Quantitative Methods - SOCI504

McGill University, Winter Term 2018

Mondays 11:30am to 2:30pm

1 Overview

This course is designed to introduce you to quantitative social science methodology. The primary focus of the course will be on regression models with special attention paid to least squares linear regression and logistic regression models. These models are a good starting point as more complicated models can be thought of as extensions of these models.

Understanding basic regression models is important even if you don't plan to become a quantitative social science researcher. A large share of published work sociology relies on quantitative analysis. Probably two-thirds of articles in a typical ASR issue use some form of regression analysis or related technique. Understanding this work and being able to critique it will be a critical part of being a successful researcher.

Quantitative modeling can be best understood as an endeavor to approximate some social process in a stylized fashion. Our models will never be accurate - the real question is whether a particular model is close enough to reality to allow us to answer the questions we are interested in. Determining whether a particular model is adequate is an art. It takes a lot of practice and experience. For this reason we will be getting our hands dirty with real data as much as possible in this course. But this class can only provide the starting point. To truly master the material you need to gain experience analyzing a lot of data. I can't emphasize this enough. It would be a very good idea for you to spend a large portion of your free time analyzing both simulated and real data so that you gain more practical experience and intuition about theses approaches.

R

In this course we will be using the open-source statistical programming language R which is increasingly used in the social sciences and beyond. In addition to the usual regression modeling R allows for carrying out state-of-the-art computer-based simulations, it has libraries for working with "big data" and you can use it to generate publication-quality data visualizations. R under a wide array of operating systems and can be downloaded for free at http://www.r-project.org/. Learning R might be a bit challenging at first, but you will realize that it is incredibly powerful. The lab sessions will be devoted to learning data analysis techniques in R.

Readings

The following books should be available at the McGill bookstore. The Fox book is somewhat technical but a useful reference and the one I will refer to throughout the course. The book by Achen is superb and I strongly encourage you to buy and closely read it.

> John Fox. 2016. *Applied Regression and Generalized Linear Models.* Sage.
>
> Achen, Christopher. 1982. *Interpreting and using regression.* Sage

Though not required I can strongly recommend:

> Fox, John. 2002. *And R and S-Plus Companion to Applied Regression.* Sage.

Other options for textbooks are:

> Andrew Gelman and Jennifer Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models.* Cambridge. [The first half covers linear regression in a less technical fashion than Fox and is quite useful. Includes examples of R code.]
>
> Kosuke Imai. *Quantitative Social Science.* Princeton University Press. [More targeted at undergraduates but a very well thought out book.]

There will also be some emphasis on data visualization and graphical methods for data analysis. Two classic works in that field that I cannot recommend highly enough are:

> Cleveland, William S. 1993. *Visualizing Data.* Summit, NJ: Hobart

Tufte, Edward. *The Visual Display of Quantitative Information*

Two great online resources for visualizing data using R:

Healy, Kieran. *Data Visualization for Social Science: A practical introduction with R and ggplot2.* forthcoming with Princeton University press but available for free online at `http://socviz.co/`

Grolemund, Garrett and Hadley Wickham (2016). R for Data Science `http://r4ds.had.co.nz/data-visualisation.html`

In addition I will assign articles and chapters from books.

2 Components of the course and evaluation

This course is scheduled once a week for three hours. We will likely not use the whole time every week but we'll see. For about 1 to 1.5 hours every I will give a lecture presenting material. Some weeks we will also discuss and article that was assigned to read. Then we will have a lab session for 1 to 1.5 hours where you will work through some problem which will likely be very similar to the problem-set you have as homework. In addition we will schedule an open lab one evening during the week where the TA for the course will provide additional instruction and assistance. Although not mandatory, attending that lab session is highly recommended.

2.1 Problem Sets (40%):

The only way to learn quantitative analysis is by doing. Thus most weeks there will be a problem set / homework assignment. The problem-set is due at the beginning of next class in hardcopy. In addition you will have to upload your R-code to the assignment folder on the course website. We will figure out a way to get them back to you as fast as possible. You can (and should) consult your colleagues if you get stuck on the problem set. But you have to write your own code! Copying somebody else's code is cheating.

Problem-sets will be graded on a very basic scale (TBD). If you are not happy with the grade you received you have the chance to re-do/correct that part of the problem set once. You will only be able to re-do problem sets that you attempted at the first due date.

2.2 Participation (10%):

This means coming prepared to class and labs and being a good colleague. The learning in this class should be a collaborative endeavour. We will be available for consultation during labs and office hours and also over e-mail. To facilitate the helping each other out you should post any questions you have on the discussion board on the my-courses website.

2.3 Replication paper (50%):

The main requirement for this course is a research paper. This process will work in two steps. First you will find a recently published (approx. last 5 years but certainly no more than 10 years) research paper in your field of interest. You will then replicate the analysis in that paper. For the final paper your task is to extend upon or improve the paper you replicated.

It is imperative that you start working on this NOW.

2.4 Presenting statistical results:

One learning objective of this course is how to professionally present statistical analysis. Again the only way of learning this is practice and developing good habits. As thus I will be **very** strict in enforcing this throughout the course. This includes all homework assignments and the replication paper. Sloppy presentations will not be accepted (see re-do policy above). Generally this will mean journal quality tables, graphs and writeup of your result. Take a look at tables in the American Sociological Review or the American Journal of Sociology. This will be the minimum standard! We will establish early in the course what constitutes acceptable level of presentation and we will practice this throughout the class. Suffice it to say for now that copy-pasting raw regression output will not pass.

