

# POLS 508: Data Analysis

Emory University

Fall 2011

Meeting room: Tarbuton 120A  
Meeting time: Monday, 3:00–6:00pm

Instructor: Drew Linzer  
Email: [dlinzer@emory.edu](mailto:dlinzer@emory.edu)  
Phone: 404-727-0697  
Office: Tarbuton 102  
Office hours: Thursday, 10:00am–12:00pm

## Overview

The field of political science is concerned with understanding and explaining systems of government, power, authority, political behavior, public policy, and international affairs. While much political science research is motivated by *normative* concerns—that is, questions of right and wrong, and how the political world *should* operate—the actual study of political phenomena is dominated by *empirical* questions—asking what *is* the nature of the political world? The research methods available to political scientists to answer these questions have traditionally been divided into two camps: qualitative methods of descriptive and case study research; and quantitative methods, which take a more intensively mathematical approach.

This course is a graduate-level introduction to the philosophy and techniques of quantitative empirical political science. Students will gain basic skills in conceiving and conducting statistical analyses for their own research, in addition to becoming more informed readers of the quantitative research published in the major social science journals. Although some statistical theory will, by necessity, be covered, the emphasis of this course will remain primarily applied: how to prepare a dataset, choose an appropriate statistical procedure, estimate a model, and interpret and present your findings. Topics will span both descriptive and inferential statistics, including measures of central tendency and dispersion, probability, tabular analysis, and bivariate and multiple linear regression models. Students will also receive training in both the Stata and R statistical software packages.

This course, POLS 508, is mandatory for political science PhD and BA/MA candidates at Emory University. It is also a prerequisite for the more advanced graduate statistics offerings, The Linear Model (POLS 509), Limited Dependent Variable Models (POLS 570), Longitudinal Data Analysis (POLS 571), and Bayesian Statistical Modeling (POLS 585). It is taught concurrently with POLS 507, Research Design and Data Collection, and complements POLS 506, Qualitative Methods.

## Grading and Evaluation

Everyone will receive a grade at the end of the semester, but more important than the grade is that you make a commitment to do the reading, complete the assignments in full and on time, participate in class, perform well on the exams, think critically, write well, and generally demonstrate that you are serious about learning how to do high-level academic research.

Expect to devote a significant amount of time each week to problem sets and class preparation. I will trust you to complete all of the readings listed for each week prior to coming to class. Problem sets will be handed out most weeks, and will be due at the start of class the following week. In all, these will comprise 25% of your grade. Collaboration between students is allowed, but please hand in your own work.

There will be two exams: a midterm on October 24 covering everything taught in the first six weeks, and a final exam on December 12, which will be cumulative with an emphasis on the second half of the course. The midterm will contribute 25% and the final 25% to the overall grade.

The remaining 25% of the grade will come from a medium-length, quantitatively-oriented research paper which will allow you to apply statistical methods to a problem of substantive political interest. This paper will be due at the end of the semester. Further guidelines will be handed out later.

## Computer Software

Computing is an essential component of modern applied statistical analysis. In this course, we will learn to use two different software packages, each with its own advantages and disadvantages. The first is Stata, which is perhaps the most widely used statistical software in the field of political science today. Stata combines a simple point-and-click menu and button user interface with a vast array of built-in statistical models, from the most basic to the highly advanced. With Stata, the user can “jump right in” and run various models without needing any additional computer programming. For more complex analyses, Stata also has a command-line interface and a limited programming language.

R, in contrast, is not a statistical program *per se*, but rather a powerful statistical programming language and “environment” that has been optimized for modeling, simulation, and data visualization. Using R requires learning R’s command syntax, which, while more intuitive than Stata’s, still takes a significant initial investment of time and effort. However, once you understand the principles of translating your research questions into commands R can understand, R gives you much greater flexibility and control over your data analysis. It is also completely free to download at <http://www.r-project.org>. For these and other reasons, R is becoming increasingly popular across the social sciences, and many quantitative researchers (myself included) use R almost exclusively in their research.

