Preliminary Syllabus for Spring 2017
ADVANCED ANALYTIC/QUANTITATIVE TECHNIQUES (G4018)

Day/Time: Friday 10:10am-12:00pm
Location: 602 Hamilton Hall

Gregory M. Eirich, QMSS
gme2101@columbia.edu
Office Hours in 807A IAB: TBA

Teaching Assistant: Xiaoyu Zhang, xz2479@columbia.edu

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 -- creating and analyzing text as data, including "bag of words" analysis, contextual analysis and topic modeling. Three -- ways of better approximating experimental designs with observation data will be highlighted, like instrumental variables, propensity score matching and regression continuity. Finally, four -- modeling of multilevel data, like panel data and geographic data, will also be practiced.

Another important goal of the course is to teach students how to manipulate, analyze and visualize network data themselves using statistical software. We will mainly use the program R for most of the software work. Lab assignments will be given out, and we will aim to have weekly lab meetings (**which will be completely optional**) right after class, but only if a space can be found. Regardless, there will be copies of the code used in lab for students to practice at their convenience.

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.

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:

1. *Introductory Econometrics: A Modern Approach,* 4th Edition, by Jeffrey Wooldridge (South-Western College Pub) ISBN-13=9780324581621

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

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

Two Exams = 50%
Computer Labs = 35%
Attendance and Participation = 15%

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 20 – **Introduction**

Part I: Limited Dependent Variables

Jan 27 - **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 3 – **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: Text as Data

Feb 10 – **Getting Started:** Where to get texts; formatting texts; organizing texts; capturing meta-data; units of analysis; arranging text through stemming, stop-words, and other pre-processing. (Jockers 2014, Chs. 1-3; all Jockers can be accessed here via CU Libraries as an e-book). **"Bag of Words"** (Francis, Louise, and Matt Flynn. "Text mining handbook." *Casualty Actuarial Society E-Forum*, Spring 2010. 2010). **Sentiment Analysis** via automatic dictionary-based methods, including LIWC, RID, and the Harvard IV-4. (Ryan C. Black, Sarah A. Treul, Timothy R. Johnson, and Jerry Goldman. "Emotions, oral arguments, and Supreme Court decision making." *The Journal of Politics*, 73(2):572–581, April 2011. **--- & ---** Golder, Scott A., and Michael W. Macy. "Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures." *Science* 333.6051 (2011): 1878-1881.)

Feb 17 - **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))

* [**Lab #1 Due ~ February 17**]

Feb 24 - **Sensitizing Models to Context and Semantics.** Investigating n-grams, tokens, parts of speech, emoticons, and vocabulary richness and diversity. (Davies, Mark. "Making Google Books n-grams useful for a wide range of research on language change." *International Journal of Corpus Linguistics* 19.3 (2014): 401-416. **--- & ---** Soper, Daniel S., and Ofir Turel. "An n-gram analysis of Communications 2000--2010." *Communications of the ACM* 55.5 (2012): 81-87). **Topic Modeling** (Jockers 2014, Ch. 13; **--- & ---** DiMaggio, Paul, Manish Nag, and David Blei. "Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of US government arts funding." *Poetics* 41.6 (2013): 570-606. **--- & ---** Levy, Karen EC,

and Michael Franklin. "Driving Regulation: Using Topic Models to Examine Political Contention in the US Trucking Industry." *Social Science Computer Review* (2013): 0894439313506847.)

Mar 3 - **Machine Learning Algorithms, Classification and Text Analysis:** Methods for assessing classifier performance, feature weighting, and classification accuracy. (Jockers 2014, Ch. 12; **--- & ---** D'Orazio, Vito, et al. "Separating the Wheat from the Chaff: Applications of Automated Document Classification Using Support Vector Machines." *Political Analysis* 22.2 (2014): 224-242. **--- & ---** Monroe, Burt, Michael Colaresi, and Kevin Quinn. 2008. "Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict". *Political Analysis* 16(4))

Part III: Quasi-Experimental Techniques

Mar 10 – **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")

* [**Midterm Due ~ March 10**]

Mar 17 - Spring Break!

Mar 24 – **Propensity Score Matching** (Austin. 2011. "An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies." *Multivariate Behav Res*. 2011 May; 46(3): 399–424.)

* [**Lab #2 Due ~ March 24**]

Part IV: Time-Ordered Data Structures

Mar 31 - **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)

Apr 7 – **Difference-in-Differences Analysis** (Wooldridge p. 435-445)

Apr 14 – **Growth Curve Analysis** (Curran, Patrick J., Khawla Obeidat, and Diane Losardo. "Twelve frequently asked questions about growth curve modeling." *Journal of Cognition and Development* 11.2 (2010): 121-136.)

Part V: Multilevel Models

Apr 21 - **Multilevel Models** or **Hierarchical Linear Models** (Diez-Roux, Ana V. "Multilevel analysis in public health research." *Annual review of public health* 21.1 (2000): 171-192. **--- & ---** Duncan, Craig, Kelvyn Jones, and Graham Moon. "Context, composition and heterogeneity: using multilevel models in health research." *Social science & medicine* 46.1 (1998): 97-117. **-- & --** John Huber 2005. "Religious belief, religious participation, and social policy attitudes across countries." Working paper.)

* [**Lab #3 Due ~ April 21**]

Apr 28 - **Last Class:** Miscellaneous, FAQ + Presentations**?**

* [**Final Due ~ May 8**]