

# 17.804: Quantitative Research Methods III

Fall 2022

Instructor: Teppei Yamamoto

TA: Licheng Liu

Department of Political Science

MIT

## 1 Contact Information

	Teppei	Licheng
Email:	tepei@mit.edu	liulch@mit.edu
Office Hours	by appointment	Fri 2:10–3:30 pm
Office	E53-401	E53 3rd floor lounge
Office # (Zoom)	550 928 2449	–

## 2 Logistics

- Lectures: Mondays and Wednesdays **9:30–11am** at E53-438
- Recitations: Friday 10:00 - 11:00 am (E53-438)
- Office hours: See “Contact Information” above

All instructions are planned to be provided in-person but possibly online via Zoom meetings and other means of digital communication. All Zoom sessions will be password-protected. The password for each virtual room will be posted on the Canvas course website (see “Course Website” below).

Note that the first class meets on September 7, and the last class meets on December 14. There will be no class on the following days: October 10 (Columbus Day), and November 23 (Thanksgiving week).

Please also note that enrollment is capped at 15 students due to capacity constraints and priority is given to political science graduate students. Therefore we cannot guarantee a spot for students from other departments or outside MIT. In case there is excess demand, students will be notified whether they are permitted to remain enrolled in the class by the end of the second week of class.

## 3 Course Description

This course is the third course in the quantitative research methods sequence at the MIT political science department. Building on the first two courses of the sequence (17.800 and 17.802), this class covers advanced statistical tools for empirical analysis in modern political science. Our focus

in this course will be on techniques for *model-based inference*, including various regression models for cross-section data (e.g., binary outcome models, discrete choice models, sample selection models, event count models, survival outcome models, etc.) as well as grouped data (e.g., mixed effects models and hierarchical models). This complements the methods for *design-based inference* primarily covered in the previous course of the sequence. This course also covers basics of the fundamental statistical principles underlying these models (e.g., maximum likelihood theory, theory of generalized linear models, Bayesian statistics) as well as a variety of estimation techniques (e.g., numerical optimization, bootstrap, Markov chain Monte Carlo). The ultimate goal of this course is to provide students with adequate methodological skills for conducting cutting-edge empirical research in their own fields of substantive interest.

## 4 Prerequisites

There are three prerequisites for this course:

1. Mathematics: Intermediate college-level calculus and linear algebra.
2. Probability and statistics covered in 17.800 and 17.802, including linear regression and basic causal inference.
3. Statistical computing: familiarity with at least one statistical software, preferably R (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 pdf file under the "Resources" module as 'Math Camp II Notes.pdf').

## 5 Course Requirements

The final grades are based on the following items:

- **Problem sets (40%):** Weekly problem sets will be given, beginning in the second week of class (week of September 12) until the week before Thanksgiving. In addition, an optional problem set will be assigned after Thanksgiving. Each required problem set is due one week after it is assigned. Problem sets will contain analytical, computational, and data analysis questions. Each problem set will be counted equally toward the calculation of the final grade. The following instructions will apply to all problem sets unless otherwise noted.
  - Problem sets should be submitted electronically on Canvas *by the beginning of class (9:30 am) the day (Wednesday) it is due*. Neither late submission nor printed problem sets will be accepted unless you ask for special permission from the instructor in advance of the deadline. (Permission may be granted or not granted, with or without penalty, depending on specific circumstances.)
  - Working in groups is encouraged, but each student must submit their own writeup of the solutions. In particular, you must not simply copy and paste someone else's answers or computer code. *Violation of this policy will be considered an academic integrity issue and processed accordingly to MIT's rules and procedures for such violations.* 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. We will allow handwritten answers to some analytical problems *as long as the write-ups are readable*. If you choose to handwrite your answer to these questions, we ask that you upload a scanned copy of your work and include it in your problem set submission. For data analysis questions, include annotated code as part of your answers. All results should be presented so that they can be easily understood.
- **Quizzes (15%):** Three in-class, closed-book quizzes will take place on October 3, November 7 and November 30 during the regular class time.
- **Final project (35%):** 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. The paper should be approximately 10 pages in length and 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. You should also submit a copy of your analysis code. Co-authoring is generally encouraged, though political science Ph.D. students should be mindful that a co-authored seminar paper cannot be used as the basis of their second-year paper. Replication papers are also accepted as long as they methodologically go beyond the original analysis in some significant manner.

