Advanced Quantitative Research Methodology, Gov2001, Gov1002, and E-2001

Gary King, Iain Osgood, and Maya Sen

Class: 2–4pm Mondays (CGIS K354); Section: 6–7:30pm Thursdays (CGIS K354).

Gary King
King@Harvard.edu, http://GKing.Harvard.Edu
Phone: 617-495-2027, Administrative Assistant: 617-495-9271
Office: 313 CGIS-K

Iain Osgood, Teaching Fellow
osgood2@fas.harvard.edu
Office: 219 CGIS-K
Office hours: 10–noon Mondays (CGIS HMDC Basement Lab), or by appointment

Maya Sen, Teaching Fellow
msen@fas.harvard.edu
Office hours: 10–noon Wednesdays (CGIS HMDC Basement Lab), or by appointment

Also see the Class Web Site (`http://isites.harvard.edu/k66011`), some detailed lecture notes (`http://gking.harvard.edu/g2001syl/notes.shtml`), and a PDF version of this document (`http://gking.harvard.edu/g2001syl/g2001syl.pdf`).

Who Takes This Course? Do I have to take it for a grade? Can I sit in? Following Gov 2000 or the equivalent course, Gov 2001 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 under class Gov 1002, which is taught with this class. Non-Harvard students and others may also take this course by distance learning videos made available through the Harvard extension school, for which you can get course credit if you desire (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 through via Harvard extension school course E-2001 (See the course web site for information).

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 tend to get 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 ok with us too.

Description. Building on the analytical and theoretical background of Gov 2000, this course gives you the tools to build statistical models and useful in real social science research.

The course covers how to develop new approaches to research methods, data analysis, and statistical theory. More advanced statistical theory is not required when data and variables fit standard assumptions. Since this is not usually the case in political science and related disciplines, we often cannot use ready-made statistical procedures developed elsewhere and for other purposes. Once a underlying theory of inference is understood, it is easy to "reinvent" known statistical solutions to accommodate social science data, or to conceive original approaches and new statistical estimators when required.

Upon finishing the course, students should be able to read an original scholarly article describing a new statistical technique, implement the computer code, estimate the model with relevant data, understand and interpret the results, and an explain the results to someone unfamiliar with statistics. A substantial portion of those who complete the course publish a revised version of their class paper in a scholarly journal.

The syllabus, detailed lecture notes, and other course materials are available at `http://GKing.Harvard.Edu/class.shtml`.

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

Evaluation The main assignment is a research paper that applies some advanced method to, or develops one for, a substantive problem in your field of study. The goal of the paper is to write a publishable article, and in fact most graduate and undergraduate 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/. There will be no final exam.

Weekly readings and class assignments are the norm. Students are expected to do reading and computer work beyond that required for the class assignments and will be graded based on the additional knowledge and understanding gained from these sources. We will make suggestions about extra reading on an individual basis, depending on the specific data and statistical problems encountered in student research.

You must choose a co-author and a paper to replicate by Thursday, March 4, at 5pm, by which point you should submit via email a PDF copy of the paper along with a brief paragraph explaining your choice. On Thursday, March 25, 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 turn in a CD with all the data and information necessary to replicate the results of your analysis and reproduce your tables and figures. On that day, we will give your paper and disk to another student we choose, and give you 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 first day of Reading Period, Thursday, April 29, at 5pm. If you need an extension, you do not need to ask permission: We will accept papers until Monday, May 3, at 5pm, but since you will have had more time, papers turned in after the 29th 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.

Final grades are assigned as a weighted average of the research paper, weekly assignments, and our assessment of what you learned in the class.

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 lectures that you will view are from the Harvard FAS course, Government 2001, and this meets once a week throughout the term. While these are recorded lectures, the other aspects of the course are "live". This means you are responsible for homework, exams and all other work. There will also be weekly on-campus section meetings and office hours for students who are able to attend. Please see the Harvard Extension School distance education web site for information on the distance ed program, details on how to view lectures and for technical support. The sections will also be videotaped.

Students taking the class through the extension school will be assigned a final project instead of the class paper. They will, however, participate in the replication assignment by replicating others' work. An extension school student who prefers to do the paper assignment instead of the final project must get permission from the instructors by the third week of the semester and make satisfactory progress with weekly assignments during the semester.

Course Plan 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 or impossible formal mathematical proofs. This intuitive technique will make it much easier to understand and to implement new statistical methods.

The best way, and often the only way, to learn new statistical procedures is by doing. As such, we will make extensive use of a flexible (open-source and free) statistical software program called R, which we do not expect you to know prior to this class, although if you want to get a head start this is the place to start. Most class assignments will involve some use of R and a companion package called Zelig. We expect you to try out ideas developed each week in class by writing short programs to implement different statistical procedures. 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 Zelig. You are also welcome to use the HMDC computer labs, 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 HMDC. Having an RCE account allows you access to HMDC's servers, which are fast and well-equipped to handle large data sets or time-intensive procedures. In addition, they supply a persistent desktop environment that is accessible from any computer with an internet connection.

What to do today In order to meet your deadlines, you will need to find a coauthor and begin working on your paper *very* soon. See http://gking.harvard.edu/papers/ for details.

Help If you have any questions about homework, your paper, or anything else related to the course, we recommend that you email the class list at gov2001-l@lists.fas.harvard.edu. Since all three of us and all students will be reachable via the class list, it's the most efficient way to get answers. In addition, please do respond to inquiries if you happen to know the answer. (If you

don't want to receive all the mailings, use a mail filter to put them in a separate folder, although we do not recommend this for this mailing list.)

Outline and Lecture Notes My teaching strategy is to go through the material as fast as possible under the constraint that everyone in the class understands what is going on. Because the time necessary for the latter cannot always be predicted, I find it unhelpful or impossible to set out fixed dates for lectures on specific topics. As such, you should expect me to speed up and slow down at a moment's notice, depending on how easily the class understands different concepts.

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 at `http://gking.harvard.edu/g2001syl/notes.shtml`. Here's a general outline.

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 electiohttp://gking.harvard.edu/papers/ns

10. Properties of maximum likelihood estimation (finite sample, asymptotic, etc.)
11. Precision of likelihood estimates

http://gking.harvard.edu/papers/

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

References You have access to the *Current Index to Statistics* (CIS), an excellent electronic bibliography of statistical literature, now available through lib.harvard.edu.

Books required for this course are available in the Coop. Most of the readings after the start of the course will be based on articles that are available on the web.

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.