DATA ANALYSIS WITH PYTHON

Summer Session A
Thursdays and Fridays: 10.10am-12pm
Location: 501 Schermerhorn Hall

Gregory M. Eirich
gme2101@columbia.edu
Office Hours via Zoom TBA

Teaching Assistants:

Bengusu Ozcan <bo2297@columbia.edu>
Xintong Maxxie Tang <xt2249@columbia.edu>

Credits: 3 points

Course Goals

This course is meant to provide an introduction to regression and applied statistics for the social sciences, with a strong emphasis on utilizing the Python software language to perform the key tasks in the data analysis workflow. 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 some more advanced math will be introduced as needed.

For this course, a critical goal is to teach students how to manipulate and analyze data themselves using statistical software. We will focus almost exclusively on Python for this class (although, there will be a few cases where we will run R through Python because R can more readily do some things than Python). There will be Python write-up assignments nearly each week, tied to hands-on data analysis lab sessions. These weekly assignments will be devoted to using Python to practice commands and to develop a better intuition about social science research. TAs will hold additional weekly 1-hour lab sessions, where they will lead students through labs and/or homework and/or lecture review (which will be recorded and uploaded to Courseworks for asynchronous learners too).

Course Expectations

Format of the Class: The plan for the class is for it to operate in a “hyflex” classroom. This classroom will be set up with cameras and microphones, along with access to live Zoom. This will allow students to be in the classroom during class, and for remote students to have access to the class live and to be able to participate by asking questions and interacting with their peers. **The exact number of students who will be present during each class will depend upon the actual specific room that we are provided with and the official rules and regulations governing social distancing within that room**. As a first approximation, the hope would be that a live in person option will be available for students to come every week or every other week, depending upon the ultimate size of the class and classroom.

Synchronous Participation vs. Asynchronous Participation (and Attendance): It is my strong hope you will participate in this class synchronously so that you can benefit fully from your peers and the live instruction happening. That said, I completely understand your circumstances may make that very difficult, at least on some occasions. In which case, this course can be done asynchronously as well. Classes and lectures and labs will be recorded and available on Courseworks. Likewise, assignments and some forms of participation can also be done remotely and asynchronously.

Expectation of Regular Participation and Utilization of Courseworks: We will be monitoring student participation and completion of assignments throughout the semester. We want to make sure that students are consistently engaged, and if that becomes difficult, that students alert us to their situations. There will also be opportunities for asynchronous participation too, where students will respond to a prompt/question from me or each other. We will mainly use Campuswire for this (you will be invited to join our class page).

Exams. We will have a final take-home exam. They will include short answer and longer answer questions. This will make up the bare majority of your total grade.

Data Analysis Portion of the Class. We will have Python data analysis 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 some data.

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 assistants.

Textbook. We will be using as our textbook, Wooldridge, Jeffrey. 2008. *Introductory Econometrics: A Modern Approach*. South-Western College Pub; 4th Edition, ISBN=9780324581621. The book is on reserve at numerous libraries around the University. Feel free to use newer versions, if you prefer.

There are three very helpful Python books too:

- Downey, Allen B. *Think stats: exploratory data analysis*. O'Reilly Media, Inc., 2014, with the 2nd edition available for free download here.
- McKinney, Wes. *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython*. O'Reilly Media, Inc., 2017. 2nd edition. Information here.
- Sheppard, Kevin. "Introduction to Python for econometrics, statistics and data analysis." *University of Oxford, version* 4 (2019). Here for free download.

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:

Final Exam = 35%
Python Lab Reports = 55%
“Attendance” and “Participation”= 10%

Note: I follow this grade rubric, but I do not generally give out A+s, regardless of your grade:

Letter	Numeric
A	93 - 100
A-	90 - 92.99
B+	87 - 89.99
B	83 - 86.99
B-	80 - 82.99
C+	77 - 79.99
C	73 - 76.99
C-	70 - 72.99
D	60 - 66.99
F	0 - 59.99

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

(unless otherwise noted, the references are to Wooldridge, 4e)

May 6 – **First Day -- Introduction to the Class**: **Ch.19**-"Carrying Out an Empirical Project." **Statistics**, **Data Structures, and Basic Descriptives**: **Ch. 1**-"The Nature of Econometrics and Economic Data"; **Appendices A.1**-"The Summation Operator and Descriptive Statistics," **A.3**-"Proportions and Percentages," **B.1**-"Random Variables and Their Probability Distributions," and **B.3**-"Features of Probability Distributions"

