Political Science GU4710x: Principles of Quantitative Political Research I
Fall 2019, Section 1

Robert Y. Shapiro, Rm.730 IAB
and
Tel: (212) 854-3944, email: rys3@columbia.edu

Office hours: Monday 11-12 p.m.
by appointment on most days

Teaching Fellows: George Georgarakis, gng2109@columbia.edu
Jorge Mangonnet, jmg2159@columbia.edu

Principles of Quantitative Political Research

This course examines the basic methods of data analysis and statistics that political scientists use in quantitative research that attempts to make causal inferences about how the political world works. The same methods apply to other kinds of problems about cause and effect relationships more generally. The course will provide students with extensive experience in analyzing data and in writing (and thus reading) research papers about testable theories and hypotheses.

The course will assume that students have no mathematical background beyond high school algebra and no experience using computers for data analysis. Some recent U.S. survey and other data will be used for course assignments, along with (optionally) additional data sets that have been assembled by Columbia University's Digital Social Science Center (DSSC, Lehman Library, 2nd floor IAB). Any student wishing to use other data should consult first with a teaching assistant (TA) as soon as possible (and, as needed, the DSSC), in order to enter the data into a useable computer file. Students should also consult with the instructor, as needed, about available sources for different kinds of data.

The course is structured by five required short research papers, one data documentation exercise, and one introductory data analysis exercise. The papers should be no more than 5 pages each (not counting tables and computation), and they should be typed and double spaced. All computations done by computer must include all computer programming and procedure commands that were executed (this will be discussed in class). Students should submit PAPER copies of all assignment and are responsible for keeping back-up copies (I may also ask for the electronic copies). There is no midterm exam nor final exam for this course during exam week.

Students are required to use the University's computer system e-mail accounts (user id's/"Uni") through Columbia University Information Technology (CUIT). Course data and documentation will be made available online through "CourseWorks" and/or through the CUIT lab network, and online using Internet Explorer or other Web browser computer software. Various important instructional materials will be made available through CourseWorks.

All but one of the assignments require computerized data analysis. You will be taught how to use the microcomputers and computer programs in CUIT's computer labs (e.g., in the Libraries and elsewhere). There may also be an additional fee for classroom instructional materials. For computer storage students will need USB storage sticks/keys/ drives, CDs, or

other devices depending on the data and the computers which students use. We will be using mainly a set of statistical programs called **Stata and R**. The course has a one hour weekly discussion section that students are required to register for that will assist in learning the statistical programs, and there will be additional lab sessions during the first weeks of the semester for this as well.

While the course readings will cover most of the data analysis and statistical methods of interest, it is not possible to do the assignments without attending the lectures. The lectures will go into the data analysis and statistical methods in greater depth, and they will cover matters related directly to the assignments which are not covered fully in the readings. Therefore: *Class attendance is required* and is necessary for your success in the course. We will begin taking attendance during the third week of class. You are allowed to miss up to four (4) class meetings without penalty. Missing more than four (4) days of class will require you to speak with the instructor, and it will affect your course grade. Grades will be determined by how well you are able to write reports using statistical analysis; consequently, the later assignments will be weighted more heavily than the earlier ones. Students are required to turn in their papers on time; there will be a two-day grace period, but after that papers will lose one grade for each day late.

The **required textbook** is available at the Columbia University Bookstore and will be on reserve at Lehman Library in the International Affairs Building:

Paul M. Kellstedt and Guy D. Whitten, The Fundamentals of Political Science Research, Third Edition, 2018, or Second Edition, 2013. The chapters and pages listed below are for the third edition. Lehman Library Reserve has the second edition, if not the third.

There is no other required reading. Students might be interested in referring to any standard statistics textbook that covers mathematical statistics and applications, such as the two below (or any others):

D. Moore, G. McCabe, and B. Craig, Introduction to the Practice of Statistics, Eighth Edition (or later or earlier editions).

D. Knoke, G. Bohrnstedt, and A.P. Mee, Statistics for Social Data Analysis, Fourth Edition (or earlier editions)

Other Readings of Interest

J. Davis, The Logic of Causal Order (a useful basic reading on causal thinking)
M. Lewis-Beck, Applied Regression (good treatment of basic regression analysis)
H. Asher, Causal Modeling, Second Edition (further coverage of causal modeling)
C. Achen, Interpreting and Using Regression (topics in regressions)
W. Shively, The Craft of Political Research, Sixth, Fifth, Fourth, Third or later editions
S. K. Kachigan, Statistical Analysis, Chapters 1-9

Lawrence C. Hamilton, Statistics with STATA (any edition)
STATA ("help--command" menus available on-line in STATA itself)

Supplemental/Recommended: A.H. Studenmund, Using Econometrics: A Practical Guide, Sixth Edition or later, or D. Gujarati, Basic Econometrics, Fifth Edition or later, highly recommended for those planning to do more advanced data analysis; these have been alternative texts for GU4712 Principles of Quantitative Political Research II, which is offered during the Spring semester. See also, J.M. Wooldridge, Introductory Econometrics, 4th Edition. Related Reading: G. King, R. Keohane, and S. Verba, Designing Social Inquiry: Scientific Inference in Qualitative Research; H. Brady and D. Collier, eds., Rethinking Social Inquiry: Diverse Tools, Shared Standards.