Students are expected to adhere to the following deadlines:

- September to early October: **Start** thinking about possible topics, exploring data sources, and running simple analyses on acquired data sets. Run your ideas by the TA and instructor during their office hours and after classes/recitations to obtain their reactions.
- October 17: Turn in a **brief description of your proposed project**. By this date you need to have found your coauthor, acquired the data you plan to use, and completed a descriptive analysis of the data (e.g. simple summary statistics, crosstabs and plots). Meet with the instructor to discuss your proposal during his office hours. You may be asked to revise and resubmit the proposal in two weeks from the meeting.
- December 7, 12, and 14: 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).
- December 14: **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. If you need an extension, contact the instructor by email well in advance.
- **Participation and presentation (10%):** Students are strongly encouraged to ask questions and actively participate in discussions during lectures and recitation sessions (and piazza) .

In addition, there will be recommended readings and lecture notes. Students are strongly encouraged to complete readings 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/15755>

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 on your Canvas website or directly at

<https://piazza.com/mit/fall2022/17804/home>

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 instructor or TA* (unless they are of personal nature) — we will not answer them!

## 8 Recitation Sessions

Recitation sessions will be held in person on Fridays 10:00-11:00 AM (E53-438). Sessions will cover various topics, including review of lecture material, hints on problem set questions, and help with computing issues. The TA will run the sessions and can give more details. Attendance is very 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 and writing your own program.

In addition to the materials from the department's math camps (see above), there are many resources for statistics and data science using R that are targeted at both introductory and advanced levels. Examples:

- Fox, John and Sanford Weisberg. 2010. *An R Companion to Applied Regression*. Sage Publications.
- Venables, W. N. and B. D. Ripley. 2002. *Modern Applied Statistics with S*, 4th ed. Springer.

- Wickham, Hadley and Garrett Golemund. 2017. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. O'Reilly Media.
- For specific questions about R, searching the CRAN website or Stack Overflow with appropriate keywords will often yield satisfactory results.
- There are a lot of other useful online resources, especially about newer-generation packages oriented for data science applications. Check out materials on RStudio's website (documentations, cheatsheets, videos, webinars, etc.).
- As a last resort, you can post your question to the R help e-mail list, but be sure to read the posting guidelines before doing so, and follow exactly what they say. The list is run by a very busy group of people (you will frequently get answers from R Core team members) and they can be nasty if you are not respectful of the norms.

## 10 Books

- Recommended books: We will read chapters from these books throughout the course. We strongly recommend that you at least purchase (1) either one of the first two books, (2) Gelman and Hill, and (3) Gelman et al. These books can be purchased at online bookstores (e.g. Amazon) and they will be on reserve in the library.
  - Wooldridge, Jeffrey. 2010. *Econometric Analysis of Cross Section and Panel Data*, 2nd ed. MIT Press. (A standard reference for applied econometricians for most topics covered in the first part of the course.)
  - Cameron, Colin and Pravin Trivedi. 2005. *Microeconometrics: Methods and Applications*. Cambridge University Press. (Slightly less standard, but covers most of the topics throughout the course.)
  - Gelman, Andrew and Jennifer Hill. 2006. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press. (A standard, non-technical textbook for Bayesian hierarchical models.)
  - Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari and Donald B. Rubin. 2014. *Bayesian Data Analysis*, 3rd ed. Chapman and Hall/CRC. (A standard textbook on applied Bayesian statistics.)
- Optional books: These books are standard references for specific topics covered in this course. We will assign a chapter or two from them. Those chapters will be on electronic reserve. Nice books to have for advanced students, but no need to purchase only for this course.
  - McCullagh, P. and J. A. Nelder. 1989. *Generalized Linear Models*, 2nd ed. Chapman and Hall/CRC. (Generalized linear models)
  - Efron, Bradley and Robert J. Tibshirani. 1994. *An Introduction to the Bootstrap*. Chapman and Hall/CRC. (Bootstrap)
  - Kalbfleisch, John D. and Ross L. Prentice. 2002. *The Statistical Analysis of Failure Time Data*, 2nd ed. Wiley. (Survival analysis)

## 11 Tentative Course Outline

### 11.1 Generalized Linear Models and Extensions

#### Binary Outcome Models

1. Binary Logit and Probit Models

*Recommended:*

- Wooldridge Ch.15 or Cameron & Trivedi Ch.14

2. Theory of Maximum Likelihood Estimation

*Recommended:*

- Wooldridge Ch.13 or Cameron & Trivedi Ch.5, 7.2–7.4
- Buse, A. 1982. “The Likelihood Ratio, Wald, and Lagrange Multiplier Tests: An Expository Note.” *The American Statistician*, 36(3), 153–157.

