SOCIOLOGY 6707

INTERMEDIATE DATA ANALYSIS

Winter 2018

Blair Wheaton
Department of Sociology

Time: **Tuesday 2-5,** tutorial Fridays 1-2

Place: Room 240, Dept. of Sociology, 725 Spadina Avenue

Overview

This course functions as a follow-up to a first graduate level statistics course. The obvious goal is to develop the student's skills as both a producer and a consumer of data, but the less obvious, but equally important, goal is to develop an understanding of the fit between ideas and models, i.e., how ideas are expressed in models.

The course is intended not just for the specialized student, but for a range of students with different needs. Some of these needs could be:

- increased reading breadth in areas of interest, or in key areas of the discipline or the social sciences in general.
- development of data analytic skills and awareness of available choices in data analysis situations.
- comprehension and application of specific techniques to be used in ongoing and future personal research, dissertation research, and research grants.
- learning a language and a thinking framework that gives access to a wider range of the discipline and the social sciences and thus facilitates the generalization of future audiences in one's own research..

Some Themes

- ***The matching of ideas to their representation in models***. A common problem in many areas of research is the lack of fit between the ideas stated in a theory and the way the theory is tested. We emphasize the issue of fit and representation of ideas, as "embodied" in the techniques included in the course.
- ***Practice in data analysis techniques.*** The course has been designed and run for a number of years with using exercises on the computer, primarily using SAS but also other programs (such as HLM, LISREL). These exercises each involve data analysis problems that the student articulates. My role in this part of the course is to be available to students to help with execution of computer problems, and to discuss

problems in analyzing and interpreting data. I generally encourage students to help each other with the analytic phases of each project.

- ***Choosing the appropriate technique given the analytical situation and type of data.*** As the course progresses, an increasingly important issue will be choice of technique, as a function of: type of data, sample size, distributions of variables, nature of underlying concepts (continuous vs. categorical), preferred modes of interpretation, and the nature of the question.
- ***Methods and theory closely linked***. Different techniques frame specific forms of theorizing that are not possible outside of those methodological frameworks.

Topic Details

Part 1: Generalizing the Regression Model

- Interactions and their Interpretation
- Nonlinear Regression (functional forms, splines)
- Logistic Regression (binomial, multinomial, ordinal)
- Regression for Nonnormal Variables (Poisson and Negative Binomial)

We will consider the nature and interpretation of interactions (multiplicative effects of variables), nonlinearity (both functional forms and spline regression), generalizations to categorical outcomes (logistic regression), including both dichotomous and multiple category outcomes, and regression for rare and highly skewed outcomes.

Part 2: Structural Equation Models

- The Transition from Equations to Models
- Structural Equation Models: A Basic Introduction

This year we will include a very abbreviated section on structural equation models, including a section on the point of specifying models and process, and a section on the basics of structural equation modeling. This section will also discuss identification, reciprocal causality, and panel data.

Part 3: Hierarchical and Growth Curve Models.

- Basic 2-level and 3-level HLM models
- Growth Curve Models
- The Generalized HLM Model (Poisson, Logistic).

These methods address the classic theoretical problem of effects across levels of social reality, specifically targeting the effects of social context on individuals. Social context refers to the effects of any shared context with collective membership. The concept is closely related to idea that layers of social reality can be seen as "nesting units" of increasing size and complexity. Thus, you can study the effects of schools on students, of neighbourhoods on families, of family structure on children, of community on individual opportunities, of social structure on individuals, etc. We will also consider hierarchical models for discrete outcomes. The growth curve model is a direct extension of the general hierarchical linear model, used to track trajectories of change over lives as a function of time. This technique is especially useful for specifying sources of disparities or inequalities in developmental, social, or other life outcomes over time, with an emphasis on the timing in the life course of the appearance of disparities.

Part 4: Combined Cross-Section/Time Series Analysis: Fixed and Random Effects Regression for Panel Data

- Fixed and Random Effects Regression for Panel Data
- Fixed Effects Models for Other Techniques in the Course.

This section considers cross-section / time series models, with an emphasis on fixed-effects and random-effects models and what they do and do not accomplish. Fixed effects models are emphasized, primarily because they claim to take into account a broad class of unmeasured variables left out of the regression which may overlap with the effects of the independent variables in the equation. Fixed effects stand for stable individual differences in all forms , e.g., biological givens, ascriptive social statuses, and family background. We conclude this section by applying the fixed effects model to techniques discussed earlier in the course, including structural equation models, logistic regression, and Poisson models.

