Fall 2016 | Tue., 2–4pm | Classroom: CGIS K354

Gov 1000/2000/2000e/Stat E-190: Quantitative Research Methodology

Matthew Blackwell

Office: CGIS K305
Office Hours: Wed., 2–4pm
mblackwell@gov.harvard.edu
http://www.mattblackwell.org

TF: Mayya Komisarchik
Office Hours: TBD
email: mkomisarchik@fas.harvard.edu

TF: David Romney
Office Hours: TBD
email: dromney@fas.harvard.edu

NOTE: Syllabus is under construction and will change before the start of term.

General Information

How can we detect voting irregularities? What causes individuals to vote? In what sense (if any) does democracy (or trade) reduce the probability of war? Quantitative political scientists address these questions and many others by using and developing statistical methods that are informed by theories in political science and the social sciences more generally. In this course, we provide an introduction to the tools used in basic quantitative social science research. The first part of the course cover introductory univariate statistics, while the remainder of the course focuses on linear regression models. Furthermore, the principles learned in this course provide a foundation for the future study of more advanced topics in quantitative political methodology. We will cover both the

theoretical and computational aspects of statistics, proving important theorems and learning how to analyze real data.

While the tools of statistical inference are worth studying in their own right, another goal of this course is to provide graduate students (and some undergraduates) with the necessary skills to critically read, interpret, and replicate the quantitative content of many political science articles. As such, the statistical methods covered in this course will be presented within the context of a number of articles. Throughout the term, we will reanalyze the data and revisit the conclusions from various prominent papers in the social sciences.

Who should take which course number?

We have designed the class with a great deal of flexibility in mind and have various course numbers that correspond to students with different backgrounds. Note that all sections of the course will use the R statistical computing environment.

Gov 2000 This is the default course number for all graduate students who will be doing *any* empirical research in political science or the social sciences more generally. Even if you think you're going to use only use qualitative methods in your research, you should still take this course to give yourself a solid footing and understanding of quantitative methods. Talk to older graduate students and faculty—you'll need to know more methods than you think both for the dissertation and the job market. This section will teach you to be flexible data analysts, capable of tailoring standard methods to the unique situation of each task. You will control the tools, not the other way around. You will learn to write and adjust code to replicate and critique results from the literature. You will also learn how to work with basic statistical theory, working with proofs of canonical results and building the foundation for working at a higher methodological level. *This level will require some calculus and matrix algebra knowledge at the level of the Harvard Gov Math Prefresher (see prerequisites section below).*

Gov 2000e This course number designed for those students who plan to do *absolutely no empirical work* in political science. Students in this section will focus on the analysis and critique of methods and empirical work. You will still do some data analysis, but the coding aspect of the class will be less emphasized. In its place, you will be expected to produce a higher and more competent level of analysis/criticism in all assignments (no free lunches). Furthermore, there will be less use of technical mathematical statistics in this section.

Gov 1000 This is the default course number for undergraduate students and roughly covers the same material as Gov 2000 with special tailoring for Gov undergraduates. This means fewer technical computational or mathematical questions on assignments.

Stat E-190 This is the course number for Harvard Extension School students. Those taking the course for graduate-level credit will do work corresponding to the Gov 2000 level while those taking it for undergraduate-level credit will do work corresponding to the Gov 1000 level. Lectures and sections will be taped and made available to all students within 1 to 2 days.

Prerequisites

The most important prerequisite is a willingness to work hard on possibly unfamiliar material. Statistical methods is like a language and it will take time and dedication to master its vocabulary, its grammar, and its idioms. This presents a challenge for us as instructors to give you the best intuition and a challenge for you as a student to work hard to internalize that intuition.

Formally, the prerequisites vary for different types of students. For graduate students in the Government Department there are no course prerequisites except the completion of the Math Prefresher (or the equivalent). For other graduate students, undergraduate students, and Extension School students, the prerequisite for all course numbers is a basic course in statistics such as Gov 50, Gov E-1005, Stat E-100, or the equivalent. For graduate-level enrollment (either in Gov 2000 or Stat E-190), some previous experience with probability, calculus, and matrix algebra is strongly recommended. Working knowledge of basic algebra is assumed for all course numbers.

For any student who meets the prerequisites yet is concerned with his or her preparedness for the course, we strongly encourage the following in advance of the semester. First, we recommend reading and working through the exercises in David Freedman, Robert Pisani, and Roger Purves, Statistics, 2007 (any of the older editions should suffice as well). Next, we encourage familiarization with the R for the section of the course the student intends on taking. Moreover, if the student plans on typesetting problem set answers in LaTeX, familiarity with the LaTeXmarkup language would be helpful. Resources on R and LaTeXare available under the "Resources" tab on the class website. Finally, it may be helpful to review the material from the Government Department Math Prefresher, available here: http://projects.iq.harvard.edu/prefresher/home.

