PSY 5939: Categorical Data Analysis
Spring 2023 Syllabus

Last updated January 05, 2023

Instructor Information

- **Instructor**: Dr. Stefany Coxe
  - Ph.D., Quantitative Psychology from Arizona State University
    * I **evaluate** and **apply** advanced *statistical methods* in *behavioral research*. I am especially interested in regression models for **categorical outcomes** and **statistical graphics**, but I also do a lot of other things.
  - **Email** (stefany.coxe@fiu.edu) is **always** the best way to contact me! If you don't hear back from me in a couple days, send me another email
  - **Student Hours** (a.k.a. "office hours") are time for **you to meet with me** and talk about class (or other things like non-class analyses or papers or anything else). Email me to arrange a time to meet on Zoom (link also on Canvas)

Course Information

Time and Location

- **Thursday** from *9:30am - 10:45am* in **SIPA 103**

Learning Goals

This course covers topics related to statistical analysis of **categorical outcome variables**, focusing on methods used in the social sciences. Topics include the **generalized linear model** (GLiM, including *logistic regression* and *Poisson regression*) and **repeated measures extensions** of GLiM (such as *generalized estimating equations* and *generalized linear mixed models*). You will be able to analyze, interpret, and write up results using these methods.

Learning Objectives

- **Develop** research questions about categorical outcome variables
- **Select** the appropriate analysis approach for the research question
- **Analyze** categorical outcomes with regression-based statistical models appropriate to the research question
- **Interpret** statistical analysis output from common statistical software packages
- **Create** a written report of your findings
- **Make conclusions** about your research question(s) based on those results

Prerequisites

- Graduate coursework in analysis of variance and linear regression (PSY 5939: Quant 1 and PSY 5939: Quant 2) and multivariate statistics (PSY 5246C).

Software and Technology

- **Canvas**

Course materials will be posted on Canvas. Lecture videos will be posted on **Playposit**, with links in Canvas. You will submit all assignments via Canvas.

- **Statistical software**

We will use **R** for this course. I will sometimes provide syntax for **SPSS** as well, but all assignments should be completed in R. I hope that you're able to do things like open datasets, transform variables, and conduct linear regression in R, but don't worry if not – I'll do some intro material the first week. I will provide information about the specific procedures you will need to know for this course.

Course Structure

- **This is a hybrid course**
  - We will meet in person for 1 hour 15 minutes each week.
  - You will complete other tasks each week both *before* and *after* the in-person meeting
- **Each week will follow a similar structure:**
  - Before class (Monday through Wednesday)
    * Watch lecture **videos**
    * Start reading the week's **article** (if applicable)
    * This will get you prepared to **participate** during class
  - During class (Thursday)
    * Answer **questions** you had while watching lecture
    * Run **analyses** to see how the methods work and **interpret** the results
  - After class (Thursday through Sunday)
    * Complete **homework** (4 weeks) or **article discussion** (other weeks)

Assessments

Your work in this course will be regularly assessed using a variety of methods.

- **Lecture Quizzes (10%)**

Questions will be embedded within the lecture video. These questions will assess *whether you understand some key points* in the lecture. You will have **unlimited** attempts to answer each question. Getting at least 80% of the questions correct will give you full credit for the assignment.

- **Article Discussions and Reflections (15%)**

We will use *Perusall* to conduct group article discussions. I will provide prompts / direction; you will annotate the article with questions and comments. These are designed to give you some experience reading and understanding *quantitative methods articles* (which can be really hard to read) and to get you to think about the topics in more detail. You will also write up a short (~250 word) **reflection** on the article.

- **Homework (40%)**

There will be four (4) homework assignments covering each of the broad topics we will cover. The assignments involve running *analyses*, making some *decisions* based on the analyses, *interpreting* output, and presenting the *results*.

- **Final Project**

You will propose and conduct a project using your own dataset or a publicly available dataset, culminating in a short paper. I want you to focus on developing research questions about **categorical outcomes** and mapping them on to **appropriate analyses**.

- **Proposal (5%)**

You will turn in a 1 to 2 page proposal for your project. The purpose of the proposal is to get you to **select a dataset**, start to **solidify your ideas**, and **get feedback** and additional resources from me. *You can change the direction of the project later in the semester as you learn more.*

- **Presentation (10%)**

A short presentation about your final project. There are two main purposes to the presentation. First, while I expect that your analyses should be complete (or nearly so) at this point, preparing the presentation should help you **organize your thoughts for the paper**. Second, the presentation will give you **practice presenting your analysis findings in a group setting** and give you a chance to get feedback. Approximately 15 to 20 minutes per person, to be recorded and uploaded to Canvas.

