Syllabus

SOC 681: Categorical Data Analysis
(and other Nonlinear Models)

Fall 2021

MW 12:30 – 1:45pm
STON 345

Professor Mize
tmize@purdue.edu

Updated: 2021-11-09

Syllabus Contents (click hyperlinks below to jump to a section)

Course Information

- SOC 681: Categorical Data Analysis
  - CRN: 15552
  - MW 12:30 – 1:45pm
  - 3 credit hours
- An updated copy of this syllabus and all other course materials is always available on Brightspace

Instructor Information

- Dr. Trenton Mize, Assistant Professor of Sociology & Core Faculty Member of Advanced Methodologies of the Behavioral, Health, and Social Sciences (AMAP)
  - tmize@purdue.edu
  - Office: Stone Hall (STON) 326B
  - Office hours: Fridays 12:30 – 1:45pm
    - Book an office hours appointment here
      - Specify in-person or Zoom when booking
  - Preferred mode of contact: (1) When possible, attend my office hours to ask questions. If that is not possible, (2) please email me.
  - Preferred pronouns: he/him/his
  - Preferred name: Trent, Professor Mize, or Dr. Mize

Advanced Methodologies in the Behavioral, Health, and Social Sciences at Purdue (AMAP) Graduate Certificate

- AMAP has an interdisciplinary methods certificate
- This course is an approved quantitative course you can apply towards completion of the certificate

Attendance Policy

- Purdue's classroom attendance policies are back to normal operations for Fall 2021.
- You need a valid excuse to miss class. To ask for an excused absence, email me before the class period. I will excuse absences for:
  - Illness (doctors note required in most circumstances)
  - Mandated quarantining
  - Religious observances
  - Traveling for academic events (e.g. a conference)
  - *Other valid reasons will be approved on a case-by-case basis

Course Description

Many—perhaps even most— behavioral, health, and social science questions include outcome variables that are categorical. E.g. Which political candidate will win the next election? What social class does a person belong to? How many publications does it take to receive tenure? How many drinks does the average person consume per week? Answering these—and countless other—questions cannot be adequately accomplished via the linear regression model and instead require the more advanced techniques covered extensively in this course.

Categorical Data Analysis is a course in applied statistics that primarily deals with regression models in which the dependent variable is binary, nominal, ordinal, or count. In addition, some flexible methods for nonlinearities within the linear regression framework will be briefly covered. Many common statistical issues encountered by social scientists require different methods when the dependent variable is not continuous. E.g. Interpretation of coefficients, calculation of predictions, testing of interaction effects, testing for mediation, assessing model fit, and many other techniques require a different approach for models of categorical dependent variables compared to the methods used for linear regression. The focus of the course is on interpretation and learning to deal with the complications introduced by the nonlinearity of the models. While we will cover the mathematical details of the models, students will be primarily assessed on their ability to use, present, and interpret the various models we cover.

Specific models considered include: probit and logit for binary outcomes; ordered logit/probit and the generalized ordered logit model for ordinal outcomes; multinomial logit for nominal outcomes; Poisson, negative binomial, and zero inflated models for counts; and fractional response, LOWESS, and local polynomial smoothing methods for continuous and quasi-continuous outcomes.

Learning Objectives

1. Identify the key issues and consequences of treating categorical dependent variables as continuous in statistical models
2. Be able to produce and interpret statistical models for categorical dependent variables, including those for binary, ordinal, nominal, and count dependent variables
3. Understand how to apply categorical data analysis to substantive research questions and describe and present them to general interest academic audiences

Required Prerequisites

- Course
    - A course in linear regression or ANOVA is a required prerequisite; e.g. HDFS 590, POL 501, PSY 606, PSY 631, SOC 680, or STAT 512. Contact the instructor if you are unsure whether your previous coursework has prepared you for this course.

Required Texts, Readings, and Software

Books

- **Required book:**
    - Lewis-Beck, Colin and Michael Lewis-Beck. 2015. *Applied Regression: An Introduction.* 2nd edition. Sage.
        - To ensure you are prepared for the course, this is a required reading as a prerequisite. You must read this book either before the course or by the end of the first week of classes; there will be a graded assignment on the book's content.
        - I have chosen this book as it is a short and accessible overview of everything you need to know prior to the beginning of the course. I teach the course assuming all students are familiar with the information covered in this short book.
            - Note you can access a free eBook version through Purdue's library.
- **Recommended book:**
    - Long, J. Scott and Jeremy Freese. 2014. *Regression Models for Categorical Dependent Variables Using Stata*. 3rd Edition. College Station, TX: Stata Press.
        - $54 Kindle version available
- **Optional/supplemental books:**
    - Agresti, Alan. 2012. *Categorical Data Analysis*. 3rd Edition. New York: Wiley.
        - Note there is free online access available through Purdue's library
    - Long, J. Scott. 1997. *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, CA: Sage.

