

**Political Science W4910x: Principles of Quantitative Political Research**  
Fall 2012, Section 1

Mark Lindeman

Email: [taxshift@gmail.com](mailto:taxshift@gmail.com) [for best results, mention 4910 in the subject]

Cell: 845/399-0133

This course *intensively* examines the principles and basic methods of quantitative social science research. Its purpose is to provide experience in analyzing social science data and in writing (and reading) research papers about testable *substantive theories*.

The course assumes that students have little mathematical background beyond high school algebra and no experience using computers for data analysis. Most social scientists do not need intensive training in the mathematical underpinnings of analysis – but they need to understand what the methods do, how to apply them, how to interpret the results, and what to avoid.

Some recent U.S. survey data will be used for course assignments; additional data sets that have been assembled by Columbia University's Electronic Data Service (EDS in Lehman Library) are also available for use. You are encouraged to use other data sources if you like, but you should consult with a teaching assistant (TA) about the logistics of using your data with your chosen statistical software. If you have a topic in mind but not a dataset, you should consult with the instructor and/or TAs (and/or subject experts) about possible sources.

The course is structured by five required short research papers and two introductory computer-training exercises. (There may be optional problem sets along the way, if the instructor sees a need for them or if students wish to do some in consultation with the teaching assistants; this is encouraged for first year Ph.D. students who will be taking PS W4912.) Papers will be submitted electronically (details to follow). The papers should be no more than 5 pages each (not counting tables and computation), and they should be typed and double-spaced. To document computations done by computer, you must include all computer programming and procedure commands that were executed (this will be discussed in class).

Course data and documentation will be made available through the CUIT lab network in the "Shapiro" folder (since I am filling in for Bob Shapiro); the data are also available on external Web sites. Most instructional materials will be posted on the class "Courseworks" site.

All but one of the assignments require computerized data analysis. You will be taught how to use the computer programs in CUIT's computer labs (e.g., in Lehman Library and elsewhere). In order to use the labs, you may be required to pay for an "extended" Cunix account if it is not already part of your school's tuition, which it is for nearly all students (also available through CUIT). There may also be a small additional fee for classroom instruction materials. For computer storage, students will also need USB storage sticks/keys/drives, CDs, or other devices depending on the data and computers students use. We will be using mainly a set of statistical programs called STATA (for Windows), and there may be additional lab sessions covering other computer applications. You may obtain your own copy of STATA. Some students may prefer to use R, a free open-source environment for statistical computing.

We do not expect anyone to complete and understand all of the readings listed below (they often overlap considerably, so compare the readings and pick the ones that you find most helpful), but *it is not possible to do the assignments without attending the lectures*. The lectures will frequently cover matters related directly to the assignments which are not covered fully in the readings. Also, students should pay attention to the general concepts in the course outline below and make sure that they understand them fully; they should consult the assigned readings and with the teaching assistants, as needed, to review them in the corresponding weeks. Your grades will be determined by how well you have learned the methods of the course; consequently, the later assignments will be weighted more heavily than the earlier ones. Students are urged to turn in their papers on time; no grades of "Incomplete" will be given (except in cases of personal emergencies).

All of the required ("Readings") and most of the supplemental/"Recommended" ("Rec.") readings are available at the Columbia University Bookstore and are on reserve at Lehman Library. The main assigned books are as follows:

Main Texts (normally read chapters in at least one of the two texts listed here)

- D. Moore, G. McCabe, and B. Craig, *Introduction to the Practice of Statistics*, Seventh Edition (or earlier).
- D. Knoke, G. Bohrnstedt, and A.P. Mee, *Statistics for Social Data Analysis*, Fourth Edition (or chapters in earlier eds.), alternative and supplement to MMC.

Other Readings (including required and supplemental/recommended; see below)

- J. Davis, *The Logic of Causal Order* (a central reading for the course on causal modeling)
- W. Shively, *The Craft of Political Research*, Sixth, Fifth, Fourth, or Third Editions
- S. K. Kachigan, *Statistical Analysis*, Chapters 1-9
- M. Lewis-Beck, *Applied Regression* (good treatment of regression analysis)
- C. Achen, *Interpreting and Using Regression* (topics in regressions)
- H. Asher, *Causal Modeling*, Second Edition (further coverage of causal modeling)
- STATA ("help--command" menus available on-line in STATA itself; Lawrence C. Hamilton, *Statistics with STATA*, Updated for Version 10, 2009 (earlier editions updated for at least Version 8 may be useful).

Supplemental/Recommended: A.H. Studenmund, *Using Econometrics: A Practical Guide*, Sixth Edition, or D. Gujarati, *Basic Econometrics*, Fifth Edition, highly recommended for those planning to do more advanced data analysis; these have been alternative texts for W4911y Analysis of Political Data, which is offered during the Spring semester. See also, J.M. Wooldridge, *Introductory Econometrics*, 4th Edition.

Related Reading--highly recommended though not required: 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 (subject to change)

### Weeks 1-2. Theory Construction and the Evaluation of Evidence.

- Concepts and variables
- Causal theories and models (flow graphs/path diagrams).
- “Recursive” models, one-way causation (versus “nonrecursive” models).
- Unit of analysis.
- Levels of measurement (nominal, ordinal, interval/ratio; and the special case of “dichotomous variables”)
- Data and measurement (operationalization).
- Validity and reliability
- Explanation, covariation, hypotheses, and inferences.

