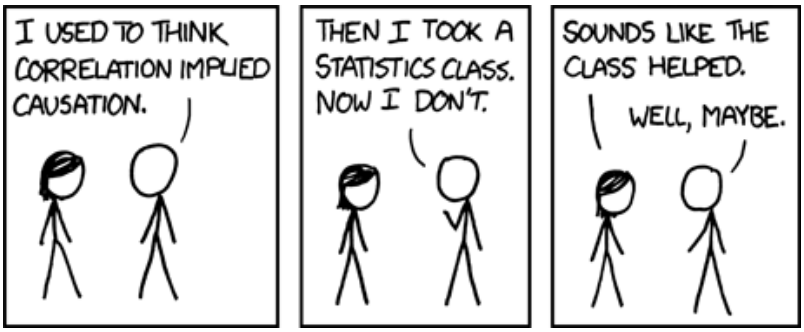
PSC 782: Advanced Research Methods in Political Science

Class time: Tuesday, 5:30 - 8:15 PM
Location: Mack Social Science 236

Instructor: Kevin Banda
Office: Mack Social Science 237
Office hours: Tuesdays and Thursdays, 12:30 - 1:30 PM and by appointment
Email: kbanda@unr.edu

Figure 1: XKCD comics #552, "Correlation". http://xkcd.com/552/



A series of medical studies in the 1990s used observational data to conclude that women who take supplemental estrogen had far lower rates of heart disease. On this recommendation, lots of women began taking supplemental estrogen. Then in 2004, a randomized experimental trial found that this sort of hormone replacement therapy actually slightly *increases* the risk of heart disease.[1] What gives? It turns out that the women who decided to take supplemental estrogen tended to be wealthier, and also tended to belong to gyms, eat healthier foods, and so on. The observational studies failed to include socio-economic status as a control variable, and as a result, they drew a false conclusion that was directly harmful to the health of patients.

Does this mean that observational studies are inherently flawed? No, it simply means that the researchers relied on correlations to draw conclusions. Linear regression is a tool that moves beyond the myopic inferences provided by correlations by allowing us to control for confounding variables in observational data.

This course is an overview of the techniques that make serious statistical modeling in the social sciences possible. Statistical mediation – the inclusion of controls – is an essential practice for deriving accurate results. In addition, regression can be used to make predictions outside a sample with measures of certainty regarding these predictions. Regression results have direct substantive meaning beyond the "rejected, not rejected" dichotomy of simple hypothesis tests: not only can we describe whether or not we expect a political outcome to change in a particular direction, we can say how much we expect the outcome to change. We can condition these statements on other factors. And we can create a model that allows the outcome to change in a myriad of nonlinear ways. By the halfway point of this course, you will be running and interpreting models that do all these things.

This course will depart from traditional courses in linear regression in a few important ways. Our focus will be very applied, with a larger emphasis on data, intuition, and interpretation than is typical. Data management – the steps one must take before any statistics can be used at all – tends to be the most time consuming part of research, and mistakes at the data management stage can corrupt any statistical model, no matter how carefully constructed. So we will take the entire first month of the course to focus solely on techniques and strategies for data management.

We will strongly emphasize the substantive as well as the technical interpretation of regression results. We will also discuss many of the common bad practices involving linear regression in political science, and develop strategies for avoiding these pitfalls.

[1] Debbie A Lawlor, George Davey Smith, and Shah Ebrahim. 2004. "Commentary: The Hormone Replacement–Coronary Heart Disease Conundrum: Is this the Death of Observational Epidemiology?" *International Journal of Epidemiology.* 33(3): 464-467.

By the end of the course, you will be able to

- acquire, clean and prepare data for analysis quickly, correctly, and in a way that is easy to replicate and document,
- describe the meaning of coefficients in language that speaks to a substantive audience as well as to a technical one,
- specify, graph and interpret any interactive or curvilinear relationship,
- and explain exactly what a p-value means and what the limitations of these inferential statistics are.