COURSE OUTLINE AND ASSIGNMENTS

Approx. Weeks 1-2. Scientific Study, Theory Construction, Research Design, Measurement and the Evaluation of Evidence. Causal theories ("recursive" models, one-way causation, versus "nonrecursive" models/reciprocal causation), representations of causal models (flow graphs/"path diagrams," and later writing out equations), concepts and variables, independent versus dependent variables, units of analysis, experimental versus observational research and the "comparative method," "levels of measurement" ("categorical" versus "continuous" variables; nominal, ordinal, interval, and ratio levels, and the special case of "dichotomous variables"), data and measurement (operationalization), validity and reliability. Explanation, covariation, hypotheses and inferences based on "samples". Readings: Kellstedt and Whitten, Ch.1-5, start 6.

Paper 1. As a thought exercise (just your own speculation) or based on any political science research you are familiar with or wish to read about in any book or journal article, write about a theory concerning the causal relationships between and among a dependent variable and particular independent variables (you may have one or more that one independent variables of interest at the outset). Assuming you had unlimited research funding, propose a research design and measurement plan for a study that would use cross-sectional (observational) data and statistical analysis to examine the theory. You could propose to replicate a past study that you have read about. Your paper should be clear on the following: What are the conceptual variables (especially the dependent one that is explained by the independent variable(s))? What is the unit of analysis? Draw and discuss the path diagram (flow graph) for the theory, including both causal and "noncausal" effects. What is the measurement, data collection, and sampling strategy? How are the variables to be measured? What are the categories of the variables/measures? What are the hypotheses concerning the relationships among the variables/measures? Explain how this study would offer evidence bearing on the causal relationship(s) of interest.

Weeks 3-4. Univariate Analysis. Statistics and population parameters: means, proportions, variance, standard deviation. Reading in advance of later lectures on sampling and sampling

distributions. "Sampling error," random error. The central limit theorem. The standard error of a mean or proportion -- or of any estimate. Confidence intervals. Hypothesis testing, Type I and Type II errors. Readings: Kellstedt and Whitten, Ch.6-7; **Stata and R** documentation for discussion and lab sessions. It is imperative that student attend class since the sequence of topics covered in class, will deviate from the readings: sampling and statistical inference after bivariate analysis (below).

Exercise #1. Data Documentation (Instruction sessions and materials will be provided for all computer work). Explore the codebooks and/or other documentation that are available on-line (through the Internet on in the computer lab) for the data that you plan to use for **Paper 2** (below). Obtain and print out the relevant documentation (including question wordings in the case of survey questions) for at least 4 variables in the data set that you plan (tentatively) to use.

Exercise #2. Data Analysis with Stata or R. Begin the data analysis for **Paper 2** (below). Select the measures that you will use in the assignment. Report the frequency distributions and circle the relevant univariate statistics for the original measures (i.e., the relevant nominal or ordinal level statistics). Then dichotomize the measures, coding them 0-1 (Note: it is best to do this by creating a new variable) and report their frequencies and again circle the relevant univariate statistics. Students should pay special attention to saving their "command" syntax or "log" or "do" files as instructed.

Weeks 5-6. Bivariate Analysis and Causal theories. Estimating simple bivariate relationships -- two variable theories and systems. Cross tabulations: percentaging tables and percentage differences. Differences in means. Measures of association and hypothesis testing (descriptive and inferential statistics). Sampling distributions and statistical inference. How to write a (quantitative) research paper. Statistical versus substantive significance? Readings: Kellstedt and Whitten, Chapter 8 and review 7. **Stata and R** documentation.

Paper 2. Write up a bivariate causal analysis (with a theory and everything else) involving two polytomous ordinal or nominal-level variables or one polytomous variable and one dichotomous variable (the original variables can actually be any variable -- nominal, ordinal, interval, or ratio level -- but you should collapse them so that they have a small number of categories). Do the analysis in the following ways: (1) First, describe the relationship, if any, that appears in the bivariate cross-tab. Just refer to the relevant percentages and interpret the relationship substantively. (2) Next, dichotomize both variables and analyze them using a path diagram and a percentage difference. (3) Then repeat the analysis, treating the dependent variable as if it were interval-level and comparing the means of it for each category of the independent variable; if, however, your dependent variable is a polytomous nominal-level variable, you must dichotomize it for this analysis to make any sense (preferably coding it 0-1; or you must justify treating the original nominal-level variable as interval-level, which is normally not possible to do).