May 7 – **Regression 1 -- Simple Regression Model**: **Ch. 2.1**-"Definition of the Simple Regression Model," **2.2**-"Deriving the Ordinary Least Squares Estimates", and **2.4**-"Units of Measurement and Functional Form" (through "The Effects of Changing Units of Measurement on OLS Statistics"); **Correlation**: **Appendices B.2**-"Joint Distributions, Conditional Distributions and Independence," and **B.4**-"Features for Joint and Conditional Distributions"; **Goodness-of-Fit and R^2**: **Ch. 2.3**-"Properties of OLS on Any Sample of Data"; and **Hypothesis Testing**: **Ch. 4.1**-"Sampling Distributions of the OLS Estimators", **4.2**-"Testing Hypotheses about a Single Population Parameter", and **4.3**-"Confidence Intervals", Appendices **B.5**-"The Normal and Related Distributions" (through "Additional Properties of the Normal Distribution"), **C.5**-"Interval Estimation and Confidence Intervals" and **C.6**-"Hypothesis Testing"

Recommended Other Readings: * Abbott, Andrew. "Transcending general linear reality." *Sociological theory*(1988): 169-186. * Berk, Richard A. "Chapter 1: Statistical Learning as a Regression Problem." *Statistical Learning from a Regression Perspective*. Springer International Publishing, 2016. 1-53.

May 13 – **Regression 2 -- Multiple Regression Analysis**: **Ch. 3.1**-"Motivation for Multiple Regression", **3.2**-"Mechanics and Interpretation of Ordinary Least Squares", **Ch. 3.3**-"The Expected Value of the OLS" (but only from "Assumption MLR.4" on), **Ch. 6.1**-"Effects of Scaling on OLS", and **Ch. 6.3**-More on Goodness of Fit and Selection of Regressors"; and **Hypothesis Testing in Multiple Regression: Ch. 4.4**-"Testing Hypotheses about a Single Linear Combination of Parameters", and **4.6**-"Reporting Regression Results"

Recommended Other Readings: * Let's Put Garbage Can Regressions and Garbage Can Probits Where They Belong by Christopher H. Achen * Westfall, Jacob, and Tal Yarkoni. "Statistically controlling for confounding constructs is harder than you think." *PloS one* 11.3 (2016): e0152719.

May 14 – **Regression 3 -- Log Transformations: 2.4**-"Units of Measurement and Functional Form" (only "Incorporating Nonlinearities in Simple Regression"), **Ch. 6.2**-"More on Functional Form" (only "More on Using Logarithmic Functional Forms"); **Categorical-by-Continuous Interactions**: **Ch. 7.1**-"Describing Qualitative Information", **7.2**-"A Single Dummy Independent Variable", **7.3**-"Using Dummy Variables for Multiple Categories", and 7.4-"Interactions Involving Dummy Variables"

Recommended Other Readings: * *Interactions*: Friedrich, Robert J. "In defense of multiplicative terms in multiple regression equations." *American Journal of Political Science*(1982): 797-833. * *Log-Transformations*: Gustavsson, Sara, et al. "Regression models for log-normal data: comparing different methods for quantifying the association between abdominal adiposity and biomarkers of inflammation and insulin resistance." *International journal of environmental research and public health* 11.4 (2014): 3521-3539.

May 16 -- Lab #1 due

May 20 – **Regression 4 -- Continuous-by-Continuous Interactions**: **Ch. 6.2**-"More on Functional Form" (only "Models with Interaction Terms"); **Quadratics**: **Ch. 6.2**-"More on Functional Form" (only "Models with Quadratics"); and **F-Tests**: **Ch. 4.5**-"Testing Multiple Linear Restrictions: The *F* Test"

May 21 – **Regression 5 -- The Gauss-Markov Assumptions and Asymptotics**: **Ch. 2.5**-"Expected Values and Variances of the OLS Estimators", **Ch. 3.3**-"The Expected Value of the OLS", **3.4**-"The Variance of the OLS Estimators", and **3.5**-"Efficiency of OLS: The Guass-Markov Theorem", and **Ch. 5.1**-"Consistency", **5.2**-"Asymptotic Normality and Large Sample Inference", and **5.3**-"Asymptotic Efficiency of OLS"; **More Specification and Data Issues**: **Ch. 8**-"Heteroskedasticity", **Ch. 9.1**-"Functional Form Misspecification", **9.5**-"Missing Data, Nonrandom Samples and Outlying Observations", and **9.6**-"Least Absolute Deviations Estimation"

