

Advanced Topics in Quantitative Data Analysis

Winter 2011

Last update: 12 December 2010

Visit the e-learning site of the course to find the most recent version of the syllabus.

Department of Political Science
Central European University

Instructor: Gábor Tóka (room FT 804, email: tokag at ceu.hu)

Classes: Thursday, 1:30-3:10 pm

Office hours: FT804, Wednesday 1-3 pm and by appointment

Credits: 2 CEU credits, 4 ECTS credits

Goals

This is a two-credit course that is open to both MA and PhD students that aims at enabling students to carry out computer-assisted quantitative analyses on their own on how micro behavior of individual actors is shaped by context, and to understand better the statistical apparatus and tools used in contemporary research in policy, international relations and political science research. Yet this is not a statistics course as such. Instead, the emphasis is on the practical know-how that a practicing empirical researcher employs in coping with various kinds of problems in everyday data analysis: from filling in gaps in the data and finding appropriate computer programs through developing scripts automating a complex analysis to understanding and tackling technical problems with further theorizing about the political reality that generated them. All classes are hands-on computer lab exercises using the R or STATA softwares (possibly SPSS or HLM at the junctions where this is appropriate and justified by popular demand in the class), i.e. you can try the methods instantly as we learn about them. The prerequisite to effective participation in the course is a sound understanding of basic measures of dispersion (i.e., sum of squares, variance, standard deviation), the notion of standard errors, and multivariate linear regression analysis, i.e. the completion of an introductory course in statistics. What we plan to achieve in this course is to make you comfortable with and thoughtful about actually using these procedures and competently cope with complications that go beyond what you studied in class before, but commonly occur when you are using real data in your professional career. In other words, the course is meant to assist developing skills and habits that will allow you to conduct and present statistical analyses professionally and keep up independently with developments in the ever-changing field of statistical methodology according to your own needs.

We start off with quickly revisiting the assumptions behind ordinary least squares regression and reviewing the substantive reasons why modellers of social processes (say the flow of foreign direct investment in democratic and non-democratic countries) may encounter serious violations of some or most of these assumptions. We then learn to diagnose with graphs and formal tests if such violations occur in our data and to set up simple Monte Carlo simulations and bootstrap procedures to study the possible impact of those violations on the results of our analysis. Next we look for methods that can appropriately adjust our data and analysis to take into account these problems. We first examine how simple recoding of the variables given in a data set may enable us to more appropriately take into account the actual mechanisms of the social and political reality that we study, and how such adjustments in an of themselves can solve seemingly technical problems of statistical estimation. We will also consider various distributions, i.e. situations when the outcome of interest is not normally distributed but is, for instance, dichotomous (like war or not war), and what adjustments of standard regression analysis can respond to the dependent variable displaying them, introducing logit, probit, ordered logit, and poisson regression, as well as graphical methods of understanding results obtained with these estimators. Building on what we learnt about simulations and bootstrapping, by then we will also be ready to understand state-of-the-art methods of multiple imputation (as well as when it can be used) to deal with missing data problems in our analyses, and the use of similar methods in establishing the reliability of new data, e.g. about the policy preferences of different actors derived from text available on the internet.

The second third of the course examines contingency and functional equivalence, i.e. situations of interaction when the impact of some variables (say your gender) is dependent on the presence/absence of another factor (say the political institutional context in which you act). This is particularly common in politics because complex multi-level and multi-actor systems provide ample space for multiple conjunctural causation (i.e., the same outcome can be caused by different causes in different contexts), feedback (e.g., actors provide information to each other about the consequences of each other's actions), learning (e.g., once actors understand some causal sequences, they adjust their behavior accordingly), tit for tat (actors may condition their strategies on the strategy followed by other actors in previous rounds of the same game), and similar interactive processes in generating sometimes disastrous and some times beneficial outcomes. We will learn about various simple, commonly used methods of displaying and understanding the impact of interaction effects, and simple techniques for estimating so-called multilevel (i.e., "hierarchical") models quickly, reliably and accurately.

