SOCI/DEMG 612: Categorical Data Analysis

Professor Xi Song

Spring, 2022

[version 1/11/2022]

E-mail: xisong@upenn.edu
Office Hours: Tuesday/Thursday 5–6pm
Tuesday 12-1pm
Office: McNeil 271

Web: **https://canvas.upenn.edu/courses/1476358**
Class Hours: Tuesday/Thursday 3:30–4:50pm

Class Room: Education 120
https://xisong-spring2022.youcanbook.me/

Alexander Adames
Office Hours: Monday/Wednesday 11am-12pm
Lab Hours: Friday 10am-11am

TA Email: adames@sas.upenn.edu
Location: McNeil Building Atrium
Lab location: TBA

Course Overview

This course teaches statistical methods for analyzing categorical data, with an emphasis on practical applications rather than statistical theories. The goal of this course is to teach sociology students to learn from categorical data. The course stresses the use of various statistical methods to explain the phenomena and test models in order to address social science and policy questions, broadly defined. Familiarity with multivariate linear regression models for continuous dependent variables is assumed. Portions of textbooks and selected articles in the current literature will be assigned as Readings. There will also be a weekly tutorial taught by the teaching assistant.

Prerequisites

A prior statistics course—SOCI 536, or the equivalent—is required.

Contacts

You can reach me via email; however, I do not respond to email between 9 pm and 9 am (and neither do the teaching assistant) or over the weekend. If I don't respond within 24 hours, please

feel free to send me a polite reminder. I don't intend to be unavailable, but sometimes I get quite a lot of email and/or I simply get swamped. Reminders do not offend me.

I will respond to most of the emails regarding the course, and this is the best way to work through simple questions. Please check your email and Canvas several times a week. Email is one of the best ways to keep in touch with our class when we are not in class. More complex questions would likely require more time, and for these, I recommend my office hours.

Textbooks

- **Required**
  1. Long, Scott. 1997. *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, CA: Sage. (Hereafter Long)
  2. Powers, Daniel and Yu Xie. 2008. *Statistical Methods for Categorical Data Analysis*, Second Edition. Howard House, England: Emerald. (Hereafter Powers & Xie)
- **Optional**
  1. Agresti, Alan. 2007. *An Introduction to Categorical Data Analysis*. Wiley.
  2. Agresti, Alan. 2013. *Categorical Data Analysis*. 3rd Edition. Wiley.
  3. Burger, Scott. 2018. Introduction to Machine Learning with R: Rigorous Mathematical Analysis.
- **Recommended**
  1. Long, J. Scott. 2009. *The Workflow of Data Analysis using Stata*. Stata Press.
  2. Miller, Jane E. 2005. *The Chicago Guide to Writing about Multivariate Analysis*. University of Chicago Press.

Required Software

- **R**
- **RStudio** is an integrated development environment for **R**, a programming language for statistical computing and graphics.
  - **R** tutorial courses on **DataCamp**
  - More tutorials in TA sessions

Class Requirements and Evaluation

1) Weekly Problem Sets (30% of your final grade)

Problem Sets will be due in class the week after they are assigned on **Thursday** at **3:30** pm. Please submit both a paper copy (hand in to TA) and upload an electronic version to Canvas. Any programming language is accepted for the sexercises. **If students have any questions on Problem Sets they should first ask TA and only ask the professor if the TA is unable to help.**

There are **9** assignments in the semester (except for advanced topics). It is important that you do each set of weekly assignment completely and on time; **late submissions will not be accepted**. If for some reason you do not complete your assignment on time, I encourage you to complete it on your own, but we will not accept it for credit. **To compensate for this strict policy, I will drop the lowest grade you receive on an assignment when we tabulate your overall grade.** In the first few weeks, you will be doing analyses using a major U.S. national sample survey (e.g., NORC's General Social Survey). As the quarter progresses, however, for most of the assignments you will be able to substitute data of your own, focusing on topics that interest you and/or that pertain to your term paper.

