Gov 1006: Models

David Kane

Description

Statistical models help us to understand the world. This class explores the use of models for analysis in the social sciences broadly, and in political science specifically. Does a history of slavery in a county influence contemporary political views? Does perceived demographic change impact policy preferences? Does having daughters affect a judge's rulings? We use the R programming language, RStudio, Git and GitHub. Each student will complete a *replication* as their final project, an attempt, successful or not, to reproduce the results from a published article in the academic literature. This class provides an introduction to data science and is designed to lay the groundwork for an empirical senior thesis.

Prerequisites: Gov 1005. If you took Gov 1005 prior to the fall of 2019, you will also need a course on statistics (Gov 50, Stat 104 or the equivalent). Special accommodations will be made for juniors concentrating in Government and planning on writing a senior thesis, the traditional constituency of Gov 61. You must have a laptop with R, RStudio and Git installed.

Logistics: Class meets from 12:45 to 2:45 on Wednesdays in Harvard Hall 202. You should be available for work with your classmates on at least one evening prior to class, presumable either Monday or Tuesday.

Course Metaphor

By taking this class, you seek entrance to the metaphorical School of Athens, the community of scholars stretching across the centuries. Admittance is not easy.

Course Goals

This course has three main goals. First, we teach you how to replicate and critique published academic articles in political science. This sets the stage for advanced course work and for an empirical senior thesis. Second, we emphasize a professional approach to data science, using tools like Git and GitHub. Third, we prepare you for the Data Science Program in the Government Department.

Course Staff

- Preceptor David Kane; dkane@fas.harvard.edu; IS South 310; 646-644-3626; office hours Thursday from 8:30 to 11:00, generally held in Fisher Commons. Please address me as "Preceptor," not "David," nor "Preceptor Kane," nor "Professor Kane," nor "Mr. Kane," nor, worst of all, "Dr. Kane." I respond to e-mail within 24 hours. If I don't, e-mail me again.
- Teaching Fellow: Alice Xu (alicexu@g.harvard.edu).

Figure 1: *The School of Athens*, 1511, by Raphael. "The School of Athens represents all the greatest mathematicians, philosophers and scientists from classical antiquity gathered together sharing their ideas and learning from each other. These figures all lived at different times, but here they are gathered together under one roof."

- Course Assistants: Jack Luby (jluby@college.harvard.edu) and Enxhi Buxheli (ebuxheli@college.harvard.edu). Jack's Study Hall will be Monday from 6:00 to 9:00 PM at Beren Hall room B102 in Winthrop. Enxhi's Study Hall will be Tuesday from 4:00 to 7:00 PM at Lowell House.

Textbooks

Regression and Other Stories (RAOS) by Andrew Gelman, Jennifer Hill, and Aki Vehtari is the required textbook for the class. It is unpublished but photocopies are available at Gnomon Copy for $50.

Course Policies

Philosophy: Same as Gov 1005.

Workload: The course should take about 10 to 15 hours a week, outside of class meetings, the midterm and the final project. This is an expected average across the class as a whole. **It is not a maximum.** Some students will end up spending less time. Others will spend **much, much more**.

Section: We do not have normal sections. Instead, you will spend 30 minutes each week meeting with a TF, either alone or in a small group.

Use your Harvard e-mail: Please use your official Harvard e-mail address for all aspects of this class, especially things like signing up for services like DataCamp, GitHub, and so on. Doing so makes it much easier for us to figure out who is doing what. This may not be easy if you already connect with these services but, even in that case, you should be able to add your Harvard e-mail address to your account.

Piazza: All general questions — those not of a personal nature — should be posted to Piazza so that all students can benefit from both the question and the answer(s).

Plagiarism: If you plagiarize, you will fail the course. See the Harvard College Handbook for Students for details.

Tools: You must use R, RStudio, Git and GitHub for this class. You are responsible for installing and updating all necessary tools on your laptop. We are not your tech support.

Missing Class: You expect me to be present for class. I expect the same of you.

Major Emergencies: We are not monsters. If you are hit with a major emergency, we will be sympathetic. A signed letter (not an e-mail) from your Resident Dean solves most problems.

