POL 604: APPLIED DATA ANALYSIS III

Fall 2019 **Instructor:** Vittorio Mérola
Wednesdays 10.00 am – 1.00 pm **Office:** SBS N721
Room: SBS S740 **Office Hours:** Wednesday 3-4 & 7-8 pm, and by appointment
Stony Brook University **Email:** vittorio.merola@stonybrook.edu

Course Description

This course is an advanced required course for PhD students in the Department of Political Science. It moves beyond the standard linear regression model, generally estimated using ordinary least-squares, and introduces the theory and practice of likelihood inference for statistical models. Several different maximum likelihood estimators will be discussed, covering a variety of the different types of data that are often encountered in the social sciences. In presenting these models, we will discuss theory, estimation, interpretation of findings, as well as emphasize software techniques using R.

As such, the course builds upon previous methodology courses. Students are expected to have a background in basic linear regression, elementary probability theory and a working knowledge of multivariate calculus and basic linear algebra. While some experience with statistical computing in R is ideal, it is not required.

By the end of the course, students should be able to:

- Understand and estimate the various parametric models discussed during the semester.
- Interpret the various models, and present results using predictions, substantive interpretations, simulations, etc.
- Learn and employ new statistical models not covered in this course.
- Gain an appareciation for the challenges of statistical methodology.
- Incorporate these insights and concepts into their original research.

Requirements and Grading

The class meets once a week for almost three hours, which will provide us with sufficient time to think about and discus the issues at hand. Active classroom participation will be essential to making the course a success. Completing the recommended readings before class is recommended. Statistics is learned through repitition and seeing things with new eyes, so the more time spent thinking about the material, the more likely it is to really sink in.

The student's final grade in the course will be based on the following requirements:

- **Class Participation (10%)**. Students are expected to attend class. While these are lectures, students are expected to ask questions and interact with the material as appropriate. Students are allowed to miss one class during the semester, without a valid reason. Each additional unexcused absense will count against their participation grade.
- **Homeworks (30%)**. There will be 6 homework assignments during the semester. These assignments will cover techniques discussed in lecture, and build on the codes and datasets presented during class. You are free to work with your classmates on these assignments, although you are required to write up your own answers and run your own computer code. Each homework is worth 5% of the final grade.
- **Replication Project (30%)**. Students are required to complete a replication project for this course. They must find a published paper or a working paper that uses some of the techniques covered in class, obtain the data, and then replicate the (relevant) results in that article. In the write-up, students should briefly describe the theory being tested, describe the data set, and measures used. Next, students should describe what steps were required to replicate the data, and then compare the results obtained with those in the article. Most importantly, students should provide an interpretation of the results that goes beyond that in the article, along with additional analyses probing the results. Before starting, see the following website as it offers good advice on replication projects: http://gking.harvard.edu/papers. Some journals maintain replication archives that have the data for that article. There is also a replication archive at the ICPSR website (https://www.icpsr.umich.edu/icpsrweb/deposit/pra/) and Harvard's dataverse cite (https://dataverse.harvard.edu/). If there is a particular data set that is of interest, students can also check the author's website before contacting the author to request the data. Replication papers should be approximately 15-20 (double-spaced) pages, and are due by midnight (end of the day) **December 10**.
- **Final Exam (30%)**. During the last class of the semester (12/4), students will complete a final exam. At around 10 am, students will be emailed an exam, along with the relevant datasets. Students will have until 5 pm that same day to complete the exam and return it by email. They can take the exam from anywhere, but they are required to work on it alone. All answers should be their own, and any evidence of cooperation will be penalized. That said, the exam is completely open notes. The purpose of the exam is to test students' knowledge of the material presented, as well as their ability to use the techniques and software learned in class. The exam will be designed to give students a rough sense of the type of questions that might appear on the qualifying exam, from this particular class.

Texts and Material

Recommended Texts

The following textbooks are recommended. I will post the necessary chapters from these books (and others) for the various weeks of this course, although this is to serve as background

material, since all the required concepts and applications will be covered in lecture.

- King, Gary. 1997. *Unifying Political Methodology: The Likelihood Theory of Statistical Inference*. Ann Arbor, Michigan: University of Michigan Press.
- Long, J. Scott. 1997. *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, CA: Sage Publications.
- Smithson, Michael, and Edgar C. Merkle. 2014. *Generalized Linear Models for Categorical and Limited Dependent Variables*. Boca Raton, FL: Chapman and Hall/CRC.

Additional Resources

- Cameron, A. Colin, and Pravin K. Trivedi. 2005. *Microeconometrics*. New York: Cambridge University Press.
- Faraway, Julian J. 2016. *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models, second edition*. Boca Raton: Chapman Hall/CRC.
- Fox, Jonathan. 2015. *Applied Regression Analysis and Generalized Linear Models, 3rd edition*. Thousand Oaks, CA: Sage Publications.
- Gaubatz, Kurt Taylor. 2015. *A Survivor's Guide to R*. Los Angeles, CA: Sage Publications.
- Gelman, Andrew, and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. New York: Cambridge University Press.
- Greene, William. 2012. *Econometric Analysis. 7th edition*. New York: Prentice Hall.
- Maindonald, John and John Braun. 2007. *Data Analysis and Graphics Using R: An Example Based Approach*. New York, NY: Cambridge University Press.
- McCullagh, Peter, and John A. Nelder. 1989. *Generalized Linear Models, 2nd edition*. London: Chapman and Hall.
- Snijders, A. B. Tom, Roel J. Bosker. 2011. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling, 2nd edition*. Sage Publications.
- Wooldrige, Jeffry M. 2010. *Econometric Analysis of Cross Section and Panel Data, 2nd edition*. Cambridge: MIT Press
- Wooldridge, Jeffrey M. 2013. *Introductory Econometrics: A Modern Approach, 5th edition*. Mason, OH: South-Western Cengage Learning.

