Global School in Empirical Research Methods

Categorical Data Analysis
28 August-1 September 2017
Shawna N. Smith
University of Michigan

Course Overview:
Many variables of interest to social, political and behavioral scientists are non-continuous, either through nature or through measurement. Outcomes like vote choice, social class, condom use, and/or number of Facebook friends necessarily violate key assumptions of the simple linear regression framework and require other model estimation strategies. Although advances in software have made estimation of these models trivial, model non-linearities make post-estimation interpretation difficult and require investigators to make choices about which aspects of the data space best represent underlying social dynamics.

The course begins by considering the general objectives for interpreting the results of any regression-type model and then considers why these objectives are more complicated within nonlinear models. Basic concepts 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.

Course Schedule:

August 28 (Monday):	9:30am-12:15pm & 1:15-3pm
August 29 (Tuesday):	9:30am-12:15pm & 1:15-3pm
August 30 (Wednesday):	9:30am-12:15pm & 1:15-3pm
August 31 (Thursday):	9:30am-12:15pm & 1:15-3pm
September 1 (Friday):	9:30am-12:15pm & 1:15-3pm

Location: Seminar room: ______________________
Computer lab: _______________________

Software:
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 14 for model estimation and interpretation; I encourage you to do the same. *N.B.:* While the course assumes familiarity with the linear regression model, it does not assume familiarity with Stata.

We have a computer room available during all class meetings. Some class sessions will be held here so participants can work through lecture notes and do-files while we discuss.

Determinations as to how when such sessions will be held will be determined by course participants and instructor.

Required Text
Lecture Notes and Lab Guide for Categorical Data Analysis. These notes contain copies of all lecture notes and materials used in the computing lab. Be sure to bring these notes to all class sessions.

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**

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**
28 August 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 • **Long** Ch. 3; **P&X** Ch.1	A1: Math Review
29 August 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	

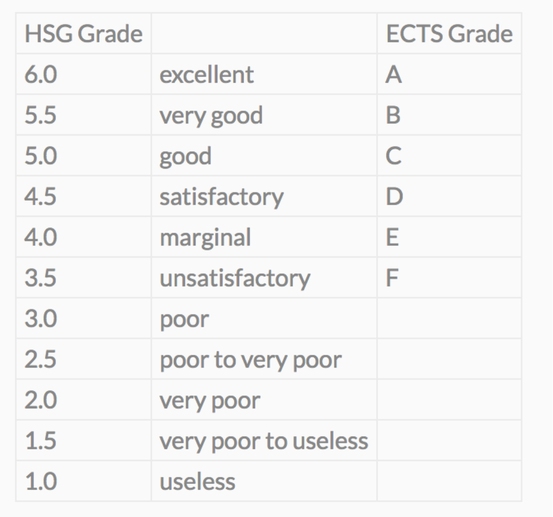
30 August Wednesday	• BRM redux: Group differences & interactions • Ordinal variables; a latent variable model • Estimation of ORM; latent variable interpretations; Pr(y=k)	• **Long** Ch. 5; **P&X** Ch. 7	
31 August Thursday	• Odds ratios; parallel regression assumption and proportional odds • Multinomial 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
1 Sept 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	
15 Sept Friday		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:



HSG Grade		ECTS Grade
6.0	excellent	A
5.5	very good	B
5.0	good	C
4.5	satisfactory	D
4.0	marginal	E
3.5	unsatisfactory	F
3.0	poor	
2.5	poor to very poor	
2.0	very poor	
1.5	very poor to useless	
1.0	useless	

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.