Course Details

Lectures

Lectures will be held weekly and will cover the broad theoretical topics of the course. In addition, we will work through example problems, computation in R, and canonical or insightful proofs of key results. Lectures will be taped and made available to both the extension school students and those in the College and GSAS.

Reading

There are readings for each topic and they mostly cover the theory of the method. Obviously, read the required readings and any others that pique your curiosity. In addition, though, engage with the readings: take notes, re-derive expressions, write down your impressions or confusions, talk with your classmates, preferably through Canvas (see below for more details). All of your classes should be pushing your research forward and you will be more creative the more you actively read. Some weeks there may be reading quizzes that are optional, but will count toward your participation grade if completed.

Grading

- weekly homework assignments (50% of final grade)
- a midterm exam (10% of final grade)
- cumulative take-home final exam (30% of final grade)
- participation (10% of final grade).

Homeworks

Methods are tools and it isn't very instructive to read a lot about hammers or watch someone else wield a hammer. You need to get your hands on a hammer or two. Thus, in this course, you will have homeworks on a weekly basis. They will be a mix of analytic problems, computer simulations, and data analysis. For all sections, the homework will be due before the start of section (Thursdays at 5:00pm). Solutions will be posted on Thursday night after section. Students have the option to "self correct" one homework over the course of the term on the basis of the solution key (due the following Thursday at 5:00pm). These corrections should take the form of an updated homework with comments added to indicate where mistakes were made and that demonstrate an understanding of those mistakes.

These homeworks should be typed and well-formatted, with tables and figures incorporated into the text. We will grade on a $(+, \checkmark, -)$ basis (including half grades between these categories).[1] No late homework will accepted except in the case of a documented emergency.

[1] All sufficiently attempted homework will be typed and well organized with all problems attempted, and all sufficiently corrected homework will include typed and well organized comments integrated into the original homework. The instructor will determine sufficiency in borderline cases.

Midterm

The midterm will be a checkout exam, that should only take a few hours to complete, and only involves short analytical problems. You will have five hours to complete exam, but it should take less time than this. This exam will be available for checkout one week after we finish the material on univariate statistics, and it is designed to ensure that all students understand the foundational material before we move to regression. Both FAS and Extension School students will upload the completed exam to the course Canvas site. The **tentative** schedule for the midterm is that it will be distributed on roughly October 13th and due back on October 20th. *This schedule is subject to change depending how fast our course is moving.*

Take-home Final

The take–home final exam will be handed out on Friday, December 2, one week before the last day of reading period. It will be due at 5:00pm on Friday, December 9, the last day of reading period. The take–home final primarily involves data analysis and interpretation. Note that the format and goals for the take–home exam are very different from the format and goals for the midterm exam. Both FAS and Extension School students will upload the completed exam to the course Canvas site.

Collaboration Policy

We encourage students to work together on the homework assignments, but you must write your own solutions (this includes computer code), and you must write the names of your collaborators on your assignment. I also strongly suggest that you make a solo effort at all the problems before consulting others. The midterm and the final will be very difficult if you have no experience working on your own. **There is no collaboration allowed on either the midterm or the final exam.**

Participation

Ten percent of the grade will be awarded for class participation, quality of presentation on the homework, and reading comments. A preliminary version of the lecture notes will be posted on Friday evening with references to pages of the textbook on the notes. Posting questions on Canvas about the assigned reading or the lecture notes will count towards class participation. These comments and questions provide feedback for tailoring the Tuesday lecture to the needs of the students in the course.

Sections

There will be two sections for this course, the time and location of which are to be determined. These sections will be taped and made available to all students. These sections will focus on reviewing material from class that is useful for the problem sets. The first two sections will cover an introduction to R and a review of basic probability for those who would like one.

Course Canvas Site & Discussion Board

We will be using Canvas to host the course website this year. You can find the site at the following URL: https://canvas.harvard.edu/courses/4533. On Canvas you will find a Discussion Board for class-related discussion. The quicker you begin asking questions on Canvas, the quicker you'll benefit from the collective knowledge of your classmates and instructors. This is an ideal forum for posting questions regarding the course material and/or computing. I encourage students to reply to each other's questions, and a student's respectful and constructive participation on Canvas will count toward his/her class participation grade.

Office Hours and Availability

My office hours are 2-4pm Wednesdays and really any time that I'm in my office with the door open. The office hours for the TFs are posted above and will be held in the CGIS Cafe, known as the Fisher Family Commons. If you have questions about the course material, computational issues, or other course-related issues please do not hesitate to set up an appointment with either any of us.

