifying new observations), or finally some combination of these three issues. The course will provide students with the tools to design observational studies and experimental interventions into large and unstructured data sets at increasingly massive scales and at different degrees of computational complexity.

How will we go about learning these tools? In this class, we will learn to program and program to learn. What do I mean? First, we will use the R program environment to learn the building blocks of programming. As we develop these programming skills, we will use them to help us understand how different types of data analysis tools work. We will start very small and learn how to scale up. In the beginning of the course, we will not make use of many packages other than the base packages available by default in R. As we proceed, we will learn how models for data work before then investigating the functions that exist in the large, always increasing catalogue of packages available for you to use in R. The development of new functions in R is advancing rapidly. The tools you learn in this class will help you improve as a programmer and a data scientists but learning how to program and using your programming skills to learn how to analyze data.

Course Objectives

Students will lean how to design studies to take advantage of the wealth of information contained in new massive scale online datasets such as data available from twitter and many newly digitized document corpuses now available online. The focus of the course is on learning to program in R and designing studies in such away as to maximize the validity of inferences obtained from these complex datasets.

Course Prerequisites

Students should have some familiarity with concepts from research design and statistics. Generally, exposure to these concepts occurs during the first year course at a typical PhD program in political science. Students should also have some familiarity with the R computing environment. The more familiarity with R the better.

Course Details

- We will begin each class period with a "programming challenge" (approximately 20-25 minutes).
- I will then each give a short lecture over the class material (approximately 30-45 minutes).
- The remaining portion of class (approximately 1.5-2 hours) will be devoted to hands on learning with R, simulated data, and the large scale datasets provided by the instructor.
- The course schedule section, which is below, provides even more details about the topic of the lecture for each class day, citations for the discussion readings, and chapter entries from the text books for the lab portions of the class.

Required Readings

1. Efron, Bradley and Trevor Hastie. 2016. *Computer Age Statistical Inference* Cambridge University Press. https://web.stanford.edu/~hastie/CASI/
2. Gelman, Andrew and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.
3. Matloff, Norman. 2011. *Art of R Programming: A Tour of Statistical Software Design*. no starch press.

Suggested Readings

4. Bolker, Ben. 2007. Ecological Models and Data in R. Princeton NJ: Princeton University Press.
5. Stan Development Team. 2018. "Stan Modeling Language: User's Guide and Reference Manual. Version 2.19." http://mc-stan.org/manual.html
6. Trochim and Donnelly — Trochim, William and James P. Donnelly. 2007.*The Research Methods Knowledge Base*, 3rd Edition. Cincinnati, OH, Atomic Dog Publishing. http://www.socialresearchmethods.net/kb/
7. Additional suggested articles are listed below.

Day 1: Introduction to Inference and Programming

Readings:

1. Rubin, Donald B. 2008. "For Objective Causal Inference, Design Trumps Analysis." *Annals of Applied Statistics* 2(3):808-840. https://doi.org/10.1214/08-AOAS187

Suggested Readings:

2. Gelman and Hill (Ch.2, "Concepts and methods from basic probability and statistics").
3. Lin, Winston, Donald P. Green, and Alexander Coppock. "Standard operating procedures for Don Greens lab at Columbia." Version 1.05: June 7, 2016. https://github.com/acoppock/Green-Lab-SOP
4. Shadish, William R. 2010. "Campbell and Rubin: A Primer and Comparison of Their Approaches to Causal Inference in Field Settings." *Psychological Methods* 15(1):3-17. https://doi.org/10.1037/a0015916

Day 2: Data Management and Large Scale Data Structures

Readings:

1. Matlofff 2011 (all chapters).

Day 3: Text as Data

Readings:

1. Grimmer, Justin and Brandon M. Stewart. Forthcoming. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21(3):267-297.

Suggested Readings:

1. Barberá, Pablo. 2015. "Birds of the Same Feather Tweet Together. Bayesian Ideal Point Estimation Using Twitter Data." *Political Analysis* 23(1):76-91.

2. Barberá, Pablo and Thomas Zeitzoff. 2017. "The New Public Address System: Why Do World Leaders Adopt Social Media?" *International Studies Quarterly* 62(1):121-130. https://doi.org/10.1093/isq/sqx047.

3. Steinert-Threlkeld, Zachary C. 2018. *Twitter as Data*. Elements in Quantitative and Computational Methods for the Social Sciences. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108529327.

