Advanced Quantitative Research Methodology, Gov2001, Gov1002, and Stat-E200

Gary King, Stephen Pettigrew, and Solé Prillaman

Class: Mondays 2–4pm (CGIS-S010 Tsai Auditorium).
Section: Wednesdays 6:00-7:00pm & 7:00-8:00pm (CGIS-K354).
Course website: http://projects.iq.harvard.edu/gov2001

Gary King
king@harvard.edu, http://gking.harvard.edu
Phone: 617-500-7570, Administrative Assistant: 617-495-9271
Office: CGIS-K313

Stephen Pettigrew, Teaching Fellow
pettigrew@fas.harvard.edu, stephenpettigrew.com
Office hours: TBA.

Solé Prillaman, Teaching Fellow
soledadartiz@fas.harvard.edu, http://scholar.harvard.edu/sprillaman
Office hours: TBA.

Who Takes This Course? Do I have to take it for a grade? Can I sit in? Following Gov 2000 or the equivalent course in linear regression, Gov2001 is the second in the methods sequence for Government Department graduate and undergraduate students. While not required, most Government graduate students doing empirical work take the course. Graduate students in other departments and schools at Harvard (and in the area) also take the course. Undergraduates preparing to write quantitative theses are especially welcome to take Gov 1002, which is taught along with this class. Non-Harvard students and others may also take this course by registering through the Harvard extension school, for course credit or as an auditor (see course number E-2001).

If there are seats in the room you're welcome to attend even if you're not formally registered, but if possible we would appreciate if you would sign up formally (as our teaching fellows get paid more!). If you are not a Harvard student, you can easily do this via Harvard extension school course E-2001.

If you need cross-registration papers signed, please bring them to the first class. We observe that students who take the course for a grade participate more and get far more out of the experience (even among many of those who think or say it will be otherwise), but pass/fail and formal auditing are okay with us too.

Overview. For students who've taken a course in linear regression (such as Harvard's Gov2000), this course gives you the tools to learn new statistical methods or build them yourself. We focus on methods practically useful in real social science research. We aim to give you two types of skills

First, we who how to develop new approaches to research methods, data analysis, and statistical theory. More advanced statistical theory is not required when data and variables follow standard assumptions. Since this is not usually the case in most of the social sciences, we often cannot use ready-made statistical procedures developed elsewhere and for other purposes. We teach the underlying theory of inference (which, at its most fundamental is merely using facts you know to learn about facts you don't know); once understood, we can easily reinvent known statistical solutions to accommodate social science data, or even invent original approaches when required. Students will learn how to read an original scholarly article describing a new statistical technique, implement it in computer code, estimate the model with relevant data, understand and interpret the results, and present and explain the results to someone unfamiliar with statistics.

Second, students will learn how to make novel substantive contributions to the scholarly literature. In the past, some students who completed the course published a revised version of their class paper in a scholarly journal. For most of these students, this was their first professional publication

The syllabus, detailed lecture notes, and other course materials are available at the course website: http://projects.iq.harvard.edu/gov2001.

Prerequisites Gov 2000, a course in linear regression (with matrices), or the equivalent.

What to Expect from the Class Most in-class experience will be lecture-based, but some parts are designed as a collective experience. This means that other students will be counting on you (and you on them), and so please come to class prepared. If you don't understand something, that's perfectly fine; we'll figure it out together and make sure no one is left behind. But if you don't put in the effort, it will hurt what everyone gets out of the class.

We have redesigned the course again this year, including several new teaching and learning technologies. We expect you to make a genuine effort to participate in the following activities prior to the class for which they are assigned:

1. Watch a short video of Gary King giving part of a class lecture from a previous year.

2. Complete the assigned readings

3. For the readings, our web site offers separate but integrated collaborative annotation tools. This means that if you find a portion of the lecture difficult or confusing, you should post a question about it. You can pause the video and do this right there as you watch. Similarly, if you have a question about one of the readings, post a question on the discussions page on Canvas. If you think you may know an answer to a query another student posted, or have a suggestion, please try to answer it. In fact, if you merely have an interesting idea related to the video or text, please contribute that as well.

Being prepared by having watched the videos and done the reading enables us to devote class time to difficult, confusing, or interesting ideas that arise. We will also be able to make more detailed connections to student projects and interests.

