

**SOCIOL 401-2: STATISTICAL ANALYSIS OF SOCIAL DATA (II)**  
**SPRING 2009**

Class: Tuesday/Thursday, 9:30-10:45pm, University Library 3622  
Lab: Monday, 4-5pm

Professor: Jeremy Freese  
1810 Chicago Avenue, Rm 211  
e-mail: [jfreese@northwestern.edu](mailto:jfreese@northwestern.edu)  
Office Hours: Thurs, 11-12:30 (and by appt)

TA: Ben Ruiz  
1810 Chicago Avenue, Rm 108  
e-mail: [b-ruiz@northwestern.edu](mailto:b-ruiz@northwestern.edu)  
Office hours: Tues, 11:30-1:30

**OVERVIEW**

This course will continue the project begun in your earlier statistics courses: developing your ability to draw substantively meaningful and accurate inferences from quantitative social data. We will proceed by considering the extension of the familiar linear regression model to several different classes of outcomes: censored, binary, ordinal, nominal, and count. 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 these models is intended to help cultivate an understanding of fundamental principles of statistical inference that extend beyond any specific set of models.

**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**

The most important readings for the course are the lecture notes I provide on Blackboard. Except for the possibility of addenda in response to issues raised in class, I will provide the long version of notes at least three days prior to the lecture in which they will be discussed. You are expected to have reviewed these prior to the pertinent lecture and to bring them to class. In all, the handouts will be voluminous and you should get yourself a binder and three-hole punch (or whatever works for you) now. The lecture slides will often be brought as handouts to class and might not be posted until afterwards.

The book for the course is:

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

As the name of the second author of the above might suggest, the book has nontrivial overlap with the lecture notes that will be provided, although the book focuses more on the practice of the analysis in Stata.

Other readings for the course will be made 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.

## **DELIVERABLES**

*Final paper.* You will write a final paper for this that makes prominent use of one of the models for binary, ordered, nominal, or count data considered in class. The paper should concisely set up a research question, describe the data and analytic strategy that will be used to answer the question, present and interpret the results, and briefly conclude. The paper will be due **June 8<sup>th</sup>** at 11am. *Incompletes can be noxious to the careers and mental health of graduate students and so will only be given in unusual circumstances.*

*Homework assignments.* There will be regular assignments corresponding to the different units of the course that focus on the appropriate estimation of models and interpretation of results.

*Weekly check-in.* You are required to e-mail me ([jfreese@northwestern.edu](mailto:jfreese@northwestern.edu)) one comment, question, or suggestion regarding the course each week. 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 (11:59:59.9 PM) each week.<sup>1</sup>

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

---

<sup>1</sup> Q: Does Jeremy actually care if I do this?  
A: Yes! He does!

*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) You must highlight numbers in your output that correspond to numbers in your assignment.
- (b) **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, either use a sufficiently-small-but-still-readable font.
- (c) 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.

## **GRADING**

Two-thirds of your final grade for the course will be based on the assignments, and one-third on the final paper. 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 any problems there will also be handled by deduction from the baseline grade.

## **COMMUNICATION**

Students are presumed to be members of this century and therefore to check e-mail regularly. We will use e-mail to send announcements to the class as needed. *I strongly prefer e-mail to the telephone as a means of contact regarding the course.*

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

## SCHEDULE OF TOPICS AND READINGS

The schedule 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 3/31	Orientation	Syllabus
	R 4/2	Linear regression model	Long, Chapter 2
2	T 4/7	Maximum likelihood	
	R 4/9	Censored outcomes	Long, Chapter 7
3	T 4/14	Censored outcomes	
	R 4/16	Binary outcomes	Long and Freese, Chapter 3-4
4	T 4/21	Binary outcomes	
	R 4/23	Binary outcomes	
5	T 4/28	Binary outcomes	
	R 4/30	Propensity score models	Morgan and Winship reading
6	T 5/5	TBA	
	R 5/7	Ordered outcomes	Long and Freese, Chapter 5
7	T 5/12	Ordered outcomes	
	R 5/14	Ordered outcomes	
8	T 5/19	Nominal outcomes	Long and Freese, Chapter 6
	R 5/21	Nominal outcomes	
9	T 5/26	Nominal outcomes	Long and Freese, Chapter 7
	R 5/28	Nominal outcomes	
10	T 6/1	Count outcomes	Long and Freese, Chapter 8
	R 6/3	Count outcomes	