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: An course on computer programming (Gov 1005, CS 50 or the equivalent), a course on statistics (Gov 50, Stat 104 or the equivalent) and experience with R and Git. You must have a laptop with R, RStudio and Git installed.

Logistics: Class meets from 12:45 to 2:45 on Wednesdays in K-031 in CGIS Knafel. You should be available for work with your classmates on at least one evening prior to class, presumable either Monday or Tuesday.

Course Metaphors

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

In Marine Corps recruit training, the instructors tell you what to do, precisely, at every moment of every day. You may have found Gov 1005 to be a similar experience. The training to become a Force Recon Marine is very different. *You must do everything yourself.* The same is true in Gov 1006. John Ripley, a veteran of Force Recon, was prepared that Easter morning, not because anyone had told him what to do, but because his training and experience had prepared him for almost anything.

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, specifically the math requirements of the graduate courses (2000/2001/2002/2003).

Course Staff

- Preceptor David Kane; dkane@fas.harvard.edu; CGIS 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."
- Teaching Fellow Mark Hill; markhill@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."

Figure 2: *Ripley at the Bridge*, by Col Charles Waterhouse, USMCR (Ret.). "John Walter Ripley was a United States Marine Corps officer who received the Navy Cross for his actions in combat during the Vietnam War. On Easter morning 1972, Captain Ripley repeatedly exposed himself to intense enemy fire over a three-hour period as he prepared to blow up an essential bridge in Dong Ha. His actions significantly hampered the North Vietnamese Army's advance into South Vietnam."

Textbooks

Data Analysis Using Regression and Multilevel/Hierarchical Models by Andrew Gelman and Jennifer Hill is the required textbook for the class. Ignore any references to BUGS software. We will be using Stan instead.

Resources

R for Data Science by Garrett Grolemund and Hadley Wickham
Data Visualization: A practical introduction by Kieran Healy
Happy Git and GitHub for the useR by Jenny Bryan
R Markdown: The Definitive Guide by Yihui Xie, J. J. Allaire, Garrett Grolemund
R Packages by Hadley Wickham

Course Policies

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**.

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.

Slack: All general questions — those not of a personal nature — should be posted to Slack 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 all problems.

Computer Emergencies: We are very unsympathetic to 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: An assignment is a day late if it is turned in anytime 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* which 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.

DataCamp: We make extensive use of lessons from DataCamp. All DataCamp courses are graded pass/fail. Each week's course(s) are due by Tuesday at 10:06 AM (except in the cases of holidays or exams).

Priorities: The most important reading each week will be the replication article. I will expect you to have studied this closely. I will ask precise questions about it and expect you to know the answer. I think that GH is a wonderful textbook. I recommend that you read it thoroughly. But I will warn you ahead of time about any close quizzing I plan on doing in class.

Grading

Participation: 10 points. I expect you to participate, both in class and online. Helping your fellow students, especially on Slack, is the best form of participation. Be a good class citizen. Missing class 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. Each day late (beyond the five allowed) results in -1 point from the 5 points total allocated for DataCamp assignments. If you use up more than 5 points, further days late will make a negative contribution to your final grade.

Replications: 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. Replications are due at 9:00 AM on Wednesdays. Your Github repo must include an R markdown file and the knitted html (or pdf). Please e-mail a GitHub link *and* copies of the Rmd and html/pdf files to Mark and to Preceptor.

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. Extensions are possible, if requested.

Final Project: 40 points. You will replicate and extend a published paper from the academic literature. See here for details. Due May 10 at 10:06 PM. Any remaining Late Days may be used at this time. You will also present your interim work during one of the last two class sessions.

Replications

We spend two weeks on each replication exercise. In the first week, you will write a "Replication Report." In the second week, you will write a "Replication Commentary." Both times, you will e-mail us a link to your GitHub repo along with a copy of your Rmd and your pdf (or html) files. The title of these documents should be "Replication Report (or Commentary) on Author (year)" with the author(s) and year of publication of the replicated article.

There are two different types of writing involved in these documents. First, you make comments on the original code. These are often casual, sometimes obscure and occasionally speculative. Their audience is people who want to work directly with the code which supports this research, especially the most important audience: *future you*. Note that you are creating a new file, not simply adding comments to the code you got from the authors. There is nothing wrong with copying/pasting large chunks of code from their file to yours. (Of course, you will cite the original location of their code.) But this new file is your work. You understand what it does. (Or you are explicit when you don't understand something.) You explain the approach clearly. You own the work.