Readings: Davis, *The Logic of Causal Order*; Moore, McCabe and Craig, , p.xxiii-xxv, 148-152, 129-130, 170-171; Knoke, Bohrnstedt, and Mee, Ch.1, Ch.11 p.371-377; Shively, Ch.1-6,10; Kachigan, Ch. 1-2; M. Rosenberg, *The Logic of Survey Analysis*, Ch.1. Recommended (supplemental, not required): H. Blalock, *Theory Construction*, Ch.1-3; A. Stinchcombe, *Constructing Social Theories*, Ch.3 (esp. 130-148); King, Keohane, and Verba, *Designing Social Inquiry: Scientific Inference in Qualitative Research*, Ch.1-3.

**Assignment 1.** Find a theoretical statement from a piece of research in a social science book or journal article. What are the conceptual variables (and 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. What are the hypotheses concerning the relationships between variables? How are the variables measured? What are the categories of the variables/measures. Discuss the substantive meaning or interpretation(s) of each causal effect or noncausal effect (correlation without a causal explanation), emphasizing the theoretically important ones. Discuss the article's findings and conclusions.

### Weeks 3-4. Univariate Analysis.

- Statistics: means, proportions, variance, standard deviation.
- 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: Moore, McCabe, and Craig, Ch.1, 3-6, 7.1, 8.1; Knoke, Bohrnstedt, and Mee, Ch.2-3, Appendix A; Kachigan, Ch.3-9; STATA.

**Computer/Documentation Training Exercise #1.** (Instruction sessions and materials will be provided for all computer work). Explore the NORC General Social Survey Codebook, one of the National Election Study codebooks, or other data sources that are available on-line (through

the Internet) and that you plan to use for Assignment 2 (below). Obtain, printout, and submit the question wordings and other available information for at least 4 variables in the data set.

**Computer Training/Statistics Exercise #2.** Begin Assignment 2 (below). Select the measures that you will use in the assignment. Report the frequency distributions and circle the useful univariate statistics for the original measures. Then dichotomize them, coding them 0-1 (NB: it is usually best to do this by creating a new variable) and report their frequencies and 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, percentage differences
- Differences in means.
- Measures of association and hypothesis testing (descriptive and inferential statistics).
- How to write a (quantitative) research paper.

Readings: Moore, McCabe, and Craig, McCabe, Ch.7.1, 7.2, 8.1, 8.2; Knoke, Bohrnstedt, and Mee, Ch.4-5; Shively, Ch.8 (p.116-123 only) and Ch.9; Kachigan, Ch.7-9; STATA. Rec: J. Davis, "Analyzing Contingency Tables with Linear Flow Graphs: D-Systems," in D. Heise, ed., *Sociological Methodology 1976*.

**Assignment 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 distributions for interval/ratio level (continuous) variables.
- The ordinary least squares (OLS) regression model.
- The Gauss-Markov assumptions and theorem (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.

- Functional form.
- Analysis of residuals.
- Further discussion of nominal and ordinal level statistics in crosstabulations.

Readings: Moore, McCabe, and Craig, Ch. 2, 10, 12, and 9; Knoke, Bohrnstedt, and Mee, Ch.6, 4-5; Studenmund, 1-3, 5, and 4 (on Gauss-Markov assumptions); Lewis-Beck, Ch.1-2; STATA. Rec. Shively, Ch.7.

**Assignment 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 (cross-tab) 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.

Readings: Knoke, Bohrnstedt, and Mee, Ch.7; Rosenberg, Ch.2-9 and Appendixes A & B; Davis, *The Logic of Causal Order*. Rec.: Davis, "...D-Systems"; Blalock, "Four-variable Causal Models and Partial Correlations," *American Journal of Sociology* 68 (1962): 182-194.

- Multiple regression analysis, interpreting coefficients, goodness of fit, analysis of residuals.
- The problem of specification, functional form.
- Dummy variables.
- Systems of equations, recursive systems, path analysis. Uses of unstandardized versus standardized coefficients.
- Testing further for interactions.

- Analysis of covariance.
- Potential problems with regression models.

Readings: Moore, McCabe, and Craig, Ch.11, compare 13; Knoke, Bohrnstedt, and Mee, Ch.8,11; Lewis-Beck, Ch.3; Asher; Achen; Shively, Ch.8, p.125-131; STATA. Rec.: Studenmund, Ch.2-4,6 7; Berry and Feldman, *Multiple Regression in Practice*. G. King, R. Keohane, and S. Verba, *Designing Social Inquiry: Scientific Inference in Qualitative Research*, Ch.4-6; Brady and Collier, skim.

**Assignment 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.

**Assignment 5.** This is the same as Assignment 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 Assignment 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.

- A look back at cross tabulations. Its strengths and weaknesses. (Readings: previously assigned; Rec.: skim Moore, McCabe, and Craig, Ch.14-15; Knoke, Bohrnstedt, and Mee, Ch.9,12,10; Studenmund, Ch.7-15; Shively, p.123-125).
- Limitations of multiple regression analysis -- caution! Violations of assumptions of regression analysis (Gauss-Markov; see Studenmund, Ch.4). The cure: e.g., Political Science W4911y: Analysis of Political Data; Political Science W4912: Multivariate Analysis; International Affairs U6815: Statistical Analysis for Policymaking; or other courses that cover how to diagnose and remedy problems encountered when these assumptions are violated.

(Optional) **Assignment 6.** (For those who want it or for whom 5 papers just weren't enough.) Read and critique a research article (from a book chapter or journal) that uses quantitative analysis. Or, alternatively, do a final but brief research paper using whatever statistical methods you think are warranted. (Students may work in groups of two on this project if they assemble their own data sets.)