SOCIOL 401-2: STATISTICAL ANALYSIS OF SOCIAL DATA, II
SPRING 2013

Class: Monday/Wednesday, 12:30-1:50pm, Parkes Hall 222
Lab: 1-1:50 in Library B183

Professor: Jeremy Freese
1810 Chicago Avenue, Rm 211
e-mail: *jfreese@northwestern.edu*
Office hours: Thurs, 2-3:30 (and by appt)

TA: Justin Louie
e-mail: *justin-louie@kellogg.northwestern.edu*
Office hours: Tues, 2-3 (and by appt.)

OVERVIEW

We will continue the project begun in your earlier statistics courses: developing your ability to draw substantively meaningful and accurate inferences from quantitative social data, as well as to evaluate quantitative evidence presented by others. We will proceed by considering a variety of extensions to the familiar linear regression model. These models are frequently used in quantitative social science and so an understanding of them is valuable in its own right. More importantly, however, our close and practical consideration of modeling strategies for these outcomes is intended to help cultivate an understanding of fundamental principles of statistical inference, data analysis, and modeling that extend beyond any specific set of techniques.

COURSE GOALS

If you put in a fair effort, your instructor and TA will endeavor mightily to advance your training on all of the following fronts by the end of the course:

1. Your understanding of the basic logic, broad flexibility, elegant beauty of maximum likelihood estimation.

2. Your familiarity with a variety of models that are used frequently in contemporary social science.

3. Your knowledge of the essential importance of interpretation and presentation in data analysis.

4. Your appreciation of both the work and fun of applied data analysis.

5. Your cognizance of the gap between "statistics" and "causal inference." This will include understanding the limitations of covariate adjustment of observational data as a strategy for causal inference.

6. Your ability to use statistical software—specifically, Stata—to analyze quantitative data to answer social research questions.

And yet: Sociology's required statistics sequence, of which this is the final course, does not suffice to impart the requisite skills for a social science research career that is based primarily on analysis of quantitative training. For that, you will need additional training while in graduate school, more first-hand experience with the craft of data analysis than what you will get from this sequence, and commitment to staying fresh with training over your career. Methodological competence, much less expertise, is an ongoing project for as long as you are engaged in research.

Also: For students who do not intend to engage in quantitative research after the conclusion of this sequence, there is sometimes the idea that data analysis courses might be directed more toward a capacity for literate consumption of quantitative research, apart from the craft of knowing how to do it. *"Please. I just want to be able to read quantitative articles and feel like I understand what's going on."* Of course the cultivation of such a capacity is all to the good, and we will take up the evaluation of quantitative evidence. But, software mastery aside, your instructor has never been sure if there really exists a competence base for quantitative-research-consumption that is separate from the competence base for quantitative-research-production. The analogy would be to the idea of foreign language instruction that is directed only to being able to read the language and not to being able to write it. In addition, evaluation of quantitative evidence often turns on key issues—surrounding measurement, sampling, and general logic of causal inference—that will receive mention in the course but not with anything like the thoroughness they deserve (only so much can be done in a quarter).

PREREQUISITES

This course follows Sociology 400 and Sociology 401-1. Accordingly, I presume familiarity with linear algebra, with the material covered in a standard introduction to social science statistics for undergraduates, and with the basics of linear regression. All computer work in the course will be conducted using Stata. Some work in the course requires add-on packages to Stata to be used; more details about this will be provided.

READINGS

Blackboard will be used as a repository for this syllabus, any updates to it, lecture slides, and readings. Please notify me by e-mail of any technical or other problems with materials provided via Blackboard.

The readings for this class are in transition. Lecture slides will be brought as handouts to class and posted on Blackboard. These will be augmented in some cases by an extended set of lecture notes.

There is also a book that has been associated with this course, for which I am a co-author.[1] The book is currently under revision to take advantage of current Stata features. Some

[1] Long, J. Scott and Jeremy Freese. 2006. *Regression Models for Categorical Dependent Variables Using Stata, Second Edition*. College Station, TX: Stata Press.

chapters from the new version will be provided, and I will have to figure out what to do about the others.

Other readings for the course are available in Blackboard. Sources used for additional course reading include:

Long, J. Scott. 1997. *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: Sage.

Morgan, Stephen L., and Christopher Winship. 2007. *Counterfactuals and causal inference: Methods and principles for social research*. Cambridge, UK: Cambridge University Press.

Both these books have much to commend them and are readily available from online booksellers.

WHAT YOU WILL BE DOING

Exercises. We will have regular assignments corresponding to the different units of the course that focus on the appropriate estimation of models and interpretation of results. You will have at least one week to complete any exercise. Exercises will be worth a particular number of points.

