Global School in Empirical Research Methods

Categorical Data Analysis
4-8 June 2018
Shawna N. Smith
University of Michigan

Course Overview:
By nature or by measurement, dependent variables of interest to social and behavioral scientists are frequently non-continuous. Outcomes like vote choice, condom use, and/or number of Instagram followers necessarily violate key assumptions of the linear regression framework and require other model estimation strategies. Categorical Data Analysis (CDA) is an applied statistics course that focuses on the estimation and interpretation of nonlinear effects for non-continuous outcomes. While nonlinear models are often conceptualized as extensions of the linear framework (i.e., 'generalized linear models'), common analytic questions related to both inference and predictive analytics (e.g., forecasting, machine learning) require different methods when the outcome is not linear—e.g., interpretation of coefficients, calculation of posterior predictions/classifications, assessing model fit, testing for significance, comparing coefficients across models, and interpreting interaction effects, to name a few. In fact, using techniques that assume a linear effects framework when effects are nonlinear puts the scientist at risk of drawing incorrect conclusions and/or interpretations.

CDA is designed to ensure you don't make these mistakes. We begin by considering the general objectives for interpreting the results of any regression-type model and then considering why these objectives are more complicated with nonlinear models. Basic concepts (including the idea of identification) and notation are introduced through a short review of the linear regression model, and a short overview of the method of maximum likelihood estimation. From there, we will derive the logit and probit models for use with binary outcomes, and also introduce a variety of post-estimation tools for interpreting nonlinear models. We will then extend these models and methods of interpretation from binary outcomes to ordinal outcomes using the ordinal logit and probit models, and the multinomial logit model for nominal outcomes. Finally, the course will conclude by introducing a series of models for count data, including Poisson regression, negative binomial regression, and zero-modified variant models.

CDA is designed for the applied analyst; its focus is on teaching you the tools you need for effective model presentation and interpretation. Our primary focus is on regression models for binary, nominal, ordinal, and count dependent variables, however several flexible techniques for evaluating nonlinearities within the linear regression framework will also also covered. Support and code for model estimation and post estimation is provided for both Stata and R.

Course website: http://www.shawnasmith.net/gserm/

Course Schedule:

June 4 (Monday):	9:15-12pm & 1:30-3:15pm
June 5 (Tuesday):	9:15-12pm & 1:30-3:15pm
June 6 (Wednesday):	9:15-12pm & 1:30-3:15pm
June 7 (Thursday):	9:15-12pm & 1:30-3:15pm
June 8 (Friday):	9:15-12pm & 1:30-3:15pm

Location: Seminar room: ______________________
Computer lab: _______________________

Software & Computing:
Models for this course are presented in broad strokes; however, a major component of this course is application through model estimation, post-estimation and interpretation. For pedagogical purposes, I will use Stata 15 in course lectures for model estimation and interpretation, but Stata versions 12 and higher will suffice. Course support will be provided for both Stata and R. While Stata—and most popular statistical software packages—includes native estimation (and even post-estimation) commands for categorical models, we will also use a set of ado files written for Stata by Scott Long & Jeremy Freese that facilitate the (at times complicated) interpretation of categorical models within Stata. This suite of commands is called SPost. These post-estimation commands can also be emulated in R although this will require more investigation on the part of the student. A variety of packages now exist in R relevant to the course that we are happy to provide guidance on when possible. *If you are taking this course for credit, **you will need to complete assignments using either Stata and SPost 13 commands or R with appropriate post-estimation commands***.

N.B.: While the course assumes familiarity with the linear regression model, it does not assume familiarity with Stata or R.

