SOCIOLOGY 6707

INTERMEDIATE DATA ANALYSIS

Winter 2022

Blair Wheaton
Department of Sociology

Time: **T**uesday 2-5***, or*** Tuesday 2-4 and Friday 2-3

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

TAs: **Rachel Meiorin (rachel.meiorin@mail.utoronto.ca)**
TBA

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 quantitative findings and 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, including:

- 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..

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. 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: An Introduction
- Structural Equation Models: Testing and Fitting
- Cross-Group Structural Equation Models

We will include a 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. We will also consider the testing and fitting of SEM models, and the comparison of models across groups.

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 here 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.
- Grouped Continuous Time Models

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 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. All of the necessary tools for programming are taught in tutorials, using template programs.

Required Work

There is a mid-term test in class, discussed below. Beyond that, you will do three assignments, coupled with a final test where you will have choice to do **two** of four questions.

Both the TA's and I are available throughout assignment work to answer questions about computer issues and the interpretation of assignment questions. *Both the second and third assignments involve choice of a specific technique from a list of topics* (structural equation models, hierarchical models, fixed effects 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.

There will be a final test one week after the last class in the same time period. On this test, you will do the two questions ***not*** on the topics chosen for assignments 2 and 3.

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 text is allowed, and so are your notes, and assignments.

Weights for Required Work

Work	Weights
1. Assignment #1	20%
2. Term Test	20%
3. Assignment #2	20%
4. Assignment #3	20%
5. Final Test (in class)	20%

Data Used for Assignments

This course is an overview of a series of techniques. *I require 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 (although it may be necessary for HLM); 3) in general, data has to be at least three-wave panel data, or sufficiently clustered to allow for hierarchical models.

We will be using the NLSY79, the National Longitudinal Survey of Youth, 1979, and and/or the NLSY79 Children and Young Adults, for assignments, except for HLM. The NLSY79 is one of the most widely used data sets in the world: it is a continuing longitudinal study of youth 14-22 in 1979 in the United States. There were a total of 12, 686 individuals in the original sample. A sub-sample of individuals was followed every year until 1994, and then every two years after that, until 2018. The age of the sample at that time varied from 49 to 58 years old. You will be able to set up your own account on the NLSY website and download data directly. The NLSY79 Child/YA data set is a separate study of the children of the NLSY79 cohort. There are 11,521 children that have been followed since birth.with the oldest now reaching 45 by 2018. Both data sets can be reviewed at the NLSY web site, including the design, topics, measures, and codebooks. You can download the data directly from this site.

We will also have access to specially extracted and anonymized data from the Toronto Study of Neighbourhoods and Well-Being (2011) for the HLM assignment. If you do this assignment, you must sign a release form promising not to use your results for public presentation in any form or for publication. In fact, you can apply for that approval after the course, if results look promising.

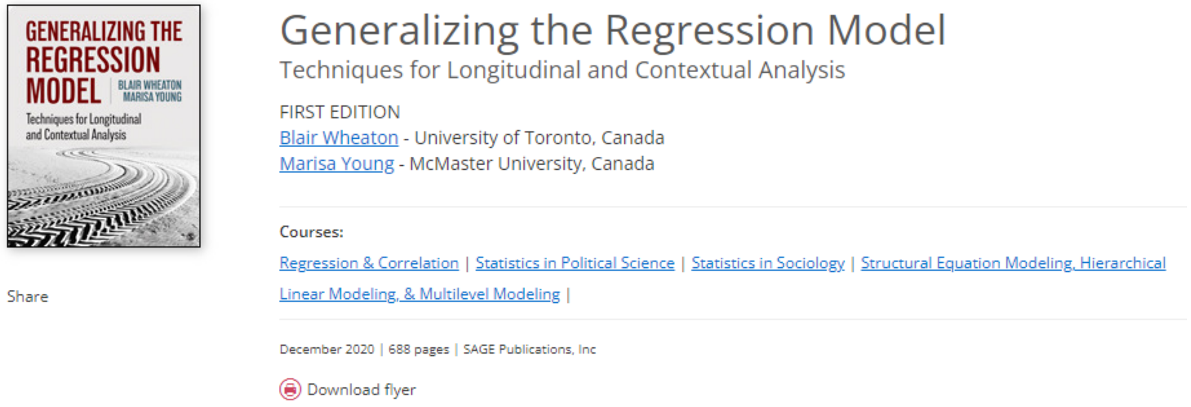
Software

My approach to software this year is to allow students to use SAS, which I teach as part of the class, STATA, if you already have it and know how to use it, or R. TA resources will be available for SAS and STATA especially, but also R if possible.

