17.806: Quantitative Research Methods IV

Spring 2015

Instructor: In Song Kim
TA: Dean Knox

Department of Political Science
MIT

1 Contact Information

	In Song	Dean
Office:	E53–407	E53–434
Email:	insong@mit.edu	dcknox@mit.edu
Phone:	617–253–3138	
URL:	http://web.mit.edu/insong/www	

2 Logistics

- Lectures: Tuesdays and Thursdays 3:00–4:30pm, E53–485
- Recitations: Fridays 10:30–11:30am, E53–438
- In Song's office hours: Make an appointment
- Dean's office hours: Tuesdays 10:00am–12:00pm

3 Course Description

Empirical research in political science is entering a new era of "Big Data" where a diverse range of data sources have become available to researchers. How can we take advantage of these new data sources and improve our understanding of politics? This course introduces various techniques available for social scientists that can automatically collect, visualize, and analyze massive datasets. It is the fourth course in the quantitative research methods sequence at the MIT political science department. Building on the first three courses of the sequence (17.800, 17.802, and 17.804), this class covers a set of methods for both predictive and descriptive learning using the tools of probability theory.

4 Prerequisites

There are three prerequisites for this course:

1. Mathematics: multivariate calculus and linear algebra.

2. Probability and statistics covered in 17.800, 17.802 and 17.804, including linear regression, Bayesian statistics
3. Statistical computing: proficiency with at least one statistical software. We will use `R` in this course (more on this below).

For 1, refer to this year's math camp materials to see the minimum you need to know; see

Math Camp 1: https://stellar.mit.edu/S/project/mathprefresher/
Math Camp 2: https://stellar.mit.edu/S/project/mathcamp2/

This class will assume that you have already had some prior exposure to the material covered and go through many concepts relatively quickly.

5 Course Requirements

The final grades are based on the following items:

- **Problem sets** (45%): Five problem sets will be given throughout the semester. Problem sets will contain analytical, computational, and data analysis questions. Each problem set will contribute equally toward the calculation of the final grade. The following instructions will apply to all problem sets unless otherwise noted.
  - All answers should be typed. Students are strongly encouraged to use LaTeX, a typesetting system that has become popular in the field. Please make sure that your code follows Google's `R` Style Guide rules (here is the URL).
  - Neither late submission nor electronic submission will be accepted unless you ask for special permission from the instructor in advance. (Permission may be granted or not granted, with or without penalty, depending on the specific circumstances.)
  - Working in groups is encouraged, but each student must submit their own writeup of the solutions. In particular, you should not copy someone else's answers or computer code. We also ask you to write down the names of the other students with whom you solved the problems together on the first sheet of your solutions.
  - For analytical questions, you should include your intermediate steps, as well as comments on those steps when appropriate. For data analysis questions, include annotated code as part of your answers. All results should be presented so that they can be easily understood.
- **Final project** (50%): The final project will be a poster which applies a method learned in this course to an empirical problem of your substantive interest.

  Students are expected to engage in automated data collection, and collect their own data related to an empirical problem of own interest. To help this, we are going to cover (1) the basics of Python, and (2) web-scraping at the beginning of the semester. Students who do not have particular target online data sources should consult with the instructor by February 13th. Replication papers are allowed, but you must go beyond the original analysis in some significant way by collecting additional data *and* applying techniques learned in the course. If you have any doubts, please consult with the instructor or TA.

Collaboration: We encourage you to collaborate with another student (a group should not consist of more than 2 students). Note that most cutting-edge research is collaborative (see any recent issue of *APSR* or *AJPS*), and collaboration is more likely result in a good, potentially publishable paper (multiple brains are usually better than one).

Deadlines: Please be aware of the following deadlines. Late submission will be penalized.

- **March 17 (Data acquisition):** By this date, you should acquire the data to be analyzed (e.g., by scraping webpages, requesting the replication data from the original authors). Please upload your data to the Stellar webpage with a one page description. You will be giving in-class presentations on data on this date.
- **April 2 (Descriptive data analysis):** By this date, you should finish descriptive data analysis. Please upload a brief memo to the Stellar webpage with the following components.
  - Data description
  - Main theoretical/empirical contributions/motivations
  - Figures/tables
- **April 14 (Initial analysis):** By this date, you should finish initial data analysis. Meet with the instructor to get feedback on your analysis (schedule a meeting with the instructor in the week of April 13).
- **April 30 (Draft write-up):** By this date, you should finish your empirical analyses, with all the figures and tables ready. You should submit a write-up to the Stellar webpage by midnight (20% of course grade). The write-up should consist of
  - Title
  - Abstract (150 words)
  - Introduction (2 pages max)
  - the figures and tables with informative captions
- **May 7 (Written feedback):** You should submit your comments on the write-ups of other students by midnight. Your feedback will be graded based on its quality (10% of course grade).
- **May 12 (Final Poster):** The poster should summarize the theoretical/empirical contributions, the methods you utilized, and the results (figures and tables). The poster should be posted to the the Stellar webpage by midnight (20% of course grade). The size of the poster should be A0 (33.1 × 46.8 inches).
- **May 15 (Poster Presentation)**

- **Participation and presentation** (5%): Students are strongly encouraged to ask questions and actively participate in discussions during lectures and recitation sessions.

In addition, there will be recommended readings for each section of the course which students are strongly encouraged to complete prior to the lectures in order to get the most out of them.