- **Getting Access to Stata/R:**
  - ***Stata:*** Access to Stata is available through the GSERM labs. Several versions are also available to purchase at different price points; I am happy to provide guidance as to which would be most appropriate for your needs.
  - ***R:*** R is free. You can download it at https://www.r-project.org/
    - **R Studio** is a free program that greatly upgrades R's user-interface and can be downloaded at https://www.rstudio.com/

- **Getting Started using Stata:** New to Stata? No worries—this course will catch you up quickly. However, I strongly suggest working through the "Getting Started using Stata" document available on the course website (http://shawnasmith.net/gserm/) prior to Day 1 of class. Feel free to get in touch if you have questions.
  - *New to R?* As R has a steeper learning curve, I would not recommend attempting to learn R solely for the purposes of this course. However, I am happy to recommend resources for those of you so inclined:

        - The two textbooks recommended above provide good introductions for 'getting started' with R, as well as lots of *in situ* code.
        - Mike Marin (UBC Public Health) has a great series of videos introducing R online at http://www.statslectures.com/index.php/r-stats-videos-tutorials/getting-started-with-r.

- **Downloading SPost:** If you will be using Stata on a personal computer, then you will need to install the current SPost suite of commands. Here's the step-by-step:
    - *Pre-reqs:* Internet access & administrative privileges
    - In Stata, type `search spost13` into the command line.
    - In the viewer window that appears, click the link for "spost13_ado from http://www.indiana.edu/~jslsoc/stata"
    - Follow directions to install
    - Double-check install by typing `help mchange` in the command line. If a help window pops up, SPost 13 has been installed correctly.

- **Accessing course data and `usecda`:** Course data will be available for download through the course website. It is also available for us in Stata with the `usecda` command. `usecda` is a command written specifically for this course to expedite access to course datasets & examples. It is currently only available for download through Shawna's Github account. To download on any computer:
    - Tell Stata where to download the file from by using the following command in the Stata command line or in a do-file:
      ```
      net from "https://shawnana79.github.io/data"
      ```
    - Install the program by either: (a) clicking on the blue `usecda` link that appears in the output following the previous command; or (b) using the command: `net install usecda`
    - Check out the help file by typing `help usecda` in the Stata command line

Required Text

Lecture Notes and Lab Guide for Categorical Data Analysis. This coursepack contain copies of the overheads for the lectures and materials used in the computing lab. It will be provided at the beginning of our first class session. Be sure to bring these notes to all lecture and lab sessions.

- For participants that prefer electronic versions, component parts are also available on the course website.

Recommended Texts

Long, J. Scott. 1997. *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, CA: Sage. *Hereafter:* **Long**

Powers, Daniel A. & Yu Xie. 2008. *Statistical Methods for Categorical Data Analysis*. 2nd Edition. Bingley, UK: Emerald Press. *Hereafter:* **P&X**

For the Stata devotees: Long, J. Scott & Jeremy Freese. 2014. *Regression Models for Categorical Dependent Variables Using Stata*. 3rd Edition. College Station, TX: Stata Press. *Hereafter:* **L&F**

Or if you like R: I'm still searching for my favorite here, but a couple of good ones are:

- Monogan, James E. III. 2015. *Political Analysis Using R*. New York, NY: Springer. *Hereafter:* **Monogan**.
- Fox, John & Sanford Weisberg. 2010. *An R Companion to Applied Regression*. Thousand Oaks, CA: Sage. *Hereafter:* **F&W**.

Course Schedule, Suggested Readings & Assignment Due Dates

N.B.: The exact content of the course will vary depending on the background & interests of participants. In other words, this schedule is subject to change.

