17.806: Quantitative Research Methods IV

Spring 2019

Instructor: F. Daniel Hidalgo
TA: Sean Liu

Department of Political Science
MIT

Contact Information

<table>
<thead>
<tr>
<th>Danny</th>
<th>Sean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Office: E53–403</td>
<td>E53–406</td>
</tr>
<tr>
<td>Email: <a href="mailto:dhidalgo@mit.edu">dhidalgo@mit.edu</a></td>
<td><a href="mailto:ssliu@mit.edu">ssliu@mit.edu</a></td>
</tr>
<tr>
<td>URL: <a href="https://www.dhidalgo.me/">https://www.dhidalgo.me/</a></td>
<td><a href="https://ssliu.weebly.com">https://ssliu.weebly.com</a></td>
</tr>
</tbody>
</table>

Logistics

- Lectures: Mondays and Wednesdays 1:30–3:00pm, E53–438
- Recitations: Fridays, 9:00am–10:00am, E53–485
- Danny’s office hours: Wednesdays 3:00pm–5:00pm
- Sean’s office hours: Mondays 11:15am–1:15pm, E53-438.

Course Description

This course is the fourth and final course in the quantitative methods sequence at the MIT political science department. The course covers various advanced topics in applied statistics, including those that have only recently been developed in the methodological literature and are yet to be widely applied in political science. The topics for this year are organized into five broad areas: (1) gathering data, where we introduce various techniques for automated data collection, (2) exploring data, where we focus on the visualization and summary of large datasets; (3) statistical learning, where we provide an overview of machine learning algorithms for predictive and descriptive inference; (4) natural language processing, where we cover various techniques for summarizing and analyzing text, and (5) statistical learning as applied to causal inference.
Prerequisites

There are three prerequisites for this course:

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 \texttt{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

\begin{itemize}
  \item \textbf{Math Camp 1}: \url{https://stellar.mit.edu/S/project/mathprefresher/}
  \item \textbf{Math Camp 2}: \url{https://stellar.mit.edu/S/project/mathcamp2/}
\end{itemize}

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

Course Requirements

The final grades are based on the following items:

- **Problem sets** (45%): Six 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\ or \texttt{R} Markdown. Please make sure that your code follows Google's \texttt{R} Style Guide rules (here is the [URL]).
  
  - No late 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.) Electronic submission of problem sets will only be accepted via Stellar, and late submissions will automatically be flagged.
  
  - Working in groups is encouraged, but each student must submit their own write-up of the solutions unless otherwise instructed. 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 and/or simulation questions, include annotated code as part of your answers. All results should be presented in a single document so that they can be easily understood. \texttt{R}Markdown is strongly encouraged.

- **Final project** (50%): The final project will be a paper which applies methods learned in this course to an empirical problem of your substantive interest.
  
  1. **Data** (10%)
Students are expected to collect their own data related to an empirical problem of own interest. Attending the Political Methodology Lab Workshops is strongly encouraged.

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 and TA.

2. Paper and Poster (35%)

You will write a short 4 page paper describing your data, research design, and findings. Your paper should be formatted as follows:

- Title
- Abstract (150 words)
- Introduction (2 pages max, double spaced): Introduction must contain the following.
  (a) The problem/puzzle to be solved
  (b) Explain why previous work and methods leave the problem unresolved
  (c) Your contribution, i.e., the solution to the problem/puzzle. You need to give the reader a clear sense of how you will solve the problem.
  (d) Brief summary of your findings
- Data section (2 pages max, double spaced)
- Figures and tables with informative captions

You will make a poster that will be part of a public presentation to the department. More details can be found below.

3. Presentation (5%) Students will give presentations in front of the class during the regular class time (early May). Presentations should last about 10 minutes (determined based on the class size, but time limits will be strictly enforced).

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. No late submission will be accepted without special permission in advance.

