

Course Syllabus for Sociology 73994
Categorical Data Analysis
Fall 2011

Instructor Richard Williams
741 Flanner (574-631-6668)
Email: Richard.A.Williams.5@ND.EDU
Office Hours: 1:45-2:30 MW and by appointment.
Immediately after class is also good.
Skype: rw120555. Video appts are welcome but should be scheduled in advance.
Personal Web Page: <http://www.nd.edu/~rwilliam/>

Time and Place MW 11:45-1:00, Main 404

Course Web Page

<http://www.nd.edu/~rwilliam/xsoc73994/index.html>

Notes, readings, etc. will be placed on the course web page. Some of the readings will be scanned so you'll probably want to be working from a high-speed connection when downloading and printing them. I sometimes have trouble printing scanned material, so let me know if you have any problems.

Overview. This course discusses methods and models for the analysis of categorical dependent variables and their applications in social science research. Researchers are often interested in the determinants of categorical outcomes. For example, such outcomes might be binary (lives/dies), ordinal (very likely/ somewhat likely/ not likely), nominal (taking the bus, car, or train to work) or count (the number of times something has happened, such as the number of articles written). When dependent variables are categorical rather than continuous, conventional OLS regression techniques are not appropriate. This course therefore discusses the wide array of methods that are available for examining categorical outcomes. As we will see, many of these are special types of *generalized linear models*.

Heavy use will be made of Stata. You are welcome to use other programs like SAS but my ability to help you will be greatly limited. If you aren't familiar with Stata, don't worry; the text provides an excellent discussion and I have various handouts to help you.

While underlying theory will be discussed, the greatest emphasis will be on application and interpretation of models and results. Course requirements will include writing a quantitative paper using one or more of the methods discussed. Sociology 63992 and 63993 (Graduate Statistics I & II) or their equivalents are prerequisites for the course. Students from outside of Sociology are welcome.

You can get a pretty good feel for what the course will be like by looking at the book review I wrote for the main text we will be using, [Regression Models for Categorical Dependent Variables Using Stata, Second Edition](#), by J. Scott Long and Jeremy Freese. The description of

the book states “This book discusses how to fit and interpret regression models for categorical data with Stata and includes some commands written by the authors. Hypothesis testing and goodness-of-fit statistics are also discussed...This book is ideal for students or applied researchers who want to know how to fit this type of model and understand its output.”

For the most part, in the early part of the course I plan to work our way through the book. However, I will also often provide supplemental information and (especially later in the semester) I will cover additional topics.

Required Readings

Regression Models for Categorical Dependent Variables Using Stata, Second Edition, by J. Scott Long and Jeremy Freese (NOT available in bookstore; I understand it is cheaper if you order direct from Stata Press at <http://www.stata.com/bookstore/regmodcdvs.html>).

Online readings packet (compiled by Richard Williams)

The Long and Freese book really really really is required. I’ll expect you to have read the required chapters before class and will often ask you questions about it.

The book often assumes background knowledge that you don’t necessarily have. Also, I will be covering some topics that are not in the book. Therefore, there will also be several other required or recommended readings that I will make available on the web and/or distribute in class.

Those who want a more advanced treatment are encouraged to read Regression Models for Categorical and Limited Dependent Variables, also by J. Scott Long. Another good advanced book is Statistical Methods for Categorical Data Analysis, by Daniel A. Powers and Yu Xie.

Grading. Student performance will be evaluated in the following ways.

- Empirical research paper (90%). You are to use one or more of the methods we go over in this class (or a related relevant technique if approved by me). *You are required to use the material covered in adjusted predictions/marginal effects and/or one of the advanced methods covered after the first few weeks.* I have seen a lot of papers in recent years that, while good, did not really go that much beyond the methods covered in Stats II, so this year I am requiring that papers employ some of the new techniques covered in this course. Start thinking about this soon. Classes will occasionally be devoted to discussing the current status of your project, and the last few classes will be used to present your papers. The presentation of the paper will be considered in this part of the grade. *The paper is due Friday, December 9.*
- Homework/discussion questions/class participation (10%). I will periodically give you discussion questions and some problems to work. The emphasis will often be on getting you familiar with the material before we cover it, rather than afterward.

