Course Syllabus for Sociology 73994
Categorical Data Analysis
Spring 2022 (Revised April 12, 2022)

Instructor

Richard Williams
4058 Jenkins Nanovic (but I will rarely if ever be there)
Office: 574-631-6668, Mobile: 574-360-1017
Email: rwilliam@nd.edu
Personal Web Page: https://www3.nd.edu/~rwilliam/

Canvas

We will make extensive use of Canvas in this course. The course Canvas page will include the most critical links for the course, e.g. for the class Zoom links and the course web page. All assignments should be submitted through Canvas unless you are told otherwise. I may also try a few other features at times, e.g. make announcements. If you've never used Canvas, you can check out https://community.canvaslms.com/t5/Student-Guide/tkb-p/student .

TA

Junrong Sheng
4042 Jenkins and Nanovic Hall
Email: jsheng2@nd.edu
Phone: 574-383-6459
Office Hours: Tuesday 10:00-12:00

Time and Place

MW 9:30 – 10:45 AM Jenkins and Nanovic Hall B032
Lab: Debartolo 228, 3:30-5:00 PM Friday

Office Hours

MW 1:00-2:00 and by appointment. I want to do office hours and appointments by Zoom most of the time but in-person is also possible and sometimes necessary. I am generally very accessible via phone, voicemail, email, and Zoom – including on nights and weekends if necessary. If you want a meeting with me just send a list of convenient times for you and we can work out something.

Course Web Page

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

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

Also, there is always somebody who wants to use a method we don't cover until the closing weeks. I therefore keep all of last year's notes at

https://www3.nd.edu/~rwilliam/stats3/index.html

I usually make at least minor tweaks to the notes every year but the old notes (along with the readings) will probably be enough to get you started. I can also help you individually if you want to rush ahead on some method or maybe even use a method we won't go over at all.

A Note on the Treatment of Gender & Race. Versions of some of my statistics notes first saw life thirty years ago. There have been many changes in statistical practice and preferred wording over those 30 years. For example, traditionally, most studies have (and perhaps still do) treat gender as binary and fixed at birth. These practices are unfortunate, not only because they are inaccurate, but because they can perpetuate harm experienced by transgender and non-binary communities. Preferred race-related terminology and statistical practices have also evolved over that time. My notes (and for that matter probably most of my publications) reflect common past practices. As I update handouts, I am making changes to reflect current, more inclusive terminology and methods. For more details (and also a list of resources for those interested in doing research involving alternative gender identities that have not been widely studied in the past) see https://www3.nd.edu/~rwilliam/stats/RaceGenderNote.pdf.

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 or R 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 16 or 17 is ideal but slightly older versions are probably 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 if they have the necessary background.

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/ . There may be other cheap alternatives.

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/ . You can also get a less aesthetically pleasing but fully functional and free version from the ND Library at http://methods.sagepub.com.proxy.library.nd.edu/book/fixed-effects-regression-models.

Online readings packet (compiled by Richard Williams). This will be the most critical source, as it includes most of my lecture notes and handouts.

You should read the required chapters and/or the corresponding online handouts before class and come prepared with any questions you have. I think the Long & Freese book is fantastic and goes into much more detail than I can in class but some students have said my online notes are sufficient. The Allison book will be especially helpful if you are analyzing longitudinal data.

The books sometimes assume background knowledge that you do not 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 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 more advanced CDA methods covered after the first few weeks, e.g. Ordinal Regression, 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 European 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/
- https://lucyinstitute.nd.edu/services/cssr/resources/#Databases

- Notre Dame's Center for Social Science Research (CSSR) can also help students with acquiring data sets, data management, statistical analysis, and Stata. For more, see https://lucyinstitute.nd.edu/services/cssr/.

- 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 1 (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 3*. (This will also be HW #6.) 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 around March 30 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

while 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 20th. The final paper itself is due on May 3rd.* You will present on your work-in-progress 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 think these will help you to understand the material better and produce better papers, and also force you to be familiar with methods besides the ones you use in your paper. 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. I try to time the assignments so that they are due about a week after we have covered the relevant material.

  - *The first seven assignments are required*. After that, *you have to do one of the three assignments on advanced topics*. If you do more than eight assignments the lowest scores will be dropped. What advanced assignment you choose to do may be affected by the methods you are using in your paper, e.g. if you have panel data you'll probably want to do the panel data homework.

  - Homeworks will usually be graded on a 5 point scale, with 0 = terrible/didn't do it to 5 = very good, got most things right. The TA also has the option to make a small part of the HW grade be dependent on supplementary exercises given in lab.

  - 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 will likely suffer quite a bit. For example, two missed assignments would keep you from getting an A for the course even if everything else you did was excellent.

  - Note: Homeworks should be handed in on time!!! Much of the material is cumulative, and if you fall behind by a week or two it may hurt you for the rest of the semester. If you desperately feel the need for a short extension (perhaps because of illness or other problems) talk to the TA. If you have problems that your fear may make you chronically late, you should talk to me and we'll see what we can work out. You have to work at a steady pace in this course, but I don't think the workload is much greater than it is in most graduate classes; it is just different.

