Syllabus: Quantitative Social Science Methods, I (Gov2001, Gov1002, and Stat E-200)

Gary King
Harvard University*

July 12, 2020

This year, the class will be online only. The lectures will be recorded and made available through our learning platform Perusall. In addition, we will have weekly class meetings on Zoom to discuss the course material. Section information is posted Canvas, and all current information about the course and an updated version of this document can be found on the class web page **j.mp/G2001**.

1 What's it about?

This is a first course in political methodology — a version of data science or applied statistics. Political methodology is the methodological subfield of the discipline of political science, akin to econometrics within economics, psychometrics within psychology, sociological methodology within sociology, biostatistics within public health and medicine, and dozens of others. These methodological subfields are increasingly interconnected across disciplines and are often known together under broader monikers, such as data science, applied statistics, or computational social science. Political science is an unusually diverse discipline, welcoming of an exceptionally broad array of approaches, substantive questions, theories, and scholars. As such, learning methods in political methodology will tend to give you experience with a broader array of specific methods and, a focus on unifying perspectives that can help you integrate an understanding of approaches originating in many other areas.

Our goal in this course is to help you do high quality research. We discuss three broad, interrelated subjects, and make some progress on each every week.

1. *Understanding Statistical inference* — which is simply using facts you know to learn about facts you don't know. We do this from (a) a conceptual perspective so that you truly understand the methods we discuss, and ultimately (b) so that you feel appropriately and completely comfortable using these methods in your own

*Albert J. Weatherhead III University Professor, Director of the Institute for Quantitative Social Science, 1737 Cambridge Street, Harvard University, Cambridge MA 02138; GaryKing.org, King@Harvard.edu, (617) 500-7570.

work. With this knowledge, you should be able to easily digest articles about new methods invented after this class ends, implement the methods, apply them to your data, interpret the results, and explain them to others.

2. *Making novel substantive contributions to a scholarly literature*. This sounds hard, but almost everyone gets there and numerous graduate *and* undergraduate students in previous years have published revised versions of their class papers in scholarly journals as their first professional publication. Large numbers of class papers have also turned into books, senior theses, dissertations, blog posts, and conference presentations, and many have won awards and have been reported in the media.

3. *Learning how to choose a topic for research, especially how to solve hard problems by changing rather than answering the question, and how to identify a big idea.* Almost every assignment you've had since nursery school involved doing the best you could to answer an immutable question posed by the instructor. Yet, when conducting research, you get to choose what question to study, and that question has a bigger impact on how good the answer is than almost any other factor. This critical (and rarely taught) skill has a massive effect on careers of successful academics, politicians, startup entrepreneurs, consultants, advocates, and others.

Several important parts of this class are designed as *collective experiences* — specifically designed on the basis of social science research to help you learn more, understand the material faster, and remember it for longer. This means that other students will be counting on you (and you on them), and so not meeting a deadline or not attending class hurts you *and* all those around you. Please come to class prepared and meet all assignments as listed on the class web site. As you will see below, not meeting one assignment can have cascading effects on your fellow students. If you don't understand something, that's perfectly fine; we'll figure it out together and make sure you're not left behind. Your main job is to try and to stay engaged.

2 Who Takes It?

A diverse array of students have taken this course — including undergraduates; graduate students from every Harvard school; graduate students, postdocs, others from nearby universities; and many others via distance learning. Unexpected connections with students in fields you might not know about or at different levels, experiences, or ages tends to be a valuable part of the class experience. Here's a summary:

This course, as Gov2001, is the first course in the methods sequence for *Harvard Government Department graduate students*. Almost all Government graduate students take this course. *Graduate students in other departments and schools at Harvard* (and in the area) also frequently take the course.

Undergraduates should sign up for this class as Gov1002. Integrating you into this course, and treating you like a graduate student can be a great experience, and not as scary as it sounds. Unlike most of college, research is a team sport, with everyone contributing what they are capable of. Sometimes undergraduates have better math or coding skills, and graduate students may have more wisdom about the scholarly literature (and sometimes the reverse); together they make great research teams.

