Sociology 952
MATHEMATICAL AND STATISTICAL APPLICATIONS IN SOCIOLOGY
Topic: Models for Categorical and Limited Dependent Variables
Fall, 2016

Instructor: John Allen Logan, 3452 Social Science, 262-0995, logan@ssc.wisc.edu.

Prerequisites: Sociology 361 and 362, or consent of instructor.

Time and place: Thursday, 2:45PM - 4:40PM, Sewell Social Sciences Bldg., room 2435.

Office hours: by appointment.

SUBJECT MATTER

Researchers often collect data in categorical form, either because categories are inherently appropriate for the theoretical variable, or because measurement difficulties prevent the collection of quantitative values. Though categorical independent variables present little difficulty, categorical dependent variables are more challenging.

The course covers the analysis of data where the dependent variables are categorical or partly categorical and partly continuous. Such dependent variables can be dichotomous, polytomous, counted, ordered, censored, and/or subject to selection. (It will be assumed students have had some exposure to the simplest logit and probit models covered in Sociology 362, or equivalent.)

In addition to the usual cross-sectional data in which observations are mutually independent, we will also study models for clustered data, especially panel data in which a single subject is observed several times. *Random effects models* are a natural development for such situations, as we will see. Because categorical panel data models quickly become complex, we also look at *marginal models* that integrate over detailed relationships among observations to provide average predictions.

Characteristic of the models we will study is kind of "unpacking" of the categorical responses we find recorded in datasets. That is, the models don't just treat the categories as found objects, but use particular interpretations of their origins to make sense of them. These interpretations can presume the categories result from the collapsing of unobserved quantitative variables, or from choices made by individuals, for example.

One focus of this course is the sociological context of categorical data; that is, understanding, and if possible adapting methods to, the social processes that determined the responses. A variety of methods involving censored responses, self-selection, and observed choices take this approach (and are covered in the class). The instructor has a research interest in two-sided models that treat both sides of a job or marriage market (or other social system) simultaneously, treating the entire dataset as a collection of mutually dependent observations. One session will be devoted to this kind of model and its estimation, including instruction in the use of a Stata program package, tsmreg. As preparation for this, Bayesian inference via MCMC methods will be covered in a prior session.

Throughout the course we will emphasize the distinction between probability models and estimation methods, and the advantages of learning to express social scientific ideas in probabilistic terms. Empirical analysis will use various forms of estimation (e.g., maximum likelihood, generalized estimating equations, simulated maximum likelihood, Bayesian analysis with Gibbs sampling and other Metropolis algorithms). Depending on students' interests, we may explore the estimation of custom, rather than prepackaged, models using the programming features underlying Stata (which are quite advanced).

The course emphasizes understanding the logic of methods, obtaining familiarity with software, and practical exercises. Students should have some enthusiasm for mathematics as a means of expressing ideas, but are not expected to do complex derivations, etc. Calculus as an active skill is not required, but a basic familiarity with the ideas of derivatives and integrals is desirable.

REQUIRED TEXTS:

Agresti, Alan. An Introduction to Categorical Data Analysis. 2nd ed. 2007. Wiley. Available as a free pdf file here: https://mregresion.files.wordpress.com/2012/08/agresti-introduction-to-categorical-data.pdf .

Kenneth Train, *Discrete Choice Methods with Simulation.* Published by Cambridge University Press. Second edition, 2009. Available as a set of free pdfs here: http://eml.berkeley.edu/books/choice2.html

Breen, Richard. 1996. *Regression Models: Censored, Sample Selected, or Truncated Data*. Quantitative Applications in the Social Sciences 111. Thousand Oaks, CA: Sage.

RESEARCH PAPER

The main course requirement is an empirical research paper that applies some of the methods covered in the seminar to a substantive problem in social science. The paper should include formulation of a research problem, linkage to substantive literature, analysis of suitable data, and discussion of findings. In essence, it should be ready for submission to a journal.

Prohibited paper topics. These types of papers are not permitted:

1. Purely methodological papers;
2. Loglinear mobility model papers, except as part of comparative estimations;
3. Papers where treatment of the dependent variable as categorical is just an artifice to satisfy the course requirement, and treatment as a continuous variable would have been better;
4. Excessively simplistic papers, such as one applying a simple logistic regression model to cross sectional data.

The following article is recommended as a guide to producing a good statistical analysis and research paper: Leland Wilkinson, et al. 1999. "Statistical Methods in Psychology Journals: Guidelines and Explanations." American Psychologist 54: 594-604, which is available online via MadCat. This paper is recommended for advice on organizing your research data: Svend Juul, "Take Good Care of Your Data." www.epidata.dk/downloads/takecare.pdf , but other systems also work well (e.g., J.S. Long's book).

Paper proposal: Students should submit a proposal for the term paper that outlines the problem to be addressed, the data to be used, and the empirical analyses to be undertaken. **Due: October 6, 2014. All proposals must be submitted in PDF format in provided dropboxes at the Learn@UW course web site.**The instructor will respond to proposals with comments if the proposals are turned in on time.

