17.804: Quantitative Research Methods III
Fall 2016

Instructor: Jonathan Hersh
TA: Ignacio Puente

1 Contact Information

	Jonathan	Ignacio
Office:	E53-401	E53-408
Phone:	617-253-6959	
Email:	hershj@mit.edu	puente@mit.edu

2 Logistics

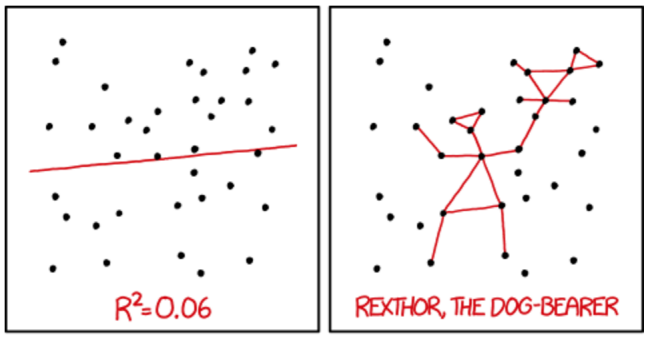
- Lectures: Tuesdays and Thursdays **1:00 - 2:30pm**, **E53-438**
- Recitations: Fridays time TBD
- Jonathan's office hours: TBD
- Ignacio's office hours: TBD
- First recitation session will take place on September 16th. No recitation session on September 9.

Please note:

- There is no lecture on 10/11 (Columbus Day) or 11/24 (Thanksgiving break).
- There is no recitation on 9/23 (Student holiday), 11/11 (Veteran's Day) or 11/25 (Thanksgiving Holiday). Reschedule?

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**, **event count 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.



I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER TO GUESS THE DIRECTION OF THE CORRELATION FROM THE SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.

4 Prerequisites

There are three prerequisites for this course:

- Mathematics: Basic college-level calculus and linear algebra.
- Probability and statistics covered in 17.800 and 17.802, including linear regression and basic causal inference.
- Statistical computing: familiarity with at least one statistical software. We will use R and JAGS 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; see

https://stellar.mit.edu/S/project/mathcamp2/

5 Course Requirements

The final grades are based on the following items:

- **Problem sets** (40%): Weekly problem sets will be given throughout the semester. 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.
  - Neither late submission nor electronic submission 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 the specific circumstances.)
  - Working in groups is encouraged, but each student must submit their own write-up of the solutions. In particular, don't just 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. For data analysis questions, include commented/annotated code as part of your answers. All results should be presented so that they can be easily understood. Consider this as practice for how to efficiently format your tables and figures for publication. Make use of both ggplot2, stargazer, and other helpful packages.

- **Quizzes** (15%): Three in-class, closed-book quizzes will take place on Mondays (October 4, November 3 and December 1st) during the regular class time. These will be short (30 mins) and are intended to motivate you to keep up with the material.

- **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 strongly encouraged. 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 13: **Turn in a 1-page 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 8 and 13th: Students will give presentations in front of the class during the regular class time. Presentations should last about 10 minutes (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. Students should prepare electronic slides to accompany their presentation. Performance on this presentation will be counted toward the class participation grade (see below). Make final revisions to your paper based on the feedback.
  - December 13: **Paper due**. Please turn in one printed copy of your paper by the end of the day, and email electronic copies to the instructor and TA.

- **Participation and presentation** (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 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 Stellar website for this course at:

https://stellar.mit.edu/S/course/17/fa16/17.804/

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 Stellar course website:

piazza.com/mit/fall2016/17804/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 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 ??? on days and times to be determined in the first week of class. Sessions will cover a review of the theoretical material and also provide help with computing issues. Ignacio will run the sessions and can give more details. Attendance is strongly encouraged.

9 Notes on Computing

In this course we use http://www.r-project.org/, 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 (Python, Stata, Matlab), 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 R targeted at both introductory and advanced levels, including:

- Fox, John and Sanford Weisberg. 2010. *An R Companion to Applied Regression*. Sage Publications. (focused on regression analysis)
- Venables, W. N. and B. D. Ripley. 2002. *Modern Applied Statistics with S*}, 4th ed. Springer. (general statistics)

- Wickham, Hadley. 2014. *Advanced R.* CRC Press. (R language)
- Lander, Jared. 2013. *R for Everyone: Advanced Analytics and Graphics.* Addison-Wesley. (readable and entertaining introduction to statistical data science)
- For specific questions about R, searching the CRAN website or Stack Overflow with appropriate keywords will often yield satisfactory results.
- Does anyone read this far into the syllabus? Email me before Sept 15th that you saw this easter egg and I'll add 5% to your problem set grade.
- 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.

For Bayesian statistical modeling, we also use http://mcmc-jags.sourceforge.net/, a cross-platform, open-source software for Markov chain Monte Carlo (MCMC) via Gibbs sampling. JAGS uses syntax similar to R and comes with an easy-to-use interface with R.