2) Midterm Exam (30% of your final grade)

The mid-term exam will take place on **March 31, 2022** from 3:30 pm to 5 pm. The exam is open-book.

3) Final Paper (and Term Paper Proposal) (30% of your final grade)

The course will culminate in a term paper on a topic of your choosing in which you will carry out a quantitative analysis of some substantive issue using the technical and analytic skills developed by doing the assignments. It is not uncommon for course term papers to lead to or revise master's papers or chapters of Ph.D. dissertations and/or publications.

With instructor's prior approval, you may write co-authored papers with **no more than two** authors. Both authors must be students in the class. In the case of co-authorship, the paper should detail what each author contributes to the project and include a separate paragraph or document detailing what each author contributed.

Your final term paper will be due at the end of the semester on **May 10, 2022, 5 pm**. Late papers will not be accepted. More information on this project will be distributed over the semester.

4) Class Presentation (10% of your final grade)

Each student is required to present their final paper in the student presentation conference during the last week of the class. In addition, each student needs to either present on an advanced topic or serve as a discussant in the student presentation conference.

Course Policies

During Class

I understand that the electronic recording of notes will be important for class and so computers will be allowed in class. Please refrain from using computers for anything but activities related to the class. Phones are prohibited as they are rarely useful for anything in the course. Eating is not allowed in class due to covid; drinking is allowed but you will need to keep your face covering.

After Class

I expect you to spend **3** hours in class and at least **6-8** hours after class each week on the course subjects.

Office Hours

The scheduled office hours are on Tuesday from 12 pm to 1 pm and Tuesday and Thursdays from 5 pm to 6 pm. However, due to my travel schedules, I have to cancel the office hours on **February 8, 15, April 5, 19**. For those weeks, I will email the class about my new office hours or you can send me an email to schedule separate meetings.

Policies on Incomplete Grades and Late Assignments

Late assignments will not be accepted. See the policy discussed about weekly problem sets.

Grading Policies

The typical UPenn grading scale will be used. Normally, grading will not be on a curve. You can access your personal grades on the course web page as we move along in the course. Your final course grade will be figured according to the following cutoffs:

A = 94 – 100	C = 73 – 76
A- = 90 – 93	C- = 70 – 72
B+ = 87 – 89	D+ = 67 – 69
B = 83 – 86	D = 63 – 66
B- = 80 – 82	D- = 60 – 62
C+ = 77 – 79	F = 59 and Below

However, if no one receives higher than 90+, I reserve the right to curve the scale dependent on overall class scores at the end of the semester. Any curve will only ever make it easier to obtain a certain letter grade.

Canvas

You can download all course materials from the course Canvas website:

https://canvas.upenn.edu/courses/1637391

Plan of Lecture

- Basic Concepts and Introduction to R
- The Logit Model for Binary Outcomes
- Simple Models and Association Measures for Two-Way Contingency Tables
- Models and Association Measures for Multi-Way Tables
- Models for Ordinal Dependent Variables
- Multinomial Logit Model
- Mid-Term Examination (Open Book)
- Tobit Regression for Censoring and Truncation
- Poisson Models for Count Data
- Conditional Logit Model
- Group-Based Trajectory Models
- Sequence Analysis
- Tree-Based Methods and Random Forest

Class Schedules (Subject to Change)

Topic 1 (January 13 & 18): *Basic Concepts and Regression Review*

- Chapter 1 Introduction in Powers & Xie.
- Chapter 2 Review of Linear Regression Models in Powers & Xie

Topic 2 (January 20 & 25): *The Logit Model for Binary Outcomes*

- Chapter 3 Models for Binary Data in Powers & Xie.
- For sociological examples, see Chapter 13 in Treiman

Special Topic (January 27): ***Introduction to R and Social Science Computing***

- Guest speaker: Hunter York (PhD student, Princeton University)

Topic 3 (February 1 & 3): ***Simple Models and Association Measures for Two-Way Contingency Tables***

- Chapter 4.1–4.4 in Powers & Xie

Topic 4 (February 8 & 10): ***Models and Association Measures for Multi-Way Tables***

- Chapter 4.6 in Powers & Xie
- Mare, Robert D. 1991. "Five Decades of Educational Assortative Mating." *American Sociological Review* 56:15-32.