Additional Readings

- All readings will be provided on Brightspace as PDFs

Lecture Notes

- All lecture notes will be provided on Brightspace as PDFs
    - I prefer students to take notes by hand so I will bring printed versions of the notes for each student to each class. Those who prefer to view the notes electronically on a computer or tablet should inform me during the first week of class.

Statistical Software

- The focus of the course is on the substance of the models. However, we will also extensively cover estimation and application. The examples in the lecture notes use Stata 17. It is helpful—but not necessary—to be familiar with Stata before the course. I will provide an "Introduction to Stata" guide the first week of class as a crash course for those who are new to Stata.
- ***Stata 17***
    - Completing the assignments will be easiest if done in Stata as the examples in the lecture notes will use Stata; I also provide template Stata do-files for the assignments.
    - You can download Stata 17 for free from Purdue's software site.
- ***R***
    - Those who are experts in R may use it for their assignments if they would like. Note that although I am familiar with R and will provide some code and examples, students are mostly "on their own" if they wish to use R for assignments in terms of software issues. That is, I am happy to help with Stata issues and feel confident that I can help solve any computing problems—I cannot make this same guarantee for issues using R.

Assignments and Weekly Readings

Applied Data Analysis Assignments

Overview

- The primary source of grades for the course is a series of applied data analysis assignments. These will typically follow the types of models we are covering in class. For example, one assignment will focus on estimating, interpreting, and presenting models for count outcomes.
    - Assignments will be available on the course website before we begin covering the relevant topic. I encourage you to read over the relevant assignment before we begin covering those topics so that you can ask questions in class about any specific problems you may have as we cover that material.

Working Together

- I encourage you to work with your classmates as you complete the assignments. Seeing the unique findings and issues across different examples will be helpful as you learn the models. In addition, discussing interpretations and presentation strategies with others will improve the clarity of your understanding and writing. However, your work must be original—meaning that the specific models you estimate and interpret in the assignments cannot be identical (or even very close) to those of another student in the class.

Data for Assignments

- The assignments will be applied—meaning you will be using real health/social science data to estimate models and then interpret and present the findings. Several semi-cleaned datasets will be available on the course website for your use (General Social Survey, Add Health, Health and Retirement Study, and Pew Politics 2016). If you wish to use a different dataset than the ones available for your assignments, you must get my approval. Any dataset for the assignments must have a sample size of at least 500. Unless you already have a dataset cleaned and organized, I recommend you use the ones on the course website.

Weekly Readings and Discussion

- Most weeks there will be required readings. In general, there will be no more than two readings in a given week.
    - Readings may supplement course material, show a published example of research using the methods we are covering, etc.
    - We will discuss the readings in class; all students are expected to participant in the discussions.

Grading

Course grades refer to the number of points towards your final grade you earn from each assignment or reading discussion. The course has 1,000 total points. Therefore, an assignment worth 100 points is worth 10% of your final grade.

- In-class discussion of readings
  - Regular participation in the in-class discussions of the readings is required
  - Participation is worth 100 points
- Assignments
  - Each assignment is worth 100 points
    - 9 assignments * 100 points = 900 total points

Revising Assignments

Optional Revisions

- All assignments, regardless of your initial grade, can be revised within 10 days from when the assignment is graded. Revisions are optional; you do not have to revise any of your assignments that received a grade on the initial submission.
  - Revised assignments are eligible for a 10-point bump or half the original lost points back—whichever is greater. E.g. A 90 can become a 100. A 70 can become an 85 (30 lost points / 2 = 15 possible points on revision).
  - If you choose to revise an assignment, you only have to revise your answers that were incorrect; make sure to be attentive to the comments you received on your original assignment. When revising an assignment, please use the Track Changes feature in Word so the changes made are clear. Upload revised assignments to Brightspace.

Reject & Resubmits: Required Revisions

- Occasionally, I will not give a grade but instead give a "Reject & Resubmit" which is a required revision. Reject & Resubmits require a revision in order to receive a non-zero grade for the assignment.
  - I assign Reject & Resubmits when the quality of the initial submission is well below course standards. This can be due to major problems but can also sometimes be a minor issue (e.g. miscoded dependent variable) that nonetheless populates the rest of the assignment with errors.
  - Your grade will not be penalized on your revision submission—i.e. you can still receive up to a full 100 points on the revision.