- March 20 (Descriptive data analysis): By this date, you should acquire the data to be analyzed and conduct preliminary descriptive data analysis. Please upload a brief memo to the Stellar webpage with the following components.
  * Main theoretical/empirical contributions/motivations
  * Data description (why better than previous data)
  * Up to three Figures/tables with informative captions
- May 13 (Final Paper): By this date, you should submit your final paper to the Stellar webpage by midnight.
- May 15 (Poster Presentation) TBD
• **Participation (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.

**Course Website**

You can find the Stellar website for this course at:

http://stellar.mit.edu/S/course/17/sp19/17.806/

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

**Questions about Course Materials**

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/spring2019/17806

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!

**Recitation Sessions**

Weekly recitation sessions will be held in at TBA on Fridays. 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.

**Notes on Auditing**

In order to audit this course, one must

- Obtain the course instructor's permission
- Register officially as a listener
- Complete all problem sets
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](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://www-i6.informatik.rwth-aachen.de/dreuw/latexbeamerposter.php).

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.

- If your project requires intensive computational resources, I recommend using the political science computing cluster.

Books

- Recommended books: We will read chapters from these books throughout the course.
  - Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning*. Springer. [PDF](http://example.com)
  - Daniel Jurafsky and James Martin. 2018. *Speech and Language Processing*. Prentice Hall. [PDF](http://example.com)
  - Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2014. *An Introduction to Statistical Learning*. Springer. [PDF](http://example.com)

Tentative Course Outline

Introduction

1. Big Data in Political Science

  Recommended Reading:

**Getting Data**

1. Web Scraping, Regular Expressions

*Recommended Reading:*

- Jurafsky and Martin 2.1
- For a basic tutorial on HTML, consult 3 sources linked from this blog post: [Three great places to start learning HTML](#)
- Data Camp Course: [Working with Web Data in R](#)

**Dimension Reduction**

1. Principal Component Analysis

*Required Reading:*

- James, Witten, Hastie, and Tibshirani 10.2
- Hastie, Tibshirani, and Friedman 14.5.1

*Recommended Reading:*


2. Factor Analysis

*Recommended Reading:*


3. t-sne
Recommended Reading:


Mixture Models

1. Probability Distributions

Required Reading:

- Bishop Ch.2, Appendix B

2. EM Algorithm

Required Reading:

- Bishop Ch.9

Recommended Reading:


Supervised Learning

1. Over-fitting (Model Selection), Cross-validation

Required Reading:

- Hastie, Tibshirani, and Friedman 7.

Recommended Reading:


2. Linear Methods (Ridge Regression, LASSO)

Required Reading:

- Hastie, Tibshirani, and Friedman 3.1–3.4

Recommended Reading:

3. Tree-Based Methods

**Required Reading:**
- Hastie, Tibshirani, and Friedman 9, 15

**Recommended Reading:**

4. Ensemble Methods

**Required Reading:**
- Hastie, Tibshirani, and Friedman 10

**Recommended Reading:**

5. Interpretable Machine Learning

**Required Reading:**
- Documentation for the DALEX package: URL

**Text Analysis**

1. Text as Data: corpuses, document term matrices, N-gram models

**Required Reading:**
- Jurafsky and Martin 2.2-2.6, 3.1-3.3

**Recommended Reading:**

2. Classifying Text

*Required Reading:*

• Jurafsky and Martin 4, 5

*Recommended Reading:*

• Brandon M. Stewart and Yuri M. Zhukov. 2009. “Use of Force and Civil–Military Relations in Russia: An Automated Content Analysis.” *Small Wars & Insurgencies* 20, no. 2 (June 1): 319–343

3. Words and Votes: Scaling with Text

*Required Reading:*


*Recommended Reading:*


4. Topic Models

*Recommended Reading:*


5. Embeddings

*Recommended Reading:*

• Jurafsky and Martin 6
Causal Inference and Machine Learning

1. Machine learning for Causal Inference

Required Reading:


Recommended Reading:

- Susan Athey and Guido Imbens. 2016. “Recursive Partitioning for heterogeneous causal effects.” *Proceedings of the National Academy of Sciences*, 113(27), 7353–7360

2. Machine learning for Panel Settings

Required Reading: