POL 2504: Statistics for Political Scientists

Michael J. Donnelly*

Updated August 27, 2024

This course is designed to expose PhD students to the basic tools of quantitative analysis in political science. At the end of the course, you will be prepared to do three things:

1. Read and critique quantitative political science
2. Conduct basic analyses, including both descriptive inference and linear regression
3. Take POL 2507 and move forward in the broader quantitative sequence in our department

Item 1 is absolutely essential to your development as political scientists. Every subfield in political science other than political theory has a rich literature of both quantitative and qualitative research designs, and having the ability to read and understand the strengths and weaknesses of an article or book based largely on quantitative methods is required to grasp the full gamut of claims made about your dissertation topic. This class will give you the tools to interpret most quantitative articles and books and identify the (often implicit) assumptions on which their evidence is based. By identifying these assumptions, you will be able to decide for yourself whether to trust the results of a piece of scholarly work.

While you may or may not include substantial quantitative work in your dissertation itself, having that option greatly improves your ability to craft a dissertation that answers the questions you want to ask. This course relies on R, open-source software that is widely used both in the social sciences and industry. Facility with a statistical software is essential to the practice of quantitative political science. In our department, there are faculty and graduate students who use R, Python, Stata, SPSS, and probably others. We focus on R in this course because it is free and because I believe it is easier to pick up additional languages having already learned the basics of R than to pick up R based on knowledge of other software.[1]

The third item is similarly important. Though not all of you will go on to take more quantitative courses, especially if you are early in the program, you want to keep your options open. Graduate school — and your coursework-heavy first and second year especially — is the best time to obtain skills like quantitative analysis, which are often very hard to learn outside the classroom.

*Note the "J." There is another (retired) UT political scientist named Michael Donnelly. If you email him, I will not respond.

[1]Though it is not the primary goal of the course to prepare you for non-academic jobs, those of you who think that might lie in your future would do well to take this aspect of the course especially seriously, as skill with statistical software is one of the most oft-cited 'transferrable' skills that political scientists obtain in graduate school.

1 Contact Information

Professor: Michael J. Donnelly
Office: SS 3105
Office Phone: 416-978-0344
Email: mj.donnelly@utoronto.ca
URL: http://www.MichaelJDonnelly.net

Grader/TA: Jesslene Lee
Office hours: TBA
Email: jesslene.lee@mail.utoronto.ca

2 Prerequisites

This course has no prerequisites, though I assume you have some background knowledge of basic algebra. If your math skills are extremely rusty, contact me and I can offer some pointers on getting up to speed.

The course is designed for PhD students in political science. Students who do not fit this description should contact me in advance to discuss whether this is the right course for them.

3 Logistics

- Lectures: Tuesdays 12-2, SS 561
- Problem set review sessions: TBA
- Michael's Office hours: 12:30-2:30pm on Mondays. Appointments by email are also available.
- Email policy: I will respond to all emails within **two working days**. If I have not gotten back to you by then, feel free to send a reminder. If you email me about an assignment fewer than two working days before it is due, I cannot guarantee that I will respond in time for you to use my comments.

4 Course Requirements

Your final grade is based on participation, two types of assignments, and two tests.

- **Participation** (total of 10%): This is a lab-based course, mixing lecture, discussion, and group activities. Participation is therefore an essential part of the learning process. Attendance is mandatory. If you expect to miss class, or if you miss class unexpectedly, I expect you to let me know quickly.
- **Problem sets** (35%): There will be ten weekly problem sets of varying length. See the calendar below for weeks with particularly challenging problem sets (to help you plan ahead).
- **Problem set comments** (10%): All problem sets will be peer graded. I will then grade your comments.

- **Midterm** (15%): There will be an in-class midterm exam comprising both written and programmed responses.
- **Final exam** (30%): The final exam is cumulative and covers both statistical and programming questions.

5 General rules for problem sets

Submission Problem sets are due on Quercus at 12:00pm on the day of class. In order to allow your colleagues to grade them, you must get them in on time. See below for details on how to format them.

Plagiarism & cheating Cooperation on the problem sets is encouraged, but all students should understand the answers they submit (i.e., be able to explain them) and write up the answers separately (i.e., don't copy and paste). If you copy someone's code, even if you make changes, it will be obvious to a more experienced coder.

Plagiarism and other violations of academic integrity will not be tolerated. See the university policies[2] for more details.

