DATA ANALYSIS FOR THE SOCIAL SCIENCES  
(G4015)

Wednesday 11:00am-1:00pm  
Location: 403 International Affairs Building

(Optional Lab:  Wednesday 1:00pm-2:00pm, Location: 270B International Affairs Building)

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

Teaching Assistant: Xiaoyu Zhang  
xz2479@columbia.edu

Teaching Assistant: Lurgeon Vaval  
lv2301@tc.columbia.edu

Course Goals

This course is meant to provide an introduction to probability and social statistics, tailored to the types of analyses and data issues encountered by QMSS students. The chief goal is to help students generate and interpret quantitative data in helpful and provocative ways. The hope is that by trying to measure the social world, students will see their thinking become clearer and their understandings of concepts grow more complex. They will also become competent at reading statistical results in social science publications and in other media.

Only basic mathematics skills are assumed, but it is hoped that students will become more facile with numbers, functions and their relationships.

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. There will be an optional lab section every other week, which will be devoted to using these software programs to practice commands and to develop a paper using the General Social Survey, World Values Survey or another dataset of the student’s choosing.

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 and longer answer 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. We will have optional bi-weekly R data analysis lab assignments to write up and hand in. Students will be graded in terms of your ability to operate the program, select the most appropriate statistics for each type of analysis, interpret the statistics generated, and write brief summaries about what they have learned. In short, you will develop your own “social theory” using the GSS or some other data.

The lab will be focused on using the R software. It is free, on all the computers at Columbia and we will help students download it if need be.

Independent Data Analysis and Interpretation. There will be a semester-long independent project. Unless the student selects another dataset, it will be based on the General Social Survey. This final exercise will require the students to integrate many of the skills and lessons learned throughout the semester into a final research report, but more information will be given about this assignment as the semester progresses. The bi-weekly R data analysis lab assignments are designed to move the project along. Depending on time, students may present their final results to the class.

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.


Other Readings. In some weeks, there will be additional readings from other sources.

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%
Independent Project and 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 of Classes, for Tuesday’s Section
(unless otherwise noted, the references are to Wooldridge, 4e)

Sept 7 – **First Day -- Introduction to the Class**: Ch.19

Sept 14 – **Statistics, Data Structures, and Basic Descriptives**: Ch. 1; Appendices A.1, A.3, B.1, and B.3

Sept 21 – **Regression 1 -- Simple Regression Model**: Ch. 2.1, 2.2, and 2.4 (through “The Effects of Changing Units of Measurement on OLS Statistics”); **Correlation**: Appendices B.2 and B.4; **Goodness-of-Fit and R^2**: Ch. 2.3; and **Hypothesis Testing**: Ch. 4.1, 4.2, and 4.3, Appendices B.5 (through “Additional Properties of the Normal Distribution”), C.5 and C.6

Sept 28 – **Regression 2 -- Multiple Regression Analysis**: Ch. 3.1-3.2, Ch. 3.3 (but only from “Assumption MLR.4” on), Ch. 6.1, and Ch. 6.3; and **Hypothesis Testing in Multiple Regression**: Ch. 4.4-4.6

Oct 5 – **Regression 3 -- Categorical-by-Continuous Interactions**: Ch. 7.1-7.4; **Continuous-by-Continuous Interactions**: Ch. 6.2 (only “Models with Interaction Terms”); and **Quadratics**: Ch. 6.2 (only “Models with Quadratics”); and **Log Transformations**: 2.4 (only “Incorporating Nonlinearities in Simple Regression”), Ch. 6.2 (only “More on Using Logarithmic Functional Forms”)

Oct 12 – **Regression 4 -- The Gauss-Markov Assumptions and Asymptotics**: Ch. 2.5, Ch. 3.3-3.5, and Ch. 5.1-5.3; **More Specification and Data Issues**: Ch. 8, Ch. 9.1, 9.5, and 9.6

Oct 19 – **Regression 5 -- Catching up on (Inevitable) OLS Loose Ends**

Oct 26 – **Models for Binary Outcomes -- Linear Probability Model**: Ch. 7.5; **Binary Logistic Regression**: Ch. 17.1; Agresti & Finlay Ch. 15--*On Courseworks*; Appendix C.4 (only “Maximum Likelihood”); **Logistic Regression vs. a Naive Bayes Classifier**: Material TBA; and **Additional Goodness-of-Fit Measures (esp., BIC and AIC)**: Material TBA

Nov 2 – **Logistic Models for Ordered and Categorical Data**: Agresti & Finlay Ch. 15--*On Courseworks*; and **Generalized Linear Models**: Agresti and Finlay (p. 458-462; 468-469)--*On Courseworks*

Nov 9 – **First Differences Analysis**: Ch. 13.3-13.5

Nov 16 – **Data Reduction Techniques -- Scales, Factor Analysis, Cluster Analysis, and LASSO**: All Material TBA

Nov 23 – No class! (Thanksgiving Holiday)

Nov 30 – **A Chance to Extend Any of the Material Since the Midterm**

Dec 7 – **Last Class -- Miscellaneous, FAQ, Review, (Perhaps) Student Presentations, and My Final Words of Wisdom**