The bottom line is that it is worth knowing how to use both, and you will use both as you progress through the department’s quantitative methods courses. Using either software package well—meaning, knowing how to get the computer to perform the analyses that you want to perform, and documenting and archiving your analyses for future replication—will require practice, but part of this class is to guide you through that process.

Stata and R are both available in the computer lab. You may also wish to obtain your own copies for personal use. R is free, but Stata is not. Student “GradPlan” copies of Stata can be purchased online at <http://www.stata.com/order/new/edu/gradplans/gp-campus.html>. For programming in R, a decent text editor is a must. For Windows users, I recommend WinEdt version 5.5, which you can download at <http://www.winedt.com>. A student license is \$30. There are also a variety of options for Mac users, including jEdit and Emacs, which are free, and TextMate, which costs about \$50 depending upon the current value of the euro.

## Required Texts

Please purchase each of the following texts. We will read Agresti and Finlay almost cover-to-cover. The Hamilton and Verzani books provide specific guidance on implementing the techniques we will be learning in Stata and R, respectively. There is a fair bit of redundancy between the three books, but this is done by design to aid in your understanding of the course material.

Agresti, Alan and Barbara Finlay. 2009. *Statistical Methods for the Social Sciences*, Fourth Edition. Prentice Hall.

Hamilton, Lawrence C. 2009. *Statistics with Stata: Updated for Version 10*. Brooks/Cole.

Verzani, John. 2004. *Using R for Introductory Statistics*. Chapman & Hall/CRC.

It is an unfortunate fact of life that statistics books are expensive. Think of these books as an investment in yourself and your graduate training. You will find that you use and refer back to them for many, many years. For additional guidance on the R programming language, there is also a free online manual at <http://cran.r-project.org/doc/manuals/R-intro.pdf>.

## Class schedule

**August 29:** Introduction. Why and how do we quantify political phenomena?

Gould, S. J. and R. C. Lewontin. 1979. [The Spandrels of San Marco and the Panglossian Paradigm: A Critique of the Adaptationist Programme](#). *Proceedings of the Royal Society of London. Series B, Biological Sciences*. 205(1161): 581-598.

Gould, Stephen Jay. 1978. [Morton's Ranking of Races by Cranial Capacity](#). *Science*. 200(4341): 503-509.

Moe, Terry M. 1979. [On the Scientific Status of Rational Models](#). *American Journal of Political Science*. 23(1): 215-243.

**September 5:** No class; Labor Day holiday.

**September 12:** Data acquisition and management;

Presentation by Rob O'Reilly, Emory Electronic Data Center.

Agresti and Finlay, Chapter 1.

Verzani, Chapter 1 and Appendices A and B.

Hamilton, Chapters 2 and 16.

King, Gary. 1995. [Replication, Replication](#). *PS: Political Science and Politics*. 28(3): 444-452.

Herrnson, Paul S. 1995. [Replication, Verification, Secondary Analysis, and Data Collection in Political Science](#). *PS: Political Science and Politics*. 28(3): 452-455.

**September 19:** What is a variable? Description and visualization.

Agresti and Finlay, Chapters 2-3.

Verzani, Chapter 2.

Hamilton, Chapters 3-4.

Mayhew, David R. 1974. [Congressional Elections: The Case of the Vanishing Marginals](#). *Polity*. 6(3): 295-317.

Hofstadter, Douglas R. 1985. [On Number Numbness](#). In *Metamagical Themas*. Basic Books, 115-135. Originally in *Scientific American* 246(5): 20-34, as "Number Numbness, or Why Innumeracy May Be Just as Dangerous as Illiteracy."

**September 26:** Randomness: probability distributions.

Agresti and Finlay, Chapter 4.

Verzani, Chapters 5-6.