Part 5: Event History/ Survival Models

- The Discrete-Time Event History Model.

The course will consider cases where event history models should be used rather than logistic regression, including the many situations where the timing of an event is as important as its occurrence. Often we study events (marriage, promotion, entry into the labour force, childbearing, etc.) which occur at different times for different people. The event history model takes into account both the occurrence of an event and its timing, while logistic regression can only study the occurrence of the event. We will only consider basic discrete-time models this year, but these models are very flexible.

Prerequisites

The course assumes you know the basics of linear regression, including multiple regression. There will be a voluntary review class for basic regression held in the first week or two. Some of the tutorials on Fridays will provide examples of software used in the course.

Required Work

I usually require three assignments, each involving at least some computer work in SAS or HLM. Both the TA(s) and I are available throughout assignment work to answer questions about computer issues and the interpretation of assignment questions. This year the third assignment will be optional, vs. writing a final exam.

The first assignment is standard for everyone with a fixed due dates, assignment two will allow for variable due dates, and the third assignment will be due the last day of the last week of class.

The first assignment considers generalizations of the regression model discussed in the first three weeks; the last two assignments ask you to choose ***two of four*** possible questions on structural equation models, hierarchical models, panel regression, and event history models.

I expect students to work in groups, formed voluntarily and by mutual agreement among students. This is encouraged for three reasons: 1) to distribute the workload; 2) to encourage collective learning and communication of skills and knowledge among students; and 3) to avoid isolating students with specific computer problems. *All grades from these assignments are assigned equally to students within groups.* Groups ***must*** be from 2 to 3 in size.

Finally, there is a term test and an optional "final", scheduled the week after class ends, not in the last class. ***You must select doing either the final or the third assignment.*** For those with mainly theoretical interests in the material, the final may be preferred. Because of this option, you will be allowed to do the last assignment on your own if that is preferable. But this is not required.

The first test is designed to provide a review of notes at a crucial point in the course. I have found in the past that students gain in their ability to understand material in the later phases of the course due to this review of material and consolidation of their knowledge in the middle of the course. All tests are “open-book”, i.e., my notes are allowed. ***The final is non-cumulative***. Students will have the choice of two of four questions on the final.

Weights for Required Work

Work	Weight
1. Assignment #1	25%
2. Term Test	25%
3. Assignment #2	25%
4. Assignment #3 or Final	25%

Data Used for Assignments

This course is an overview of a series of techniques. For reasons discussed below, I re quire that everyone use the same class data for assignments. This requirement is based on my experience with more flexible approaches I used in the past, which led to a number of problems. The most important computer issue in a course such as this is *not* running a procedure – it is manipulating the data. This ***is*** a course in data analysis. As a result, I make your coding part of the issue in grading assignments, because this is where so many of the problems in producing credible findings occur.

I require using class data for these reasons: 1) I need to understand myself the structure of the data, the nature of the sample, and whether variables exist that conform to what you want to do, so that I can give advice during assignments; 2) there are few data sets that can be used for *all* of the techniques in this course, and we do not want to change data sets across assignments (especially because of HLM, fixed effects, and event history); 3) data has to be at least three-wave panel data, and sufficiently clustered to allow for hierarchical models.

I have permission to attach neighborhood data to the 3-wave National Survey of Families and Households in the U.S., and these data have extensive life history information. I am exploring the possibility of using two other data sets instead: the Panel Study of Income Dynamics (PSID) and its off-shoot studies, the National Longitudinal Survey of Youth, 1979 and 1997 (NLSY-79 and NLSY-97), or the The National Longitudinal Study of Adolescent to Adult Health (ADD Health). This last study is a study of adolescents at 12-14, followed through to 28 years old, and *does* have some publicly-accessible contextual data.

Reading

There is *no* required reading beyond the set of notes I have developed specifically for this course. The material is intended to be a relatively friendly but rigorous discussion of each technique or type of analysis. The notes are sold on an individual basis, and will be available at Three Cent Copy across the street.

I also hand out a reference list for each topic in the course, printed in a separate document. This list can be downloaded from Blackboard, or you can get a copy in class. Under each topic of the reference list, I will include both basic introductions and more complete overviews of each technique, as well as a list of recent substantive applications of the technique in journals.