Non-Harvard students and others may take this course by registering through the Harvard extension school, for course credit or as an auditor (see course number Stat E-200). Students taking this course for graduate credit will have additional requirements for the final in comparison with students taking the course for undergraduate credit.

If you need *cross-registration papers* signed, please bring them to the first class. If you're willing to put the effort in, you're welcome in our class. No special permission is necessary.

Auditing: We observe that students who take the course for a grade participate more and get far more out of the experience (even among those who say they are different!), but pass/fail and formal auditing are okay with us too. Be careful not to fool yourself, however: This is a nearly universal empirical regularity.

3 Prerequisites

Students taking Gov2001 should have the same background as provided by undergraduate political science (or social science) programs at major colleges and universities, including the typically required undergraduate methods class (i.e., with topics such as introductory probability and statistics, data analysis, regression, and probability). Most importantly, you should have a basic understanding of linear regression and probability distributions. Some experience with the statistical programming language R would be helpful but not necessary (see the course website for how to get up to speed).

4 Weekly Assignments

Expect reading and problem set assignments every week. Readings will be announced in the last few minutes of every class; problem sets will be released on Canvas and Perusall after every lecture.

Readings and Video Lectures The readings and video lectures have a collaborative, social component that involves annotations and Slack-like chat channels. It is essential for you to participate in these activities outside of class. This ensures that you obtain answers to questions whenever they occur to you; that you learn more because you have the opportunity to teach others ("teaching teaches the teacher," often more than the student); and that you remember better as happens when learning during social interactions. Read on for how we will do this operationally.

Problem sets and assessment questions Find a regular group of 3–4 classmates to work on *Group Problem Sets* (if you're having trouble finding a group, please ask the TFs). For these, "cheating" is encouraged in a specific sense following these ground rules: You must give each question a try on your own before your group meeting and, after your meeting, write up your work on your own (i.e., by yourself, without having anyone check your work before you hand it in). You must also communicate the membership of problem set groups to the TFs before the first assignment and if it changes. We will also have *Individual Problem Sets*, which may not be discussed with anyone else. Think of

them as a take home final, distributed in a less stressful (but more real world way) over the course of the entire semester.

Write up your responses to all questions in Group Problem Sets and Individual Problem Sets *as a teaching document*. Write a couple sentences for each response to explain, to provide adequate context for a classmate who will read your problem set, the TFs, and me. This will help us understand you and, most importantly, will help you understand the material better. Problem sets are graded on a 0 to 5 point scale, and 1 point of 5 is devoted to "understandability."

So what are the groups for? Reaching an impasse when working alone on difficult problems is more common than when in a group. Research also shows that social connections motivate learning, and helps keep us focused longer than we are able on our own. It's a lot more fun too. *So then why try it first, and write it up, by yourself?* Actual learning occurs when you figure something out yourself (that "Ah ha" feeling also tends to be highly memorable) and then especially when you explain it to others — and much less so when someone tells you the answer. Thus, you will find that our problem sets have questions that often come in (unidentified) sets, since in case you wound up being told the answer to the first question in the set you will have other chances to have that deep, memorable feeling of having figured it out yourself for the second one.

Technology We will acquire, read, and annotate reading materials at `Perusall.com`. Most readings are available for free but some need to be purchased, which should be done through Perusall. The readings and videos can also be accessed in Perusall by starting from Canvas, Harvard's learning management system. *All questions about problem sets or assessment questions* must be asked (only) in Perusall by annotating the assignment so that everyone receives the same information and benefits equally. (Perusall is a project my collaborators and I designed to help you learn; if you have suggestions for features, I'd love to hear them.) We will use Perusall to post announcements. You can also communicate in chat channels in Perusall privately with instructors, privately with other students in your group (without having to share personal information like cell phone numbers with each other), one-on-one with any other student in class, or with the whole class or section.

If you have a question about one of the readings, a problem set, a solution set, or something else, add an annotation or comment in Perusall. If you think you may know an answer to a query another student posted, or have a suggestion, please try to answer it. If you merely have an interesting idea that might be of interest to others in class, please contribute that as well. It doesn't matter at all if you say something wrong, or get the wrong answer in these discussions, but it does matter a lot that you be engaged with the material and your fellow students.