- 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 February 23rd (shortly 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 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, papers, and any supplementary materials the TA wants to provide. I may occasionally take over the lab to cover additional material or to keep us on-schedule.

Special adaptations when/if using online learning

- I expect most classes will be in-person. However, if I am sick, some of you are sick, weather conditions are terrible, etc., we may conduct class via Zoom. Check your email before coming to class in case we switch to Zoom at the last minute.
- I will usually hold my office hours and individual appointments via Zoom. If you want a face-to-face meeting, we can arrange that. If I need to be on campus anyway we can meet either in-person or via Zoom.
- You will benefit from attending an online class if you are fully awake and present. Please do whatever is necessary for you to get to that state.
- I expect you to keep your camera on most of the time – it helps me if I can see you and I think it will help you to concentrate better. If that creates problems for you please let me know. You can use a Zoom background if you don't want people to see how messy your place is!
- Please use your full name when Zooming. It is fine to use your preferred first name, e.g. "Beth" instead of "Elizabeth."
- I suggest you tell Zoom to use a nice (not weird or unprofessional) picture of yourself. This is what people see when your video is muted. You can set this up in your Zoom profile.
- I will often cold-call on people (both in online and in-person classes). I am not trying to embarrass anyone – it is fine to say you do not know – but low-stakes cold calling can be a good way to keep students involved and paying attention. Some of my undergraduate students said they did not like cold-calling at first but then found that it was very helpful for staying focused.
- Some material may be presented asynchronously, e.g. instead of having a regular class you will be asked to watch or read something on your own.

- It will be nice if you can Zoom from a place where you are not required to wear a mask, e.g. your apartment. But, you are required to comply with University rules wherever you are Zooming from.
- I do NOT intend to record most sessions. I want people to feel free to share their thoughts. Since class is online, you should be able to make most classes even if you are sick or quarantined. If this policy creates problems for you for some reason, let me know why and we will see if we can work something out.
- You may not always be seeing me in person, but Zoom is great for one-on-one meetings. Screen-sharing especially is really great when talking about statistics. If you want to meet with me, send me a list of good times (including possibly nights or weekends) and we should be able to work out a mutually agreeably time.

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. ologit, oprobit, gologit, interval regression, and heterogeneous choice/ location scale models).

Specific Methods & Models to be discussed/Tentative Schedule.

Following is the likely 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 optional 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.

- *January 10, 12 – 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.

    - Homework #1 is due on January 20, 2022. It does not actually rely on the above material. Instead, it requires you to do preliminary work on the data set you hope to use. Both the TA and I will look at it it and comment..

- *January 19, 24, 26 – Models for Binomial Outcomes: Basics of Logistic Regression* – e.g. lives/dies, gets married/does not 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

    - Homework # 2 is due February 3, 2022.

- *January 31, Feb 2 – 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
  - Homework #3 is due February 10, 2022.

- *February 7 – 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.

- *February 9 – 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.

  - Homework #4 is due February 17, 2022.

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

- *February 14 – 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

- *February 16 – 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
  - Homework # 5 is due February 24, 2022.

- *February 21 – Ordinal Independent Variables.* We often want to use ordinal variables as independent/explanatory variables in our models. Rightly or wrongly, it is very common to treat such variables as continuous. We will discuss when it is appropriate to do so. We will also discuss other possible strategies that can be employed with ordinal independent variables, such as the use of Sheaf coefficients.

  - Readings: Any readings will be on the course web page.
  - Note: You may want to read this earlier because it may be useful for whatever type of analysis you are doing.

- *February 23 – Discuss paper proposals in class.*

  - Homework # 6 (Your Paper proposal) is due March 3, 2022. Both the TA and I will look at it.

- *February 28 – 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.

- *March 2 – 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.

  - Homework # 7 is due March 17, 2022

March 5 to March 13 – Easter Break

III. Advanced Topics (Subject to Change or Re-Ordering). Here we will talk about other commonly used CDA methods or advanced methods, including count models, panel data/multilevel methods and advanced models for ordinal regression. The ordering of topics may

change depending on what methods seem most needed for the papers students are working on. You have to do at least one of the three homeworks on advanced topics.

- *March 14, 16, 21 – 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.

    - Readings: Allison's book will be invaluable here. Stata has entire manuals on XT (cross-sectional time series) and ME (multilevel mixed effects) commands. Any other readings will be on the course web page.

    - Homework #8 is due March 31, 2022.

- *March 23, 28 – 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 articles of mine that I will recommend you read.

- *March 30 – 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.

    - Homework # 9 is due April 7, 2022.

- *April 4, 6 – 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
- Homework # 10 is due April 14, 2022.

IV. Advanced Special Topics (time permitting)

- *April 11 – 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.
- *April 13 – 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: Any readings will be on the course web page.

V. End of Semester Wrap Up

- *April 20, 22 (during lab), 25, 26 – Presentations on Papers*. Papers should be well underway but do not need to be finished yet. You can use the feedback you receive to make your final revisions.
- *Final Paper is due Tuesday, May 3, 2022, at 10 am.*