Second, you write prose that appears in the knitted html/pdf replication documents. These comments are much more formal, with the same style and cadence as an academic article.

In the Replication Report, you will replicate the published work and clean up the associated code. First, you need to get their code to run. This is harder than you might expect! Second, you will clean up their code, making it consistent with the Tidyverse Style Guide. You should also add lots of comments. It is fine — encouraged, in fact! — to admit that you don't understand why they have made some choices. We don't

expect you to, immediately, understand the finer points of the latest academic research. But we want to see you engage with their work, explain in your own words what is going on. Why have they made the choices they made? What other approaches might be worth trying? What don't you understand? The Replication Report html/pdf may very well include very little prose. (It must have an abstract and a bibliography which cites the original article and the data/code location.) It will simply show the tables and figures from the original article, with, perhaps, some brief comments. You will generally be adding hundreds of words of comments in the code itself, comments which won't appear in the Report. It is OK if you skip some tables and figures, but you must state this fact explicitly in the Report. Also, you should not just skip something without discussing it on Slack. Perhaps we can help you replicate it!

In the Replication Commentary, you will focus the analysis. Throw away all the random tables/figures you don't care about. Pick one or two key tables/figures. (Ensure that you can still replicate the published results for those figures/tables!) Then try to discuss/extend/expand their results in some fashion. Some weeks, we will provide precise guidelines as to what we want you to do, generally in accordance with important concepts from the Gelman/Hill readings. Other weeks, you will have much more freedom. In the Commentary, you will be providing the very rough framework of an academic article. You will have an abstract, for example. You will have a bibliography. Your (new) figures and tables will have clear captions. And so on.

Imagine that you will be working as a research assistant for a Harvard Professor of Government next summer. We want to prepare you to shine in that role. These replications are the best preparation possible.

Guidelines

As you work through your Replication Reports, Replication Commentaries and Final Project, keep the following guidelines in mind:

- Your tables should look like the tables you see in published academic articles. The gt package is the best tool to use for tables while the stargazer package is good for regressions. For the Reports, you should do your best to mimic the format of the original article. This will exercise your table-making muscles. You can make different choices than the published article as long as you justify those choices in your code comments.
- Your figures should look like the figures you see in published academic articles. R, itself, is the best tool for making figures and there are code chunk options, like `fig.cap` which are useful to know.
- Your bibliography should look like the bibliographies you see in published academic articles. The Replication Report will generally have only two entries in its bibliography: the original article and the data/code repository. The Replication Commentary will have those two plus a couple of others. For starters, you should cite R itself. See `citation()`. The Final Project will feature a full scale academic bibliography.
- Your code should follow the Tidyverse Style Guide.

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.

Milestones

All milestones are due at 10:06 PM

- Mid-March: Discuss your final project idea with either Preceptor or Mark. (2 points)
- March 29: Public GitHub repo with your data and a non-trivial README. (2 points)

- April 5: Initial pass at replicating the data analysis in your target paper. Does not need to be complete but you must have made an honest effort at getting everything to run. (2 points)
- April 12: Replication (or a documented failure to replicate) of all results in your target paper. Place all figures and tables in an Appendix to your Rmd. Begin write up. (2 points)
- April 19: Initial draft of your project. Must include a bibliography and an extension of the analysis from your target paper. (2 points)
- April 26/May 1: Class presentations. Must included a completed draft, which will be graded, of your paper. (10 points)
- May 10: Final project due. (20 points)

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

- Consider installing SourceTree to examine the details of your git repository. This is probably overkill for what we have done so far but might prove useful when you starting working on group problem sets.
- Keep up on the latest news in data science. We recommend subscribing to R Weekly.

Key Concepts

- Logistic regression
- Multilevel/Hierarchical models
- Bayesian inference
- Fake data simulation
- Predictive simulation
- Posterior predictive model checks

Key Manuscript Skills

- Bibliography and references
- Footnotes
- PDF creation
- Tables (`gt` package)
- Regression tables (`stargazer`, `kable` and `kableExtra`)
- Use of inst/extdata directory

Schedule

Week 1: January 30.

Bring your laptops. We will have a full class meeting, replicating some of the data analysis in:

McDermott, Rose, Dustin Tingley, and Peter K. Hatemi. 2014. "Assortative Mating on Ideology Could Operate Through Olfactory Cues." American Journal of Political Science 58 (4): 997–1005. article and supporting information

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.

Week 2: February 6.

