Course Goals

This course is meant to train students in advanced quantitative techniques in the social sciences. We will look at four main areas of interest. One -- modeling of limited dependent variables, like Poisson, tobit and gamma-distributed will be discussed. Two -- modeling of multilevel data, like panel data and geographic data, will also be practiced. Three -- ways of better approximating experimental designs with observation data will be highlighted, like instrumental variables, propensity score matching and regression continuity. Four -- creating and analyzing text as data, including “bag of words” analysis, contextual analysis and topic modeling.

Another important goal of the course is to teach students how to manipulate and analyze data themselves using statistical software. We will focus mainly on the program R. The last hour of each class will be devoted to using this software program to practice commands and to complete lab assignments.

Students ought to be familiar with regression models from other courses, but only basic math will be presumed.

Course Expectations

Attendance and Class Participation. Your attendance and participation are necessary at every meeting.

Exams. We will have two take-home exams. They will include short answer, longer answer and multiple choice questions. They make up the bare majority of your total grade.

Homework. Homework problems will be assigned as the semester progresses. It is expected that you will do your homework. It will be graded.

Data Analysis Portion of the Class. There will be a separate lab portion of the course, where we will focus on learning R.

Lab Assignments. Students will have 3 large lab reports due throughout the semester. They will be based on writing up the results of performing the commands learned from the lectures. Specific instructions,
format and deadlines will be given as the semester progresses.

Plagiarism and Academic Dishonesty: Students must do all their work within the boundaries of acceptable academic norms. See the Academic Honesty page of the CU website regarding college policy on plagiarism and other forms of academic dishonesty - http://www.columbia.edu/cu/history/ugrad/main/handbook/academic_honesty.html. Students found guilty of plagiarism or academic dishonesty will be subject to appropriate disciplinary action, which may include reduction of grade, a failure in the course, suspension or expulsion. This includes lab reports – if they are copied from another student, severe penalties may be applied.

Late Assignments. Students will lose points for handing in late assignments, at the discretion of the instructor and teaching assistant.

Textbooks. We will be using one textbook:


For individual weeks, other resources will be given throughout the semester.

Suggested Additional Readings. For more advanced students, additional possible readings can also be suggested, to see the concepts and methods in action in actual research articles and books – those references will be given out separately in a few weeks.

Grade Distribution. The distribution of the parts for your grade is as follows:

- Two Exams = 50%
- R Labs = 30%
- Attendance, Participation, and HW = 20%

Changes: There may be adjustments in the scheduling of assignments, exams, and classrooms. Changes will be posted on Courseworks along with other announcements.

Proposed Schedule for the Course Lectures

Jan 22 – Introduction

Part I: Limited Dependent Variables

Jan 29 - Review of Multiple Regression/Linear Regression (Wooldridge, Chs. 3-5); Review of Logistic Regression: Binary (Ch. 17.1; Park 2013; Appendix C.4 (only “Maximum Likelihood”): Ordinal (Bender & Grouven 1997; Norusis v.13; Greene, Ch. 18); Multinomial (Moutinho and Hutcheson forthcoming; Greene, Ch. 18); Interactions and Predicted Probabilities (Carina Mood. "Logistic
regression: Why we cannot do what we think we can do and what we can do about it.” European Sociological Review 2010 26(1): 67-82 [not on Courseworks])

Feb 5 – **Generalized Linear Models, including Poisson and Gamma** (Fox, Ch. 15, Wooldridge p. 587-594); **Tobit Regression** (Wooldridge p. 595-600); & **Censoring and Truncation** (Wooldridge p. 600-608)

**Part II: Time-Ordered Data Structures**

Feb 12 - **First Differences Analysis** (Wooldridge p. 455-465); **Fixed Effects** (Wooldridge p. 481-489); & **Random Effects** (Wooldridge p. 489 – 493); & **Lagged Dependent Variable** (Wooldridge p. 310-312)

Feb 19 – **Difference-in-Differences Analysis** (Wooldridge p. 435-445)


**Part III: Quasi-Experimental Techniques**

Mar 4 – **Instrumental Variables and Two Stage Least Squares** (Wooldridge 506-529); & **Natural Experiments** (Wooldridge 506-529); & **Regression Discontinuity** (Lee & Munk. 2008. “Using Regression Discontinuity Design for Program Evaluation”)


Mar 18 - Spring Break!

**Part IV: Multilevel Models**


**Part V: Text as Data**


Apr 8 - **Clustering and Comparison of Texts:** Quantitative methods for comparing texts via concordances, co-occurrences, and keyword ratios; complexity and readability measures; and dissimilarity measures (Jockers 2014, Ch. 11; --- & --- Light, Ryan. "From Words to Networks and Back: Digital Text, Computational Social Science, and the Case of Presidential Inaugural Addresses." *Social Currents* (2014))


Apr 29 - **Last Class:** Miscellaneous, FAQ + Presentations?