Day	**Topic**	**Suggested Readings**	**Due**
4 June Monday	• Overview of class; Introduction to models • Review of linear regression; Identification; Maximum Likelihood Estimation (short overview) • Linear probability model; Identification of Pr(y=1); Two philosophies: transformational and latent variable approach for binary outcomes • Estimation of BRM; Odds ratios	• **Long** Ch. 1 • **Long** Ch. 2; **P&X** Ch. 2; **L&F** Ch. 1-2 **F&W** Ch. 1-2 or **Monogan** Ch. 1-2 (R) • **Long** Ch. 3; **P&X** Ch.1	A1: Math Review
5 June Tuesday	• Using Pr(y=1) to interpret the BRM (pt. 1): tables & plots; discrete change • Using Pr(y=1) to interpret the BRM (pt. 2): plots; difference at means vs. mean of difference; partial change/margins • Internal measures of fit; Hypothesis testing; Wald and LR tests; Confidence intervals • Scalar measures of fit: pseudo-R2, AIC, BIC	• **Long** Ch. 4	
6 June Wednesday	• BRM redux: Group differences & interactions • Ordinal variables; a latent variable model • Estimation of ORM; latent variable interpretations; Pr(y=k)	• **Long** Ch. 5; **P&X** Ch. 7	

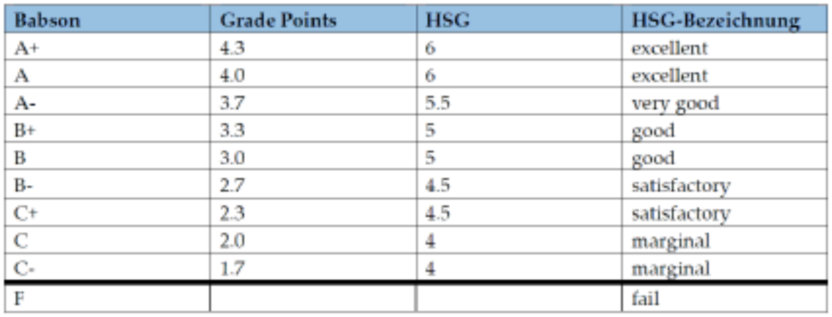
7 June Thursday	• Odds ratios; parallel regression assumption and proportional odds • Multinomial logit as a set of BLMs; IIA • Tests for the MNLM; Calculating predicted probabilities; Interpretation using Pr(y=k) • Odds ratio plots; Discrete change plots • Putting it all together	• **Long** Ch. 6; **P&X** Ch. 8	A2: BRM + T&F
8 June Friday	• Count variables (conceptually) • Poisson process; estimation of PRM; assessing fit; the big idea of heterogeneity; interpretation • Adding unobserved heterogeneity; estimation of NBRM • With-zeros models; zero-modified and zero-inflated models; interpretation • Comparisons among count models • Extensions of models (as requested) • Course wrap-up	• **Long** Ch. 8	
1 July Sunday		A3: ORM & MNLM and A4: Count *due via email to* **shawnana@umich.edu**; include "**GSERMCDA:**" in subject line	

Grading

Participant's overall grades are based on completion of four assignments weighted as follows:

- **A1 Math Review:** 1/6
- **A2 BRM + T&F:** 1/3
- **A3 ORM & MNLM:** 1/3
- **A4 Count:** 1/6

The standard University of St. Gallen (HSG) grading system applies:



Babson	Grade Points	HSG	HSG-Bezeichnung
A+	4.3	6	excellent
A	4.0	6	excellent
A-	3.7	5.5	very good
B+	3.3	5	good
B	3.0	5	good
B-	2.7	4.5	satisfactory
C+	2.3	4.5	satisfactory
C	2.0	4	marginal
C-	1.7	4	marginal
F			fail

Getting Help

I am available to provide feedback or answer questions during lunch breaks & after course hours. Due to the compressed natue of this course (& my desire for you all to digest as much material as possible!), I encourage you to bring up questions or concerns early & often. If you would like to discuss questions or concerns related to categorical data analysis for a particular paper or thesis, I would encourage you to make an appointment to meet before or after lecture one day, or during the lunch break.

- I can also be reached by email both during & after this course at shawnana@umich.edu. Ensure a prompt response to your email by prefacing your subject with **"GSERMCDA:".**

Academic Integrity

It is not possible for us to have an intellectual community without honor. I expect that you demonstrate respect by recognizing the labor of those who create intellectual products. Academic dishonesty (including cheating and plagiarism) will not be tolerated and will be dealt with according to university policy.