10 Books

- **Recommended books:** We will read chapters from these books throughout the course. We strongly recommend that you at least purchase (1) Cameron and Trivedi, (2) Jackman and (3) Gelman and Hill. These books will be available for purchase at COOP and online bookstores (e.g. Amazon) and on reserve in the library.
  - Cameron, Colin and Pravin Trivedi. 2005. *Microeconometrics: Methods and Applications*. Cambridge University Press. (Slightly less standard than Wooldridge, but subjectively more readable.)
  - Gelman, Andrew and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press. (A standard, non-technical textbook for Bayesian hierarchical models.)
  - Jackman, Simon. 2009. *Bayesian Analysis for the Social Sciences*. Wiley. (Introduction to Bayesian statistical modeling with political science applications. Good mix of theory and practice.)
- **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.
  - Wooldridge, Jeffrey. 2010. *Econometric Analysis of Cross Section and Panel Data*. MIT Press. (Standard panel data textbook.)
  - Dobson, Annette J. and Adrian G. Barnett. 2008. *An Introduction to Generalized Linear Models*, 3rd ed. Chapman and Hall/CRC. (Generalized linear models)
  - 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)
- Hastie, Trevor, Tibshirani, Robert, and Friedman, Jerome. 2010. *Elements of Statistical Learning.* Springer. (Classic statistical data science text.)
- James, Gareth, Witten, Daniela, Hastie, Trevor, and Robert Tibshirani. 2013. *Introduction to Statistical Learning*. Springer. (Covers many of the topics as the above text with more intuition and less mathematical depth.)
- Gelman, Andrew, John B. Carlin, Hal S. Stern and Donald B. Rubin. 2012. *Bayesian Data Analysis*, 3rd ed. Chapman and Hall/CRC. (A standard textbook on applied Bayesian statistics.)
- Kruscheke, John. 2014. *Doing Bayesian Data Analysis*, 2nd ed. Academic Press. (Verbose, funny, and not very mathy.)
- Hoff, Peter. 2010. *A First Course in Bayesian Statistical Methods*. Springer. (Mathematical with simulations.)

11 Tentative Course Outline

11.1 Generalized Linear Models and Extensions

Binary Outcome Models

1. Binary Logit and Probit Models

   Recommended:

   - Cameron & Trivedi Ch. 14

2. Theory of Maximum Likelihood Estimation

   Recommended:

   - 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

   - Cameron & Trivedi Ch.10

Discrete Choice Models

1. Multinomial Logit and Probit Models
2. Ordered Logit and Probit Models

   Recommended:

   - 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.

GLM & Event Count Models

1. Theory of Generalized Linear Models

   Recommended:

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

2. Event Count Models

   Recommended:

   - 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:

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

   Optional:

   - Imai, Kosuke and Kim, In Song. 2016. "When Should We Use Linear Fixed Effects Regression Models for Causal Inference with Longitudinal Data?"

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 Statistical Data Science Methods

Resampling Methods

1. Cross-Validation and Bootstrap

   Recommended:

   - James, Witten, Hastie, and Tibshirani, Ch. 5.1-5.2

Shrinkage Methods and Regularization

1. Lasso, Ridge, and Elastic-net

   Recommended:

   - Hastie, Tibshirani and Friedman, Ch. 3.4
   - Belloni, Alexandre, Chernozhukov, Victor, and Hansen Christian, 2014. "High-Dimensional Methods and Inference on Structural and Treatment Effects," *Journal of Economic Perspective* 28, No 2: 29-50.

   Optional:

   - James, Witten, Hastie, and Tibshirani, Ch. 6.1-6.2
   - Tibshirani, R., 1996. "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society*. Series B (Methodological), pp.267-288.
   - Kim, In Song. 2016. "Political Cleavages within Industry: Firm-level Lobbying for Trade Liberalization".

11.3 Text Analysis (Date TBD)

Introduction to Text Analysis

1. Guest lecture: Richard Nielsen

   Recommended:

   - Lucas, C., Nielsen, R. A., Roberts, M. E., Stewart, B. M., Storer, A., & Tingley, D. (2015). Computer-assisted text analysis for comparative politics. Political Analysis, mpu019.
   - Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, mps028.

11.4 Bayesian Statistical Modeling

Introduction to Bayesian Statistics

1. Basic Concepts of Bayesian Statistics

   Recommended:

   - Jackman, Ch.1, 2

2. Markov Chain Monte Carlo

   Recommended:

   - Jackman, Ch.4 (skim), 5 and 6

   Optional:

   - Kruschke, Ch. 7 (more intuitive explanation of MCMC)

- 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.
- Resnik, Philip and Hardisty, Eric. 2010. "Gibbs Sampling for the Uninitiated."
- 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:

   - Jackman, Ch.7
   - Gelman & Hill, Ch.14, 15
   - Kruschke, Ch. 9 (more intuitive explanation of MCMC)
   - Shor, B., Bafumi, J., Keele, L., & Park, D. (2007). "A Bayesian multilevel modeling approach to time-series cross-sectional data." *Political Analysis*, 15(2), 165-181.
   - Warshaw, C., & Rodden, J. (2012). "How should we measure district-level public opinion on individual issues?" *The Journal of Politics*, 74(01), 203-219.
   - Caughey, D., & Warshaw, C. (2015). "Dynamic Estimation of Latent Opinion Using a Hierarchical Group-Level IRT Model." *Political Analysis.*
   - Caughey, D., & Warshaw, C. (2015). "The Dynamics of State Policy Liberalism, 1936–2014". *American Journal of Political Science.*

11.5 Tree-based Methods (Time Permitting)

1. Regression Trees, Bagging and Boosting

   Recommended:

   - Hastie, Tibshirani and Friedman, Ch. 9.2, 10,

2. Random Forests

   Recommended:

   - Hastie, Tibshirani and Friedman, Ch. 15

   Optional:

   - Wager, S. and Athey, S., 2015. "Estimation and inference of heterogeneous treatment effects using random forests." arXiv preprint arXiv:1510.04342.

3. Bayesian Additive Regression Trees

   Recommended:

- Chipman, H.A., George, E.I. and McCulloch, R.E., 2010. "BART: Bayesian additive regression trees." *The Annals of Applied Statistics*, pp.266-298.

Optional:

- Green, D.P. and Kern, H.L., 2012. Modeling heterogeneous treatment effects in survey experiments with Bayesian additive regression trees. *Public opinion quarterly*, p.nfs036.