What we talk about in class depends largely on the content of your out-of-class annotations and interactions. Participating in Perusall will help make our class better for you, me, and everyone in class.

How Assignments Work

1. You must prepare assignments in R Markdown. The TFs and I will know who you are, but the PDF you upload will be anonymized.

2. The problem set will be distributed on Perusall. Clarifying questions must be asked by annotating the problem set so all can see. (Obviously, you should not give away the answer even if you know it at this point.)

3. Turn your assignment in by the date and time indicated on the class web site.

4. A few minutes after it is due, a solution set will be released on Perusall. (For this reason, we are unable to give any credit for late assignments, although we will still give you comments.) Any questions about the problem set solutions must be asked in Perusall by annotating the solution set distributed there.

5. You will be randomly and anonymously assigned the problem set written by another class member. By a specific date and time given on the course web page, your job is to annotate this problem set by explaining as clearly, helpfully, and concisely as you can how each answer could have been improved, using the distributed solution set as guidance. Your goal is to teach your classmate so they truly and deeply understand how to solve this problem and problems like it. When you find a mistake, you should try to go beyond referring your classmate to the solution set and identify where your classmate made a mistake. Answers that are wrong obviously require more explanation from you, but may also be able to help your classmate learn how to improve a question, or connect it to related information you know or on the web, even if they got it right.

6. The instructors will assign a grade to your classmate for their problem set (which is our ultimate responsibility of course) and, if possible, improve the explanations in your annotations. The grade you receive for this assignment is the grade for what you turned in plus 1 extra point if you do a good job helping your fellow student and 2 points if you do unusually great job. (Some of the difference between 1 and 2 points will depend on the quality of the student's work you are assigned. On average over the semester, our random assignment will equalize these differences.) If you do not complete your assigned peer-grading in the time allotted, your total grade for your assignment and peer grading will be 0.

7. Read the grades and annotations on your assignment and try to understand how you might have improved your answers. Any questions you might have on grades can be sent privately to us through a Perusall chat channel designed for this purpose.

Section Section will cover various topics, including review of class materials, hints for working through the problem sets, and help with computing issues. Attendance is *strongly* encouraged. There will be two sections held per week: time information is posted on Canvas. The material covered will be identical in the two sections happening in a given week, so feel free to come to whichever is best for your schedule. Additionally, our TFs will each have 2 hours of office hours every week for more help in a less structured setting. We will send out a poll to find the times that work best for the group and send out confirmed office hours times in the first week of class.

5 Replication Paper Assignment

For graduate students enrolled in Gov2001, the main class assignment is to write a research paper that replicates and extends an existing scholarly work, while applying some advanced method to a substantive problem in some substantive field of study. Most undergraduate students enrolled in Gov1002 and all non-Harvard students enrolled in Stat E-200 complete a final exam instead of the replication paper (more information for these students in Section 10).

Your goal for this assignment is to produce a paper that is publishable in a scholarly journal — something we assume you have never done before and do not presently have any idea how to do. We will show you! Detailed information on the assignment can be found in an article I wrote about it called "Publication, Publication," at GaryKing/papers along with continuing updates. For initial versions of some papers from recent years, see the class Dataverse at j.mp/G2001dv.

As this assignment involves carefully choreographed hand-offs and interactions that connects you to everyone else in class, please respect these deadlines. Everyone is depending on you. Here's the list of requirements; the exact date and time of the deadline for each is given on the class web site, j.mp/G2001.

1. Identify your coauthors. All papers must be coauthored with one or two other members of this class. If you have problems, or would like suggestions, talk to the TFs.

2. Identify a scholarly article to replicate that meets our specific criteria. Upload to Canvas a PDF copy of the article, along with a brief paragraph (of less than about 200 words) explaining your choice. This paragraph must also list a classmate (outside your group) willing to testify that your article choice met all the criteria listed in "Publication, Publication".