Missed or late assignments; incomplete final grade

- Late assignments will be accepted at my discretion and will include a penalty of 10 points per class period late
    - If you expect to have trouble meeting an assignment deadline because of a foreseeable event (e.g. you will be out of town for a conference or you will be observing a religious holiday)—please contact me as early as possible and we can arrange a plan for completing the assignment without penalty.
- A grade of incomplete (I) will be given only in unusual circumstances. To receive an "I" grade, a written request via email must be submitted prior to December 1, and approved by the instructor. The request must describe the circumstances, along with a proposed timeline for completing the course work. Submitting a request does not ensure that an incomplete grade will be granted. If granted, you will be required to fill out and sign an "Incomplete Contract" form that will be turned in with the course grades. Any requests made after the course is completed will not be considered for an incomplete grade.

Grading Scale

Your final grade will be based on the sum of your grade on all assignments and discussion participation which can range from 0 to 1,000 points. Your final point total will be converted to a letter grade using the following category criteria:

- ≥ 970.00 A+
- 930.00 – 960.99 A
- 900.00 – 920.99 A-
- 870.00 – 890.99 B+
- 830.00 – 860.99 B
- 800.00 – 820.99 B-
- 770.00 – 790.99 C+
- 730.00 – 760.99 C
- 700.00 – 720.99 C-
- 670.00 – 690.99 D+
- 630.00 – 660.99 D
- 600.00 – 620.99 D-
- ≤ 590.99 F

Note your final grade percent will not be rounded. I do not change a student's final grade for any reason.

Asking for Help

There are a few basic guidelines to follow when asking for help that will greatly aid my ability to help you with computing problems such as debugging code, recoding a variable, getting a model to estimate correctly, convergence problems, or other common issues. The most important principle is that you should provide me with the files necessary to reproduce the problem you are having. That is, instead of only sending problematic output, also send the data and coding files necessary to reproduce the problem. Keep in mind that these guidelines apply both when you are asking for help over email and when asking for help in person. Specifically:

1. Include basic descriptive statistics of your model variables. If you use Stata, **`desctable`** is an easy way to do this.
2. Send a small version of the dataset you are using that includes all necessary model variables, but excludes extraneous variables in the dataset. E.g. If you are using the GSS, do not send along a dataset with thousands of variables; if you are only using 10 variables for your analysis—only include those in the dataset you send.
3. Send a coding file (e.g. do-file for Stata) that includes only the necessary code to reproduce the problem you are encountering. I.e. This file should load the appropriate dataset and estimate any models and post-estimation necessary *for the specific issue you are asking for help with*. If your problem is with data management, include only the code relevant to recoding/manipulating the specific variable causing the problem. That is, you should not send the files for your entire assignment or project—only send what is necessary to reproduce your issue.
4. Send the relevant output that illustrates the problem and any errors messages you are encountering. E.g. This can be a Stata log-file or a text file. Again, only include output relevant to the problem you are experiencing.
5. To ensure that my email does not reject your message, do not send large files as email attachments. Instead, create a folder in Dropbox, Box, or Google Drive and include the necessary files. All of these services allow you to send a "shareable link" that will allow me to access the files.

Academic Integrity

- Purdue's Honor Pledge states: "As a boilermaker pursuing academic excellence, I pledge to be honest and true in all that I do. Accountable together - we are Purdue."
- I encourage you to study with other students in the class, to have them read over and comment on your notes or assignments, or to provide other feedback.
    - **However, any work you turned in for a grade needs to be completed solely by the student receiving the grade.**
    - For example, while it is acceptable—and encouraged—for you to ask another student to read an assignment or to share code for suggestions and comments, the assignments themselves must be written entirely by the student receiving the grade.
- Incidents of academic misconduct in this course will be addressed by the course instructor and referred to the Office of Student Rights and Responsibilities (OSRR) for review at the university level. Any violation of course policies as it relates to academic integrity will result minimally in a zero grade for that particular assignment. In addition, all incidents of academic misconduct will be forwarded to OSRR, where university penalties, including removal from the university, may be considered.

Additional Policies

Welcome statement

In this course, each voice in the classroom has something of value to contribute. Please take care to respect the different experiences, beliefs and values expressed by students and staff involved in this course. I support Purdue's commitment to diversity and welcome individuals of all ages, backgrounds, citizenships, disabilities, sexes, education levels, ethnicities, family statuses, genders, gender identities, geographical locations, languages, military experiences, political views, races, religions, sexual orientations, socioeconomic statuses, and work experiences.