Fienberg, Stephen E. 1971. [Randomization and Social Affairs: The 1970 Draft Lottery](#). *Science*. 171(3968): 255-261.

**October 3:** Sampling distributions and confidence intervals.

Agresti and Finlay, Chapter 5.

Verzani, Chapter 7.

Siegfried, Tom. 2010. [Odds Are, It's Wrong: Science fails to face the shortcomings of statistics](#). *Science News*, March 27. 177(7): 26.

**October 10:** No class; Fall break.

**October 17:** Statistical significance and hypothesis testing.

Agresti and Finlay, Chapters 6-7.

Verzani, Chapter 8.

Gill, Jeff. 1999. [The Insignificance of Null Hypothesis Significance Testing](#). *Political Research Quarterly*. 52(3): 647-674.

Gigerenzer, Gerd. [Mindless statistics](#). 2004. *The Journal of Socio-Economics*. 33(5): 587-606.

**October 24:** Midterm Exam.

**October 31:** Analyzing categorical variables.

Agresti and Finlay, Chapter 8.

Verzani, Sections 3.1, 9.1, and 9.2.

Licklider, Roy. 1995. [The Consequences of Negotiated Settlements in Civil Wars, 1945-1993](#). *American Political Science Review*. 89(3): 681-690.

**November 7:** Bivariate correlation and regression.

Agresti and Finlay, Chapter 9.

Verzani, Chapter 3 and Appendix D.

Hamilton, Chapter 6.

Cleveland, William S. and Robert McGill. 1984. [The Many Faces of a Scatterplot](#). *Journal of the American Statistical Association*. 79(388): 807-822.

Tufte, Edward R. 1969. [Improving Data Analysis in Political Science](#). *World Politics*. 21(4): 641-654.

**November 14:** Linear regression: fit and diagnostics.

Agresti and Finlay, Chapter 9. (Please re-read)

Verzani, Sections 10.1-10.2.

Hamilton, Chapters 7-8.

Choi, Seung-Whan. 2009. [The Effect of Outliers on Regression Analysis: Regime Type and Foreign Direct Investment](#). *Quarterly Journal of Political Science*. 4(2): 153-165.

**November 21:** Multiple regression and confounding.

Agresti and Finlay, Chapters 10-11 and 13.

Verzani, Chapter 4 and Section 10.3.

Leamer, Edward E. 1983. [Let's Take the Con Out of Econometrics](#). *The American Economic Review*. 73(1): 31-43.

Braumoeller, Bear F. 2004. [Hypothesis Testing and Multiplicative Interaction Terms](#). *International Organization*. (58)4: 807-820.

Friedrich, Robert J. 1982. [In Defense of Multiplicative Terms in Multiple Regression Equations](#). *American Journal of Political Science*. 26(4): 797-833.

**November 28:** Interpreting and communicating model results.

King, Gary. 1986. [How Not to Lie with Statistics: Avoiding Common Mistakes in Quantitative Political Science](#). *American Journal of Political Science*. 30(3): 666-687.

Ansolabehere, Stephen, James M. Snyder, Jr. and Charles Stewart, III. 2001. [Candidate Positioning in U.S. House Elections](#). *American Journal of Political Science*. 45(1): 136-159.

Gelman, Andrew. 2011. [Why Tables Are Really Much Better Than Graphs](#). *Journal of Computational and Graphical Statistics*. 20(1): 3-7.

**December 5:** Model selection, forecasting, and cross-validation.

Agresti and Finlay, Chapters 14-15.

Verzani, Chapter 12.

Hamilton, Chapter 10.

Luskin, Robert C. 1991. [Abusus Non Tollit Usum: Standardized Coefficients, Correlations, and R<sup>2</sup>s](#). *American Journal of Political Science*. 35(4): 1032-1046.

King, Gary. 1991. ["Truth" Is Stranger than Prediction, More Questionable than Causal Inference](#). *American Journal of Political Science*. 35(4): 1047-1053.

**December 12:** Final exam.