Computational Tools The best way, and often the only way, to learn new statistical procedures is by doing. We will therefore make extensive use of a flexible (open-source and free) statistical software program called R and a companion package called Zelig. R is probably the most widely used statistical software, and Zelig is one of the most widely used packages in R. You will learn how to program in this class, if you do not know already.

For hardware, you are welcome to use your own computers. To install R and Zelig on your computer, see http://gking.harvard.edu/zelig. You are also welcome to use the HMDC computer labs (in the concourse and 3rd floor of CGI-Knafel), which have computers with R already installed on them. Harvard affiliates also have the option of registering for a Research Computing Environment (RCE) account through http://hmdc.harvard.edu. Having an RCE account allows you access to HMDC's cluster of servers, which are fast and well-equipped to handle large data sets or time-intensive procedures. In addition, these servers supply a persistent (linux) desktop environment that is accessible from any computer with an Internet connection.

Most of the probability and statistical theory in this class will be taught in the context of ”Monte Carlo simulation” (which we do not expect you to know prior to the course). We will write computer programs to verify, or substitute for, more difficult formal mathematical proofs. This intuitive technique will make it much easier to understand and to implement new statistical methods.

Problem Sets and Assessment Problems In addition to the final paper, you will have to complete weekly problem sets. We are going to be using the Quizzes tool on Canvas for assigning problem sets electronically. We strongly encourage students to work together on the problem sets, but the work that you turn in must ultimately be your own. Problem sets must be submitted each week by the beginning of section (Wednesday 6PM). The full solution key will be posted so you can review your answers. Because we will be posting answer/solution keys immediately after deadlines, late work will not receive any credit. You can still turn in late work for feedback and help learning the material. The problem sets - including looking at the solution keys - is an extremely important part of the learning process, so please keep up with the work!

Addtionally, we will be incorporating a new evaluation tool into the problem sets this year. Each week there will be a quiz on the Canvas site (separate from the problem set) which will serve as an ”assessment question.” This question will be about a topic covered in the previous week's problem set. Your scores on these questions throughout the semester will be averaged and incorporated into your final grade. Think of these questions as a final exam, which you're taking over the course of the semester. As such, you are not allowed to discuss the assessment questions with other students or seek the help of anybody else.

Replication Paper The main assignment is to write a research paper that replicates an existing piece of scholarship. The goal of the paper is to apply some advanced method to, or develop one for, a substantive problem in your field of study. You should aim to produce a publishable article, and, in fact, most students do publish their final paper in a scholarly journal. (I know it sounds hard, but that's only because you haven't learned some of the material we go over in class!) More information about the paper can be found at http://gking.harvard.edu/papers/.

You must choose a co-author and a paper to replicate by Wednesday, February 26, at 5pm, by which point you should email us a PDF copy of the paper along with a brief paragraph explaining your choice. You are also required to have one of your classmates sign off on your article choice after checking that your article meets all the criteria listed in Publication, Publication. On Wednesday, March 26, you must turn in a draft of the paper with little text but with figures and tables, and a proposed table of contents for your paper, in a relatively polished form. You should also arrange to hand over all of the data and information necessary to replicate the results of your analysis and reproduce your tables and figures. (Many students email their files; students with larger datasets often set up shared Dropboxes.) On that day, you will hand over your paper and materials to another student, and, in exchange, you will receive another student's paper. Your task for the following week is to replicate the other student's analysis and write a memo to this student (with a copy to us), pointing out ways to make the paper and the analysis better. You will be evaluated based on how helpful, not how destructive, you are.

The final version of the paper is due the day before Reading Period, Wednesday April 30, at 5pm. You must turn in a hard copy of the paper and arrange to hand over all data and code (either by email or by Dropbox). You must also follow standard academic practice and create a permanent replication archive by uploading all your data and code to the Gov2001 Dataverse: (http://dvn.iq.harvard.edu/dvn/dv/gov2001).

If you need an extension with the replication paper, you do not need to ask permission: We will accept papers until Monday, May 5, at 5pm, but since you will have had more time, papers turned in after the April 30 deadline will be graded according to proportionately higher standards. The number of incompletes we plan to give is governed by a Poisson distribution with $\lambda = 0.01$, so please plan accordingly.

Once all papers are turned in, we will turn over your replication paper to another student, and assign you a replication paper to evaluate. Your last assignment for the class will then be to read and comment on a fellow student's work and to grade this student according to certain guidelines we will provide. Your main objective is to give the student feedback on what changes and improvements need to happen in order for the paper to be published. As always, you will be evaluated based on how helpful, not how destructive, you are. Your comments on your fellow student's paper are due Monday, May 12, at 5pm.