I have an license for SAS that allows me to freely distribute it to students. You can use the full version in this class, and keep it for the future. But you can also register online for SAS OnDemand, which is free to all students, and allows a centralized access to the data and all procedures. We will show you how to use SAS OnDemand as part of the class. The advantage of SAS OnDemand is that it runs online, via servers in North Carolina, and all types of PC's and Mac's can access it.

Reading

There is *no* required reading beyond the textbook which is assigned in this course. I will also give you links to template programs, practice question solutions, and annotated slides. I will also post my "Basic Math" slides used for boot camp each year.


GENERALIZING THE REGRESSION MODEL
BLAIR WHEATON
MARISA YOUNG
Techniques for Longitudinal and Contextual Analysis
Share

Generalizing the Regression Model

Techniques for Longitudinal and Contextual Analysis

FIRST EDITION

Blair Wheaton - University of Toronto, Canada

Marisa Young - McMaster University, Canada

Courses:

Regression & Correlation | Statistics in Political Science | Statistics in Sociology | Structural Equation Modeling, Hierarchical Linear Modeling, & Multilevel Modeling |

December 2020 | 688 pages | SAGE Publications, Inc

Download flyer

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

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.

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 Software	http://math.arizona.edu/~www_main_2002/software/azmath.html
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

The following parts of the syllabus are a required template, but also express important principles, procedures, and values.

Penalty for Lateness Clause

For both undergraduate and graduate courses, instructors are not obliged to accept late work, except where there are legitimate, documented reasons beyond a student's control. In such cases, a late penalty is normally not appropriate.

In this course, assignments are only accepted up to two days beyond the due date, and at a 10% discount that applies to the entire group working on the assignment. Late assignments have historically been a rarity in this course for this reason.

Academic Integrity Clause

Copying, plagiarizing, falsifying medical certificates, or other forms of academic misconduct will not be tolerated. Any student caught engaging in such activities will be referred to the Dean's office for adjudication. Any student abetting or otherwise assisting in such misconduct will also be subject to academic penalties. Students are expected to cite sources in all written work and presentations. See this link for tips for how to use sources well: (http://www.writing.utoronto.ca/advice/using-sources/how-not-to-plagiarize).

According to Section B.I.1.(e) of the Code of Behaviour on Academic Matters it is an offence *"to submit, without the knowledge and approval of the instructor to whom it is submitted, any academic work for which credit has previously been obtained or is being sought in another course or program of study in the University or elsewhere."*

By enrolling in this course, you agree to abide by the university's rules regarding academic conduct, as outlined in the Calendar. You are expected to be familiar with the *Code of Behaviour on Academic Matters* (http://www.artsci.utoronto.ca/osai/The-rules/code/the-code-of-behaviour-on-academic-matters) and *Code of Student Conduct* *(*http://www.viceprovoststudents.utoronto.ca/publicationsandpolicies/codeofstudentconduct.htm) which spell out your rights, your duties and provide all the details on grading regulations and academic offences at the University of Toronto.

Accessibility Services

It is the University of Toronto's goal to create a community that is inclusive of all persons and treats all members of the community in an equitable manner. In creating such a community, the University aims to foster a climate of understanding and mutual respect for the dignity and worth of all persons. Please see the University of Toronto Governing Council "Statement of Commitment Regarding Persons with Disabilities" at http://www.governingcouncil.utoronto.ca/Assets/Governing+Council+Digital+Assets/Policies/PDF/ppnov012004.pdf.

In working toward this goal, the University will strive to provide support for, and facilitate the accommodation of individuals with disabilities so that all may share the same level of access to opportunities, participate in the full range of activities that the University offers, and achieve their full potential as members of the University community. We take seriously our obligation to make this course as welcoming and accessible as feasible for students with diverse needs. We also understand that disabilities can change over time and will do our best to accommodate you.