- **Presentation Discussion (5%)**

Each student should ask a question of at least 2 other students about their presentations. The original student should attempt to answer the questions. (Feel free to have further discussion as well!)

- **Final Paper (15%)**

The final written record of your project. This should be in the style of a journal article, with Introduction, Methods, Results, and Discussion sections.

Grades

Grade	Percentage
A	>=93
A-	90 - 92.99
B+	87 - 89.99
B	83 - 86.99
B-	80 - 82.99
C+	77 - 79.99
C	70 - 76.99
F	<= 69.99

Tentative Schedule

Week	Topic	L	R	H	S
Jan 09	Intro, GLiM	1	1		
Jan 16	Logistic regression	2	2		
Jan 23	Logistic regression	3	3		
Jan 30	Ordinal / multinomial	4		1	
Feb 06	Poisson regression	5	4		
Feb 13	Poisson regression	6	5		
Feb 20	GLiM wrap-up	7		2	
Feb 27	SPRING BREAK				
Mar 06	Contingency tables	8			Proposal
Mar 13	Contingency tables	9	6		
Mar 20	Repeated measures	10		3	
Mar 27	Repeated measures	11	7		
Apr 03	Repeated measures	12	8		
Apr 10	Meetings			4	
Apr 17	Presentations				Presentation
Apr 24	Finals				Paper

Due dates subject to change due to hurricane, emergency, scheduling changes, etc.

- **L**: Watch lecture videos by **Wednesday at 8pm**.
- **R** or **H**: Reflections and homework assignments are due **Sunday** by midnight
- **S**: Special assignments
  - **Proposal** due *March 12* by midnight
  - **Presentation** due *April 23* by midnight
  - **Presentation discussion** due *April 26* by midnight
  - **Final paper** *due April 28* by midnight

Course and University Policies

Attendance

Attendance is not explicitly part of your grade in this course, but activities completed during the in-person portion of the course will be **very** helpful.

If you need to miss class (such as for illness, religious event, professional activity, university-sanctioned event, or **any other reason**), please contact me to make any necessary arrangements.

Accessibility / Accomodation

Any student with a disability or other need that may require special accommodations for this course should make this known to the instructor during the first week of class. You can contact the Disability Resource Center at

- http://drc.fiu.edu
- drcupgl@fiu.edu
- 305-348-3532
- Graham Center 190

Academic Dishonesty

Please refer to your student handbook for a description of what constitutes academic dishonesty. *While you may work with other students on your homework assignments, I expect all students to complete and turn in their own work.*

Academic Misconduct

Students at Florida International University are expected to adhere to the highest standards of integrity in every aspect of their lives. Honesty in academic matters is part of this obligation. Academic integrity is the adherence to those special values regarding life and work in an academic community. Any act or omission by a student which violates this concept of academic integrity shall be defined as academic misconduct and shall be subject to the procedures and penalties set forth herein. All students are expected to adhere to a standard of academic conduct, which demonstrates respect for themselves, their fellow students, and the educational mission of the University. All students are deemed by the University to understand that if they are found responsible for academic misconduct, they will be subject to the Academic Misconduct procedures and sanctions, as outlined in the Student Handbook.

https://dasa.fiu.edu/all-departments/student-conduct-and-academic-integrity/

References

General linear models (GLMs)

Barker, L. E., & Shaw, K. M. (2015). Best (but oft-forgotten) practices: checking assumptions concerning regression residuals. The American journal of clinical nutrition, 102(3), 533-539.

Cohen, J., Cohen, P., West, S.G. & Aiken, L.S. (2003). Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences. L. Erlbaum Associates, Mahwah, N.J.

Gelman, A., & Hill, J. (2006). Data analysis using regression and multilevel/hierarchical models. Cambridge University Press.

Hickey, G. L., Kontopantelis, E., Takkenberg, J. J., & Beyersdorf, F. (2019). Statistical primer: checking model assumptions with regression diagnostics. Interactive cardiovascular and thoracic surgery, 28(1), 1-8.