Computer Software

Quantitative social science research requires the use of computers. Throughout the course, I will go over R code to perform the main techniques discussed in class. You are free to use another software for the class assignments, but I recommend to try to use R. It is the only software I will go over in class.

R Help

It takes a while to properly learn how to use R. Thankfully, there are an almost infinite amount of resources online to help you learn R. In this class, we will spend some time going through code and learning how to use R during class time However, if you still need additional help understanding R, or if you would like to deepen your knowledge of R beyond what we cover in class, the following free resources should be helpful:

- `http://www.statmethods.net/index.html` – Quick-R is a help website for problems you may encounter when using R.
- `https://cran.r-project.org/doc/contrib/Owen-TheRGuide.pdf` – A good introduction to R programming.
- `https://cran.r-project.org/doc/manuals/R-intro.pdf` – A slightly more in-depth introduction to R programming.
- `https://cran.r-project.org/doc/contrib/Verzani-SimpleR.pdf` – A free book teaching you statistics through R
- `https://cran.r-project.org/web/packages/IPSUR/vignettes/IPSUR.pdf` – A very comprehensive, but accessible, overview of R.
- `https://leanpub.com/rprogramming` – A fairly comprehensive book on the nuts and bolts of R.
- `http://www.burns-stat.com/pages/Tutor/R_inferno.pdf` – A more advanced R programming book.
- `https://r4ds.had.co.nz/index.html/` – Another more advanced R programming book.
- `https://resources.rstudio.com/` – Useful videos and tutorials, with an RStudio focus.
- `http://zevross.com/blog/2014/08/04/beautiful-plotting-in-r-a-ggplot2-cheatsheet-3/` – Great introduction to plotting in R using ggplot2.
- `https://cran.r-project.org/doc/contrib/Short-refcard.pdf` – A quick and useful reference card for commonly used commands.

- `http://scs.math.yorku.ca/index.php/R:_Getting_started_with_R` – A useful site for learning R and statistics more generally.
- `http://scs.math.yorku.ca/index.php/R:_Getting_started_with_R` – A bunch of additional resources.
- `https://stackoverflow.com/questions/tagged/r` – A good place to ask questions and find answers to questions.

Readings

Week 1 (8/28): Introduction

Week 2 (9/4): Introduction to Maximum Likelihood

Recommended readings:

- Faraway, Chapter 1.
- King, Chapter 4 (skim Chapters 2-3).
- Long, Chapter 2.

Week 3 (9/11): Generalized Linear Models

Recommended readings:

- Faraway, Chapter 8.
- McCullagh and Nelder, Chapter 2 (skim Chapter 3).
- Smithson and Merkle, Chapter 1.

Week 4 (9/18): Binary Outcomes

- Homework 1 is due

Recommended readings:

- Faraway, 4.1-4.2.
- King, 5.1-5.3.
- Long, Chapters 3-4.
- McCullagh and Nelder, Chapter 4.
- Smithson and Merkle, Chapter 2.

Week 5 (9/25): Ordered Outcomes

Recommended readings:

- King, 5.4-5.6.
- Long, Chapter 5.
- Smithson and Merkle, Chapter 4.

Week 6 (10/2): Nominal Outcomes

- Homework 2 is due

Recommended readings:

- Long, Chapter 6.
- McCullagh and Nelder, 5.3-5.5.
- Smithson and Merkle, Chapter 3.1.

Week 7 (10/16): Conditional/Nested Models

Recommended readings:

- Faraway, 4.3-4.5, 7.3-7.4.
- McCullagh and Nelder, 7.1-7.3.
- Smithson and Merkle, Chapter 3.2-3.5.

Week 8 (10/23): Censoring, Truncation, Selection Models

- Homework 3 is due

Recommended readings:

- King, Chapter 9.
- Long, Chapter 7.
- Smithson and Merkle, Chapter 7.
- Gelman and Hill, 6.7-6.8.

Week 9 (10/30): Missing Data

Recommended readings:

- Gelman and Hill, Chapter 25.
- King, Gary, James, Honaker, Anne Joseph, and Kenneth Scheve. 2001. "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation." *American Political Science Review* 95 (March): 49-69.

Week 10 (11/6): Count Models

- Homework 4 is due

Recommended readings:

- Faraway, Chapter 5.
- Gelman and Hill, 6.2.
- King, 5.7-5.10.
- Lond, Chapter 8.
- Smithson and Merkle, Chapter 5.

Week 11 (11/13): Panel Data Introduction

Recommended readings:

- Faraway, Chapters 10-11.

Week 12 (11/20): Mixed Effect Models I

- Homework 5 is due

Recommended readings:

- Faraway Chapter 13.
- Gelman and Hill, Chapters 11-12.
- Wooldridge (2010), Chapter 10.

Week 13 (11/22): Mixed Effect Models II

Recommended readings:

- Gelman and Hill, Chapters 13-15.
- Smithson and Merkle, 8.1-8.2.

Week 14 (11/27): Thanksgiving Break – No Class

- Homework 6 is due by **12/9**

Week 15 (12/4): Take Home Final Exam – No Class

- Replication Project due by **12/10**