3. Numerical Optimization

*Recommended:*

- Wooldridge Ch.12.7 or Cameron & Trivedi Ch.10

4. Bootstrap and Monte Carlo Approximation

*Recommended:*

- Efron & Tibshirani, Ch.6
- King, Gary, Mike Tomz and Jason Wittenberg, 2000, “Making the Most out of Statistical Analysis: Improving the Interpretation and Presentation.” *American Journal of Political Science*, 44(2), pp.341–355.

*Optional:*

- Wooldridge Ch.12.8.2 or Cameron & Trivedi Ch.11

#### Discrete Choice Models

1. Multinomial Logit and Probit Models

2. Ordered Logit and Probit Models

*Recommended:*

- Wooldridge Ch.16 or Cameron & Trivedi Ch.15

*Optional:*

- Alvarez, R. Michael and Jonathan Nagler, 1995, “Economics, Issues and the Perot Candidacy: Voter Choice in the 1992 Presidential Election.” *American Journal of Political Science*, 39(3), 714–744.

## Event Count Models

### 1. Theory of Generalized Linear Models

*Recommended:*

- McCullagh & Nelder, Ch.2
- Gelman & Hill, Ch.6

*Optional:*

- McCullagh & Nelder, Ch.9.1, 9.2

### 2. Event Count Models

*Recommended:*

- Wooldridge Ch.18 or Cameron & Trivedi Ch.20
- Wand, Jonathan N., Kenneth W. Shotts, Jasjeet S. Sekhon, Walter R. Mebane, Jr., Michael C. Herron and Henry E. Brady. 2001. "The Butterfly Did It: The Aberrant Vote for Buchanan in Palm Beach County, Florida." *American Political Science Review*, 95(4), 793–810.

## Models for Panel and Multilevel Data

### 1. Fixed and Random Effects Models

*Recommended:*

- Wooldridge, Ch.10 or Cameron & Trivedi, Ch.21
- Green, Donald P., Soo Yeon H. Kim and David Yoon, 2001, "Dirty Pool," *International Organization*, 55(2), 441–468.

### 2. Mixed Effects Models

*Recommended:*

- Gelman & Hill, Ch.11
- Zorn, Christopher J.W., 2001, "Generalized Estimating Equation Models for Correlated Data: A Review with Applications," *American Journal of Political Science*, 45(2), 470–490.

*Optional:*

- Cameron & Trivedi, Ch.22.8, 24.6

## 11.2 Bayesian Statistical Modeling

### Introduction to Bayesian Statistics

#### 1. Basic Concepts of Bayesian Statistics

*Recommended:*

- Gelman et al., Ch.1, 2, 3, 4 and 5.

#### 2. Markov Chain Monte Carlo

*Recommended:*

- Gelman et al., Ch.10, 11 and 12.

*Optional:*

- Jackman, Simon, 2000, "Estimation and Inference via Bayesian Simulation: An Introduction to Markov Chain Monte Carlo," *American Journal of Political Science*, 44(2), 375–404.
- Casella, George and Edward I. George, 1992, "Explaining the Gibbs Sampler," *The American Statistician*, 46(3), 167–174.
- Chib, Siddhartha and Edward Greenberg, 1995, "Understanding the Metropolis-Hastings Algorithm," *The American Statistician*, 49(4), 327–335.

### Bayesian Statistical Modeling

#### 1. Hierarchical Linear and Nonlinear Models

*Recommended:*

- Gelman & Hill, Ch.12, 13

*Optional:*

- Gelman & Hill, Ch.14, 15

#### 2. Missing Data

*Recommended:*

- Gelman et al. Ch.18

*Optional:*

- Gary King, James Honaker, Anne Joseph and Kenneth Scheve. Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation. *American Political Science Review*, Vol. 95, No. 1 (Mar., 2001), pp. 49-69

#### 3. Measurement and Item Response Theory

*Recommended:*

- Jackman, Simon. 2009. *Bayesian Analysis for the Social Sciences*. Wiley. Ch.9.



*Optional:*

- Treier, Shawn and Simon Jackman. 2008. “Democracy as a Latent Variable.” *American Journal of Political Science*, 52(1): 201–217.
- Clinton, Joshua, Simon Jackman and Douglas Rivers. 2004. “The Statistical Analysis of Roll Call Data.” *American Political Science Review*, 98(2): 355–370.