The final paper is due at 11:59 p.m., December 17, 2016 (one week after the last class). All papers must be submitted in PDF format in the course dropbox. Also provide your **supporting computer analysis files**, either as PDF or plain ASCII text appendices to your paper or by posting them where I can get access to them until grades are posted. Keep in mind the instructor does **not** use a Windows computer or Microsoft Word, and won't read your paper if it is in MS Word format.

HOMEWORK

There will generally be homework assignments for each topic. These will be assigned in class and due the next week, unless otherwise specified. There will not be a homework assignment for the last lecture's topic, so the final homework due date will be the last day of classes.

Assignments that are handed in late are penalized by at least one letter grade. No assignments may be handed in after the class session *following* the assigned due date.

COMPUTING

Stata will be the main statistical analysis package, with occasional use of other programs such as SAS as needed. These programs are available on the SSC Linux and/or Windows servers. Class members are eligible for free accounts on these systems. See the SSCC consultants for assistance.

Some class time will be devoted to use of Stata and possibly other programs for model estimation. *No* time, however, will be given to data extraction and management using general packages or dataset-specific extractors.

GRADING

The term paper will determine 60 percent of the final grade. Homework assignments count 30 percent and class participation 10 percent. The lowest homework grade (including any failure to submit a particular homework assignment) will be dropped in calculating final grades.

THE GRADE "INCOMPLETE" WILL NOT BE GIVEN.

SUPPLEMENTARY READINGS

Recommended Supplemental Textbooks

Agresti, Alan. 2013. *Categorical Data Analysis.* 3rd edition. Wiley. *A more detailed treatment of the topics of our Agresti text.*

Ben-Akiva, Moshe, and Steven R. Lerman. 1985. *Discrete Choice Analysis: Theory and Application to Travel Demand.* Cambridge, MA: MIT Press. *Beautiful, clear exposition of discrete choice models and many extensions.*

Freese, Jeremy, and J. Scott Long. Regression Models for Categorical Dependent Variables Using Stata. 3rd. ed., to be published Sept. 30, 2014. *Useful details and special Stata programs for interpreting the results of categorical regression models.*

Hosmer, David W., and Stanley Lemeshow. 2000. *Applied Logistic Regression*. 2nd. Ed. New York: Wiley. *Excellent on diagnostics for this model and its extensions.*

Long, J. Scott. 1997. *Regression Models for Categorical and Limited-Dependent Variables*. Thousand Oaks, CA: Sage. *More theoretical than Freese and Long, and not oriented to particular software, this book is very clear and less dense than our text for the models it covers.*

Maddala, G.S. 1983. *Limited-Dependent and Qualitative Variables in Econometrics.* Cambridge: Cambridge University Press. *Classic development of selection models.*

Molenberghs, Geert, and Geert Verbeke. 2006. *Models for Discrete Longitudinal Data*. Springer. More advanced treatment of all types of clustered categorical data.

Simonoff, Jeffrey S. 2003. *Analyzing Categorical Data*. Springer. *A good supplement to Agresti, but less advanced. Takes a normal regression model as the point of departure.*

Other Readings.

The following works can be consulted for additional treatment of some of the course topics. Some readings may be assigned as required reading for particular topics during the semester.

Altman, Micah, Jeff Gill, and Michael P. McDonald. 2004. *Numerical Issues in Statistical Computing for Social Scientists.* Wiley. *A bookfull of cautions about computationally intensive methods.*

Amemiya, Takeshi. 1985. *Advanced Econometrics.* Cambridge, MA: Harvard University Press.

Arminger, Gerhard, Clifford C. Clogg, and Michael E. Sobel (eds.). 1995. *Handbook of Statistical Modeling for the Social and Behavioral Sciences*. New York and London: Plenum.

Bergstrom, C. T., and Real, L.A. (2000), "Towards a Theory of Mutual Mate Choice: Lessons from Two-Sided Matching," *Evolutionary Ecology Research*, 2, 493–508.

Berk, Richard A. 1983. "An Introduction to Sample Selection Bias in Sociological Data." *American Sociological Review* 48:386-98.

Boyd, Donald, Hamilton Lankford, Susanna Loeb, and James Wyckoff. 2013. "Analyzing the Determinants of the Matching of Public School Teachers to Jobs: Disentangling the Preferences of Teachers and Employers." *Journal of Labor Economics* 31(1):83-117.

Breen, Richard. 1994. "Individual Level Models for Mobility Tables and Other CrossClassifications." *Sociological Methods and Research* 23(2):147-173.

Cameron, Stephen V., and James J. Heckman. 1998. "Life Cycle Schooling and Dynamic Selection Bias: Models and Evidence for Five Cohorts of American Males." *Journal of Political Economy* 106(22):262-333.

Congdon, Peter. 2005. *Bayesian Models for Categorical Data.* Wiley.

Davidson, Russell, and James G. MacKinnon. 2004. *Econometric Theory and Methods*. Oxford.