Late assignments Late assignments will receive deductions of 20% per day. When submitting to Quercus, recognize that it can be slow, and that can sometimes push your submission past the deadline. Similarly, it is sometimes down for maintenance. I will not grant extensions for normal Quercus delays, so make sure to leave yourself a time cushion.

Grade appeals must be made within two weeks of receiving the grade. They must include a 100-200 word written statement of why the assignment deserves to be re-graded. The grade will change only in cases where the second grading is more than 10 points different from the first (i.e. a 60 will not be changed unless the second grading produces a score of 70+ or 50-). Grades can go up or down on the second grading.

Lost assignments Keep a backup of everything. Have you backed up your computer to the cloud or to an external hard drive this week? No? Go do that now and then come back and finish reading this.

Use of generative AI You may use ChatGPT or other generative artificial intelligence tools to help you with your code. I suggest not using it for a first draft of the code, but when you are stuck or when you have bugs, it is quite good at suggesting solutions. If you do so, document it. Each time you use it, put a comment in your code of the following form:

```
## Typo solved by pasting code into ChatGPT.
```

or

```
## Two level for loop structure corrected with ChatGPT.
```

Norms about citation of such tools are still developing, but for now, this is how we will do it for problem sets. In-class assessments will **not** allow the use of generative AI.

[2] http://www.utoronto.ca/academicintegrity

6 Using R

Learning R is required for this course. Though it is not required, I will conduct classes on the assumption that you work using RStudio, a front-end interface for R. It is freely available for any of the major operating systems. **Before** the first class, you should download

- R from https://cran.r-project.org/
- RStudio from https://www.rstudio.com/products/RStudio/ (note that the free desktop version is the one you're looking for).
- The textbook's package, for which instructions can be found at https://kosukeimai.github.io/qss-package/

If you have any trouble downloading and installing these, please let me know before the first class.

7 Readings

The main textbooks for this course, which I will refer to as QSS, is

- Kosuke Imai. *Quantitative Social Science: An Introduction*. Princeton: Princeton University Press, 2017.

This book covers both basic statistical concepts and an introduction for R. It does not assume any substantial statistical or programming background. Though the author is a political scientist, the examples come from political science, sociology, and economics. This is a good thing, as it forces us to take a step back from the details of our own literatures and examine key assumptions in our statistical approaches. Published in the US, it unsurprisingly uses many American examples, but also includes examples relevant to comparative politics and international relations. Where possible, I have inserted additional Canadian examples.

There is also a "tidyverse" version of the book[3] that you might prefer. You are welcome to use either. I go back and forth between base R and the tidyverse regularly, and you probably will too as you get used to R.

If you haven't done much math recently, I strongly urge you to get a copy of Will H. Moore and David A. Siegel. *A Mathematics Course for Political and Social Research*. Princeton (N.J.): Princeton university press, 2013 (M&S below), which is a very readable primer on basic math concepts, and also offers an intuitive take on probability.

Students may find the following books to be useful supplementary reading, and any student who anticipates using quantitative methods regularly should acquire at least a couple of them. Starred readings below are recommended, but not required.

- Scott Ashworth, Christopher R. Berry, and Ethan Bueno de Mesquita. *Theory and Credibility: Integrating Theoretical and Empirical Social Science*. Princeton (N.J.): Princeton University press, 2021. ISBN: 978-0-691-21383-5 978-0-691-21382-8

[3] Kosuke Imai and Nora Webb Williams. *Quantitative Social Science: An Introduction in Tidyverse*. Princeton: Princeton University Press, 2022. ISBN: 978-0-691-22229-5.

- (D&S) M H DeGroot and M J Schervish. *Probability and Statistics*. 3rd ed. New York: Addison-Wesley
- (KKV) Gary King, Robert O Keohane, and Sidney Verba. *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton: Princeton University Press, 1994
- Jeffrey M Wooldridge. *Introductory Econometrics: A Modern Approach*. 5th ed. Mason, OH: South-Western Cengage Learning, 2013
- Paul M Kellstedt and Guy D. Whitten. *The Fundamentals of Political Science Research*. New York: Cambridge University Press, 2013
- Gary King. *Unifying Political Methodology: The Likelihood Theory of Statistical Inference*. Ann Arbor: The University of Michigan Press, 1998
- Joshua D Angrist and Jörn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press, 2009

7.1 Learning R

If you aren't familiar with R, you should use both the textbook's resources and what you can find on the internet. A few good places to start:

- Hadley Wickham and Garret Grolemund. *R for Data Science: Visualize, Model, Transform, Tidy, and Import Data*. Sebastopol, CA: O'Reilly, 2017. URL: https://r4ds.had.co.nz/
- Rohan Alexander. *Telling Stories with Data*. Toronto: Bookdown, 2023. URL: https://www.tellingstorieswithdata.com/index.html
- Alex Douglas et al. *An Introduction to R*. Online: Bookdown, 2022. URL: https://intro2r.com

Here is a link to a list of more than 200 free books about R: https://www.r-bloggers.com/2021/05/big-book-of-r-has-over-200-books/. Some are good. Some are not.