3. Turn in a draft of your paper with completed figures and tables, and a proposed outline of the paper, in a relatively polished form. This draft need not have much text yet (although the more you complete, the more useful comments you will get in return). Also turn in a *replication data file*, with all of the data and information necessary to replicate your results and reproduce your tables and figures. At the same time, we will assign your paper to several other students to replicate, and you will receive another group's paper to replicate.

4. Replicate the other group's proto-paper and write a memo to them (with a copy to us), pointing out ways to improve their paper and analysis. You will be evaluated based on how helpful, not how destructive, you are. The best comments are written so fellow students can hear and learn from them rather than trying to demonstrate how smart you are.

5. Turn in the final version of the paper. By the same deadline, you must also follow standard academic practice and create a permanent replication data archive by uploading all your data and code to the Gov2001 Dataverse (j.mp/G2001dv). If you would like an *extension* with this (and only this) deadline, you do not need to ask permission: We will accept papers for a week after the due date given on the class web site (although since you will have had more time, papers turned in after the original deadline will be graded according to proportionately higher standards).

6. Once your paper is turned in, we will assign it to another student and assign you a small set of other papers to evaluate. Your last assignment for this class is to read and comment on a fellow group's final paper and to give a tentative grade to this paper according to specific guidelines we will provide. Your main objective is to give feedback on what changes and improvements need to happen in order for the paper to be published (we'll explain!). As always, you will be evaluated based on how helpful, not how destructive, you are.

6 Computational Tools

The best way to learn new statistical procedures is by doing. We will make extensive use of a flexible (open-source and free) statistical software program called R and a companion package called Zelig (another project of mine we designed for you and those in your position). R is among the most widely used statistical software packages, and Zelig is a widely used package written 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 zeligproject.org. You are also welcome to use the HMDC computer labs if they are available (in the concourse and 3rd floor of CGIS North-Knafel, 1737 Cambridge Street), which have computers with R already installed on them. Harvard affiliates also have the option of registering for a Research Computing Environment (SID) account through http://hmdc.harvard.edu. Having a SID 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.

7 Theories of Teaching and Learning

To improve this class, I often work on the theory and practice of teaching and learning. Out of this class has come a variety of projects, some of which we will use in this class. These include Learning Catalytics (LearingCatalytics.com), Perusall (Perusall.com), a collaborative video annotation system, a textbook replacement (booc.io), an article called "How Social Science Research can Improve Teaching" (j.mp/HowSSt), the data sharing and archiving project called dataverse (dataverse.org), statistical software projects Zelig: Everyone's Statistical Software (zeligproject.org) and CLARIFY: Software for Interpreting and Presenting Statistical Results, the articles "Replication, Replication" (j.mp/replrepl), and "Publication, Publication" (GaryKing.org/papers), and other articles and projects (see j.mp/gklearn). Ideas and suggestions welcome.

For 2020, we have developed a video annotation system (part of Perusall), a specially designed video format, and a new way of running the in class experience. Suggestions are always welcome.

8 How to Find Me

I do not have regular office hours (which, if you think about it, are mainly designed to prevent you from visiting the professor nearly the entire week). When not closed by a pandemic, I'm in my office at K313 CGIS with the door open; come by when it is convenient for you (and no one else is there). My administrative assistant in the office next to me can help find me if you can't.

In the pandemic, and in general, sending a message in Perusall to me is often the easiest way to find me. The TFs (and your fellow students) are also a tremendous resource. If you have a question about a problem set, please annotate the problem set in Perusall instead.

If you're registered for this course, you may join our private Facebook group designed for alumni of this course, `j.mp/G2001fb`; we communicate, distribute various job info, consulting gigs, and occasional other information.

9 Class Grades

Final grades will be a weighted average of (a) the replication assignment (or final exam), (b) weekly problem sets and assessment questions, and (c) engagement. "Engagement" includes coming to class and section, coming prepared, joining in the discussion in class and out of class (through Perusall and via assignment and paper groups). We recognize that some may prefer to participate more in person, online, or both, and compensating for one by emphasizing another is fine. However, to receive engagement credit, you must at a minimum come to class every week and be a responsible member of the class community. Finding other ways of helping your classmates learn more or to build class camaraderie is also helpful and appreciated.

