

**Pol. Sci. W4911y: Analysis of Political Data/Applied Econometrics**  
**Spring 2014, Section 1**

Professor Robert Shapiro  
730 IAB, Ph. 854-3944,  
E-mail: rys3@columbia.edu

Office Hours: Mon. 11-12 Noon  
and most days by appointment

This course will intensively examine some of the data analysis methods which deal with problems occurring in the use of multiple regression analysis. It will stress computer applications and cover, as needed, coding and processing data. Emphasis will also be placed on research design and writing research reports.

The course assumes that students are familiar with basic multiple regression analysis and have analyzed data using a computer program (e.g., any standard statistical programs on micro-computers or larger machines -- Stata, R, SPSS, SAS, etc.). Students will be instructed on the use of the microcomputers and Stata statistical software programs available in the CUIT computer labs (several campus locations) or through SIPA. In order to use the labs, students are required to have extended CUNIX accounts or SIPA computer lab access for the semester (this is normally provided through regular tuition and student registration), and there will be an additional fee for classroom instructional materials. (Students are permitted to use other statistical programs on personal or other computers, but they should consult with the instructor and a teaching assistant about this and about obtaining appropriate data sets.)

The course requires the completion of five short data analysis papers, with an "optional" 6<sup>th</sup> paper (5 pages maximum of text, plus [not counted] tables and graphs). These papers should be typed and double-spaced. Students are responsible for keeping copies of the papers which they submit. For two of the assignments students are required to use (1) some cross-sectional data (or short panel data) other than the survey data explicitly made available for the class, and (2) time series data which will not be directly provided. These data sets can be data already available in electronic form or they can be raw data assembled by the student from published or other sources. Each of these data sets should have at least one dependent variable and four independent variables for at least 25 observations (cases; fewer variables or cases may be permitted upon consultation with the instructor). Students are urged to begin as soon as possible to look for data sources in the library and elsewhere. For the remaining assignments students may use these same data or any of the other recent U.S. and foreign data, World Bank, or other older data made available in the CUIT labs, the SIPA lab, or the Electronic Data Service's (EDS, second and third floor IAB, Lehman Library) on-line "Data Gate". The data that will be made directly available for the course in the CUIT labs are recent NORC General Social Surveys (GSS), American National Election Study (ANES) surveys, and National Election Pool (NEP) "exit polls".

I do not expect everyone to understand fully some of the complex course readings, but it will be difficult to do the assigned papers without attending every class meeting. Each assignment will be weighted approximately equally in determining grades; however, for the first two assignments, I will "drop" the lower of the two grades received. **Students are required to turn in their papers as PAPER copies (not electronic unless requested further to do so) on the**

**schedule to be announced in class (the official due date will be at least one week after I have covered the relevant topic), allowing only for an additional two-day grace period for Papers #1-4. Papers submitted late will lose one grade per day late. No grades of "Incomplete" will normally be given (except in cases of emergencies). Paper #5 is due by Thursday, May 1; the optional Paper #6 is due on Monday, May 5th.**

All of the required and most of the recommended readings are available at the Columbia University Bookstore and are on reserve at Lehman Library. The assigned books are:

**Main texts:**

A. H. Studenmund: Using Econometrics, Sixth Edition, 2011.  
or D. Gujarati and D. Porter, Basic Econometrics, Fifth Edition, 2009.

**Others readings:**

C. Achen, Interpreting and Using Regression  
J. Aldrich and F. Nelson, Linear Probability, Logit, and Probit Models  
H. Asher, Causal Modeling, 2nd edition  
J. Davis, The Logic of Causal Order  
J. Kim and C. Mueller, Introduction to Factor Analysis  
M. Lewis-Beck, Applied Regression  
C. Ostrom, Time Series Analysis, 2nd ed.  
J. Sullivan and S. Feldman, Multiple Indicators  
W. Berry, Nonrecursive Causal Models

Recommended: J. M. Wooldridge, Introductory Econometrics, 4<sup>th</sup> edition, 2009 (a very good and widely used alternative text); L.C. Hamilton, Statistics with STATA, 2009 or any later edition; G. King, R. Keohane, and S. Verba, Designing Social Inquiry: Scientific Inference in Qualitative Research (not assigned but highly recommended); J. Miller, The Chicago Guide to Writing about Numbers and The Chicago Guide to Writing about Multivariate Analysis.

## COURSE OUTLINE AND ASSIGNMENTS

**Weeks 1-4 (approx.). Multiple Regression Analysis.** Introduction and overview of the course. Prerequisites and course requirements. Computer training. Levels of measurement. Sources for cross-sectional and time series data. Theory construction, causal models, hypothesis testing, evaluating evidence, and statistical inference. Review of Ordinary Least Squares (OLS) regression analysis, multiple regression analysis, unstandardized versus standardized coefficients, important uses of dummy variables, analysis of variance and covariance (i.e., how they can be

viewed as identical to variations of multiple regression analysis). Structural equations, path analysis, recursive vs. nonrecursive models. Panel data and related change designs (see Time Series Analysis below). Statistical interactions, Chow-test, multicollinearity, outliers and the importance of case studies, analysis of residuals, heteroskedasticity (Weighted Least Squares [WLS]; "White-corrected"/"robust" standard errors), simple non-linear transformations, and other problems. Data coding and data processing. Use of PCs\microcomputers, Stata. Editing and formatting data (text vs. word-processing files). Weighted data. Missing data. Measurement error. Readings: Studenmund, (Ch.17 reviews "Statistical Principles") Ch. 1-8, 10-11 (or Gujarati and Porter, Ch.1-9,10,11,13, Appendix A [B-C on the matrix algebra approach]; Lewis-Beck; Davis; Achen; Paul Allison, "Testing for Interactions in Multiple Regression," American Journal of Sociology 83 (1977): 144-153. Hamilton, Statistics with STATA, p.44-52, on importing data. Beginning here and throughout the course students can use Hamilton as a reference guide for Stata commands and procedures, and they can refer to the Miller books for writing about statistical analyses. Recommended (not required): Wooldridge, Ch. 1-9, 19; Appendixes A-E.

