Course Syllabus for Sociology 73994
Categorical Data Analysis
Spring 2017

Instructor	Richard Williams 741 Flanner (Office: 574-631-6668) (Mobile: 574-360-1017) Email: Richard.A.Williams.5@ND.Edu Skype: rw120555 Personal Web Page: http://www3.nd.edu/~rwilliam/
TA	Christopher Quiroz 813 Flanner Mobile: 619-808-5965 Office Hours: 12:30-2:00 Monday and by appointment Email:cquiroz@nd.edu
Time and Place	Class: Flanner 925, 11:00-12:15, Monday, Wednesday Lab: Debartolo 331, 3:3-5:00 Friday
Office Hours	MW 1:30-3:00 and by appointment. I am generally very accessible via phone, voicemail, email, Facetime, Skype, and probably Zoom. I may set up extra office hours when there is high demand to see me.

Course Web Page

http://www3.nd.edu/~rwilliam/xsoc73994/index.html

Notes, readings, etc. will be placed on the course web page.

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 you are on your own if you do. If you aren't familiar with Stata, don't worry; the text provides an excellent discussion and I have various handouts to help you. Stata 14 is ideal but Stata 12 or 13 are also OK.

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 63997 (or equivalent introductory statistics courses) is the prerequisite for the course. Students from outside of Sociology are welcome.

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, Third Edition, 2014, by J. Scott Long and Jeremy Freese (NOT available in bookstore; it is cheaper if you order direct from Stata Press at http://www.stata.com/bookstore/regression-models-categorical-dependent-variables/).

Fixed Effects Regression Models, 2009, by Paul Allison. (NOT Available in Bookstore, but you can get it from Amazon or elsewhere, e.g. http://www.amazon.com/Effects-Regression-Quantitative-Applications-Sciences/dp/0761924973/ .

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 Allison book will be especially helpful if you are analyzing longitudinal data.

The books sometimes assume background knowledge that you don't necessarily have. Also, I will be covering many topics that are not in the books. 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 (60%).
    - 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 CDA methods covered after the first few weeks, e.g. Count Models, Panel Data methods.* Start thinking about this soon.
    - You should get my approval, but in most cases you can use any data set that you like. Many people already have a data set they are working with. If you don't, sources that students have found helpful in the past include the General Social

Survey, ICPSR, and the Eurpoean Social Survey. For information on these & other data sets see

  - http://www.norc.org/GSS+Website/
  - http://www.icpsr.umich.edu/
  - http://www.europeansocialsurvey.org/
  - http://csr.nd.edu/available-resources/social-science-data/

- Notre Dame's Center for Social Research (CSR) can also help students with things like acquiring data sets and statistical analysis. For more, see http://csr.nd.edu/.

- People have occasionally wanted to use data sets that were not available until very late in the semester – indeed, sometimes not until the semester was over. This was always problematic, and is even more so now that the graduate school is allowing far less time to finish incompletes. You should be working on your papers throughout the semester, not just frantically scrambling to put something together in a few days near the end. *You therefore must have your complete data set available to you by March 9 (or at least have enough of it by then to write a satisfactory paper).* If that isn't going to happen, you should pick another topic. Also, the homework will often give you the opportunity to work with a data set of your choice, so the sooner you have your data set, the sooner you will be able to try out different techniques with it.

- Classes and/or labs will occasionally be devoted to discussing the current status of your project, and the last few classes/labs will be used to present your papers.

- *I want a paper proposal no later than March 9, 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!

- *By April 5th you should send me a few paragraphs updating me on the status of your paper*, e.g. let me know how the analysis is proceeding. If you are encountering any problems or have any questions this would be a good time to let me know or to schedule a meeting with me.

- I also expect everyone to meet with me outside of class at least once to discuss how your paper is going. (Skype, Zoom, phone calls and multiple emails are ok.) Occasionally major problems don't surface until late in the semester. I like to provide feedback on projects but I can't do that if you never contact me outside of class. In the past, some people have practically set up camp outside my office but others have been rarely seen, so I want to make sure everybody maintains at least some contact with me.

  - *You need to be ready to present by April 19th. The final paper itself is due on May 8th*. You will present on your papers the last few weeks of the semester. One or two of the labs may be used to allow more time for presentations. You can use feedback on your presentation to make final revisions on your paper.

- Homework & possibly other assignments (40%). I expect this to have a fairly neutral impact on most people's grades. But, I think these will help you to understand the material better and produce better papers. These assignments aren't meant to be especially challenging or grueling but they will require that you understand the major concepts behind a method. I have 10 assignments planned but that may change depending on whether we are keeping up with everything I would like to cover. I will aim for about one assignment every week. Homeworks will usually be graded on a 4 point scale, with 0 = terrible/didn't do it to 4 = pretty good, got most things right. Note that even though each HW only counts for a few points, if you consistently do poorly or don't do some of them at all your grade could suffer quite a bit.