Weeks 7-8. Bivariate Regression Analysis. Bivariate distribution for interval/ratio level (continuous) variables. The ordinary least squares (OLS) regression model. The level of measurement requirements. The "classical"/Gauss-Markov regression assumptions (to be covered further at the end of the course). "Dummy variables" and "analysis of variance" (ANOVA). Unstandardized ("b") versus standardized ("beta") regression coefficients. Measures of goodness of fit and the Pearson correlation coefficient, "r". Functional form and analysis of residuals. Further discussion of nominal and ordinal level statistics in cross-tabulations. Readings: Kellstedt and Whitten, Ch.9 and review 8 and Ch.11, pages 246-256 on "dummy variables" in regression analysis; We will be returning later to Ch. 9, pages 207-212 on regression assumptions; see below. **Stata and R** documentation.

Paper 3. Test some simple theories (three bivariate ones) by doing the following bivariate analyses: (1) Get a bivariate plot ("scatterplot") of the relationship between two interval/ratio level variables, and estimate the regression equation; write out the regression equation and interpret the coefficients and other statistics. (2) Estimate a regression equation for the relationship between two dichotomous variables (0-1 dummy variables), and compare this with the results from the 2 by 2 contingency table ("crosstab") of the two variables. (3) Compare the means of some dependent variable (any variable, including ordinal or dichotomous ones, which you are willing to treat as continuous) for each category of a polytomous nominal level variable (or any variable you wish to treat as nominal level); then replicate this "analysis of variance" using "dummy variable regression"; write out the regression equation and interpret the coefficients. (You should take note here that (2) is a special case of the more general (3), in that (2) has a dependent variable that is dichotomous (special case of an interval-type variable), and (2) happens to require only one dummy variable as the independent variable (1 = 2 - 1 categories, where the general case is: # of dummy variables required = # of categories - 1).

Weeks 9-13. Complex Theories and Multivariate Regression. Estimating multi-variable (three or more variable) models involving recursive systems. Elaborating a theory with additional variables. Interactions (specification effects/conditional relationships). Intervening variables. Spurious relationships. Suppressor variables. Partial correlations, properties of linear systems. The limits of statistical explanation. Regression with many independent variables: Interpreting coefficients, goodness of fit, and unexplained variance? The problem of specification; functional form. Dummy variables. Systems of equations, recursive models, path analysis. Uses of unstandardized coefficients versus standardized coefficients. Testing further for interactions (analysis of covariance). Multicollinearity and other potential problems with regression models. Readings: Kellstedt and Whitten, Ch.10-11; **Stata and R** documentation.

Paper 4. Using bivariate and multiple regression analysis and path analysis, examine and write up a three variable causal model. (For convenience in this assignment, to limit the number of conditional regressions/correlations to examine, you may recode the independent variables so that they have a small number of categories -- e.g., as few as 2 or 3 categories (but must be ordinal level (not nominal) to thereby treat as interval level). The dependent variable must be treatable as interval/ratio level (can assume this for ordinal or dichotomous variables). Comment

on direct and indirect effects, spurious relationships, and any statistical interactions (first-order). Decompose the important zero-order relationships (bivariate correlations).

Paper 5. This is the same as Paper 4 but for a four or more variable model. You need not estimate the conditional regressions, but you should test all first-order interactions. If you wish, you may add one or more variables to the model examined in Paper 4. In this analysis, however, you should **not** recode any of the variables into a smaller number of categories unless you have a good reason for doing so. If you have a polytomous nominal level independent variable, the variable should be treated as a set of dummy variables).

Week 14. Wrap-up. How to do the statistical analyses we have done using "**R**" statistical programs. Readings: All previously assigned readings, especially Kellstedt and Whitten, Ch.12, review Ch. 11, and see pages 207-212 and 232-233 on regression assumptions; **Stata and R** documentation. Limitations of multiple regression analysis -- caution! Violations of the assumptions of regression analysis ("Classical Assumptions," Gauss-Markov)? Outliers, influential cases? Nonlinear relationships? Heteroskedasticity? Problems with dichotomous dependent variables? Multicollinearity? Time-series analysis? Reciprocal causation/"endogeneity"? Measurement error? The cure: Political Science GU4712: Principles of Quantitative Political Research II; Political Science GU4714: or any course in Econometrics or other courses that cover how to diagnose and remedy problems encountered when the regression assumptions are violated.