May 24 -- Lab #2 due

Recommended Other Readings: * *Normality Assumption:* Lumley, Thomas, et al. "The importance of the normality assumption in large public health data sets." Annual review of public health 23.1 (2002): 151-169. * *Heteroskedasticity*: * King, Gary, and Margaret E. Roberts. "How robust standard errors expose methodological problems they do not fix, and what to do about it." Political Analysis 23.2 (2014): 159-179. * Rigobon, Roberto, and Dani Rodrik. "Rule of law, democracy, openness, and income." Economics of transition 13.3 (2005): 533-564. *Outliers*: * Ruiter, Stijn, and Nan Dirk De Graaf. "National context, religiosity, and volunteering: Results from 53 countries." American Sociological Review 71.2 (2006): 191-210. * Van der Meer, Tom, Manfred Te Grotenhuis, and Ben Pelzer. "Influential cases in multilevel modeling: A methodological comment." American Sociological Review 75.1 (2010): 173-178. * Ruiter, Stijn, and Nan Dirk De Graaf. "National religious context and volunteering: More rigorous tests supporting the association." American Sociological Review 75.1 (2010): 179-184. & * Jasso, G. (1985). Marital Coital Frequency and the Passage of Time: Estimating the Separate Effects of Spouses' Ages and Marital Duration, Birth and Marriage Cohorts, and Period Influences. American Sociological Review, 50(2):224-241. * Kahn, J.R. and Udry J.R. (1986). Marital Coital Frequency: Unnoticed Outliers and Unspecified Interactions Lead to Erroneous Conclusions.American Sociological Review, 51(5):734-737. * Jasso, G. (1986). Is It Outlier Deletion or Is It Sample Truncation? Notes on Science and Sexuality. American Sociological Review, 51(5):738-742. * *Quantile Regression:* * Budig, Michelle J., and Melissa J. Hodges. "Differences in disadvantage: Variation in the motherhood penalty across white women's earnings distribution." American Sociological Review 75.5 (2010): 705-728. * Killewald, Alexandra, and Jonathan Bearak. "Is the motherhood penalty larger for low-wage women? A comment on quantile regression." American Sociological Review 79.2 (2014): 350-357. * Budig, Michelle J., and Melissa J. Hodges. "Statistical models and empirical evidence for differences in the motherhood penalty across the earnings distribution." American Sociological Review 79.2 (2014): 358-364. * *Missing Data Imputation*: Matthew Blackwell, James Honaker, and Gary King. 2017. "A Unified Approach to Measurement Error and Missing Data: Details and Extensions." Sociological Methods and Research, 46, 3, Pp. 342-369.

May 27 – **Models for Binary Outcomes -- Linear Probability Model**: **Ch. 7.5**-"A Binary Dependent Variable: The Linear Probability Model"; and **Binary Logistic Regression**: **Ch. 17.1**-"Logit and Probit Models for Binary Response"; **Appendix C.4**-"General Approaches to Parameter Estimation" (only "Maximum Likelihood")

Recommended Other Readings: Mood, Carina. "Logistic regression: Why we cannot do what we think we can do, and what we can do about it." European sociological review 26.1 (2010): 67-82. * Allison, Paul. "In Defense of Logit – Part 1 and Part 2" MARCH 28, 2017. Statistical Horizons blog. * Park, Hyeoun. "An introduction to logistic regression: from basic concepts to interpretation with particular attention to nursing domain." *Journal of Korean Academy of Nursing* 43.2 (2013): 154-164. * Challenger accident example from Gary King

May 28 – **Bigger Issues with Hypothesis Testing, OLS and its Assumptions, with Possible Bayesian Improvements and Beyond, Part I**. Gelman, Andrew. "The failure of null hypothesis significance testing when studying incremental changes, and what to do about it." Apr 21 2017. * Gill, Jeff. "The insignificance of null hypothesis significance testing." *Political Research Quarterly* 52.3 (1999): 647-674.

May 30 -- Lab #3 due

June 3- **Bigger Issues with Hypothesis Testing, OLS and its Assumptions, with Possible Bayesian Improvements and Beyond, Part II,** Andrew Gelman, "Bayesian statistics: What's it all about?**"** on 13 December 2016, 8:47 pm. Statistical Modeling, Causal Inference, and Social Science * Rasmus Bååth. "Bayesian First Aid: Two Sample t-test" February 24, 2014. R-Bloggers.

June 4 – **First Differences Analysis**: **Ch. 13.3**-"Two-Period Data Analysis", **13.4**-"Policy Analysis with Two-Period Panel", and **13.5**-"Differencing with More Than Two Time Periods"

June 5 -- Lab #4 due

June 10 – **Data Reduction Techniques -- Scales, Factor Analysis, Cluster Analysis, and LASSO**: Material TBA**. Last Class -- Miscellaneous, FAQ, and Review**

June 11 -- Lab #5 due

June 11 – **Last Class -- Miscellaneous, FAQ, and Review**

June 14 -- Lab #6 due

June 16 - **Final Exam Due via Courseworks**