Kozak, M., & Piepho, H. P. (2018). What's normal anyway? Residual plots are more telling than significance tests when checking ANOVA assumptions. Journal of agronomy and crop science, 204(1), 86-98.

Osborne, J. W., & Waters, E. (2002). Four assumptions of multiple regression that researchers should always test. Practical assessment, research, and evaluation, 8(1), 2.

Generalized linear models (GLiMs): General

Agresti, A. (2003). Categorical data analysis (Vol. 482). John Wiley & Sons.

Agresti, A. (2018). An introduction to categorical data analysis. John Wiley & Sons.

Ai, C. & Norton, E. C. (2003). Interaction terms in logit and probit models. Economics Letters, 80 (1), 123–129. doi:10.1016/S0165-1765(03)00032-6

Dobson, A. J., & Barnett, A. G. (2018). An introduction to generalized linear models. Chapman and Hall/CRC.

Faraway, J. J. (2016). Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models. Chapman and Hall/CRC.

Fox, J. (2015). Applied regression analysis and generalized linear models. Sage Publications.

Geldhof, G. J., Anthony, K. P., Selig, J. P., & Mendez-Luck, C. A. (2018). Accommodating binary and count variables in mediation: A case for conditional indirect effects. International Journal of Behavioral Development, 42(2), 300-308.

Green, P., & MacLeod, C. J. (2016). SIMR: an R package for power analysis of generalized linear mixed models by simulation. Methods in Ecology and Evolution, 7(4), 493-498.

Halvorson, M. A., McCabe, C. J., Kim, D. S., Cao, X., & King, K. M. (2022). Making sense of some odd ratios: A tutorial and improvements to present practices in reporting and visualizing quantities of interest for binary and count outcome models. Psychology of Addictive Behaviors, 36(3), 284.

Hardin, J. W. & Hilbe, J. M. (2007). Generalized linear models and extensions. Stata press.

Long, J. S. (1997). Regression models for categorical and limited dependent variables (Vol. 7). Advanced quantitative techniques in the social sciences, 219.

McCabe, C. J., Halvorson, M. A., King, K. M., Cao, X., & Kim, D. S. (2020). Interpreting interaction effects in generalized linear models of nonlinear probabilities and counts. Multivariate Behavioral Research, 1-27.

McCullagh, P., & Nelder, J. A. (2019). Generalized linear models. Routledge.

Ng, V. K., & Cribbie, R. A. (2017). Using the gamma generalized linear model for modeling continuous, skewed and heteroscedastic outcomes in psychology. Current Psychology, 36(2), 225-235.

Norton, E. C., Wang, H., & Ai, C. (2004). Computing interaction effects and standard errors in logit and probit models. The Stata Journal, 4 (2), 154–167.

Smithson, M., & Merkle, E. C. (2013). Generalized linear models for categorical and continuous limited dependent variables. CRC Press.

Logistic regression

Allison, P. D. (2012). Logistic regression using SAS: Theory and application. SAS Institute.

Chen, K., Cheng, Y., Berkout, O., & Lindhiem, O. (2016). Analyzing Proportion Scores as Outcomes for Prevention Trials: A Statistical Primer. Prevention Science, 1-10.

DeMaris, A. (2002). Explained variance in logistic regression: A Monte Carlo study of proposed measures. Sociological Methods & Research, 31(1), 27-74.

Hayes, A. F., & Matthes, J. (2009). Computational procedures for probing interactions in OLS and logistic regression: SPSS and SAS implementations. Behavior research methods, 41(3), 924-936.

Long, J. S., & Mustillo, S. A. (2021). Using predictions and marginal effects to compare groups in regression models for binary outcomes. Sociological Methods & Research, 50(3), 1284-1320.

Menard, S. (2002). Applied logistic regression analysis (No. 106). Sage.

Mood, C. (2010). Logistic regression: Why we cannot do what we think we can do, and what we can do about it. European sociological review, 26(1), 67-82.

Ordinal and multinomial logistic regression

Bürkner, P. C., & Vuorre, M. (2019). Ordinal regression models in psychology: A tutorial. Advances in Methods and Practices in Psychological Science, 2(1), 77-101.

Hedeker, D. (2015). Methods for multilevel ordinal data in prevention research. Prevention Science, 16(7), 997-1006.

Liddell, T. M., & Kruschke, J. K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong?. Journal of Experimental Social Psychology, 79, 328-348.