Note: I want a paper proposal by Oct. 12, i.e. right before break. The proposal should summarize the highlights of your theoretical argument and discuss the methods you are planning to use in

your paper. You can use this as an opportunity to get my feedback on your proposed approach. Please try to keep this under 10 pages; if you've got a 50 page literature review you have prepared in conjunction with some other class you don't need to give all of it to me now!

Classroom Format. I will no doubt do a fair amount of lecturing and presentation. However, as noted above, I will often give you discussion questions beforehand. I may even ask you to present a small part of the material. I encourage you to bring up questions in class, and I encourage you even more to try to answer each others' questions. For those of you who have had me before, there will be a greater emphasis on knowing at least some of the material beforehand and working through things together in class.

Also, some class time will be devoted to discussing the current status of your paper. By early October (even before the proposal is due), you should be able to present to the class your general topic and the data and techniques you are tentatively planning on using. By early November, you will discuss the current status of your research. In the last 3 or 4 classes of the semester, you will give a 15-20 minute presentation on your completed work. (If necessary, we will use the final exam period to get all the presentations in.) We can expand the amount of time for group discussions of each others' work if there is a demand for that.

General format for presentation of methods. When going over each method, we will typically do some or all of the following. We will especially do this with the first method, logistic regression; having laid the groundwork, we'll see that many topics can be covered more quickly as we move on to new methods.

- Explain the method and its rationale. When and why would it be used? Why is OLS regression (or other methods) not appropriate? What assumptions does the method make?
- Interpreting results. Besides understanding what parameters mean, we will focus on the many techniques available in Stata for making sense of results. These include graphing techniques and the use of hypothetical plugged-in predicted values. The `margins` command, as well as some of Long & Freese's commands, will be critical here.
- Diagnostic procedures. How can we determine if the assumptions of the model are met, or if there are problems with model specification? This will include an examination of residuals and other diagnostic tests.
- Hypothesis testing. These include testing whether some or all coefficients equal zero; whether coefficients equal specific values; whether coefficients are equal to each other.
- Alternative methods for handling this type of data. In particular, we will consider different approaches for handling ordinal data (e.g. logit, probit, gologit, stereotype, interval regression, and heterogeneous choice/ location scale models).

Specific Methods & Models to be discussed. Following is the tentative listing of the methods that we will be covering. I don't have firm dates for each of these, but in general I anticipate spending 1 to 3 weeks on each major topic, starting with binary outcomes. It may go a little slower at first, but we should find that things go more quickly once we've established some background, e.g. hypothesis testing may take a little while at first but should then go more quickly. I list the relevant readings from Long and Freese but there will usually be additional readings available on the web page. Also, I am re-ordering the topics a bit this year. In the past, I covered several advanced topics but never got to some of the more basic methods covered by Long and Freese. This year I will make sure we cover the basics while still getting to more advanced methods later in the course.

Very Brief Review of Models for Continuous Outcomes – or in other words, OLS regression. There are some handouts on the course web page for this. I don't plan to cover this in class, but you should feel free to come to me with any questions you may have. Throughout the course, we'll note similarities and differences in the methods for analyzing continuous as opposed to categorical outcomes.

Overview of Generalized Linear Models & Maximum Likelihood Estimation – there are some very good readings on the course web page about this. I'll just say a little bit in the way of introduction, but we will return to the material throughout the course of the semester.

Models for Binary Outcomes –e.g. lives/dies, gets married/doesn't get married. This section will establish a lot of the background that we will use with other methods. Primary emphasis will be on logistic regression, although we will also mention probit and possibly other topics.

Readings: Long and Freese chapters 3 and 4 for this (3 you can just skim; it introduces material that we will return to several times.)

Interpreting results: Adjusted Predictions and Marginal effects. The results from binomial and ordinal models can often be difficult to interpret. All too often, researchers discuss the sign and statistical significance of results but say little about their substantive significance. I will expect every student paper to use the methods described in this section and/or one of the advanced methods we discuss later in the course.

Readings: Long and Freese discuss several methods and commands for making results more interpretable. You should be familiar with their commands, many of which are still useful, but I will show how the `margins` command introduced in Stata 11 is often a superior alternative.

