

SC706: LONGITUDINAL DATA ANALYSIS

Instructor: Natasha Sarkisian
Email: natasha@sarkisian.net
Phone: (617) 755-3178
Office: McGuinn 417

Class time: Tuesdays 4-6:20 PM
Class location: O'Neill 245
Office hours: By appointment
Webpage: <http://www.sarkisian.net/sc706/>

COURSE DESCRIPTION*

This applied course is designed for graduate students with a prior background in statistics at the level of SC703: Multivariate Statistics (or its equivalent). This means that students should have considerable experience with ordinary-least-squares (OLS) regression: I assume you have an understanding of multiple OLS regression and an ability to conduct such analyses using some statistical software (e.g., SPSS, SAS, Stata, etc.). The major topics of the course will include change models, fixed and random effects, GEE models, and mixed models.

The goals of the course are to develop the skills necessary to identify an appropriate technique, estimate models, and interpret results for independent research and to critically evaluate contemporary social research using advanced quantitative methods. The course will be applied in the sense that we will focus on estimating models and interpreting the results, rather than on understanding in detail the mathematics behind the techniques. I hope that the course will provide you with a solid foundation in longitudinal data analysis, which is a type of advanced quantitative skill that is in high demand in many fields, both in and out of academia. For those of you in the Sociology Department, the course can also provide a foundation for the "Advanced Quantitative Methods" area examination.

We will be using Stata for all the analyses throughout the course. No previous Stata experience is necessary: I will provide an introduction to Stata in the beginning of the course and guide you throughout the course. For your assignment, you can use Stata on Citrix: see <http://apps.bc.edu>.

COURSE POLICIES

For each topic in the course, I will give a lecture focusing on the reasoning behind the technique, and provide a review of the syntax used to do analyses and the output generated by Stata. Throughout that process, you will get a chance to practice conducting the analyses and interpreting the results. We will also discuss and critically evaluate published research based on the various techniques. Make sure that you carefully read these examples of published research before class and be prepared to discuss them. The course is based on an interactive relationship between the instructor and students, as well as on collaboration among the students. You are strongly encouraged to ask questions and discuss the material in class. I also encourage collaboration among the students. Please feel free to help each other when running analyses for assignments. However, everyone must turn in their own report and statistical output.

I also would like to stress that you are always welcome to come and see me with any additional questions. Email is the best way to get in touch with me to get a quick question answered or to set up an appointment to discuss something at length. You are also welcome to call me either in my office or at home (any time between 9 AM and 10 PM); however, be prepared to leave your name and number if I am not available to pick up the phone. Also, please check your email regularly: I will let you know by email when course notes are posted on the website and send other announcements from time to time.

Finally, a note on feedback. I would like to know how I could make this course experience as useful and interesting as possible. Therefore, every class in the end of class I will ask you to submit a sheet of paper

* This syllabus draws upon ideas presented in syllabi by a number of people, including Robert Kunovich, John Williamson, Joya Misra, and Doug Anderton.

with the date and at least one sentence of reaction to that class meeting, indicating what you learned, or something you liked or did not like, found interesting or controversial, found clear or too simplistic, or found confusing and in need of further (or better) explanation. You may also submit comments on the course in general.

COURSE REQUIREMENTS AND GRADING

All the required readings will be available on electronic reserve in the library: see <http://www.bc.edu/reserves>

The main assignment for this course will be to write a solid draft of a journal article based on panel data analyses. In preparation for that, you will submit a proposal (10% of your grade), the first draft of your data analysis (10% of your grade), the first draft of your article (10% of your grade), and the final draft of your data analysis (40% of your grade) and article (30% of your grade).

Proposal. The proposal will involve identifying a research question and conducting preliminary literature review, selecting a dataset, and identifying relevant variables. You should also identify your target journal. Please consult me early on if you need help to locate appropriate data or advice on journals. I would recommend selecting a continuous rather than categorical variable as your outcome; please consult with me if you prefer working with a categorical outcome.

Data Analysis. For the first draft of your data analysis, you will conduct data management, run all the necessary analyses, conduct diagnostics and apply remedies, and write a brief interpretation of your findings (you will also use graphs to assist your interpretation). For this component of your assignment, you will submit an annotated log that will contain the output (with your brief comments) for all of the tasks that you will perform for this assignment. You will have to assemble that log file manually (in your Word processor) and provide brief comments; make sure to paste graphs to the corresponding locations in the log. There is no page limit for your annotated log but please edit it to contain only the relevant syntax, output, and graphs (i.e., omit any unproductive steps). The main steps will be as follows:

- 1) Drop the variables you do not plan to use and recode the variables for your analysis. Make sure to keep all ID variables in the dataset and decide on your strategy to deal with the missing data.
- 2) Examine your variables using `xtsum` (or `xttab` and `xttrans` if appropriate). Use `xtreg` to examine the amount of variance in your dependent variable that is due to level 2 variation. Examine univariate normality, bivariate linearity, and univariate outliers and apply remedies if needed.
- 3) Fit an OLS regression model and examine residuals for heteroscedasticity, both graphically and using `hettest` command. Briefly describe what you see. Estimate OLS with robust standard errors and with cluster correction; compare the results.
- 4) Estimate a fixed-effects (FE) model. Estimate a between-effects (BE) model and compare the results to the FE model. Estimate a random-effects (RE) model, use `xttest0` to test the hypothesis that all country-specific residuals are zero, and `hausman` test to test that RE model is correctly specified. Based on your visual inspection of FE and BE and `hausman` test, decide whether FE or RE model is more appropriate. If RE model is not appropriate, estimate a model examining separately within and between effects in a random effects model. Estimate your final model with robust standard errors and adjusting for clustering. Generate residuals for that model (level 1 and level 2) and examine them for normality, linearity, and outliers; if necessary, modify your model. Evaluate additivity and multicollinearity as well.
- 5) Test for autocorrelation of residuals using `xtserial`, and, based on the decision made in #4, estimate either a FE or RE model allowing for the autoregressive error term (using `xtregar`). Evaluate the strength and direction of autocorrelation by examining the estimate of ρ as well as the modified Durbin-Watson and LBI statistics.
- 6) Estimate population-averaged model using `xtreg`, `pa` as well as `xtgee`. Examine correlations among residuals you obtained from the OLS model (to calculate correlations across time points, you have to

reshape wide the file containing predicted residuals). Decide what kind of covariance structure would be most appropriate and estimate the corresponding model using xtgee.

7) Use GLS models to assess whether there is heterogeneity of residuals; if necessary, introduce adjustments for autocorrelation as well.

8) Present the results of various models generated by OLS (with and without adjustments for clustering and robust SE), FE and RE (with and without adjustments for clustering and robust SE), xtreg with AR, population averaged xtreg, xtgee with the covariance structure you selected, and xtgl in one table as multiple columns (but omit those models that are clearly inappropriate) and discuss differences in findings. Identify which model or models you would choose for presentation in a journal article and discuss why. Briefly interpret the results of those models.

9) Graphically examine trajectories for your outcome variable for different countries and conclude if you think slopes vary across countries. Estimate mixed effects model allowing the slope of time to vary. Assess whether the linear trend for time is appropriate and modify your model if necessary. If no significant variance in the effect of time is discovered, remove that random component from the model. Introduce time-varying predictors, allow their slopes to vary as well and assess if there is significant variance for each. If there is significant variance in any of the slopes, create level 2 variables by generating averages of your time-variant predictors and use those to explain variance in the intercept as well as slopes (by using these level 2 variables as predictors as well as by generating cross-level interactions). Finally, reduce the number of included predictors and cross-level interactions by keeping only those that are statistically significant, and test whether such a reduction is appropriate using LR test and BIC. For your final model, calculate % variance explained at each level as well as total. Generate residuals (level 1 and 2) and assess normality, linearity, and outliers; if necessary, modify your model.

10) Present the results of the model with random slopes for time (with no other variables) as well as your final mixed model in a separate table. Make sure to present both the coefficients and the variance components (random components). Briefly interpret the results.

Article. For your article, you will select one or more models from your data analysis. Your article will include an introduction, literature review, data and methods, results, and conclusion, all written in journal format. Your introduction will provide a short substantive description of your theoretical argument and your research questions (typically about 2 pages). Your literature review will discuss any relevant theoretical literature as well as prior empirical research on this issue, and identify your main hypotheses (typically, this section should not exceed 8 pages). Your data and methods section will describe the dataset, variables, and your analytic methodology. Make sure to include a discussion of diagnostics and variable modifications in this section. Also, include a table with descriptive statistics for your sample. Next, your results section will discuss your findings. Please include tables (in journal format) and graphs assisting in the interpretation of results. Finally, please include a conclusion summarizing your findings, linking them to prior literature and to theory, and addressing study limitations and main contributions.

All components of this assignment will be submitted electronically (by email or using MyFiles). When you submit the first drafts of your analysis and article, I will provide you with detailed feedback that you will then use to prepare your final drafts.

The letter grades for the assignments will be determined as follows:

93-100	A
90-92	A-
87-89	B+
83-86	B
80-82	B-
60-79	C
0-59	F

COURSE OUTLINE

January 18. Introduction to Longitudinal Data Analysis.

January 25. Longitudinal Data Management using Stata.

Menard, Scott. 2002. Chapter 5 from: *Longitudinal Research*. Beverly Hills, CA: Sage Publications.

February 1. Missing Data in Longitudinal Research.

Twisk, Jos, and Wieke de Vente. 2002. Attrition in Longitudinal Studies: How to Deal with Missing Data. *Journal of Clinical Epidemiology* 55:329–37.