The final third of the course is devoted to the study of dynamics in international affairs, policy change and politics via so-called panel data analysis, i.e. when we can observe change in multiple variables over regular time intervals in several contexts (e.g. in multiple countries or in multiple dyads of countries). We engage with the basics of time-series analysis, learn tests and solutions for common problems in political research like non-stationary dependent variables, serial and spatial correlation, fixed versus random unit effects, and pay particular attention to how one can clarify the possible impact of variables that probably do not change too frequently (say constitutional design or some aspects of culture).

The empirical examples in our analyses will come from a variety of applications regarding, for instance, the impact of democracy on direct foreign investment around the contemporary world; the impact of the candidates' gender and previous incumbency on their share of personal preference

votes in the party list-proportional systems of the Czech Republic, Estonia, Poland, and Slovakia; the impact of macro-factors like within-EU trade, domestic political performance deficits and the net EU-level transfer of budgetary resources between EU member states on support for European integration; and how the design of political institutions reduces or enhances the impact of citizens' political ignorance on election outcomes

Requirements

Ten percent of your grade will depend on your active in-class participation, fifty percent on your solution to the weekly take-home exercises, and 40 percent on your final paper. Every week, you will receive take-home computer exercises related to the tasks discussed in class and will upload your solutions to the e-learning site of the course's 24 hours before the next class so that we can all learn from it. Some of these tasks will involve you in team work and they will be meant to provide enjoyable opportunities to learn. What will matter for the grade is not whether you can come up with perfect solutions but to show evidence that you seriously engaged with the task, tried to solve it making use of what you had already known before, what you learnt in this class, and what simple searches of easily available resources provide in the way of tips. In addition, you will write up the statistical analysis for an original research paper of yours in max. 4,000 words and submit it together with all the computer codes and data that you used to generate the results.

Final paper

This will have to look like the section on data analysis in a journal article and be a serious exercise in writing such things. Therefore it must not be longer than 4,000 words, i.e. half the usual maximum length of such an article. You choose your data and research topic but will need to get my approval for these choices by 15 February. The point of this approval is merely to make sure that you do something that is appropriate for the course and allows you to get a good grade at the end. Your paper must address a theoretically significant question, present your own quantitative analysis of the issue, and be written in an academic journal format. However, this time you will omit any literature review as well as the theory section, and merely state your research question, discuss the design of your statistical analysis, i.e., how you chose to do things in the statistical analysis and why, and then go on to present your own findings and the results. You are obviously expected to assess the merits and drawbacks of alternative solutions to your estimation problems and present a methodologically sound interpretation of the results that you need to present in professionally prepared tables and charts. A reference must be formally cited any time the ideas, research findings, or data of someone else is mentioned or otherwise utilized. A list of references has to be provided at the end of the paper - this, of course, must list no more and no less than every work actually referred to in the paper. You will also need to submit your data and computer codes for the computations and make sure that your results can be seamlessly replicated with them. The final version of the paper must be uploaded to the e-learning site of the course one week before the grades are due for the Winter semester. Plagiarism of any sort will lead to failing the course and be referred to the relevant disciplinary bodies for further sanctions.

Readings

A course packet will be available from the departmental office. The recommended readings will all be available from the CEU library in hard copy and/or electronic form through JSTOR, Ebsco, or the e-learning site of the course. Since both the pace with which we will progress and the amount of material that we can cover depends to a degree on the participants' background, I do not break

down readings by week below. Generally, however, the readings cover two sorts of things. Some give background on the examples that we are using and you will be asked to read these before a given example is coming up in class. The others give more general help in conducting your analyses, and your throughout the course you will receive advance advice on what chapters of key readings will be particularly relevant for the next class. Some further web links, manuals, and textbooks with help on the software that we are going to use will be provided through the e-learning site of the course. As introductory readings to the three most commonly used general statistical software in contemporary political and policy studies, you can also consult:

Everitt, Brian S., and Torsten Hothorn. 2006. *A Handbook of Statistical Analyses Using R*. Boca Raton, FL: Chapman & Hall/CRC.