Models for Ordinal Outcomes I – e.g. high/medium/low. At first, we will talk about the more basic models, like ordered logit and interval regression. Much of my own recent research involves ordinal models, so I will provide a lot of supplementary material later on.

Readings: Long & Freese, ch. 5, section 6.8

Models for Count Outcomes – Count variables indicate how many times something has happened; for example, how many articles has a professor published? Note that such variables

are not really continuous, e.g. you can't have 4.3 articles. Nonetheless, OLS regression is often used with such variables. OLS will sometimes work well, but models specially designed for count outcomes often work better. Long and Freese discuss several models for these types of data.

Readings: Long & Freese, ch. 8

Models for Multinomial/Nominal Outcomes –nominal dependent variables with more than 2 categories, e.g. votes Republican/Democrat/Other. We'll talk about multinomial logit models and possibly the conditional logit model. Multinomial logit models examine how individual-specific variables affect the likelihood of observing a given outcome, e.g. how education and experience affect a person's occupation. In conditional logit models, alternative-specific variables that differ by outcome and individual are used to predict the outcome that is chosen. For example, in a multiparty race, we can examine how the distance on issues between each candidate and the individual affects voter choice. There is a lot of material in Long & Freese about these topics that I probably won't cover in class, but you should go over it yourself if it addresses some of your research needs.

Readings: Long and Freese, chs. 6, 7

ADVANCED TOPICS (Subject to Change, depending on student interest and amount of time available.)

Categorical Data Analysis with Complicated Survey Designs – Most statistical techniques assume the data were collected via simple random sampling. However, sampling designs are often much more complicated than that, e.g. clustering and/or stratification will sometimes be used. Some individuals will be more likely to be interviewed than are others, e.g. a survey might deliberately oversample blacks. Stata has a whole set of commands for survey data called the `svy` commands. Once you understand the basic principles, they aren't all that hard to use, but there are a few key differences between them and their non-svy counterparts (in particular, hypothesis testing is somewhat different). This won't take long to cover but you should know the basics.

Readings: Again, nothing from Long and Freese, but I'll give you some excerpts from Stata's Survey Data manual and possibly other readings.

Models for Ordinal Outcomes II: Generalized Ordered Logit Models – The assumptions of the ordered logit model are often violated. The generalized ordered logit model (estimated by `gologit2`) sometimes provides a viable but still parsimonious alternative.

Readings: The course web page will have the readings on this.

Models for Group Comparisons; Heterogeneous Choice/ Location-Scale Models. We'll spend some time here talking about concerns Allison (1999) raised about comparing logit and probit coefficients across group, and two papers I wrote (Williams 2009, 2010) suggesting ways in which Allison's proposed solution could be improved upon. In particular, we will talk about how heteroskedasticity can be especially problematic in logit and ordered logit models, and what

you can do about it using my `oglm` program. Some recent research by Hauser & Andrew (2006) on determinants of educational transitions may also get worked in here.

Readings: All the readings for this will be on the course web page.

Fractional Response Models. Sometimes the dependent variable is a proportion, e.g. the percent of a firm's employees that participate in the company pension plan. Logit and probit models can easily be adapted to deal with such situations.

Readings: Readings will be on the course web page.

Panel Data. Sometimes the same individuals (or nations, or companies) are measured at multiple points in time. The statistical technique used needs to reflect the fact that the different measurements are not independent of each other. This is a big topic and goes well beyond Categorical Data Analysis, but a few basic commands, e.g. `xtlogit`, will be discussed, time permitting. (I've actually never gotten to this topic, but students who need these methods have covered them on their own.)

Readings: Stata has an entire manual on XT (cross-sectional time series) commands. If you have Stata 11 or 12 this manual is available in PDF format. Any other readings will be on the course web page.

Other Advanced Topics – Time permitting, there are several other topics involving categorical data that we might cover. For example, some types of event history analysis can be done via logistic regression models. There are some interesting things that can be done with the analysis of cross-classified data, e.g. there has been a lot of work done analyzing mobility tables. Many of these techniques require a lot of background that isn't specific to categorical data analysis, so we'll have to see what the time situation is like and what student interests are. If you want to use CDA methods we don't otherwise cover I may be able to help you find appropriate sources.

Readings: Readings (if any) will be on the course web page.