Computer Emergencies: We are unsympathetic about computer emergencies. You should keep all your work on GitHub, so it won't matter if your computer explodes. If it does explode, you will only lose the work since your last push. You can restart your work on a public computer (the basement of CGIS Knafel has machines with R/RStudio installed) or on your roommate's computer.

Late Days: Assignments (DataCamp, Problem Sets and Final Project Milestones) are always due at 11:55 PM, unless specified otherwise. An assignment is a day late if it is turned in any time after it was due (even 5 minutes after) but within 24 hours. After that, it is two days late, and so on. You have 5 *late days* in total. Late days may be used for any assignment except for the midterm and final project presentation. **You should save your late days.** If you use them early in the semester for no particularly good reason and then, later in the semester, have an actual emergency, we will not be sympathetic. We will not give you extra late days in such a situation. (That isn't fair to your classmates, and we are all about fairness.) We will just, mentally, move the late days you wasted so that they cover your actually emergency. You will now be penalized for being late earlier in the semester, when you did not have a good reason for tardiness. You may only use one late day on a given assignment. Hand it in after more than 24 hours and you get a zero on that assignment. **But you still must hand it in!** Everything must be completed. Late days accrue until you do. Each day late (beyond the five allowed) results in -1 point to your final score. This decrement is a *point*

not *percentage* penalty. In other words, each additional late day used outside of the allotted five will drop your final class grade by one grade point out of 100.

DataCamp: We make extensive use of lessons from DataCamp. All DataCamp courses are graded pass/fail. They are due by Monday at 11:55.

Grading

Participation: 10 points. I expect you to participate, both in class and online. Helping your fellow students, especially on Piazza, is the best form of participation. Be a good class citizen. Missing class or TF meetings will cost you points.

DataCamp Lessons: 5 points. Grades are pass/fail only. **These are free points!** Given the level of the questions and the hints provided, it is essentially impossible not to get full credit as long as you make an honest effort.

Problem Sets: 25 points. Most of these will be individual. Others will be completed in groups, which we will assign. All students in a group receive the same grade. Problem sets are due at 11:55 on Tuesdays. Your Github repo must include an R markdown file and the knitted html. Submit the html to Canvas.

Midterm: 20 points. The midterm is take-home. It is open-book and open-web. Because students have different schedules, you can complete the midterm any time within a 7-day window starting after midterm distribution.

Final Project: 40 points. You will replicate and extend a published paper from the academic literature. There are 8 Milestones, each worth one point. Each Milestone requires the submission of an html (or PDF) document via Canvas. The presentation (and draft of your code/paper at presentation time) is worth 12 points. You then have till May 8 to make changes and submit a final version (worth 20 points) which will be public forever.

Final Project

The inspiration for the final project comes from "Publication, Publication" by Gary King. PS: Political Science and Politics, Vol. 39, No. 1 (Jan., 2006), pp. 119-125. See also "Teaching Replication to Graduate Students" by Dragana Stojmenovska, Thijs Bol, and Thomas Leopold. Teaching Sociology, 2019, Vol. 47(4) 303 –313.

All Milestones are due at 11:55 PM on Fridays.

Useful references include writing articles and reproducible documents R and dissertation tips.

Conclusion

If you had tried to complete a replication exercise for a published academic article before taking this class, you would have done X well. Now that you have taken the class, you will do Y well. The success (or failure) of the class can be measured by comparing Y with X.

Miscelaneous

- Keep up on the latest news in data science. We recommend subscribing to R Weekly and Data is Plural.

Schedule

Week 1: January 27.

Bring your laptops. We will have a full class meeting.

Khan Academy offers Statistics. Especially useful are the chapters on probability, random variables, sampling distributions and confidence intervals. All of this material should have been covered in your previous statistics class. Those of you who have not studied multivariate calculus or linear algebra may find these classes useful: Multivariable calculus and Linear algebra.

Week 2: February 3.

RAOS, chapters 1 to 3.
The Math Prefresher, chapters 1 – 7.