Students seeking support must have an intake interview with a disability advisor to discuss their individual needs. In many instances it is easier to arrange certain accommodations with more advance notice, so we strongly encourage you to act as quickly as possible. To schedule a registration appointment with a disability advisor, please visit Accessibility Services at http://www.studentlife.utoronto.ca/as, call at 416-978-8060, or email at: accessibility.services@utoronto.ca. The office is located at 455 Spadina Avenue, 4th Floor, Suite 400.

Additional student resources for distressed or emergency situations can be located at distressedstudent.utoronto.ca; Health & Wellness Centre, 416-978-8030, http://www.studentlife.utoronto.ca/hwc, or Student Crisis Response, 416-946-7111.

Equity and Diversity

The University of Toronto is committed to equity and respect for diversity. All members of the learning environment in this course should strive to create an atmosphere of mutual respect. As a course instructor, I will neither condone nor tolerate behaviour that undermines the dignity or self-esteem of any individual in this course and wish to be alerted to any attempt to create an intimidating or hostile environment. It is our collective responsibility to create a space that is inclusive and welcomes discussion. Discrimination, harassment and hate speech will not be tolerated.

Additional information and reports on Equity and Diversity at the University of Toronto is available at http://equity.hrandequity.utoronto.ca.

JAN 2023

SUN	MON	TUE	WED	THU	FRI	SAT
01	02	03	04	05	06	07
08	09	10 Introduction Interactions in Regression	11	12	13 NO CLASS	14
15	16	17 Nonlinear Regression; *Downloading NLSY data*	18 Review: Basic Regression	19	20 Intro to SAS	21
22	23	24 Logistic Regression: Binomial, Multinomial Continue SAS	25	26	27 NO CLASS	28
29	30	31 Logistic Regression: Ordinal			GLM: Poisson and Negative Binomial Regression	

JAN 2023

SUN	MON	TUE	WED	THU	FRI	SAT
01	02	03	04	05	06	07
08	09	10 Introduction Interactions in Regression	11	12	13 NO CLASS	14
15	16	17 Nonlinear Regression; *Downloading NLSY data*	18 Review: Basic Regression	19	20 Intro to SAS	21
22	23	24 Logistic Regression: Binomial, Multinomial Continue SAS	25	26	27 NO CLASS	28
29	30	31 Logistic Regression: Ordinal				

FEB2023

SUN	MON	TUE	WED	THU	FRI	SAT
			01	02	03 GLM: Poisson and Negative Binomial Regression	04
05	06	07 SEM 1: Equation to Models	08	09	10 SEM1: Basic SEM	11
12	13	14 SEM 3: Fitting and Testing Models	15	16	17 Finish SEM Test Review Exercise 1 Due	18
19	20 Reading Week	21 Reading Week	22 Reading Week	23 Reading Week	24 Reading Week	25
26	27	28 HLM 1: Introduction				

MAR2023

SUN	MON	TUE	WED	THU	FRI	SAT
			01	02	03 Term Test: 1.5 hours	04
05	06	07 HLM 2: Examples	08	09	10 HLM 3: Generalized HLM	11
12	13	14 Generalized HLM Growth Curves	15	16	17 Growth Curve Example Intro to Panel Models	18
19	20	21 Panel Regression	22	23	24 Variations of Fixed Effects Models Exercise 2 Due	25
26	27	28 Event History Intro	29	30	31	

APR2023

SUN	MON	TUE	WED	THU	FRI	SAT
						01
02	03	04 Event History Example	05	06	07	08
09	10	11 Final In-Class Test	12	13	14	15
16	17	18	19	20	21 Exercise 3 Due	22
23	24	25	26	27	28	29
30						

READING SCHEDULE

Date	Reading in Wheaton and Young
Tuesday, Jan. 10	Chapter 2
Tuesday, Jan. 17	Chapter 3
Tuesday, Jan. 24	Sections 4.1 – 4.5
Tuesday, Jan 31	Section 4.6-4.9
Friday, Feb. 3	Chapter 5
Tuesday, February 7	Chapter 6
Friday, February 10	Chapter 7
Tuesday, February 14	Chapter 8
Tuesday, February 28	Chapter 10
Friday, March 10	Chapter 11
Tuesday, March 14	Chapter 12
Tuesday, March 21	Chapter 13
Friday, March 24	Sections 14.1-14.4
Tuesday, March 28	Chapter 15