Class Schedule

The attached schedule shows the topics covered class-by-class, as well as due dates for all required work.

Web Sites with Basic Mathematical and Statistical Help.

I also strongly encourage use of online sources for learning SAS. The UCLA site for SAS is one of the best and publicly accessible here:

http://www.ats.ucla.edu/stat/sas/

Or on You Tube, an introduction to SAS done by SAS itself:

https://www.youtube.com/watch?v=r1Yy_sYbfy0

Or the introductory online course run by Boston University:

https://support.sas.com/edu/schedules.html?ctry=us&crs=PROG1#s1=1

I will also upload to Blackboard free programming SAS guides you can also download online. These include the "famous" Little SAS Book, a general programming guide:

http://www.dermepi.eu/wp-content/uploads/2017/04/Little.SAS_.Book_.A_Primer.Third_.Edition.pdf

The Handbook of Statistical Analysis Using SAS:

http://fidy.andrianasy.free.fr/SAS%20Books/++!++%20A%20Handbook%20Of%20Statistical%20Analyses%20Using%20SAS.pdf

and the SAS programming skills website at Northwestern:

http://www.kellogg.northwestern.edu/researchcomputing/docs/SAS_Programming_Skills.pdf

PLEASE NOTE: I do require assignments in SAS so that I can give advice and grade them properly. If you want to use STATA for assignment #1, please see me. I do not encourage it, because later assignments will have to be in SAS.

See below for a more complete list of web sites that provide help for basic math concepts and some of the statistical techniques discussed in this course.

Web Sites with Basic Mathematical and Statistical Help.	
Algebrahelp.com	http://www.algebrahelp.com/index.jsp
Derivatives Defined	http://web.mit.edu/wwmath/calculus/ispath/unit02.html
Internet Resources for Math	http://www.langara.bc.ca/mathstats/resource/onWeb/
Linear Algebra Calculator	http://www.compute.uwlax.edu/lin_alg/
Logarithms Definition	http://www.purplemath.com/modules/logs.htm
Logarithms Rules	http://www.purplemath.com/modules/logrules.htm
Arizona Mathematical	http://math.arizona.edu/~www_main_2002/software/azmath.html

Software	
Probability and Statistics	http://www.ability.org.uk/probstat.html
S.O.S. Math	http://www.sosmath.com/
Derivatives: Rules and Examples	http://people.hofstra.edu/Stefan_Waner/RealWorld/tccalcp.html
Online Statistical Test	http://www.stat.ucla.edu/~dinov/courses_students.dir/Applets.dir/Normal_T_Chi2_F_Tables.htm

January 2018

Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
	1	2	3	4	5	6
7	8	9 Introduction Interactions	10	11	12 Interactions	13
14	15	16 Nonlinear Regression	17	18	19 Intro to SAS	20
21	22	23 Program Example Logistic Regression	24	25	26 Poisson Regression	27
28	29	30 Program Example SEM 1: Equations to Models	31			

February 2018

Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
				1	2	3
					SEM 2: Basic SEM	
					Exercise 1: 1st Question Due	
4	5	6	7	8	9	10
		SEM 3: Basic SEM, Fitting and testing models			Test Review	
					Exercise 1: 2nd Question Due	
11	12	13	14	15	16	17
		HLM 1: Introduction			Term Test (1.5 hours)	
18	19	20	21	22	23	24
	Reading Week	Reading Week	Reading Week	Reading Week	Reading Week	
25	26	27	28			
		HLM 2: Examples and Interpretation				

March 2018

Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
				1	2 **HLM 3: Generalized HLM**	3
4	5	6 **Growth Curve Models**	7	8	9 **Growth Curve Examples**	10
11	12	13 **Panel Regression: Intro**	14	15	16 **Fixed Effects Panel Models**	17
					Exercise 2 Due: SEM or HLM	
18	19	20 **Fixed Effects in SEM**	21	22	23 **Event History Introduction**	24
25	26	27 **Event History Example**	28	29	30 **Event History Continuous Time**	31

April 2018

Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
1	2	3 **Event History Final Review**	4	5 **Final Exam (3 hours)**	6	7
				Exercise 3 Due: Fixed Effects or Event HIstory		
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30					