GH, chapters 1 to 5. The goal this week is to review multiple regression and to introduce logistic regression. Chapter 5 is the most important part of the reading.

DataCamp: Multiple and Logistic Regression. The material on multiple regression should have been covered in your prior statistics course. Logistic regression may be new.

Replication #1: "Identifying Judicial Empathy: Does Having Daughters Cause Judges to Rule for Women's Issues?" by Adam N. Glynn and Maya Sen. American Journal of Political Science, Vol. 59, No. 1, January 2015, pp. 37–54. link We use this article in week 2 for two reasons. First, many of you will be familiar with the subject from Gov 50. Second, it has perhaps the best replication code of any article we read this semester.

This week, just try to replicate the article's results.

Week 3: February 13.

GH, chapters 7, 8, Appendix A, and Appendix B. We will not cover generalized linear regression models (beyond logistic regression) in this course, so we skip chapter 6. The use of simulation is critical, so make sure to understand chapters 7 and 8 thoroughly. Questions? Ask on Slack!

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.

Clean up, focus and extend your replication of Glynn and Sen (2015).

Those of you who have not studied multi-variate calculus or linear algebra may find these Khan Academy classes useful: Multivariable calculus and Linear algebra.

Week 4: February 20.

Monday, February 18 is a holiday so DataCamp is due on Tuesday at 10:06 AM.

GH, chapters 9 and 10. Focus this week is on causal inference. 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.

DataCamp: Hierarchical and Mixed Effects Models

Replication #2: Enos, Ryan D. 2014. "Causal Effect of Intergroup Contact on Exclusionary Attitudes," Proceedings of the National Academy of Sciences of the United States of America 111 (10):3699–3704. link

This article does not use matching, but it provides an interesting dataset with which to explore how we might use matching if the data had been observational.

Week 5: February 27.

GH, chapters 11 and 12. This is our first of two weeks on hierarchical models.

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`.

Clean up, focus and extend your replication of Enos (2014).

Week 6: March 6.

GH, chapters 13 and 14. This is our second of two weeks on hierarchical models. Again, we don't go beyond logistic models in this class, so you do not need to read chapter 15.

DataCamp: Linear Algebra for Data Science in R. Also, read the Math Prefresher, chapters 1 – 6. For students not interested in the Data Science Program, these assignments are optional.

Replication #3: Avidit Acharya, Matthew Blackwell, and Maya Sen. "The Political Legacy of American Slavery." Journal of Politics, Vol. 78, No. 3 (2016): 621-641. link

Midterm distributed March 6 and due Friday March 15 at 10:06 AM.

Week 7: March 13.

Because of the midterm, there is no further empirical work due today.

Week of March 18 is Spring Break.

Week 8: March 22.

GH, chapter 25. This week is devoted to missing data. DataCamp is due Tuesday because of Spring Break.

DataCamp: Dealing With Missing Data in R

"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.

"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.

Week 9: April 3.

GH, chapters 21 and 22. This is the portion of the syllabus which we are most likely to change during the semester. The GH chapters for this week and next focus on model fitting issues, an important topic! But their

use of the BUGS software is too dated to be useful for us, so we might replace/augment these readings with something more related to `rstanarm` and/or `brms`. Also, there may be better DataCamp classes available.

Replication #4: Enos, Ryan D. 2016. "What the Demolition of Public Housing Teaches Us about the Impact of Racial Threat on Political Behavior." American Journal of Political Science 60 (1): 123-142. link This article uses matching as well as difference-in-difference approach, as discussed in Chapter 10 of GH.

Week 10: April 10.

GH, chapters 23 and 24.

Clean up, focus and extend your replication of Enos (2016).

Week 11: April 17.

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.

DataCamp: Developing R Packages

Replication #5: Kosuke Imai, Gary King, and Carlos Velasco Rivera. Forthcoming. "Do Nonpartisan Programmatic Policies Have Partisan Electoral Effects? Evidence from Two Large Scale Experiments." Journal of Politics. link Replication exercise this week requires you to turn the results from Imai et al. (2019) into an R package. You are now operating on the cutting edge of political science research!

Week 12: April 24.

Presentations of final projects. Students will do a brief presentation, followed by everyone in the class cloning their repo and then attempting to replicate their work. We will then, as a group, dive into the complexities of their chosen data/model.

Week 13: May 1.

Presentations of final projects. It is fine if your project is still rough. You certainly will not have finished the writing. The purpose of these sessions is to get detailed feedback from your peers about your code and empirical approach.