Day 4: Simulation Based Programming and Inference

Readings:

1. Efron and Hastie (Ch.4, "Fisherian Inference and Maximum Likelihood Estimation").

2. Gelman and Hill (Ch.7, "Simulation of probability models and statistical inferences").

Suggested Readings:

3. Bolker (all chapters)

Day 5: Measurement and Latent Variable Models

Readings:

1. Gelman and Hill (Ch.13, "Multilevel linear models: varying slopes, non-nested models, and other complexities").

2. Gelman and Hill (Ch.14, "Multilevel logistic regression").

Suggested Readings:

3. Adcock, Robert, and David Collier. 2001. "Measurement Validity: A Shared Standard for Qualitative and Quantitative Research." *American Political Science Review* 95(3):529–546.

4. Gelman and Hill (Ch.16, "Multilevel modeling in Bugs and R: the basics").

5. Gelman and Hill (Ch.25, "Missing-data imputation")

6. Jackman, Simon. 2008. "Measurement." In *The Oxford Handbook of Political Methodology*, edited by Janet M. Box-Steffensmeier, Henry E. Brady, and David Collier. Oxford University Press.

7. Trochim and Donnelly. Ch 3: "The Theory of Measurement."

Day 6: Simulation Based Programming and Inference continued

Readings:

1. Efron and Hastie (Ch.10, "The Jackknife and Bootstrap")
2. Efron and Hastie (Ch.12, "Cross-Validation and C_p Estimate of Prediction Error")

Suggested Readings:

3. Efron and Hastie (Ch.11, "Bootstrap Confidence Intervals")

Day 7: Observational Data and Design Choice

Readings:

1. Efron and Hastie (Ch.15, "Large-Scale Hypothesis Testing and FDRs").

Suggested Readings:

2. Imai, Kosuke, Luke J. Keele, Dustin Tingley, and Teppei Yamamoto. 2011. "Unpacking the Black Box of Causality: Learning about Causal Mechanisms from Experimental and Observational Studies." *American Political Science Review* 105(4):765-789.
3. Shmueli, Galit. 2010. "To Explain or to Predict?." *Statistical Science* 25(3):289-310.
4. Trochim and Donnelly. (all chapters)

Day 8: Neural Networks

Readings:

1. Efron and Hastie (Ch.18, "Neural Networks and Deep Learning").

Suggested Readings:

1. Cantú, Francisco. 2019. "The Fingerprints of Fraud: Evidence from Mexico?s 1988 Presidential Election" *American Political Science Review* 113(3):710-726. https:doi.org/10.1017/S0003055419000285.

Day 9: Neural Networks continued

Readings:

1. Efron and Hastie (Ch.21, "Empirical Bayes Estimation Strategies").

Day 10: Programming Wrap Up

Additional Course Information

Resources for Harassment

Title IX makes it clear that violence and harassment based on sex and gender, including violence and harassment based on sexual orientation, are a Civil Rights offense subject to the same kinds of accountability and the same kinds of support applied to offenses against other protected categories such as race, national origin, etc. If you or someone you know has been harassed or assaulted, you can find the appropriate resources here: www.bw.edu/resources/hr/harass/policy.pdf

Language and Gender

"Language is gender-inclusive and non-sexist when we use words that affirm and respect how people describe, express, and experience their gender. Just as sexist language excludes womens experiences, non-gender-inclusive language excludes the experiences of individuals whose identities may not fit the gender binary, and/or who may not identify with the sex they were assigned at birth. Identities including trans, intersex, and genderqueer reflect personal descriptions, expressions, and experiences. Gender-inclusive/non-sexist language acknowledges people of any gender (for example, first year student versus freshman, chair versus chairman, humankind versus mankind, etc.). It also affirms non-binary gender identifications, and recognizes the difference between biological sex and gender expression. Teachers and students should use gender-inclusive words and language whenever possible in the classroom and in writing. *Students, faculty, and staff may share their preferred pronouns and names, either to the class or privately to the professor, and these gender identities and gender expressions should be honored.*." For more information:
www.wstudies.pitt.edu/faculty/gender-inclusivenon-sexist-language-syllabi-statement.

Syllabus Acknowledgments

This syllabus is based on several courses that I have taken and designed over the last several years. Some of the material is based on the Research Design (PL SC 501) course that I developed at Pennsylvania State University when I began teaching there in the fall of 2013, which itself is based on similar course developed by David Lake and Mathew McCubbins at the University of California, San Diego. It is also based on material that I developed for a graduate measurement theory class (PL SC 597) and undergraduate Social Data Analysis and Design class (SO DA 308) that I also developed at Pennsylvania State University. Elements of the syllabus and other class materials created for this class are also based in part on the Bayesian Statistics class offered by Seth Hill at University of California, San Diego and the Measurement class offered by Keith Poole at UCSD and now the University of Georgia.