Topic 5 (February 15 & 17): ***Models for Ordinal Dependent Variables***

- Chapter 7 in Powers & Xie
- Chapter 5 in Long

Topic 6 (February 22 & 24): ***Multinomial Logit Model***

- Chapter 8.1–8.5 in Powers & Xie
- Chapter 6 in Long
- Anderson, John A. 1984. "Regression and Ordered Categorical Variables." *Journal of the Royal Statistical Society: Series B (Methodological)* 46(1): 1–22.
- Williams, Richard. 2006. "Generalized Ordered Logit/Partial Proportional Odds Models for Ordinal Dependent Variables." *The Stata Journal* 6(1): 58–82.

Topic 7 (March 1 & 3): ***Tobit Regression for Censoring and Truncation***

- Chapter 7 in Long
- Glass, Jennifer, and Jerry Jacobs. 2005. "Childhood Religious Conservatism and Adult Attainment Among Black and White Women." *Social Forces* 84(1): 555–579.
- Sayer, Liana C., Suzanne M. Bianchi, and John P. Robinson. 2004. "Are Parents Investing Less in Children? Trends in Mothers' and Fathers' Time with Children." *American Journal of Sociology* 110(1): 1–43.
- Steelman, Lala Carr, and Brian Powell. 1991. "Sponsoring the Next Generation: Parental Willingness to Pay for Higher Education." *American Journal of Sociology* 96(6): 1505–1529.

March 5–13 Spring Break

Topic 8 (March 15 & 17): *Poisson Models for Count Data*

- Chapter 8 in Long
- Song, Xi, Cameron Campbell, James Z. Lee. 2015. "Ancestry Matters: Patrilineage Growth and Extinction." *American Sociological Review* 80(3): 574–602.

Topic 9 (March 22 & 24): *Conditional Logit Model*

- Chapter 8.6–8.7 in Powers & Xie
- Zeng, Zhen, and Yu Xie. 2008. "A Preference-Opportunity-Choice Framework With Applications to Intergroup Friendship." *American Journal of Sociology* 114: 615–648.

Mid-Term Review & Examination (March 29 & 31): *Open Book*

- The exam question answers will be discussed in the next week's class.

Advanced Topic (April 5 & 12): *Group-Based Trajectory Models*

- Jones, Bobby L. and Daniel S. Nagin. 2007. "Advances in Group-Based Trajectory Modeling and a SAS Procedure for Estimating Them." *Sociological Methods and Research* 35:542–571.
- Nagin, Daniel S. 2005. *Group-Based Modeling of Development*. Boston, MA: Harvard University Press.

No class on April 7 (PAA).

Advanced Topic (April 14): *Sequence Analysis*

- Abbott, Andrew. 1990. "A Primer on Sequence Methods." *Organization Science* 1:373–92.
- Abbott, Andrew. 1995. "Sequence Analysis: New Methods for Old Ideas." *Annual Review of Sociology* 21:93–113.
- Abbott, Andrew and John Forrest. 1986. "Optimal Matching Methods for Historical Sequences." *Journal of Interdisciplinary History* 16:471–94.
- Abbott, Andrew and Angela Tsay. 2000. "Sequence Analysis and Optimal Matching Methods in Sociology." *Sociological Methods & Research* 29(1):3-33.
- Aisenbrey, Silke, and Anette E. Fasang. 2010. "New Life for Old Ideas: The 'Second Wave' of Sequence Analysis Bringing the Course Back into the Life Course." *Sociological Methods & Research* 38.3: 420–462.
- Frye, Margaret, and Jenny Trinitapoli. 2015. "Ideals as Anchors for Relationship Experiences." *American Sociological Review* 80(3): 496–525.

- Humphries, John Eric. 2018. “The Causes and Consequences of Self-Employment over the Life Cycle" working paper.