Collaboration One of the best ways that people learn is by teaching and collaborating with others. In Gov2001 we facilitate collaboration in three different ways:

1. In lecture we'll use the Learning Catalytics platform for you to discuss difficult questions related to the content of the lecture. Learning Catalytics will automatically assign you to small groups to discuss your answers to these questions. This gives everybody a chance to teach and learn from their peers.

2. We also encourage you to help each other out on problem sets. While the final product that you turn in must be your own individual work, you can still seek the help of your peers if you get stuck on a particular part of the problem set. Learning from and teaching your peers is a great way to master the content of the class and foster relationships with your colleagues. Note that this does not apply to the assessment questions, which must be completed independently.

3. On the replication paper you will choose your co-collaborators. This will give you the chance to write a journal-quality research paper with the help of your peers.

Special Rules for Extension and Distance Learning Students This course is being offered as part of the Harvard Extension School's Distance Education Program. The recorded class meetings that you will view are from the Harvard FAS course, Government 2001, and this meets once per week throughout the term. Even though your participation will take place online, you are responsible for homework, readings, quizzes, and all other work. There will also be weekly on-campus section meetings and office hours for students who are able to attend, or watch the videotape of the section. Please see the Harvard Extension School distance education web site for more information.

Students taking the class through the extension school will complete a final exam instead of the replication paper. They will, however, participate in the replication assignment by replicating others' work.

All students will need to have access to the course webpage, which is operated by FAS. If you do not already have a Harvard ID, please make arrangements to get one or to set up an XID.

Readings All readings are available on the course website. All are freely accessible to members of this class, except for the required text: Gary King, Unifying Political Methodology: The Likelihood Theory of Statistical Inference. (Ann Arbor: University of Michigan Press, 1998).

In addition to the required text, we will assign a wide variety of scholarly papers. We will announce at the end of every class meeting what the reading assignment will be.

Help! Help is available when you need it. If you have any questions about the homework, your paper, or anything else related to the course, please use the class discussion forum on the Canvas site. Since all three of us and all students will be reachable via this platform, it's a very efficient way to get answers to questions that do not fit as comments on the video annotation tool sites. Please also respond to inquiries if you happen to know the answer. (You can control how often the platform emails you a digest of the latest Q&A.)

We will also use Canvas to post announcements regarding course logistics, including readings, video assignments, and problem sets.

Grading Final grades will be a weighted average of the replication paper (or final exam), weekly problem sets, and participation. (There will be no final exam.)

"Participation" includes preparing for and joining in the discussion in class, making a serious effort to contribute to video annotation, answering discussion queries on Canvas if you have a suggestion, and other ways of helping your classmates learn more. Finally, since everyone learns more when more connections exist among students, finding ways to help build class camaraderie can also count as part of participation.

Weekly Schedule The timeline below gives the outline of the weekly schedule. Students are expected to:

1. Lecture preparation (before Monday)
    - Watch assigned video segments and annotate
    - Do assigned readings and discuss on Canvas
2. Attend class (Mon. 2-4PM)
3. Complete the problem set (Wed. 6PM)
4. Attend section (Wed. 6-7PM or 7-8PM)

Keeping up with the weekly schedule is extremely important not only for your learning but for the rest of the class as well.

Course Outline After the foundational material is presented (roughly the first third of the class), I will introduce a large variety of statistical models and methods. I will choose these based on what makes sense from a pedagogical perspective at first, but as the semester goes on I will choose more and more material based on students interest and class projects.

For more information on the content of the class, see the detailed lecture notes online, which gives a general outline. Here's another version of some of the material:

Foundations

1. What is statistics?
2. What is political methodology?
3. Models and a language of inference
4. The role of simulation
    (a) To solve probability problems
    (b) to evaluate estimators
    (c) to compute features of probability distributions
    (d) to transform statistical results into quantities of interest
5. Stochastic components (normal, log-normal, Bernoulli, Poisson, etc)
6. The relationship between stochastic and systematic components and data generation processes
7. Systematic components (linear, logit, etc.)
8. Uncertainty and Inference
    (a) Probability as a model of uncertainty
    (b) Probability distributions, theory, discrete, continuous, examples
9. Inference
    (a) Inverse probability problems
    (b) The likelihood theory of inference
    (c) The Bayesian theory of inference
    (d) Detailed example: Forecasting presidential elections
10. Properties of maximum likelihood estimation (finite sample, asymptotic, etc.)
11. Precision of likelihood estimates

Specific Topics We will not get to all these topics, and the list of topics we do cover will likely include others than those listed here, depending on student interest.