**Assignment 1.** Using any data you wish, examine and write up a four or more variable causal model, based on multiple regression and (to some extent) path analysis. Present the structural equations and the path diagram. Interpret the regression coefficients (focus on the usual unstandardized coefficients; it is not required to interpret the standardized coefficients nor to decompose any zero-order relationships into direct, indirect, and noncausal [spurious] effects). Test for first-order interactions (and, if you think necessary, any theoretically compelling higher order interactions). Examine multicollinearity and provide some analysis of residuals, especially heteroskedasticity, and, as needed, outliers).

**Assignment 2.** Same as Assignment 1, use a cross-sectional data set that you have collected, coded, etc. (that is, any data set other than the GSS, ANES, or NEP data). Pay additional attention to the bivariate plots of variables, outliers, patterns of residuals, tests for heteroskedasticity, multicollinearity, and the transformation of variables to deal with functional form and other problems or considerations.

**Weeks 5-6. Models with Discrete (esp. dichotomous) Dependent Variables.** Detecting and correcting for heteroskedasticity. OLS, Weighted Least Squares (WLS) "Goldbergerizing," logit and probit analyses. Method of moments vs. method of maximum likelihood (MLE). Readings: Studenmund, Ch. 13 (Gujarati and Porter, Ch. 15 [mainly],14); Aldrich and Nelson; Hamilton, Statistics with STATA. Rec.: Wooldridge, Ch.17; Aldrich and Cnudde, "Probing the Bounds of Conventional Wisdom..." American Journal of Political Science 19 (Aug. 1975): 571-608.

**Assignment 3.** Similar to assignments 1 and 3, for a dichotomous dependent variable, estimate a model using OLS and logit and/or probit analysis. Explain why the last methods are normally considered superior. Discuss and compare the results, paying special attention to any substantive discrepancies.

**Weeks 7-8. Simultaneous Equation Models.** Nonrecursive vs. recursive systems of equations. Instrumental variables, two-stage least squares (2SLS), and other methods. The "identification" problem. Assumptions about exogenous variables. Panel analysis. Readings: Studenmund, Ch.14 (Gujarati and Porter, Ch. 18-20); Asher; Berry; Rec.: Wooldridge, Ch. 15-16; Markus, Analyzing Panel Data.

**Assignment 4.** Estimate a nonrecursive model using two-stage least squares. Discuss the assumptions made to "identify" and estimate the model. Report the first stage equations. Compare the 2SLS results with those from a recursive OLS model.

**Weeks 9-11. Time Series Analysis.** The unique nature of time series analysis. Serial correlation/autocorrelation. Estimation of lagged relationships; lagged exogenous and lagged endogenous variables. Nonstationarity and spuriousness in time series models. Generalized least squares (GLS) and pseudo-GLS; stochastic process models versus structural equation models. "Unit roots"; the Dickey-Fuller test(s). Change designs, first-differences. Panel analysis. Pooled (cross-section) time series. Fixed effects versus random effects models. Readings: Studenmund, Ch.9,12,15,17 (Gujarati and Porter Ch.12,17,21,22, and 16); Ostrom; Rec.: Wooldridge, Ch.10-14,18, has very extensive coverage of time series analysis; Pindyck and Rubinfeld, Econometric Methods and Econometric Forecasts, Part III; McCleary and Hay, Applied Time Series Analysis; Cook and Campbell, Quasi Experimentation, Ch.5-6.

**Assignment 5.** Estimate a multivariate time series model using data that you have collected. Compare and contrast your final results with those from a simple OLS model.

**Weeks 12-14. Unobserved Variables, Measurement, Factor Analysis, and Further Topics.** Errors in variables/measurement. Exploratory factor analysis: principal components, orthogonal vs. oblique rotations (assumptions), maximum likelihood method. Scale construction, reliability and validity, Cronbach's alpha. More about the mathematics and statistics of the methods examined in the course. LISREL. Readings: Studenmund, p.491-494 (Gujarati and Porter, 482-486); Kim and Mueller; Piazza, "The Analysis of Attitude Items," American Journal of Sociology 86 (November 1980): 584-603; Sullivan and Feldman; Carmines and Zeller, Reliability and Validity Assessment. Rec.: Wooldridge, p.316-322; J. Scott Long, Confirmatory Factor Analysis and Covariance Structure Models.

**(Optional) Assignment 6 (or alternative).** Examine the relationships among a set of seemingly related variables using factor analysis. Justify the assumptions underlying your factor analytic model. Present the path diagram. Construct factor scales that would seem appropriate and calculate their Cronbach's alphas. Correlate the separate items used in constructing the scales (do not focus on the final scales themselves) with several other variables. What do these correlations suggest about the scales/the separate items (see Piazza, 1980)? That is, matters of reliability and validity. **Alternative Assignment 6.** An individual or joint research project, to be approved by the instructor, using methods covered during the last weeks of the course.