17.806: Quantitative Research Methods IV

Spring 2023

Instructors: F. Daniel Hidalgo & Teppei Yamamoto
TA: Benjamín Muñoz

Department of Political Science
MIT

1 Contact Information

	F. Daniel Hidalgo	Teppei Yamamoto	Benjamín Muñoz
Office:	E53–402	E53–401	E53–414
Email:	dhidalgo@mit.edu	teppei@mit.edu	benja_mr@mit.edu
URL:	http://www.dhidalgo.me	https://web.mit.edu/teppei/www/	

2 Logistics

- Lectures: Mondays and Wednesdays, 3:30pm–5:00pm (Building E53-438)
- Recitations: Fridays, 10:00am-11:00am (Building E53-438)

Note that the first class meets on February 6. The class on February 20 (Presidents' Day) will instead be held on February 21 (Monday schedule Institute-wide). No class will be held on March 27 and 29 (Spring Break), and April 17 (Patriots' Day). Last day of class is May 15.

3 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 include the fundamentals of machine learning (research computing, unsupervised and supervised learning), text analysis, survival analysis, and advanced causal inference methods (causal machine learning, methods for longitudinal data, sensitivity analysis, and mediation analysis).

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, causal inference, and 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 and 3, we expect the level of background knowledge and skills equivalent to what is covered in the department's Math Camp II (please see the "Math Camp II Notes" pdf file posted on Canvas.)

5 Course Requirements

The final grades are based on the following items.

- **Problem sets** (50%): Six problem sets will be given throughout the semester (roughly bi-weekly) of which five are required. The first four problem sets are mandatory. Students may choose either Problem Set 5 or 6 to complete. (If they choose to submit both, the best five of the six problem set grades will become their final score.) 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 (or LaTeX typesetting in RMarkdown). Please make sure that your code follows the Google and tidyverse `R` style guide rules (URLs are here and here).
  - Late submission will not 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 in a single document so that they can be easily understood. RMarkdown is strongly encouraged.

- **Final project** (40%): The final project will be a short research paper which typically applies a method learned in this course to an empirical problem of your substantive interest. Students are encouraged to either collect their own data or work with non-traditional form of data for their empirical projects. Consult the instructors if you have a different idea (e.g. a purely methodological project).

  Students are expected to adhere to the following deadlines:

  - February to early March: **Start** thinking about possible topics, strategies for data acquisition, and running simple analyses on the collected data. Run your ideas by the TA and instructor during their office hours and after classes/recitations to obtain their reactions.

- March 13: Turn in a **1–2 page description of your proposed project**. By this date you need to have found your coauthor, acquired the data you plan to use at least partially, and run preliminary analysis on the data (e.g. simple summary statistics, crosstabs and plots). Meet with the instructors to discuss your proposal during their office hours. You may be asked to revise and resubmit the proposal in two weeks from the meeting.
- May 8, 10, and 15: Students will give **presentations** during the regular class time. Presentations should last about 15 minutes (the exact length will be determined based on the class size, but time limits will be strictly enforced) and take the form much like presentations at major academic conferences such as the APSA and MPSA annual meetings. Performance on this presentation will be counted toward the class participation grade (see below).
- May 16: **Paper due.** Upload an electronic copy of your paper along with your computer code by the end of the day to the designated Canvas page. The paper should be approximately 10 pages in length, excluding the title page and appendices. It should start with a title page containing the title, author name(s), and an abstract. The body of the paper should contain a concise statement of the research question, description of the data, empirical strategy, results, and conclusions. Literature reviews, theoretical background and motivations should be either omitted or kept to minimum. Appendices can include additional tables, figures and robustness check results. If you need an extension, contact the instructors by email in advance.

- **Participation** (10%): 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 Canvas website for this course at:

`https://canvas.mit.edu/courses/18682`

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*. This is a question-and-answer platform that is easy to use and designed to get you answers to questions quickly. We encourage you to use the Piazza Q & A board when asking questions about lectures, problem sets, and other course materials outside of recitation sessions and office hours. You can access the Piazza course page either directly from the below address or the link posted on the Canvas course website:

https://piazza.com/mit/spring2023/17806

Using Piazza will allow students to see and learn from other students' questions. 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 instructors or TAs* (unless they are of a personal nature) — we will not answer them!