Preferred names and pronouns

- The overall spirit of my policy is that I respect your decision to decide how you would like to be referred to. I expect class members to be similarly respectful.
- Specifically, I will honor the names and pronouns you provide, and your request at any point to address you by your correct name and/or gender pronoun. I also expect class members to honor the names and pronouns peers provide.

Discussion etiquette

We will discuss many topics in this class—some of which may be personal for some students. When discussing topics in class, you are encouraged to comment, question, or critique an idea, but you are not to attack an individual. Our differences, some of which are outlined in the University's nondiscrimination statement above, will add richness to this learning experience. Please consider that sarcasm and humor can be misconstrued (especially in online interactions) and generate unintended disruptions. Working as a community of learners, we can build a polite and respectful course ambience. Some etiquette rules for the course:

- Do not dominate any discussion. Give other students the opportunity to join in the discussion.
- Do not use offensive language. Present ideas appropriately.
- Avoid using vernacular and/or slang language. This could possibly lead to misinterpretation.
- Keep an "open-mind" and be willing to express even your minority opinion.
- Think and edit before you push the "Send" button or make an in-class comment.
- Do not hesitate to ask for feedback from me if you are unsure about something you wish to discuss.

Counseling and psychological services (CAPS)

Purdue University is committed to advancing the mental health and well-being of its students. If you or someone you know is feeling overwhelmed, depressed, and/or in need of support, services are available. For help, such individuals should contact Counseling and Psychological Services (CAPS) at (765)494-6995 and http://www.purdue.edu/caps/ during and after hours, on weekends and holidays, or by going to the CAPS office of the second floor of the Purdue University Student Health Center (PUSH) during business hours.

Disability Resource Center

Purdue University strives to make learning experiences as accessible as possible. If you anticipate or experience physical or academic barriers based on disability, you are welcome to let me know so that we can discuss options. You are also encouraged to contact the Disability Resource Center at: drc@purdue.edu or by phone: 765-494-1247. More details are available on our course Brightspace under Accessibility Information.

Course Schedule

Note that exact topics, readings, dates, and other aspects of the course schedule are subject to change. Always reference the syllabus on Brightspace for the most up to date version of the Course Schedule.

Required readings

- All required readings are to be completed **before class on Monday**
- Readings marked as *exemplars* are substantive research articles that use the methods we cover. You can skim the theory and discussion sections of these articles; you should primarily focus on the Methods and Results sections.

Assignment due dates

Note all assignments (except Assignment 6 Part A) are due at noon before that day's class

- Assignment 1 (100 pts): (Aug 30 @ noon): Linear regression basics
- Assignment 2 (100 pts) (Sept 6 @ noon): Data management and assignments planning
- Assignment 3 (100 pts) (Sept 17 @ midnight): Nonlinear linear regression models
- Assignment 4 (100 pts) (Sept 27 @ noon): Count models
- Assignment 5 (100 pts) (Oct 18 @ noon): Binary regression models
- Assignment 6 (100 pts): Testing, model fit, and model robustness
    - Part A due Oct 22 @ 5pm
    - Part B due Nov 5 @ 5pm
- Assignment 7 (100 pts) (Nov 15 @ noon) Interactions & cross-model comparisons
- Assignment 8 (100 pts) (Nov 24 @ 5pm): Nominal regression models
- Assignment 9 (100 pts) (Dec 13 @ noon): Ordinal regression models

Section 1: Introduction & linear regression refresher

Week 1 (Aug 23 & 25)

- Mon: Introduction to course
- Wed: Non-technical overview of regression modeling; refresher on linear regression
    - Readings: (1) CDA 2021 Cheat Sheet and (2) Davis 1985 *The Logic of Causal Order* (Chapter 1)

Week 2a (Aug 30)

- Mon: Technical overview of linear regression; Stata basics
    - Reading: (1) CDA2021 – Introduction to Stata an (2) Lewis-Beck 2015 *Applied Regression: An Introduction* (entire book except skip section 4.3)
    - ***Assignment 1 due at noon***

Section 2: Nonlinear linear regression models

Week 2b (Sep 1)

- Wed: Interpretation with marginal effects; visualizing nonlinear effects; nonlinear interaction effects

Week 3 (Sep 6 & 8)

- Mon: **No class—Labor Day**
    - ***Assignment 2 due at noon***
- Wed: Data visualization; transformations; scatterplot smoothing techniques