February 8. Two-Wave Panel Analysis

Taris, Toon W. 2000. Chapters 4 in: *A Primer in Longitudinal Data Analysis*. Thousand Oaks, CA: Sage Publications.

Johnson, David. 2005. “Two-Wave Panel Analysis: Comparing Statistical Methods for Studying the Effects of Transitions.” *Journal of Marriage and Family* 67(4):1061-75.

February 15. Fixed Effects Models.

Worrall, John L. 2008. An Introduction to Pooling Cross-Sectional and Time Series Data. Chapter 15 from *Handbook of Longitudinal Research: Design, Measurement, and Analysis* edited by Scott Menard. Academic Press.

Allison, Paul D. 2009. Chapter 2 and Appendix 1 (portion related to Ch. 2), from: *Fixed Effects Regression Models*. Sage Publications.

Baum, Christopher. 2006. Chapter 9, pp.219-226, from: *An Introduction to Modern Econometrics Using Stata*. College Station, TX: Stata Press.

February 22. Random Effects Models.

Rabe-Hesketh, Sophia, and Anders Skrondal. 2005. Chapter 2 from: *Multilevel and Longitudinal Modeling Using Stata*. College Station, TX: Stata Press.

Baum, Christopher. 2006. Chapter 9, pp.226-232, from: *An Introduction to Modern Econometrics Using Stata*. College Station, TX: Stata Press.

March 1. GEE models and GLS models with Complex Error Structures.

****Proposal due****

Twisk, Jos W. R. 2003. Chapters 4, 5 from: *Applied Longitudinal Data Analysis for Epidemiology: A Practical Guide*. Cambridge University Press.

Waldfoegel, Jane. 1997. The Effect of Children on Women's Wages. *American Sociological Review*, 62, 209-217.

March 8. No class: Spring Break

March 15. Mixed Effects Models.

Rabe-Hesketh, Sophia, and Anders Skrondal. 2005. Chapter 3 from: *Multilevel and Longitudinal Modeling Using Stata*. College Station, TX: Stata Press.

Hedecker, Donald. 2004. An Introduction to Growth Modeling. Chapter 12 from: David Kaplan (Ed.), *The Sage Handbook of Quantitative Methodology for the Social Sciences*. Thousand Oaks, CA: Sage Publications.

March 22. Model Building Strategies for Mixed Effects Models

Hox, Joop. 2010. Chapters 3 and 4 from *Multilevel Analysis: Techniques and Applications*. 2nd edition. Routledge.

Farkas, George, and Kurt Beron. 2004. The Detailed Age Trajectory of Oral Vocabulary Knowledge: Differences by Class And Race. *Social Science Research*, 33, 464-497.

March 29. Diagnostics for Mixed Effects Models

Snijders, Tom A. B., and Roel J. Bosker. 1999. Chapter 9 from *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. Thousand Oaks, CA: Sage.

April 5. No class

****Data analysis first draft due****

April 12. Panel Data Models for Categorical and Count Data

Snijders, Tom A. B., and Roel J. Bosker. 1999. Chapter 14 from *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. Thousand Oaks, CA: Sage.

Litwin, Kenneth J. 2004. "A Multilevel Multivariate Analysis of Factors Affecting Homicide Clearances." *Journal of Research in Crime and Delinquency*, 41, 327-351.

April 19. Sample Selection and Endogeneity Biases

Fu, Vincent Kang, Christopher Winship, and Robert D. Mare. 2009. Sample Selection Bias Models.

Chapter 18 from *Handbook of Data Analysis*, edited by Melissa A. Hardy and Alan Bryman. Sage.

Greenland, S. 2000. An Introduction to Instrumental Variables for Epidemiologists. *International Journal of Epidemiologists*, 29, 722-729.

April 26. Age, Period and Cohort Effects.

****Article first draft due****

Menard, Scott. 2002. Chapter 2 from: *Longitudinal Research*. Beverly Hills, CA: Sage Publications.

Glenn, Norval D. 2007. Age, Period and Cohort Effects. *Blackwell Encyclopedia of Sociology*, edited by George Ritzer. Blackwell.

Herbert L. Smith. 2008. Advances in Age–Period–Cohort Analysis. *Sociological Methods & Research* 36:287-296.

May 3. Dynamic Panel Data Models.

Halaby, Charles N. 2004. Panel Models in Sociological Research: Theory into Practice. *Annual Review of Sociology*, 30, 507-544.

Baum, Christopher. 2006. Chapter 9, pp.232-236, from: *An Introduction to Modern Econometrics Using Stata*. College Station, TX: Stata Press.

Finkel, Steven E. 2008. Linear Panel Analysis. Chapter 29 from *Handbook of Longitudinal Research: Design, Measurement, and Analysis* edited by Scott Menard. Academic Press.

****Final drafts of data analysis and article due May 14****