8 Recitation Sessions

Weekly recitation sessions will be held in person 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.

9 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. We provide problem set solutions using `R`.
- We strongly encourage you to use RMarkdown. These are useful resources to learn about RMarkdown:
  - Tierney, Nicholas. *RMarkdown for Scientists* [Link].
  - Xie, Yihui, Christophe Dervieux, and Emily Riederer. *R Markdown Cookbook* [Link].
  - Xie, Yihui, J. J. Allaire, and Garrett Grolemund. *R Markdown: The Definitive Guide* [Link].
- Following reference would be useful to write clean and efficient code in `R`:
  - Google's style guide [Link].
  - Tidyverse style guide [Link] (You do not need to use the Tidyverse but chapters 1–3 are very useful for non-Tidyverse users as well).
- If your project requires large computational resources, we recommend using xvii or the MIT SuperCloud.

10 Books

Recommended books (available online for free):

- Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning.* Springer.
- James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2014. *An Introduction to Statistical Learning.* Springer.
- Jurafsky, Daniel and James Martin. 2018. *Speech and Language Processing.* Prentice Hall. PDF

11 Course Outline

11.1 Introduction

11.2 Automated Data Collection

1. Web Scraping, Regular Expressions

   Recommended Reading:

   - Chapter 17 in Wickham, Hadley, Çetinkaya-Rundel, and Garrett Grolemund. 2023. *R for Data Science*. O'Reilly.
   - 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.
   - Jackman, Simon. 2006. "Data from the Web Into `R`" *The Political Methodologist*. 14 (2): 11–15.

11.3 Dimension Reduction

1. Principal Component Analysis, Factor Analysis

   Required Reading:

   - Shlens, Jonathon. *A Tutorial on Principal Component Analysis*. PDF

   Recommended Reading:

   - Hastie, Tibshirani, and Friedman 14.5.
   - Bond, Robert and Solomon Messing. 2015. "Quantifying Social Media's Political Space: Estimating Ideology from Publicly Revealed Preferences on Facebook." *American Political Science Review* 109 (1): 62–78.
   - Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A., & Bonneau, R. (2015). Tweeting from left to right: Is online political communication more than an echo chamber?. *Psychological science*, 26(10), 1531-1542.
   - 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–S189.
   - Bai, Jushan. 2009. "Panel Data Models with Interactive Fixed Effects." *Econometrica* 77 (4): 1229–1279.

11.4 Supervised Learning

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

   Required Reading:

   - Hastie, Tibshirani, and Friedman Ch.7.

2. Variable Selection (Ridge Regression, LASSO)

Required Reading:

- Hastie, Tibshirani, and Friedman 3.1–3.4.
- Bonica, Adam. 2018. "Inferring Roll–Call Scores from Campaign Contributions Using Supervised Machine Learning." *American Journal of Political Science* 62 (4): 830–848.

Recommended Reading:

- Bloniarz, Adam, Hanzhong Liu, Cun-Hui Zhang, Jasjeet S. Sekhon, and Bin Yu. 2016. "Lasso Adjustments of Treatment Effect Estimates in Randomized Experiments." *Proceedings of the National Academy of Sciences* 113 (27): 7383–7390.
- Tibshirani, Robert. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society. Series B (Methodological)* 58 (1): 267–288.

3. Additive Models & Ensemble Methods: Generalized Additive Models (GAM), Bagging, Boosting, Random Forests

Recommended Reading:

- Hastie, Tibshirani, and Friedman Chs.9, 15, 16.
- Montgomery, Jacob and Santiago Olivell. "Tree-Based Models for Political Science Data." 2018. *American Journal of Political Science* 62 (3): 729–744.

11.5 Machine Learning for Causal Inference

1. Machine learning for Causal Inference

Required Reading:

- Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. 2014. "High-Dimensional Methods and Inference on Structural and Treatment Effects." *Journal of Economic Perspectives* 28 (2): 29–50.
- Künzel, Sören R., Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. 2019. "Metalearners for estimating heterogeneous treatment effects using machine learning." *Proceedings of the National Academy of Sciences* 116 (10): 4156–4165.