Finally, we will use calculus and linear algebra to develop and to understand the intuition of linear regression. Don't be afraid of the math: we will move through that material slowly and with as much clarity as possible. Understanding the math behind linear regression techniques will give you power over these tools, so that you can use them with more confidence and adapt them more thoroughly to your own purposes. We will also use computer simulations to demonstrate the behavior of these tools in a laboratory setting.

Course Website

Problem sets, readings, details on the term paper, and other items will be posted on the course's WebCampus page. If you are registered, you should automatically have access to the page. If you are not registered, please speak to me so I can arrange for you to have access.

Textbooks

There are two required textbooks for the class:

- A. Colin Cameron and Pravin K. Trivedi. 2010. *Microeconometrics Using Stata, Revised Edition.* College Station, TX: Stata Press.
  - Henceforth "CT"
  - http://www.amazon.com/dp/1597180734/
- William D. Berry and Mitchell S. Sanders. 2000. *Understanding Multivariate Research: A Primer for Beginning Social Scientists.* Boulder, CO: Westview Press.
  - Henceforth "BS"
  - http://www.amazon.com/Understanding-Multivariate-Research-Beginning-Scientists/dp/0813399718

CT describes some of the theory of the methods we will use while also providing examples and code for running the methods in Stata, so I believe that this textbook is a valuable guide to keep for future reference.[2] *BS* covers the theory of regression from a slightly different perspective and focuses on interpretation rather than software. Additional readings may be announced throughout the semester, and these readings will be posted on WebCampus. Please have the readings listed for each class completed before class.

Software

We will use Stata in this course. While the code we will discuss is specific to Stata, the logic of data management and analysis is not. Thus the skills you learn in this course will translate to the use of other statistics programs. You should purchase a copy of Stata. The software isn't exactly cheap, but students receive steep discounts. Do not buy Small Stata; it won't be adequate for your needs. Stata IC will likely suffice for most students, though Stata SE can handle larger data sets (I use Stata SE). Perpetual licenses for Stata IC are $198. For Stata SE, the cost is $395. See http://www.stata.com/order/new/edu/gradplans/student-pricing/.

[2] See Fox's "Applied Regression Analysis and Generalized Linear Models", Gelman and Hill's Data Analysis Using Regression and Multilevel/Hierarchical Models, and Gujarati and Porter's Basic Econometrics for much more detailed information.

Assessment

Your grade in this course will be derived from your performance on problem sets, a research note, and a take-home final exam:

- Problem sets, totaling 40% of the final grade: you will be assigned 4-6 problem sets during the semester (depending on our pace through the material). The expectation is that each problem set will take several hours of work to complete. For mathematics problems, you are required to show all of your work in order to receive credit. For coding problems, you are required to submit a copy of your do file. Please see the statements regarding collaboration and cheating below.
- Research note, 25% of the final grade: you will be required to complete an 8 to 12 page research note describing an analysis on a topic of your choice that uses the methods we discuss in this course in an effective and accurate way. You will not be required to present a complete literature review or description of theory, but you will be expected to present complete descriptions of your data, methods, and results, and you will present a discussion that interprets these results correctly. Additional details about the paper will be posted later in the semester, but **you should start thinking about a research topic and collecting data today**.
- Research presentation, 5% of the final grade: you will be required to present your research note to the class. Presentation times will vary based on the size of the class. These presentations should be professional in nature and should largely mirror academic conference presentations or shorter versions of the job talks you *should* be attending.
- Take-home final exam, 30% of the final grade: the exam will cover the entire course and may include mathematics and statistics problems, short essays, and computational work. The exam is a take-home exam because it is far more important for you to use the methods accurately than it is for you to use the methods quickly. The exam is open book, but unlike the problem sets, no collaboration is allowed for the exam.

The table below describes this course's grading scale. Grades will be rounded to the nearest percentage point and will be posted on WebCampus in a timely manner. I will not use a strict and predefined curve when assigning grades.[3] That said, I will curve grades to some extent if necessary to reflect the difficulty of the course assignments.