8 How to prepare for class

Many of you have not taken a math class in a while, and many of you would have been among the most intelligent members of your high school classes, so you might not have ever learned how to read a math textbook carefully. In this course, there is a straightforward, though time-consuming, way to do this correctly. I estimate that most weeks should require 2-3 hours of reading, $\frac{1}{2}$-1 hours of computer review (swirl) exercises, and 1-5 hours of working on problem sets. Also include a brief time reviewing previous chapters for the midterm and the final (which you should do most weeks). I realize (and you should too) that this class is going to be more time-consuming than you are used to a single class being.

Reading QSS Each week, you will be assigned to work through part of a chapter in QSS. This should be done with your computer open and RStudio running. You do not have to run every command in the blue boxes as you read, but do NOT skip over them. Unlike reading for many purposes, reading to understand these concepts requires reading things like equations and computer code. It is part of the process.

When you come across an equation, I strongly suggest you try to read it aloud with concepts, rather than variable names. That is, instead of reading "e equals m c squared", say "Energy is equal to mass times the speed of light squared."[4] As you get lots of practice, the link between $\hat{y}_i$ and "the predicted value of y for individual i, given the data" will become second nature, and you can treat variable names as real things, reading $\hat{y}_i$ as "y hat i." Until then, read it out.

Doing the swirl exercises Each chapter is accompanied by swirl exercises — short reviews of both statistics and programming that are done at the command line in R. You should complete these exercises **before** class. See page 9 of QSS for instructions on how to download and install these from the command line and for a table showing approximately when to do each exercise. It is possible to skip questions in these exercises, but I recommend that if you do so, you read the answer and figure out why you did not get to it.

Supplementary reading The supplementary readings can be very useful. QSS uses a somewhat unconventional approach (one I think is easier to pick up), so the other textbooks will phrase things quite differently. That is useful for at least two reasons. First, if something is a bit unclear, reading it from another perspective may help it click in your mind. Second, if you think you understand a concept in one context, but rephrasing it makes you confused, that is a good sign that you don't quite have it.

Doing the problem sets The problem sets draw on both the previous week's readings and the current week's. Thus, the problem set that you will turn in before class in Week 2 will require you to have fully grasped everything we discussed in Week 1 **and** mostly grasped what we will discuss in Week 2. We will grade the second part of each problem set on effort, not correctness. Most of the 'lead' exercises (those that pertain to the current week's readings) will be doable just from what is in the text book. I therefore suggest you do the reading before doing the problem set.

The fact that there can be partial credit means that it is essential to properly format and comment your code so that I and your colleagues can see what you are trying to accomplish, even if you do not get it to work. In my own work, I am bad at this, and I regret not having developed better practices earlier. Get in the habit now and you will thank me years from now.[5]

In graduate school, I found that the best way for me to learn was to try to tackle a problem on my own first, then to meet in a regular (1-2 times per week) session to work with a small number of my colleagues. Study groups are essential to learning these things, but I suggest you avoid groups of more than 4 or 5 people, as larger groups make it easier to 'hide,' thinking that you've learned something when you really haven't.

[4] Don't worry, this is just an example. You didn't accidentally stumble into a physics class.

[5] A saying of unknown origin: "Your most important collaborator is you six months ago, and that person never answers emails." There is nothing more frustrating than staring at your own code and having no idea what you were trying to accomplish.

Writing up the problem sets Most of the questions in the problem sets require both data manipulation **and** written responses or interpretations. This may seem pointless in some cases, but the questions are designed in part to get you used to writing academic work based on numbers. Part of writing such work is saying things in multiple ways. To see this, pick up a journal article using quantitative methods and look for the authors' main claim. It will often be in a regression table, in a graph, in the captions to those two items, and in the text. You do not need to be quite that complete on problem sets, but I will often ask you to do more than one of them.

As discussed above, there are three ways to submit your work. I suggest you pick one and stick with it for at least a few weeks at a time. Part of the goal of the course is to help you figure out what kinds of work-flow processes work for you when doing quantitative work. To do this, you must practice a bit before you are comfortable.