Recommended Reading:

- Grimmer, Justin, Solomon Messing, and Sean J. Westwood. 2017. "Estimating Heterogeneous Treatment Effects and the Effects of Heterogeneous Treatments with Ensemble Methods." *Political Analysis* 25 (4): 413–434.
- Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2017. "Double/Debiased Machine Learning for Treatment and Structural Parameters." *American Economic Review* 107 (5): 261–265.
- Athey, Susan, Guido W. Imbens, and Stefan Wager. 2018. "Approximate Residual Balancing: Debiased Inference of Average Treatment Effects in High Dimensions." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 80 (4): 597–623.
- Athey, Susan and Guido Imbens. 2016. "Recursive Partitioning for heterogeneous causal effects." *Proceedings of the National Academy of Sciences* 113 (27): 7353–7360.

- Wager, Stefan and Susan Athey. 2018. "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests." *Journal of the American Statistical Association* 113 (523): 1228–1242.
- Imai, Kosuke, and Marc Ratkovic. 2013. "Estimating treatment effect heterogeneity in randomized program evaluation." *The Annals of Applied Statistics* 7 (1): 443–470.

11.6 Causal Inference with Longitudinal Data

1. Fixed effects, difference-in-differences, and synthetic control methods

   Recommended Reading:

   - Imai, Kosuke and In Song Kim. 2019. "When Should We Use Fixed Effects Regression Models for Causal Inference with Longitudinal Data?" *American Journal of Political Science* 63 (2): 467–490.
   - Imai, Kosuke, In Song Kim, and Erik Wang. "Matching Methods for Causal Inference with Time-Series Cross-Section Data." *American Journal of Political Science*. Forthcoming.
   - Xu, Yiqing. 2017. "Generalized synthetic control method: Causal inference with interactive fixed effects models." *Political Analysis* 25 (1) 57–76.
   - Athey, Susan, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens, and Khashayar Khosravi. 2017. "Matrix Completion Methods for Causal Panel Data Models." `https://arxiv.org/abs/1710.10251`.
   - De Chaisemartin, Clément, and Xavier d'Haultfoeuille. 2020. "Two-way Fixed Effects Estimators with Heterogeneous Treatment Effects." *American Economic Review* 110 (9): 2964–96.
   - Imai, Kosuke, and In Song Kim. 2020. "On the Use of Two-way Fixed Effects Regression Models for Causal Inference with Panel Data." *Political Analysis*. 29 (3): 405–415.

11.7 Survival Analysis

1. Basic Concepts of Survival Analysis

   Recommended:

   - Box-Steffensmeier, Janet M. and Bradford S. Jones, 1997, "Time is of the Essence: Event History Models in Political Science." *American Journal of Political Science*, 41(4), 1414–1461.

   Optional:

   - Freedman, David A., 2008, "Survival Analysis: A Primer," *The American Statistician*, 62(2), 110–119.
   - Wooldridge Ch.20 or Cameron & Trivedi Ch.17, 19

2. Parametric Regression Models

   Recommended:

- King, Gary, James E. Alt, Nancy Burns and Michael Laver, 1990, "A Unified Model of Cabinet Dissolution in Parliamentary Democracies," *American Journal of Political Science*, 34(3), 346–871.
- Warwick, Paul and Stephen T. Easton, 1992, "The Cabinet Stability Controversy: New Perspectives on a Classic Problem," *American Journal of Political Science*, 36(1), 122–146.

3. Semiparametric and Competing Risks Models

   Recommended:

   - Warwick, Paul, 1992, "Rising Hazards: An Underlying Dynamic of Parliamentary Government," *American Journal of Political Science*, 36(4), 857–876.
   - Diermeier, Daniel and Randy T. Stevenson, 1999, "Cabinet Survival and Competing Risks," *American Journal of Political Science*, 43(4), 1051–1068.

11.8 Text Analysis

1. Text as Data: regular expression, stemming

   Recommended Reading:

   - Grimmer, Justin, and Brandon M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21 (3): 267–297.
   - Gentzkow, Matthew, Bryan T. Kelly, and Matt Taddy. 2017. "Text as Data." *Journal of Economic Literature* 57(3): 535–74.
   - Denny, Matthew J., and Arthur Spirling. 2018. "Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It." *Political Analysis* 26 (2): 168–189.