Participation in discussions, and annotations on Perusall, will be evaluated based on the quality and quantity of meaningful engagement with the material and your fellow students; how often you give the right answer is not relevant.

The number of incompletes we plan to give is governed by a Poisson distribution with $\lambda = 0.01$, so please plan accordingly.

For undergraduate students enrolled in Gov1002, final grades will take into consideration that this is a graduate-level course.

10 Notes for Harvard Undergraduates and Extension School Students

Harvard undergraduates If you are a Harvard undergraduate student enrolled in Gov1002, all aforementioned requirements and course components apply to you, with the exception

of the replication paper assignment. We encourage undergraduates to take a final exam instead of completing this paper. If you choose the final exam option, you will still contribute to the replication assignment by replicating others' work. Most find this to be a great experience.

Extension School students If you are taking this course as part of the Harvard Extension School's Distance Education Program (Course E-200), you may participate in or watch the recorded class meetings online (they are from the Harvard FAS course Government 2001, which meets once per week throughout the regular term).

You are responsible for completing all weekly assignments and keeping up with the readings. Although you will complete a final exam instead of the replication paper, you will contribute to the replication assignment by replicating others' work. Most find this to be a great experience.

All students will need to have access to the course webpage in Canvas, which is operated by FAS. If you do not already have a Harvard ID, please get one or set up an XID at `xid.harvard.edu`.

11 Outline of Topics

The name ("file") in parentheses following the main topics below refer to filenames for slide decks I use in class (file.pdf) and handouts you can print (file-handout.pdf). If you have questions about these in class please speak up and after please annotate the handouts, which will also be in Perusall.

1. Introduction (basics)
   (a) Overview and Logistics
   (b) Statistical Models
   (c) Data Generation Processes (with Simulation)
   (d) Probability as a Model of the DGP
2. Theories of Inference (inference)
   (a) The Impossibility of Inference without Assumptions
   (b) Three Theories of Inference: Overview
   (c) Likelihood: Example, Derivation, Properties
   (d) Uncertainty in Likelihood Inference
   (e) Simulation from Likelihood Models
   (f) Extending the Linear Model with a Variance Function
3. Models for Binary Outcome Variables (bmodels)
   (a) Linear Probability, Logit, Probit Models
   (b) Interpreting Functional Forms

(c) Alternative Interpretations of Binary Models

(d) General Rules for Presenting and Interpreting Statistical Results

4. Assorted Models for Single Variable Outcomes (smodels)

   (a) Ordered Dependent Variable Models

   (b) Grouped Binary Variable Models

   (c) Count Models

   (d) Duration Models and Censoring

5. Model Evaluation (evalmodel)

   (a) How Do You Know Which Model is Better?

   (b) Evaluating Binary Variable Models

   (c) Robust Standard Errors

   (d) A Better Way to Use Robust SEs: An Application

6. Research Designs for Causal Inference (matchse)

   (a) Components of Causal Estimation Error

   (b) Research Designs

   (c) Issues in Ideal Designs

7. Detecting and Reducing Model Dependence in Causal Inference (modeldepmatch)

   (a) Detecting Model Dependence

   (b) Matching to Reduce Model Dependence

   (c) Three Matching Methods

   (d) Problems with Propensity Score Matching

   (e) The Matching Frontier

8. Multiple Equation Models (mmodels)

   (a) Identification

   (b) Seemingly Unrelated Regression Models

   (c) Reciprocal Causation (Endogeneity)

   (d) [Multinomial Choice Models]

9. Models for Missing Data (missing)

   (a) Overview

   (b) Missingness Assumptions

   (c) Application Specific Methods

   (d) Multiple Imputation

(e) Computational Algorithms

(f) What Can Go Wrong

(g) Time Series, Cross-Sectional Imputations

10. Anchoring Vignettes for Interpersonally Incomparable Survey Responses (vign)

(a) Introduction

(b) A Nonparametric Method

(c) A Parametric Method

(d) Illustrations

(e) Quantities of Interest

(f) Optimal Vignette Choice