Advanced Topic (April 19): Tree-Based Methods and Random Forest for Categorical Data

- Chapter 6 in Burger.
- Brand, Jennie E., Jiahui Xu, Bernard Koch, and Pablo Geraldo. 2021. “Uncovering Sociological Effect Heterogeneity Using Tree-Based Machine Learning." *Sociological Methodology*: 0081175021993503.

No class on April 21 (Both the instructor and TA will be away for a conference).

Final project proposal due.

Student Presentation Conference (April 26 3:30-6:30 pm)

Advanced Topics

We will cover a few advanced topics on group-based trajectory models, sequence analysis, and machine learning for categorical data. Students (individuals or groups of students) will discuss a topic in depth for the entire group. The instructor will work with the students in advance of their presentations, reviewing and participating in the formal presentations.

Student Presentation Conference

Overview

All students or research teams will need to give a final presentation about their research projects during the last class on **April 26 3:30 om to 6:30 pm**. We will divide all class presentations into 3-4 sessions. Each session will include 1 session chair, 1-2 discussants, and 4 presenters. Presenters will be asked to send their final project proposal to the session chair, discussant(s) and other presenters well in advance of the session.

Presentation Preparation

It is imperative that discussants receive your proposal no later than **April 22, 2022**, so there is time to read all papers and prepare comments; some chairs/discussants may allow a later date, but this must be directly arranged and not assumed. Please be sure to send your proposal to the session chair and other presenters when you send it to the discussant(s); this will enhance the exchange of ideas within the session.

Time

Following the rule of PAA, most presentations will only be 12 minutes long. The chair of your session will inform you of the exact time allocated to your presentation; if you have not heard from the session chair, assume that you will have 12 minutes. The time you are allotted will not include Q&A unless otherwise indicated by your chair; typically, session chairs reserve 10 or so minutes at the end of each session for questions and discussion.

Plan For Your Presentation

Below are some presentation tips provided by PAA. A good conference presentation provides a clear and succinct overview of your paper. Consider the time available and the multiple learning styles of attendees (auditory, visual, etc.) to create a valuable presentation.

- *Prepare visual aids*: Most presenters use slides, either in PowerPoint or PDF, as visual aids for their presentation.
- *Type*: Use at least 20-point type so that audience members can easily read the print on your slides. Please do not include large tables in your slides: Summarize your key results rather than presenting large, dense tables.
- *Bullets*: Limit yourself to 3-4 bullets per slide and 10 or fewer words per bullet. *Number*: A rough rule of thumb is to prepare no more than one slide for every minute you will be presenting.
- Try to avoid the use of acronyms, jargon, and abbreviations: Past conference evaluations have clearly indicated that one frustration, in particular for new and international attendees, is the use of 'insider' language, acronyms, and abbreviations that make it difficult to comprehend a presentation.

- Consider livening up your slides with graphics and pictures: Graphics can be very effective in capturing the audience's attention and focusing them on the point you want to make.
- Contact information slide: Include a slide that you put up at the beginning with your presentation title, name, and contact information.
- Please proof read and spell check.
- Practice: Practice your presentation to ensure that it highlights key points, your delivery is clear, and you finish within the time allocated.
- Email your slides: Email your slides to the session chair in advance of the session in case there are any difficulties with screen sharing.

Final Paper Instruction

You will develop a research paper during the semester. Instead of just memorizing methodological concepts and techniques, you should be able to apply them to an actual research project. The paper should include the objective of the research, a short literature review and one or two hypotheses, variables for study, and measurement. During the latter part of the quarter, you will have to conduct data analyses to test your hypothesis or support your argument. You can either bring your own social science data or use one module of General Social Surveys for this research paper.

I strongly encourage you to discuss your research ideas and plan of analysis with me as early as possible during the semester. You are also encouraged to work with a classmate (no more than 2 authors on a paper) on this final project. In fact, a lot of my own work is collaborative and several are derived from my collaborations with classmates during graduate school (the papers were not published until many years later or still unpublished). In the case of co-authorship, the paper should detail what each author contributes to the project and include a separate paragraph or document detailing what each author contributed.