Poisson regression

Atkins, D. C., & Gallop, R. J. (2007). Rethinking how family researchers model infrequent outcomes: a tutorial on count regression and zero-inflated models. Journal of Family Psychology, 21(4), 726.

Blevins, D. P., Tsang, E. W., & Spain, S. M. (2015). Count-Based Research in Management Suggestions for Improvement. Organizational Research Methods, 18(1), 47-69.

Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., … & Bolker, B. M. (2017). Modeling zero-inflated count data with glmmTMB. BioRxiv, 132753.

Campbell, H. (2021). The consequences of checking for zero-inflation and overdispersion in the analysis of count data. Methods in Ecology and Evolution, 12(4), 665-680.

Coxe, S., West, S. G., & Aiken, L. S. (2009). The analysis of count data: A gentle introduction to Poisson regression and its alternatives. Journal of personality assessment, 91(2), 121-136.

Gardner, W., Mulvey, E. P., & Shaw, E. C. (1995). Regression analyses of counts and rates: Poisson, overdispersed Poisson, and negative binomial models. Psychological bulletin, 118(3), 392.

Green, J. (2020). A tutorial on modelling health behaviour as count data with Poisson and negative binomial regression.

Land, K. C., McCall, P. L., & Nagin, D. S. (1996). A comparison of Poisson, negative binomial, and semiparametric mixed Poisson regression models with empirical applications to criminal careers data. Sociological Methods & Research, 24(4), 387-442.

Yang, S. (2014). A comparison of different methods of zero-inflated data analysis and its application in health surveys. University of Rhode Island.

Contingency tables

Bradley, D. R., Bradley, T. D., McGrath, S. G., & Cutcomb, S. D. (1979). Type I error rate of the chi-square test in independence in R x C tables that have small expected frequencies. Psychological Bulletin, 86(6), 1290.

Camilli, G., & Hopkins, K. D. (1978). Applicability of chi-square to 2 x 2 contingency tables with small expected cell frequencies. Psychological Bulletin, 85(1), 163.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. Journal of the Royal Statistical Society: Series B (Methodological), 13(2), 238-241.

Tu, Y. K., Gunnell, D., & Gilthorpe, M. S. (2008). Simpson's Paradox, Lord's Paradox, and Suppression Effects are the same phenomenon–the reversal paradox. Emerging themes in epidemiology, 5(1), 1-9.

Yule, G. U. (1903). Notes on the theory of association of attributes in statistics. Biometrika, 2(2), 121-134.

Repeated measures

Aiken, L. S., Mistler, S. A., Coxe, S., & West, S. G. (2015). Analyzing count variables in individuals and groups: Single level and multilevel models. Group Processes & Intergroup Relations, 18(3), 290-314.

Archer, K. J., Hedeker, D., Nordgren, R., & Gibbons, R. D. (2015). mixor: an R package for longitudinal and clustered ordinal response modeling.

Devine, S., Otto, A. R., Uanhoro, J. O., & Flake, J. K. (2022). A Tutorial for Quantifying Within-and Between-Participant Variance in Multilevel Logistic Models.

Hedeker, D. (2003). A mixed-effects multinomial logistic regression model. Statistics in medicine, 22(9), 1433-1446.

Hedeker, D. (2005). Generalized linear mixed models. Encyclopedia of statistics in behavioral science.

Hu, F. B., Goldberg, J., Hedeker, D., Flay, B. R., & Pentz, M. A. (1998). Comparison of population-averaged and subject-specific approaches for analyzing repeated binary outcomes. American Journal of Epidemiology, 147(7), 694-703.

Hubbard, A. E., Ahern, J., Fleischer, N. L., Van der Laan, M., Lippman, S. A., Jewell, N., Bruckner, T., & Satariano, W. A. (2010). To GEE or not to GEE: comparing population average and mixed models for estimating the associations between neighborhood risk factors and health. Epidemiology, 21(4), 467-474.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. Journal of memory and language, 59(4), 434-446.

Leckie, G., Browne, W. J., Goldstein, H., Merlo, J., & Austin, P. C. (2020). Partitioning Variation in Multilevel Models for Count Data. Psychological Methods.

Lee, Y., & Nelder, J. A. (2004). Conditional and marginal models: another view. Statistical Science, 19(2), 219-238.

Stroup, W. W. (2012). Generalized linear mixed models: modern concepts, methods and applications. CRC press.