1. Discrete regression models
   - (a) Binary variables
   - (b) Interpreting functional forms
   - (c) Ordinal variables
   - (d) Grouped uncorrelated binary variables
   - (e) Event count models — Correlated and uncorrelated events; over and under dispersion.
2. Basic time series models
3. Basic multiple equation models, including identification
4. Multinomial choice models
5. Models for selection bias, censoring, and truncation
6. Models for duration
7. Hurdle models
8. Case-control designs
9. Model dependence
10. Matching as nonparametric preprocessing
11. Rare events
12. Neural network models
13. An overview of MCMC methods
14. Compositional data
15. Missing data (item and unit nonresponse) problems
16. Ecological inference (avoiding aggregation bias)
17. Models for reciprocal causation and endogenity
18. Empirical and hierarchical Bayesian analysis
19. Time series cross-sectional data
20. Models for interpersonal incomparability in surveys
21. Text analysis

Theory of Teaching For the theory of teaching behind this class see the paper ”How Social Science Research Can Improve Teaching”, or the accompanying video at the same link.

References

Required

King, Gary. 1989. *Unifying Political Methodology: The Likelihood Theory of Statistical Inference.* Ann Arbor: University of Michigan Press.

A variety of papers will be assigned as well, available on the web.

Recommended

It is also helpful to have access to a book on R/S programming. We recommend

Fox, John. 2002. *An R and S-Plus Companion to Applied Regression.* Sage Publications.

Imai, Kosuke, Gary King, and Olivia Lau. 2008. *Zelig: Everyone's Statistical Software*, Manuscript.

Ripley, Brian D. and Venables, William N. 2002. *Modern Applied Statistics with S*, Springer.

Suggested

Pawitan, Yudi. 2001. *In All Likelihood: Statistical Modelling and Inference Using Likelihood.* Oxford University Press

Barnett, Vic. 1982. *Comparative Statistical Inference.* 2nd edition. Wiley.

Chiang, Alpha. 1984. *Fundamental Methods of Mathematical Economics.* McGraw-Hill.

DeGroot, Morris H. 1986. *Probability and Statistics* Addison-Wesley. or Mendenhall, William and Robert J. Beaver. 1994. *Mathematical Statistics with Applications.* Duxbury.

Edwards, A.W.F. 1984. *Likelihood.* Cambridge University Press.

Gelman, Andrew et al. 2004. *Bayesian Data Analysis.* Chapman and Hall.

Gill, Jeff. 2008. *Bayesian Methods: A Social and Behavioral Sciences Approach*, 2nd ed, Chapman and Hall.

Harvey, Andrew C. 1990. *The Econometric Analysis of Time Series.* MIT Press.

Joreskog, Karl G. and Dag Sorbom, edited by Jay Magidson. 1979. *Advances in Factor Analysis and Structural Equation Models.* University Press of America.

King, Gary. 1997. *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data.* Princeton: Princeton University Press.

Kleppner, Daniel and Norman Ramsey. *Quick Calculus.* Wiley.

Lee J. Bain and Max Engelhardt. 1987. *Introduction to Probability and Mathematical Statistics.* Duxbury.

McCullagh, Peter and J. A. Nelder. 1993. *Generalized Linear Models* Chapman-Hall.

Mills, Terence C. 1990. *Time Series Techniques for Economists.* New York: Cambridge University Press.

Norman J. Johnson and Samuel Kotz. *Distributions in Statistics*, four volumes. John Wiley and Sons.

Rice, John A. 1995. *Mathematical Statistics and Data Analysis, 2nd Ed.* Belmont, CA: Duxbury Press.

Rubinsten, Reuven Y. 1981. *Simulation and the Monte Carlo Method*, New York: John Wiley.

Schafer, Joseph L. 1997. *Analysis of Incomplete Multivariate Data.* New York: Chapman-Hall.

Tanner, Martin A. 1996. *Tools for statistical inference: observed data and data augmentation methods*, 3rd edition. New York: Springer.