Week 4 (Sep 13 & Sep 15)

- Mon: Assessing and modeling complex nonlinearities
- Wed: Loglinear regression models
    - Readings: (1) Goldstein 2018 *American Sociological Review* (exemplar) and (2) Mize 2016 *American Sociological Review* (exemplar)
- Fri: ***Assignment 3 due at midnight***

Section 3A: Count models

Week 5 (Sep 20 & 22)

- Mon: Count distributions; Poisson models; Maximum likelihood estimation
- Wed: Negative binomial models
    - Reading: (1) Bail Brown Mann 2017 *American Sociological Review* (exemplar) and (2) Browning et al. 2021 *American Sociological Review* (exemplar)

Section 4: Binary models

Week 6 (Sep 27 & 29)

- Mon: Linear probability model and critiques; Binary logit/probit derivations; y^*
    - ***Assignment 4 due at noon***
- Wed: Identification and scaling issues in binary models;

Week 7 (Oct 4 & 6)

- Mon: y^* and odds ratios interpretations
    - Readings: (1) GBD 2018 *Lancet* (skim article) and responses: (2) Carroll *NY Times* 2018 and (3) Spiegelhalter 2018 *Medium*
- Wed: Predicted probability interpretations; marginal effects

Week 8 (Oct 11 & 13)

- Mon: **No class—October break**
- Wed: Ideal types & risk ratios interpretations
    - Readings: (1) Pescosolido et al. 2010 *American Journal of Psychiatry* (exemplar) and (2) Lagos 2019 *American Sociological Review* (exemplar)

Week 9a (Oct 18)

- Mon: Flexible functional forms in binary models
    - ***Assignment 5 due at noon***

Section 3B: Zero-inflated count models

Week 9 (Oct 18 & 20)

- Mon: Zero-inflated Poisson and negative binomial models
- Wed: Comparative model fit / selecting between various count models

Section 5: Testing, statistical significance, and model fit

Week 9b (Oct 20 & 22)

- Wed: Internal measures of model fit (residuals, outliers, and influence)
- Fri: ***Assignment 6 (Part A) due at 5pm***

Week 10 (Oct 25, 27, & 29)

- Mon: Measures of absolute model fits (R^2 and related measures); statistical significance
  - Readings: (1) Aschwanden 2016 *FiveThirtyEight* and (2) Wasserstein and Lazar 2016 *The American Statistician*
- Wed: Significance testing; hypothesis testing & multiple comparisons

Week 11 (Nov 1, 3, & 5)

- Mon: Comparative model fit and information criteria
- Wed: Model and effect robustness
- Fri: ***Assignment 6 (Part B) due at 5pm***

Section 6: Interactions & cross-model comparisons

Week 12 (Nov 8, 10, & 12)

- Mon: Nonlinear interaction effects and a framework for testing moderation
  - Reading: Mize 2019 *Sociological Science*
- Wed: Types of interaction effects: nominal X continuous; nominal X nominal; continuous X continuous. Group comparisons
  - Reading: Doan, Miller, and Loehr 2015 *Social Forces* (exemplar)

Week 13a (Nov 15)

- Mon: Statistical mediation and other cross-model comparisons
  - Reading: Mize, Doan, and Long 2019 *Sociological Methodology*
  - ***Assignment 7 due at noon***

Section 7: Nominal models

Week 13b (Nov 17)

- Wed: Derivation of the multinomial logit model; predicted probability interpretations

Week 14 (Nov 22* & 24)

- Mon: Conditional odds ratios interpretations; independence of irrelevant alternatives assumption **Online synchronous lecture (Zoom)*
    - Readings: (1) Doan, Quadlin, and Powell 2019 *Socius* (exemplar) and (2) Campbell 2020 *Du Bois Review* (exemplar)
- Wed: **No class—Thanksgiving break**
    - ***Assignment 8 due at 5pm***

Section 8: Ordinal models

Week 15 (Nov 29 & Dec 1)

- Mon: Ordinality; latent variable model derivation
- Wed: Coefficient interpretation; predicted probability interpretations

Week 16 (Dec 6 & 8)

- Mon: Ordinal model assumptions; comparing ordinal and nominal models
    - Reading: (1) Lim and Putnam 2010 *American Sociological Review* (exemplar) and (2) Villarreal 2010 *American Sociological Review* (exemplar)
- Wed: Alternative models for ordinal outcomes

Week 17 (Dec 13)

- Mon: ***Assignment #9 due Monday Dec 13th at noon***