Assignments

DataCamp: Linear Algebra for Data Science in R. Students who have not taken Gov 1005 will, **instead**, do:

Introduction to the Tidyverse; Introduction to Shell for Data Science, Chapter 1, Manipulating files and directories; Introduction to Git for Data Science, Chapter 1, Basic workflow; and Communicating with Data in the Tidyverse, Chapter 3, Introduction to RMarkdown.

Problem Set #1.

Milestone #1. Meet with your assigned TF, either this week or next week. Ensure that your computer is set up correctly. Discuss ideas for replication paper. Submit an html with a few sentences about the paper you are currently considering. Read some background on organizing a project. For milestones this semester, unless we specify otherwise, you will just be submitting an html or PDF file via Canvas. That file should include the url for the current version of your replication project. (No worries if you throw away versions and need to change that url over time.) We can then examine the repo if we choose to.

Week 3: February 10.

RAOS, chapters 4 and 5.

Assignments

Problem Set #2.

Milestone #2. Create a repo with the data for the paper which you are considering. Create an Rmd that does something with that data. (Can be a single R command like `summary()`.) Submit the html generated by that Rmd. It is OK, for now, if you don't yet have the data for your paper. But you still must submit milestone #2, which requires data. What to do? Easy! Pick a back-up paper, with data available, and submit milestone #2 (and #3 and . . .) using that second-choice paper. As soon as you get data for your main paper, you can switch back. Always include a link to the repo, perhaps in footnote, in any html milestone submission.

Week 4: February 17.

Monday, February 17 is a holiday so DataCamp is due on Tuesday at 11:55 PM.

"Publication, Publication" by Gary King. PS: Political Science and Politics, Vol. 39, No. 1 (Jan., 2006), pp. 119-125. link This article provides the inspiration for our final project.

RAOS, chapters 6, 7 and 8. Please review Causality, Chapter 2 of *Quantitive Social Science* by Kosuke Imai. You should have already read this, either in Gov 50 or in Gov 1005.

Assignments

DataCamp: Multiple and Logistic Regression. The material on multiple regression should have been covered in your prior statistics course. Logistic regression may be new, but we won't really use it for a few more weeks.

Problem Set #3.

Milestone #3. Public GitHub repo with your data and a non-trivial README. (You can still change your paper if you want.) Submit a one paragraph html to Canvas which summarizes, in your own words, your paper. Include the url of your repo so we can check it out, if we choose to. In the repo, you must include a copy of the code of the original paper, interspersed with several pages of your own comments, and reformatted to meet our coding style guidelines. This is a lot of work! You do not have to get this code to run (yet), nor do you need to change the substance of the code. You just need to demonstrate that you understand what the code is doing and why it is doing it.

Week 5: February 24.

RAOS, chapters 9 and 10.

Assignments

DataCamp: Fundamentals of Bayesian Data Analysis in R. You should have been exposed to Bayes Theorem in your prior statistics class. This DataCamp course provides an overview/refresher on Bayesian approaches to data analysis.

Problem Set #4.

Milestone #4. You will need to use a variety of more advanced tools in order to create your paper. The single best reference is *R Markdown: The Definitive Guide*. The purpose of this milestone is to confirm that you know how to use them. The final version of your paper will be a PDF. (Consider using the **tinytex** package, especially if you are on Windows.) I recommend using the pdf_document2 class which comes with the **bookdown** package. Show us that you know how to use all the relevant tools by producing a short PDF which contains the following components: a bibliography (and associated references), a footnote, a table (using the **gt** package), a regression table (using whatever packages you like – options include **gtsummary**, **stargazer**, **kable**/**kableExtra**, and **report**). All of this can be brief (i.e., just two entries in the References is fine) and fake (made up data in the tables). We just want to see that you can get things to work. Submit the PDF to Canvas. Include the url for your Github repo in a footnote.

Week 6: March 2.

RAOS, chapters 11 and 12.

Assignments

Problem Set #5.

Midterm, covering Part 1 of ROAS, distributed March 7 and due Friday March 13 at 11:55 PM.

Week 7: March 9.

RAOS, chapters 13 and 14.

Week of March 16 is Spring Break.

Week 8: March 23.