It is not required but I **strongly** recommend that you use this opportunity to learn LaTeX- or R-markdown free typesetting software platforms that will create professionally formatted papers and allow you to focus on what is important - the content. Also LaTeX and R-markdown allow you to typeset mathematical formulae much more straightforwardly than other word processors. As with $\mathbb{R}$ there is a learning curve but the payoff is well worth it. Once you got the hang of it you will realize how inferior and clumsy MS Word is for writing academic papers. LaTeX nicely integrates with reference management software (e.g. BibDesk) that will keep your journal articles

filed for you. R-markdown has some (not all) the functionality of LaTeXbut allows you integrate your R-code and writeup into one document.

2.5 Key Dates

- January 8: First course meeting
- February 12: Settle on paper to replicate.
- March 12: Submit replication of descriptive statistics
- April 16: Last class - presentation of replication and ideas for extension
- April 30: Final papers due at 4pm. Mode of delivery TBD

3 Topics

We will go through the material as fast as possible provided that everyone in the class can follow. As such the outline below should be understood as a rough marching plan not something that we will follow adhere to at any cost. Some topics will require more than one week to cover. I will most likely update and add to the list of readings for some weeks as we move along. We will, at minimum, cover regression models for continuous data (OLS) and some models for discrete data (logistic regression). If we have time remaining at the end of the course we can cover additional topics based on student interest (and instructor expertise).

4 Policies

Academic Integrity: McGill University values academic integrity. Therefore, all students must understand the meaning and consequences of cheating, plagiarism and other academic offences under the Code of Student Conduct and Disciplinary Procedures.

Submitting Written Work in French: In accord with McGill University's Charter of Students' Rights, students in this course have the right to submit in English or in French any written work that is to be graded.

Preliminary Schedule

1 Introduction

Outline of the course, what is quantitative analysis about? introduction to R and LaTeX

- King, G. (2006). Publication, publication. *PS: Political Science & Politics*, *39*(01), 119–125
- Young, C. (2009). Model uncertainty in sociological research: An application to religion and economic growth. *American Sociological Review*, *74*(3), 380–397
- Broockman, D., Kalla, J., & Aronow, P. (2015). Irregularities in lacour (2014). Tech. rep., Stanford University (Not required but certainly an interesting and disturbing story.)
- Fox 1

2 What is regression analysis?

Conditional expectations, local averaging, functional forms, interpretations

- Tatem, A. J., Guerra, C. A., Atkinson, P. M., & Hay, S. I. (2004). Athletics: momentous sprint at the 2156 olympics? *Nature*, *431*(7008), 525–525
- Fox Chapter 2

3 Bivariate regressions

Properties of OLS, estimation, interpretation...

- Fox 5.1 and 6.1

4 Visualizing and presenting data

Examining data, introduction to lattice package in R, aesthetics and best practices

- Healy, K., & Moody, J. (2014). Data visualization in sociology. *Annual review of sociology*, *40*, 105–128

- Tufte, E. R. (2001). *The visual display of quantitative information*. Cheshire, Conn.: Graphics Press, 2nd ed ed
- Cleveland, W. S. (1993). *Visualizing data*. Murray Hill, N.J.: AT&T Bell Laboratories
- Fox Chapters 3 and 4

5 Multivariate Regression ($\sim$ 2 weeks)

Estimation and interpretation, dummy variables and interactions

- Braumoeller, B. F. (2004). Hypothesis testing and multiplicative interaction terms. *International organization*, *58*(04), 807–820
- Fox Chapters 5.2, 6.2., 7
- Aachen Chapters 5 and 6
- Achen, C. H. (2005). Let's put garbage-can regressions and garbage-can probits where they belong. *Conflict Management and Peace Science*, *22*(4), 327–339

6 Discrete regression models: Logistic and ordered logistic regression ($\sim$ 2 weeks)

Motivation, estimation, quantities of interest

- Mood, C. (2010). Logistic regression:why we cannot do what we think we can do and what we can do about it. *European Sociological Review*, *26*(1), 67–82
- Fox 14

7 The Maximum Likelihood approach to statistical inference

Stochastic and systematic components, likelihood functions, optimization

- King, G. (1998). *Unifying political methodology : the likelihood theory of statistical inference*. Ann Arbor: University of Michigan Press
- Fox 15

8 Discrete regression models continued

Nominal variables, count variables, censored variables and more, presenting and interpreting results

- Fox 14, 15
- King, G., Tomz, M., & Wittenberg, J. (2000). Making the most of statistical analyses: Improving interpretation and presentation. *American journal of political science*, (pp. 347–361)

9 Assessing Model Adequacy

Statistical assumptions, substantive knowledge and model building

- King, G., & Roberts, M. E. (2014). How robust standard errors expose methodological problems they do not fix, and what to do about it. *Political Analysis*, (p. mpu015)
- Achen, C. H. (2005). Let's put garbage-can regressions and garbage-can probits where they belong. *Conflict Management and Peace Science*, *22*(4), 327–339
- Fox 11,12
- Western, B. (1991). A comparative study of corporatist development. *American Sociological Review*, *56*(3), 283–294

OPTIONAL TOPICS - TIME PERMITTING

10 Matching

Model dependency, exact matching, propensity scores...

11 Missing data

Multiple imputation