2. Topic models: Latent Dirichlet Analysis, Correlated Topic Models, Structural Topic Models

   Recommended Reading:

   - Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. "Latent Dirichlet allocation." *Journal of Machine Learning Research* 3: 993–1022.
   - Blei, David, and John Lafferty. 2006. "Correlated Topic Models." *Advances in Neural Information Processing Systems* 18: 147.
   - Roberts, Margaret E., Stewart Brandon M., and Airoldi Edo M. 2016. "A Model of Text for Experimentation in the Social Sciences." *Journal of the American Statistical Association* 111 (515): 988–1003.
   - Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G. Rand. 2014. "Structural Topic Models for Open-Ended Survey Responses." *American Journal of Political Science* 58 (4): 1064–1082.

3. Words and Votes: Scaling with Text

Recommended Reading:

- Gerrish, Sean, and David M. Blei. 2012. "How they vote: Issue-adjusted models of legislative behavior." *Advances in Neural Information Processing Systems* 25.
- Lauderdale, Benjamin E., and Tom S. Clark. 2014. "Scaling politically meaningful dimensions using texts and votes." *American Journal of Political Science* 58 (3): 754–771.
- Slapin, Jonathan B., and Sven-Oliver Proksch. 2008. "A Scaling Model for Estimating Time-series Party Positions from Texts." *American Journal of Political Science* 52 (3): 705–722.
- Kim, In Song, John Londregan, and Marc Ratkovic. 2018. "Estimating Spatial Preferences from Votes and Text." *Political Analysis* 26 (2): 210–229.

4. Word Embeddings

Recommended Reading:

- Jurafsky and Martin Ch.6.
- Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. "Efficient Estimation of Word Representations in Vector Space." URL: `https://arxiv.org/abs/1301.3781`
- Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. "Distributed Representations of Words and Phrases and their Compositionality." URL: `https://arxiv.org/abs/1310.4546`
- Rheault, Ludovic and Christopher Cochrane. 2020. "Word Embeddings for the Analysis of Ideological Placement in Parliamentary Corpora" *Political Analysis* 28 (1): 112–133.

11.9 Sensitivity Analysis for Causal Inference

1. Fundamentals (review)

Recommended Reading:

- Guido W. Imbens. 2003. Sensitivity to Exogeneity Assumptions in Program Evaluation. *The American Economic Review* 93 (2): 126–32.
- Rosenbaum, Paul R. 2002. Observational Studies. Springer-Verlag. 2nd edition. Chapter 4.

2. E-value and related alternatives

Recommended Reading:

- VanderWeele and Ding (2017, AoIM), "Sensitivity Analysis in Observational Research: Introducing the E-value."
- Ding and VanderWeele (2016, Epi), "Sensitivity Analysis without Assumptions.'
- Blackwell (2013, PA), "A Selection Bias Approach to Sensitivity Analysis for Causal Effects." (confounding function)

- Cinelli and Hazlett (2020, JRSS-B), "Making Sense of Sensitivity: Extending Omitted Variables Bias." (robustness value)
- Zhao (2019, JASA), "On Sensitivity Value of Pair-Matched Observational Studies." (sensitivity value)

3. Bayesian sensitivity analysis

   Recommended Reading:

   - Gustavson, McCandless, Levy and Richardson (2010, Biometrics), "Simplified Bayesian Sensitivity Analysis for Mismeasured and Unobserved Confounders."

11.10 Causal Mediation Analysis

1. Causal Mediation Analysis

   Recommended Reading:

   - Imai, K., L. Keele, D. Tingley and T. 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.
   - Imai, K., L. Keele and T. Yamamoto. 2010. Identification, Inference, and Sensitivity Analysis for Causal Mediation Effects. *Statistical Science*, 25(1), 51-71.
   - Robins, James M. and Sander Greenland. 1992. Identifiability and Exchangeability of Direct and Indirect Effects. *Epidemiology*, 3: 143–155.
   - Pearl, Judea. 2001. Direct and Indirect Effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, 411–420.
   - Imai, K., D. Tingley and T. Yamamoto. 2013. Experimental Designs for Identifying Causal Mechanisms. *Journal of the Royal Statistical Society, Series A*, 176(1), 5–51.
   - Zhou, Xiang, and Teppei Yamamoto. 2023. Tracing Causal Paths from Experimental and Observational Data. *Journal of Politics*, 85(1), 250–265.