If you have a general question, you can also post it on Canvas. This is almost always the fastest way to get an answer. However, you can also email me directly at mblackwell@gov.harvard.edu. If the question is of general interest, I will forward the question and my answer to the class. Make sure to tell me explicitly in your email if you would like to stay anonymous.

Required Books

The following textbook is **required** for this course:

- Wooldridge, Jeffrey M. *Introductory Econometrics.* New York: South-Western. 5th edition. (earlier/later editions are fine)

Optional Books

- Angrist, Joshua D. and Jörn-Steffen Pischke. 2008. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.
- Berksekas, Dimitri P. and John N. Tsitsiklis, *Introduction to Probability*. Athena Scientific. (Also available as lecture notes online.)
- Diez, David M., Christopher D. Barr, and Mine Çetinkaya-Rundel. 2015. *OpenIntro Statistics*. 3rd edition. https://www.openintro.org/
- Freedman, David, Pisani, Robert, and Purves, Roger. 2007. *Statistics*. W.W. Norton & Company. 4th edition.
- Gelman, Andrew and Hill, Jennifer. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.
- Ashenfelter, Orley, Levine, Philip, and Zimmerman, David. 2003. *Statistics and Econometrics: Methods and Applications*. John Wiley & Sons.

Computing

We'll use R in this class, which you can download for free at http://www.r-project.org. R is open source and available on all major platforms (including Solaris, so no excuses). You can find a virtually endless set of resources for R on the internet, including this Getting Started With R page. You may also be interested in using RStudio, an editor and development environment for R. If you are completely new to R, you should complete this online short course, Try R.

Preliminary Schedule

The following is an anticipated schedule of course topics. The plan is to cover one topic per week, but we will go as fast as needed to make sure that everyone is understanding the material. Check the Canvas site to know what we will be covering the in an upcoming lecture.

§1 Introduction

- Course details and requirements
- What are the goals of the course?
- Basic descriptive statistics

§2 Random Variables and Probability Distributions

- Random variables
- Probability distributions
- Cumulative distribution functions
- Summarizing probability distributions
- Famous distributions
- Simulating from random variables

Reading

- Wooldridge, Appendix B.1, B.3, B.5
- Bertsekas & Tsitsiklis, 2.1-2.4 & 3.1-3.3

§3 Multiple Random Variables

- Joint and conditional distributions
- Covariance, correlation, and independence

Reading

- Wooldridge, Appendix B.2, B.4
- Bertsekas & Tsitsiklis, 2.5-2.7, 3.4-3.5

§4 Sums, Means, and Limit Theorems

- Distribution of the sample mean
- Useful inequalities
- Law of Large Numbers
- Central Limit Theorem

Reading

- DeGroot & Schevrish, Ch. 6

§5 Estimation and Statistical Inference

- Populations, samples, statistical models
- Point estimation
- Properties of estimators
- Confidence intervals

Reading

- Wooldridge, Appendix C.1-C.3, C.5

§6 Hypothesis Testing

- Hypothesis testing
- Small sample testing and confidence intervals
- Bootstrap

Reading

- Wooldridge, Appendix C.6, Ch. 6 (Appendix 6A, pp. 225–6)

§7 What is Regression?

- Difference in means
- Nonparametric regression
- Parametric models and linear regression
- Bias-variance tradeoff

Reading

- Wooldridge, Ch. 1

§8 Simple Linear Regression

- Mechanics of Ordinary Least Squares
- Assumptions of the linear model
- Properties of least squares
- Gauss-Markov Theorem
- Inference with regression

Reading

- Wooldridge, Ch. 2

§9 Linear Regression with Two Regressors

- Mechanics of regression with two regressors
- Omitted variables and multicollinearity

Reading

- Wooldridge, Ch. 3, 7.1-7.3

§10 Multiple Regression I: Matrix Form of Regression

- Matrix algebra
- Mechanics of multiple linear regression
- Inference in a multiple linear regression model

Reading

- Wooldridge, Ch. 4-5, Appendix E

§11 Multiple Regression II: Interactions, Nonlinearities, F-tests

- Interactions between variables
- Logged variables, quadratic functional forms
- Inference on multiple coefficients

Reading

- Wooldridge, Ch. 4.5, 6.2, 7.4

§12 Diagnosing and Fixing Problems

- Model fit, outliers, and influential observations
- Heteroskedasticity
- Functional form

Reading

- Wooldridge, Ch. 6.2, 9.5, Ch. 8

§13 Panel Data Models

- Clustering
- Fixed effects
- Random effects

Reading

- Wooldridge, Ch. 14