Pollock, Philip H., III. 2006. *A Stata Companion to Political Analysis*. Washington, DC: CQ Press.

Field, Andy. 2005. *Discovering Statistics (and Sex, Drugs, and Rock 'n' Roll) Using SPSS*. 2nd ed. London: Sage.

Part One (approximately weeks 1-5)

Mandatory readings:

Choi, Seung-Wan. 2009. "The Effect of Outliers on Regression Analysis: Regime Type and Foreign Direct Investment." *Quarterly Journal of Political Science* 4: 153-165.

Mooney, Christopher Z., and Robert D. Duval. 1993. *Bootstrapping: A Nonparametric Approach to Statistical Inference*. Newbury Park, CA: Sage (parts).

Albert, Jim. 2009. *Bayesian Computation with R*. 2nd ed. New York: Springer (parts).

King, Gary, James Honaker, Anne Joseph, and Kenneth Scheve. 2001. "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation." *American Political Science Review* 95 (1): 46-69.

Plus: selected parts of either of these introductory texts (choice depends on your background and software of choice):

Cameron, A. Colin, and Pravin K. Trivedi. 2009. *Microeconometrics Using Stata*. College Station, TX: Stata Press (parts).

Faraway, Julian J. 2004. "Practical Regression and Anova using R." Manuscript. Available from <http://www.stat.lsa.umich.edu/~faraway/book/prs.pdf> accessed on 12 August 2005 (parts).

Recommended readings:

Lowe, Will, Kenneth Benoit, Slava Mikhaylov, and Michael Laver. 2010. "Scaling Policy Positions From Coded Political Texts." Manuscript (forthcoming in *Legislative Studies Quarterly*). Maastricht: University of Maastricht.

Aldrich, John J., and Forrest D. Nelson. 1984. *Linear Probability, Logit, and Probit Models*. Beverly Hills, CA: Sage.

Fox, John. 2002. *An R and S-Plus Companion to Applied Regression*. Thousand Oaks, CA: Sage.

Ritz, Christian, and Jens Carl Streibig. 2008. *Nonlinear Regression with R*. New York: Springer.

Rabe-Hesketh, Sophia, and Brian S. Everitt. 2007. *A Handbook of Statistical Analyses Using Stata 9*. 4th ed. London: Chapman and Hall.

Long, J. Scott, and Jeremy Freese. 2003. *Regression Models for Categorical Dependent Variables Using Stata*. 2nd ed. College Station, TX: Stata Press.

- Lowe, Will, Kenneth Benoit, Slava Mikhaylov, and Michael Laver. 2009. "Scaling Policy Positions From Coded Units of Political Texts." Data set. Version: October 13, 2009.
- Schafer, Joseph L., and John W. Graham. 2002. "Missing Data: Our View of the State of the Art." *Psychological Methods* 7 (2): 147-177.
- Horton, Nicholas J., and Ken P. Kleinman. 2007. "Much Ado About Nothing: A Comparison of Missing Data Methods and Software to Fit Incomplete Data Regression Models." *The American Statistician* 61 (1): 79-90.
- Rubin, Donald B. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons.
- Schafer, Joseph L. 1997. *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall.
- Kaufmann, Daniel, Aart Kraay, and Massimo Mastruzzi. 2008. "Governance Matters VII: Aggregate and Individual Governance Indicators for 1996-2007." World Bank Policy Research Working Paper No. 4654.
- Efron, Bradley, and R.J. Tibshirani. 1993. *An Introduction to the Bootstrap*. London: Chapman & Hall.
- Mooney, Christopher Z. 1997. *Monte Carlo Simulations*. Thousand Oaks, CA: Sage.
- Robert, Christian, and George Casella. 2009. *Introducing Monte Carlo Methods with R*. New York: Springer.