- I like to start class on time. Excessive absences or late arrivals will likely hurt your grade. If there is some compelling reason you can't make it to class on time let me know.

Classroom Format. I will no doubt do a fair amount of lecturing and presentation. However, I encourage you to bring up questions in class, and I encourage you even more to try to answer each other's questions.

Also, some class time will be devoted to discussing the current status of your paper. By late February (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. In the last 3 or 4 classes/labs of the semester, you will give a 15-25 minute presentation on your completed work. We can expand the amount of time for group discussions of each others' work if there is a demand for that.

Labs will be used to work on your assignments and papers. I may occasionally take over the lab to cover additional material. The TA may also use the lab time to cover additional topics.

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 many of Long & Freese's commands (e.g. `mtable`), 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, 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. In general I anticipate spending 1 to 2 weeks on each major topic. 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. 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. I may also reorder the topics, e.g. I may cover panel data sooner if some people plan to use panel data for their papers.

I. Foundations of Categorical Data Analysis. This section will go over the basics of logistic regression. It will also go over techniques for making results more interpretable; analyzing data sets with complex sampling schemes; and (possibly) techniques for handling missing data. I call these topics "foundations" because once you understand them it is very easy to extend them to other CDA methods, such as ordinal and count models.

- *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.

    - **Readings:** Long and Freese chapters 1 & 2 (you can skim or skip these, depending on how comfortable you are with Stata.) Chapter 3 will also include a lot of things you may already know but will probably include a few new things.

- *Models for Binomial Outcomes: Basics of Logistic Regression*– 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. This may be review for some of you and if so you may want to do some of the optional readings that are on the course web page.

  - **Readings:** Long and Freese chapter 5

- *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. Note that Long and Freese have several other useful commands that I won't discuss much in class.

  - **Readings:** Long and Freese Chapters 4, 6. Most method-specific chapters will also contain additional useful information

- *Categorical Data Analysis with Complex 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, CDA hypothesis testing is somewhat different with survey data). This won't take long to cover but you should know the basics.

- *Missing data*. I am mostly covering this here because it is an important topic and there wasn't enough time to cover it in the new Stats I! But, several of the methods do involve the use of categorical data analysis, so it isn't totally out of place. This will be largely review for those of you who have had me before.

II. Intermediate CDA Methods. Here we will talk about other commonly used CDA methods, including ordinal regression, models for multinomial outcomes, and models for count outcomes.

- *Models for Ordinal Outcomes I* – e.g dependent variables coded 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 advanced material later on.

  - **Readings:** Long & Freese, ch. 7

- *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, ch. 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. 9

III. Advanced Topics (Subject to Change or Re-Ordering). Here we will talk about other commonly used CDA methods, including ordinal regression, models for multinomial outcomes, and models for count outcomes.

- *Intermediate logistic regression*. We will talk about the latent variable model in logistic regression; standardized coefficients; alternatives to logistic regression. Long & Freese will have covered some of this earlier.

- *Comparing logit and probit coefficients across nested models.* Researchers often present a series of nested models, e.g. block 1 includes demographic variables like race and gender, block 2 includes education, block 3 includes other explanatory variables, etc. We will discuss why this is potentially problematic in CDA models and what you may want to do instead.

- *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. In particular there are two articles of mine that I will have you read.

- *Models for Ordinal Outcomes III: Heterogeneous Choice Models and Other Methods for Comparing Logit & Probit Coefficients Across Groups.* We'll spend some time here talking about concerns Allison (1999) raised about comparing logit and probit coefficients across groups, 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.

  - **Readings:** All the readings for this will be on the course web page. Besides my notes, there will be a couple of articles that I have written on this topic.

- *Panel Data and Multilevel Data.* Sometimes the same individuals (or nations, or companies) are measured at multiple points in time. Or, you might have, say, a sample of schools with multiple students within each school. 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 and melogit, will be discussed, time permitting.

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

- *Analysis of rare events.* Conventional CDA techniques can produce biased results for events that occur rarely, e.g. outbreaks of war. Political scientist Gary King has offered some solutions, but other alternatives, such as penalized maximum likelihood, may be better.

  - **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.

- *Other Advanced Topics* – There are several other topics involving categorical data that we could cover but probably won't. For example, 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. `gsem`, which was added in Stata 13, lets you do structural equation modeling with categorical dependent variables. (If you have heard of MPLUS, Stata can now do many of the things that MPLUS does.) If you want to use CDA methods we don't otherwise cover I may be able to help you find appropriate sources. Readings (if any) will be on the course web page. We may not have time to cover it but there will be a handout on how to handle ordinal independent variables.