The final paper should be typed and at most **10** pages of 12 pt, including any text, tables, and figures. **I will only read the first 10 pages if your paper exceeds the page limit.** You should also include references, R/Stata commands, and appendices, but these materials are not counted toward the page limit. Please include margins of at least one inch on all sides of the paper. **It should be submitted via the Canvas website and delivered to my mailbox in McNeil 353**.

The report should be self-contained and suitable for a non-statistician with a college level of knowledge of statistics.

Advanced modeling methods should be defined briefly in the text to the extent necessary for understanding of the results, along with a reference. By advanced I mean methods that were not prerequisite for the course.

It is recommended that the following outline be followed in preparing the report:

1. *Abstract*. This should consist of a brief statement of the results of your analysis. This should be like that of a research paper analyzing the data.

2. *Introduction*. Here includes a clear statement of the scientific questions addressed by your analysis of the data. The goal of the statistical analysis and the social context of the problem should be clear to all who read the introduction.

3. *Analysis and Results*. Describe your analysis and its results clearly and concisely. If necessary, use graphical displays and tables to convey the results. State clearly what your research and null hypotheses and underlying assumptions (in terms of variable type, distribution, sampling, target population, etc) are. Describe methods used, approaches taken to examine the underlying assumptions and so on. Explain why the methodology is appropriate.

4. *Discussion*. This section should describe the sociological and statistical issues raised by the results described in the previous section. Limitations of the study and of the analysis should be discussed here. There may be social policy issues that you would have liked to discuss with the investigator if this had been a real collaboration with a non-statistical social scientist. If so, describe these issues and why they would be relevant to the analysis and/or interpretation of results. If appropriate, provide suggestions for further analysis or collection of additional data. Summarize your conclusions about the issue of sociological concern.

5. *References*. List the key books and articles you consulted that are reflected in your report, as well as references that might be useful to the reader if they want to know more.

6. *Tables and Figures*. These should be included in the text, but can also be included in a separate section at the end. All should have clear titles/captions, and figures should have explanatory legends.

7. *Appendices*. There can be one or more appendices. You may include more technical discussion of your methodology or any theory developed to implement your models. Appendices are not required and should be included only if you feel they add something important to your report.

You should spend the majority of your time thinking about the sociological and statistical issues and writing the report rather than spending all your time carrying out the statistical analysis. You may do a wonderful job of analysis, but it is of no use unless you can communicate the results to your audience.

If you decide to write your paper in LaTeX, you can use my template.

Evaluation

Criteria used to judge performance will include the following three factors, each given equal weight:

1. *Statistical appropriateness*. Appropriateness of the analysis and models for the data and questions. Technical execution of the analysis.

2. *Scientific appropriateness*. Thoughtfulness and simplicity of your analysis. Does your analysis really answer the sociological questions of interest?

3. *Quality of the written report*. The report will be judged based upon its organization, clarity, and accuracy. Simple, concise sentences are preferred to sentences that are convoluted or otherwise confusing.

You can find information about writing and editing service provided by the university in the course syllabus.

Choice of your data set

If you decide to use your own data, the data set should contain at least 30 cases, and at least four variables measured on each case. At least one of the variables should be an outcome variable that you want to predict or model via the other variables. All data sources should be cited and described. If you have questions about the selection of data, you can come to talk with me. Do you merely use examples from the lectures—the social world is an interesting place!

Your statistical task is to model the structure in the data and describe it. To do this, you should consider the various forms of social structure considered in the course.

Here are some suggestions:

- Start by presenting visual summaries of the data, followed by numerical summary measures.
- Describe the impact of the covariates on the outcome variable(s) via simple models first.
- For any model be sure to consider diagnostics of the appropriateness of the model.
- Construct summaries of the overall quality of the fit, and a comparison between alternatives to the model finally used.

This is an open-ended question, so feel free to experiment.