6 Course Website

You can find the Stellar website for this course at:

http://stellar.mit.edu/S/course/17/sp15/17.806/

We will distribute course materials, including readings, lecture slides and problem sets, on this website.

7 Questions about Course Materials

In this course, we will utilize an online discussion board called *Piazza*. Below is an official blurb from the Piazza team:

> Piazza is a question-and-answer platform specifically designed to get you answers fast. They support LaTeX, code formatting, embedding of images, and attaching of files. The quicker you begin asking questions on Piazza (rather than via individual emails to a classmate or one of us), the quicker you'll benefit from the collective knowledge of your classmates and instructors. We encourage you to ask questions when you're struggling to understand a concept ...

See this New York Times article to learn more about their founder's story:

http://www.nytimes.com/2011/07/04/technology/04piazza.html

In addition to recitation sessions and office hours, please use the Piazza Q&A board when asking questions about lectures, problem sets, and other course materials. You can access the Piazza course page either directly from the below address or the link posted on the Stellar course website:

https://piazza.com/mit/spring2015/17806

Using Piazza will allow students to see other students' questions and learn from them. Both the TA and the instructor will regularly check the board and answer questions posted, although everyone else is also encouraged to contribute to the discussion. A student's respectful and constructive participation on the forum will count toward his/her class participation grade. *Do not email your questions directly to the instructor or TA* (unless they are of a personal nature)— we will not answer them!

8 Recitation Sessions

Weekly recitation sessions will be held in E53-438 on Fridays 10:30–11:30am. Sessions will cover a review of the theoretical material and also provide help with computing issues. The teaching assistant will run the sessions and can give more details. Attendance is strongly encouraged.

9 Notes on Poster

Poster presentation is an efficient way to get valuable feedback from a large number of people. A poster should follow the structure of your paper, and thus it is a helpful way to think about the organization of your paper before writing it. Here are some notes.

1. **Use keywords and bullet points:** You should not use full sentences—your audience will never read them. Try to use keywords (or half sentences when needed), and make sure that you use only one line to deliver each point.

2. **Use LaTeX:** There are many online templates to help you make posters easily, e.g., http://www-i6.informatik.rwth-aachen.de/ dreuw/latexbeamerposter.php

3. **Examples:** You may find it helpful to look at some of the posters presented at Political Methods conferences. It's available here: http://polmeth.wustl.edu/media.php.

10 Notes on Computing

In this course we use R, an open-source statistical computing environment that is very widely used in statistics and political science. (If you are already well versed in another statistical software, you are free to use it, but you will be on your own.) Each problem set will contain computing and/or data analysis exercises which can be solved with R but often require going beyond canned functions to write your own program.

11 Books

- Recommended books: We will read chapters from these books throughout the course. We strongly recommend that you at least purchase Bishop. These books will be available for purchase at COOP and online bookstores (e.g. Amazon) and on reserve in the library.
  - Christopher M. Bishop. 2007. *Pattern Recognition and Machine Learning*, Springer (A great introduction to machine learning).
  - Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2005. *The Elements of Statistical Learning.* Springer.
  - Kevin P. Murphy. 2012. *Machine Learning*, The MIT Press
  - Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2014 *An Introduction to Statistical Learning.* Springer.

12 Tentative Course Outline

12.1 Research Computing

1. The Basics of Python

2. Web-scraping

   - Regular expression
   - Getting Data from the Web

     Recommended:

     - Jackman, Simon. 2006. "Data from the Web Into R" *The Political Methodologist.* 14, 2. 11–15

3. Rcpp, Armadillo

12.2 Regularization

1. Over-fitting

2. Ridge Regression

3. LASSO

   Recommended:

   - Tibshirani, Robert. (1996). "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society. Series B (Methodological)*, 267-288.
   - Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2005. *The Elements of Statistical Learning.* Ch 3.1–3.4

4. Coordinate Descent Algorithm

12.3 Dimension Reduction

1. Factor Analysis

   - Heckman, James J., and James M. Snyder (1997). "Linear probability models of the demand for attributes with an empirical application to estimating the preferences of legislators." *RAND Journal of Economics 28*: S142–189

2. Principal Component Analysis

12.4 Mixture Models

1. Probability Distributions

   Recommended:

   - Bishop Ch.2, Appendix B

2. EM Algorithm

   Recommended:

   - Bishop Ch.9
   - Imai, Kosuke, and Dustin Tingley. "A statistical method for empirical testing of competing theories." *American Journal of Political Science* 56.1 (2012): 218-236.

3. Variational Inference

   Recommended:

   - Bishop Ch.10

12.5 Text Analysis

1. Text as Data: regular expression, stemming

   Recommended:

   - Grimmer, Justin, and Brandon M. Stewart. "Text as data: The promise and pitfalls of automatic content analysis methods for political texts." *Political Analysis* (2013): 28.

2. Latent Dirichlet Analysis

   Recommended:

   - Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent Dirichlet allocation." *Journal of Machine Learning Research* 3 (2003): 993-1022.

3. Correlated Topic Models

   Recommended:

   - Blei, David, and John Lafferty. "Correlated topic models." *Advances in Neural Information Processing Systems* 18 (2006): 147.

4. Structural Topic Models

   Recommended:

   - Roberts, Margaret E., et al. "Structural Topic Models for Open-Ended Survey Responses." *American Journal of Political Science* (2014).

12.6 Sequential Data

1. Hidden Markov Models

   Recommended:

   - Bishop Ch.13