A	93-100	C	73-76
A-	90-92	C-	70-72
B+	87-89	D+	67-69
B	83-86	D	63-66
B-	80-82	D-	60-62
C+	77-79	F	0-59

Collaboration

Each student is required to turn in her own original homework assignments. That said, you are encouraged to work together with your fellow students to help each other complete the problem sets. You are also encouraged to read and comment on each other's term papers. **You may not collaborate on the final exam.**

Cheating

The following actions are examples of cheating:

- Directly copying answers on problem sets from another student, directly copying computer code from another student, or allowing answers or code to be copied
  - There is a clear difference between collaboration on problem sets and copying. If you are unsure about the difference, please come speak to me *before* there are any potential problems
- Plagiarizing any part of any problem set, any part of the term paper, or any part of any exam question, including prose, graphs, tables, and equations

[3] In other words, a certain percentage of students will not receive a given grade.

- Collaborating on the final exam

In addition, fabricating the results of statistical analyses amounts to academic dishonesty and will be treated as cheating in this course. Any student who is caught cheating on an assignment will automatically receive a 0 on the assignment and will be sanctioned as outlined here.

Incompletes

I do not assign incomplete grades unless there are compelling reasons to do so.

Schedule

Part 1: Data Management

- 1/24:
  - What is data? Stata basics.
  - Reading: CT sections 1.1-1.4, 1.10, 2.4.8
- 1/31:
  - Do file management, logical and arithmetic operators, properties of variables, and summary statistics
  - Reading: CT sections 2.4 and 3.2
- 2/7:
  - More complicated variable and data properties; loops
  - Reading: CT sections 1.8, 2.2, 2.3
- 2/14:
  - Additional data management techniques, simulations, and plots
  - Reading: CT sections 1.5-1.7, 2.5, 2.6.1, 2.6.3, 4.2, 4.3; Carsey and Harden Chapter 1

Part 2: Basics of Linear Regression

- 2/21:
  - More plots and control in observational studies
  - Reading: CT section 2.6.5, 2.6.7; Achen 2005 "Let's Put Garbage-Can Regressions and Garbage-Can Probits Where They Belong," *Conflict Management and Peace Science* vol 22 no. 4, 327-339.
- 2/28:
  - Bivariate linear regression, matrix algebra, and the OLS estimator
  - Reading: CT sections 3.1, 3.3.1-3.3.3, 3.4.1-3.4.2; BS chapters 1 and 2
- 3/7:
  - The regress command and Stata output; standard errors, confidence intervals
  - Reading: CT section 3.4; BS chapters 3-5; Wasserstein and Lazar 2016, "The ASA's Statement on p-Values: Context, Process, and Purpose."
- 3/14:
  - P-values, missing data, and measures of model fit
  - Reading: Section 10.7; http://thelogicofscience.com/2015/12/28/basic-statistics-part-4-understanding-p-values/
- 3/21:
  - Spring break, **NO CLASS**
- 3/28:
  - Interpretation of coefficients on noncontinuous covariates; logged x and y variables
  - Reading: CT sections 3.3.6, 3.6.3; take a look at BS pages 33-37 again
- 4/4:
  - Interaction terms
  - Reading: CT section 10.6; BS pages 63-72

- 4/11:
  - Interaction terms continued
  - Reading: Brambor et al. 2006 "Understanding Interaction Models: Improving Empirical Analyses," *Political Analysis* vol 14 no. 1, 63-82

Part 3: Extensions and Corrections

- 4/18:
  - Gauss-Markov assumptions, things that can go wrong with OLS
  - Reading: CT sections 3.3.1-3.3.3 again
- 4/25:
  - Omitted variable bias, endogeneity, reverse causality, instrumental variables regression
  - Reading: CT sections 3.5, 6.1-6.4; Carsey and Harden chapter 5
- 5/2:
  - Model misspecification, GLMs, heteroskedasticity, autocorrelation, and multicolinearity
  - Reading: CT sections 10.1-10.3, 3.3.4-3.3.5, 3.5, 5.1-5.3; BS pages 72-80; skim Carsey and Harden chapter 5 again
- 5/9:
  - **Research notes due at the beginning of class; research presentations**

Final exams are due via e-mail by 5:30 PM on 5/10

Figure 2: XKCD comics #882, “Significant”. http://xkcd.com/882/