"Let's Take the Con Out of Econometrics," by Edward E. Leamer. The American Economic Review, Vol. 73, No. 1 (Mar., 1983), pp. 31-43. link This is the best critique of empirical work in academia, as true today as it was more almost 40 years ago.

"Making the Most of Statistical Analyses: Improving Interpretation and Presentation" by Gary King, Michael Tomz and Jason Wittenberg. American Journal of Political Science, Vol. 44, No. 2 (Apr., 2000), pp. 347-361. link This is one of the most cited articles in political science in the last 20 years. Your final project *must* make use of these techniques.

Assignments

Milestone #5. All data from your paper must be processed and available in your repo. Submit a PDF via Canvas with the following components, all of which must be present in future submissions:

1. A footnote with your repo url and some verbiage about "All analysis for this paper is available . . ."

2. A beautiful graphic which uses this data. (May be similar to or different from a graphic in the original paper.) Use King et al (2000) for inspiration. This is the portion of the submission which will be graded most harshly. Make sure that you include a thorough caption.

3. A bibliography with at least five references, one of which will be the article you are replicating.

4. A 300 – 500 word overview of your replication paper. What analysis did they run? What did they conclude?

5. An Appendix which include a replication of at least one of the tables from your paper. (It can be a simple summary table.) Also, take a screen shot of the original table and include that image in your Appendix. We want to see how closely your results match the original paper's.

Week 9: March 30.

RAOS, chapters 15 and 16.

Assignments

DataCamp: Dealing With Missing Data in R

Problem Set #6.

Milestone #6. Submit a PDF or html via Canvas. (I usually do everything in html until the very end of the process to avoid spending too much time fussing with annoying formating issues. We required you to submit that last two milestones in PDF in order to ensure that your computer was set up properly. The final submission will be in PDF.) In addition to the the elements from Milestone #5, your paper should include:

1. An Appendix in which you replicate all results — or all the important results — from your paper. As with other aspects of this project, the exact requirements will vary across students, depending on the complexity of your replication paper. If you paper only has 3 or 4 tables, we expect you to replicate it all. If it has 50 tables, we do not expect that. Use your best judgment and talk with us. You must replicate any result which you plan to use as the base of your extension.

2. A clear statement about what aspects of the paper you were able to replicate and which parts, if any, you were not able to replicate.

3. 500 words about your proposed extension. You do not have to have done the extension yet. (That comes next week.) But it is time to start thinking about what your contribution to human knowledge will be. You seek admission to the School of Athens. What do you have to offer us?

Week 10: April 6.

RAOS, chapters 17 and 18.

Assignments

DataCamp: Bayesian Regression Modeling with rstanarm. I do not expect that we will work directly with Stan in this class. There are several packages which using Stan under the hood, including `rstanarm`. We will also use `brms`.

Problem Set #7.

Milestone #7. Completed extension. Place entire project on the web, using bookdown, Shiny or a different technology of your choice. You want a way to show off your excellent work to the world, both in preparation for Demo Day and forever after. We will discuss options in class.

Week 11: April 13.

RAOS, chapters 19 and 20.

Goal this week is to learn about R packages, which is often the best way to organize and distribute your work. *R Packages* by Hadley Wickham is an excellent reference.

Assignments

DataCamp: Developing R Packages

Problem Set #8.

Milestone #8. Completed rough draft of your paper, submitted as PDF.

Week 12: April 20.

RAOS, chapter 21.

Assignments

DataCamp: Hierarchical and Mixed Effects Models

Week 13: April 27.

Presentations of final projects. Students will participate in rolling one-on-one presentations. The purpose of these sessions is to get detailed feedback from your peers about your code and empirical approach. We will still grade your code and writing.

Resources

R Packages, 2nd edition by Hadley Wickham and Jennifer Bryan.
What They Forgot to Teach You About R by Jennifer Bryan and Jim Hester.
bookdown: Authoring Books and Technical Documents with R Markdown by Yihui Xie.
Happy Git and GitHub for the useR by Jenny Bryan.
R Markdown: The Definitive Guide by Yihui Xie, J. J. Allaire and Garrett Grolemund.
R for Data Science by Garrett Grolemund and Hadley Wickham.
Data Visualization: A practical introduction by Kieran Healy.