Part Two (approximately weeks 6-8)

Mandatory readings:

- Bartels, Larry M. 1996. "Uninformed Votes: Information Effects in Presidential Elections." *American Journal of Political Science* 40: 194-230.
- Long Jusko, Karen, and W. Phillips Shively. 2005. "Applying a Two-Step Strategy to the Analysis of Cross-National Public Opinion Data." *Political Analysis* 13 (4): 327-344.
- Braumoeller, Bear F. 2004. "Hypothesis Testing and Multiplicative Interaction Terms." *International Organization* 58 (4): 807-820.
- Kastellec, Jonathan P., and Eduardo L. Leoni. 2007. "Using Graphs Instead of Tables in Political Science." *Perspectives on Politics* 5 (04): 755-771.
- King, Gary, Michael Tomz, and Jason Wittenberg. 2000. "Making the Most of Statistical Analyses: Improving Interpretation and Presentation." *American Journal of Political Science* 44 (2): 347-361.

Recommended readings:

- Alvarez, R. Michael, Geoffrey Garrett, and Peter Lange. 1991. "Government Partisanship, Labor Organization, and Macroeconomic Performance." *American Political Science Review* 85 (2): 539-556.
- Brambor, Thomas, William Roberts Clark, and Matt Golder. 2006. "Understanding Interaction Models: Improving Empirical Analyses." *Political Analysis* 14 (1): 63-82.
- Berry, William D., Jacqueline H. R. DeMeritt, and Justin Esarey. 2010. "Testing for Interaction in Binary Logit and Probit Models: Is a Product Term Essential?" *American Journal of Political Science* 54 (1): 248-266.
- Fox, John, and Robert Andersen. 2006. "Effect Displays for Multinomial and Proportional-Odds Logit Models." *Sociological Methodology* 36 (1): 225-255.
- Kam, Cindy D., and Robert J. Franzese, Jr. 2007. *Modeling and Interpreting Interactive Hypothesis in Regression Analysis*. Ann Arbor, MI: University of Michigan Press.

- Gelman, Andrew, and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.
- Wells, Jason M., and Jonathan Krieckhaus. 2006. "Does National Context Influence Democratic Satisfaction? A Multi-Level Analysis." *Political Research Quarterly* 59 (4): 569-578.

Part Three (approximately weeks 9-12)

Mandatory readings:

- Mikhaylov, Slava, and Michael Marsh. 2009. "Policy Performance and Support for European Integration." In *The Legitimacy of the European Union After Enlargement*, edited by Jacques Thomassen. Oxford: Oxford University Press, pp. 142-164.
- Beck, Nathaniel, and Jonathan N. Katz. 1995. "What To Do (and Not To Do) with Time Series-Cross Section Data." *American Political Science Review* 89 (3): 634-647.
- Wilson, Sven E., and Daniel M. Butler. 2007. "A Lot More to Do: The Sensitivity of Time-Series Cross-Section Analyses to Simple Alternative Specifications." *Political Analysis* 15 (2): 101-123.
- Bartels, Larry M., and John Zaller. 2001. "Presidential Vote Models: A Recount." *PS: Political Science and Politics* 34 (1): 9-20. URL: <http://www.apsanet.org/ps/march01/election2000.cfm>.

Recommended readings:

- Kleiber, Christian, and Achim Zeileis. 2009. *Applied Econometrics with R*. New York: Springer.
- Shumway, Robert H., and David S. Stoffer. 2006. *Time Series Analysis and its Applications: With R Examples*. New York: Springer.
- Beck, Nathaniel. 2001. "Time-Series-Cross-Section Data: What Have We Learned in the Past Few Years?" *Annual Review of Political Science* 4 (1): 271-293.
- Hoeting, Jennifer A. 2009. "Methodology for Bayesian Model Averaging: An Update." Fort Collins, CO: Colorado State University.
- Greene, William H. 2008. *Econometric Analysis*. 6th ed. Upper Saddle River, NJ: Prentice Hall.
- Baltagi, Badi H. 2005. *Econometric Analysis of Panel Data*. 3rd ed. New York: Wiley.
- Wooldridge, Jeffrey M. 2002. *Econometric Analysis of Cross Section and Panel Data*. 3rd ed. Cambridge, MA: The MIT Press.