1. You can submit using a combination of .R and .pdf files, with the latter produced via Word for the text and R for the figures.[6] The .R file should be *only* the script, not the output. If the question asks for a number or numbers as results from an analysis, include that in the write-up, not in the .R file.

2. Both text and figures can be included in a LaTeXdocument, with accompanying R files. In this case, also include a .R file.

3. You can submit using .Rmd and .pdf files. In this case, include the code and output together in the code chunks (that is, open the chunk with echo=TRUE). The text should still include the description of the output.

Number 1 is the one you will be most familiar with, but may not be the most efficient in the long run. Number 2 is my normal work-flow. Number 3 is the approach I would like to be in the habit of using, but I have an ingrained workflow from which it is hard to transition. In all cases, it should be easy for me (and the grader) to both see if you answered the question correctly and to see how you got there.

Reviewing the previous problem set You should, if at all possible, attend the problem set review sessions. Before going, you should read through the answer key carefully. These are not intended to focus on problem sets that you have not yet submitted, but rather on completed ones.

Peer grading Grades on the problem sets should be assigned using the following scale (decimals allowed):

0. Did not complete the assignment

1. Attempted some of the assignment, made little progress

2. Attempted the whole assignment, made little progress

3. Attempted the whole assignment, made substantial progress but many mistakes

4. Completed all parts of the assignment, with few mistakes

[6] If the figures are not in-text, make sure to give them informative file names.

Enter their grade in Quercus, along with detailed comments on their code and write-ups. I expect you to use the answer keys (which will be made available to you) to explain where and how your colleagues made mistakes (in the text box on Quercus). Be specific (e.g., "On problem 1 question 2, you fit a regression on DVone, when you should have fit it on DVtwo, because the question asked for a regression on the normalised dependent variable").

9 Course Outline

Required readings/tasks marked with a •, suggested readings with ∗.

Week 1: Introductions, logistics, introduction to R — September 9

- This syllabus. **No, really, read the whole thing! I put it all here for a reason.**
- QSS, Introduction and Chapter 1

Basic concepts

Week 2: Causation — September 16

- Sections 2.1 - 2.4 of QSS
- ∗ Chapter 3 of M&S
- ∗ Chapters 1 and 3 of KKV
- **Problem set 1 due**

Week 3: Measurement — September 23

- Sections 3.1 - 3.5 of QSS
- ∗ Sections 1-1.6 of M&S
- ∗ Chapter 2 of KKV
- **Problem set 2 due**

Probability

Week 4: Basic probability concepts — September 30

- Sections 6.1 of QSS
- ∗ Chapters 9 and 10 of M&S
- ∗ Chapter 1 - 2 of D&S
- **Problem set 3 due**

Week 5: Conditional probability — October 7

- Section 6.2 of QSS

- Steven H. Strogatz. "Chances Are". In: *New York Times* (Apr. 2010), Opinionator. URL: https://opinionator.blogs.nytimes.com/2010/04/25/chances-are/
- Kenneth Chang. *How Many Triangles Are There? Here's How to Solve the Puzzle.* New York, Aug. 2019. URL: https://www.nytimes.com/2019/08/21/science/math-equation-triangles-pemdas.html
- **Problem set 4 due**

Week 6: Random variables October 21

- Sections 6.3 - 7.2 of QSS
- Ellie Murray. *Having Confidence in Confidence Intervals: An Epidemiology Teaching Resource.* 2019. URL: https://medium.com/@EpiEllie/having-confidence-in-confidence-intervals-8f881712d837 (visited on 08/27/2020)
  - Chapter 11 of M&S.
  - Chapter 3 - 4 of D&S
- **Problem set 5 due**

Week 7: **.....Midterm!.....** November 4

Relationships and Prediction

Week 8: Causation in observational studies November 11

- Sections 2.5 - 2.7 of QSS
  - Chapter 1 of Wooldridge
- **Problem set 6 due**

Week 9: Reading week - catch up and review November 10

Week 10: Correlation and regression basics November 18

- Sections 3.6 - 4.2 of QSS
  - Chapter 2 of Wooldridge
- **Problem set 7 due**

Week 11: Multiple regression November 25

- Sections 4.3 - 4.4 of QSS
  - Chapter 3 of Wooldridge
- **Problem set 8 due**

Week 12: Regression uncertainty December 2

- Sections 7.3 - 7.4 of QSS
  - Chapter 4 of Wooldridge
- **Problem set 9 due**

Final exam

1. Date TBA