Examination. We are going to have an old-school, closed-book final examination at the conclusion of the course. *(I know, I know.)* The time of the exam is to be determined. The purpose of the examination is to encourage more active learning over the course of the quarter and provide an occasion to re-contemplate it all at the end. While I do not anticipate this being a problem for anyone who puts forth good effort, you must receive a B or above on the examination in order to obtain a B or above as your final grade for the course.

Weekly check-in. You are required to e-mail your instructor (*jfreese@northwestern.edu*) one comment, question, or suggestion regarding the course each week. *Check-ins are optional—though never unwelcome—in any week in which the course has only one non-lab meeting.* This can be as brief as a single sentence and otherwise as long as you like. The point is to keep us in dialogue over the quarter. Send the e-mail by the end of Friday each week.[2]

ADDITIONAL GUIDELINES FOR ASSIGNMENTS

Unique research. You are encouraged to discuss your work with your fellow students and to learn from them, but you must complete your work on your own. For those components

[2] Q: Does Jeremy actually care if I do this?
A: *Yes! He does!*

of assignments that involve estimation and interpretation of data, you must not be using the same (or virtually the same) variables as another student.

Including output. For assignments involving the estimation and interpretation of data in Stata, as well as your final paper, you will turn in Stata output along with your assignment. You need only turn in output from commands involving transformation of variables and estimation and post-estimation commands for models that provide part of the answer to parts of the assignment. But:

(a) Everything that you turn in that involves data analysis, including any number in your final paper, **must be generated from a .do file** that is submitted with the exercise. The do file must be sufficient to reproduce all the submitted numbers, from a data file which may be requested by the instructor. (Stata has an interactive mode that is great for doing exploratory work quickly, but serious data analysis requires reproducibility, and that requires working from do files.)

(b) As part of the *.do* files you use for generating results for assignments, **you must use comments** (i.e., using lines preceded by * or //) that indicate what part of the output corresponds to what.

(c) You must, in some manner, emphasize/highlight numbers in your output that correspond to numbers in your assignment.

(d) **`You must use a fixed-width font`** (like Courier or Andale Mono) and your lines must not wrap. To have lines that do not wrap, use a sufficiently-small-but-still-readable font and/or use the set linesize command in Stata.

(e) Every binary variable that you turn in for any work in this course **must** be renamed and recoded such that the values of the variable are 0/1, and the name of the variable carries mnemonic significance consistent with 1=yes and 0=no.[3]

GRADING

Two-thirds of your final grade for the course will be based on the exercise, and one-third on the final exam. The weekly check-ins are required and any failures to submit them will result in deduction from this baseline grade. I presume adequate and professional participation, etc., from everyone, and while Northwestern students have so far never posed any problem in this respect—*and I hope to God you are no different*—but theoretically if any problems did arise, they would also be handled by deduction from the baseline grade.

[3] Your instructor generally has a life philosophy that punishments should be roughly proportional to the magnitude of offenses. With respect to this particular matter, however, all bets are off. His belief is that how you handle binary variables has emblematic significance for how you approach data analysis, and, specifically, whether you approach data analysis in a way that minimizes mistakes and confusion for yourself and others. Consider it akin to a "broken windows" theory of orderly data analysis.

SCHEDULE OF TOPICS AND READINGS

The schedule below should be understood as tentative and will be adjusted according to the pace of our progress through course materials.

Wk	Date		Topic	Reading
1	T	4/2	Orientation	Syllabus
	W	4/3	Linear regression model	Long, Chapter 2
2	M	4/8	Instrumental variables and regression discontinuity	Morgan and Winship, Chapters 7, 9.1 and 9.2; Angrist & Pischke
	W	4/10	(no class)	
3	M	4/15	Maximum likelihood	
	W	4/17	Censored/interval outcomes	Long, Chapter 7
4	M	4/22	Binary outcomes	Long and Freese, Chapter 3; *Application:* Hout and Fischer
	W	4/24	Binary outcomes	
5	M	4/29	Evaluating fit and hypothesis testing for ML models	Long and Freese, Chapter 4
	W	5/1	Propensity score models	Morgan and Winship Ch2 & 4; *Application:* Harding
6	M	5/6	Event outcomes	TBA; *Application:* Haveman et al.
	W	5/8	Event outcomes	
7	M	5/13	Ordered outcomes	Long and Freese, Chapter 5 *Application:* Vaisey
	W	5/15	Ordered outcomes	
8	M	5/20	Nominal outcomes	Long and Freese, Chapter 6 *Application:* Loveman and Muniz
	W	5/22	Nominal outcomes	
9	M	5/27	(no class - holiday)	
	W	5/29	Count outcomes	Long and Freese, Chapter 8; *Application:* Messner
10	M	6/3	Count outcomes	
	W	6/5	(spillover)	