DeMaris, Alfred. 2002. "Explained Variance in Logistic Regression; A Monte Carlo Study of Proposed Measures," *Sociological Methods and Research* 31(1): 27-74.

Diggle, Peter J., Liang, Kung-Yee, and Scott L. Zeger. 1994. *Analysis of Longitudinal Data*. New York: Oxford University Press.

Gilks, W.R., S. Richardson and D.J. Spiegelhalter (1996), "Introducing Markov chain Monte Carlo," in W.R. Gilks, S. Richardson and D.J. Spiegelhalter, eds., *Markov Chain Monte Carlo in Practice*. New York: Chapman and Hall, 1-19.

Greene, William H. 2002. *Econometric Analysis*. 5th Edition. New York: Macmillan.

Hauser, R. M. (1978). ''A structural model of the mobility table,'' *Social Forces* 56, 919–953.

Hoff, Peter D. (2009). *A First Course in Bayesian Statistical Methods.* Springer.

Hoffman, Saul d. and Greg J. Duncan. 1988. "Multinomial and Conditional Logit Discrete-Choice Models in Demography." *Demography* 25: 415-427.

Hout, Michael, O.D. Duncan, and Michael E. Sobel. 1987. "Association and Heterogeneity: Structural Models of Similarities and Differences." *Sociological Methodology* 17: 145-184.

Hsiao, Cheng. 2003. *Analysis of Panel Data*, Second edition. Cambridge.

Kalbfleisch, John D., and Ross L. Prentice. 1980. *The Statistical Analysis of Failure Time Data*. New York: Wiley.

Lin, Ken-Hou, and Jennifer Lundquist. 2013. "Mate Selection in Cyberspace: The Intersection of Race, Gender, and Education." *American Journal of Sociology* 119(1): 183-215.

Lindsey, J.K. 2004. *Introduction to Applied Statistics.* 2nd ed. Oxford.

_____. 1996. *Parametric Statistical Inference*. Oxford.

Little, Roderick J.A., and Donald B. Rubin. 1987. *Statistical Analysis with Missing Data*. New York: Wiley.

Little, Roderick J.A., and Nathaniel Schenker. 1995. "Missing Data." Pp. 39-75 in *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, edited by Gerhard Arminger, Clifford C. Clogg, and Michael E. Sobel. New York and London: Plenum.

Logan, John Allen. 1996a. "Opportunity and Choice in Socially Structured Labor Markets." *American Journal of Sociology* 102(1; July):114-160.

_____. 1996b. "Rational choice and the TSL model of occupational opportunity." *Rationality and Society* 8(1; May): 207-230.

_____, Peter D. Hoff, and Michael A. Newton. 2008. "Two-Sided Estimation of Mate Preferences for Similarities in Age, Education, and Religion." *Journal of the American Statistical Association* 103 (June): 559-569..

Long, J. Scott, and Jeremy Freese. 2003. *Regression Models for Categorical Dependent Variables Using Stata*. Stata Press.

Lynch, Scott M. (2007). *Introduction to Applied Bayesian Statistics and Estimation for Social Scientists.* Springer.

Mare, Robert D. 1980. "Social Background and School Continuation Decisions." *Journal of the American Statistical Association* 75:295-305.

_____. 1990. "Five Decades of Educational Assortative Mating." *American Sociological Review* 56:15-32.

Powers, Daniel A., and Yu Xie. 2000. *Statistical Methods for Categorical Data Analysis*. Academic Press.

Pudney, Stephen. 1989. *Modelling Individual Choice*. Oxford: Basil Blackwell.

Raftery, Adrian E. 1995. "Bayesian Model Selection in Social Research." Pp. 111-163 in *Sociological Methodology 1995*, edited by Peter V. Marsden. Cambridge, MA: Basil Blackwell.

Roth, A. E., and Sotomayor, M. A. (1990), *Two-Sided Matching: A Study in Game-Theoretic Modeling and Analysis*, Cambridge: Cambridge University Press.

Schafer, Joseph L. 1997. *Analysis of Incomplete Multivariate Data*. New York, London: Chapman and Hall.

Stolzenberg, Ross M., and Daniel A. Relles. 1997. "Tools for Intuition about Sample Selection Bias and Its Correction." *American Sociological Review* 62: 494-507.

Vermunt, Jeroen K. 1997. *Log-Linear Models for Event Histories*. Thousand Oaks, CA: Sage.

Vermunt, J.K., Rodrigo, M.F., and Ato-Garcia, M. (2001) "Modeling joint and marginal distribtuions in the analysis of categorical panel data." *Sociological Methods and Research* 30:170-196.

Winkelmann, Rainer. 2000. *Econometric Analysis of Count Data*. 3d edition, revised and enlarged. Berlin: Spring-Verlag.

Winship, Christopher, and Larry Radbill. 1994. "Sampling Weights and Regression Analysis." *Sociological Methods and Research* 23(2):230-257.

Woolridge, J.M.. (2001) *Econometric Analysis of Cross Section and Panel Data.* Cambridge: MIT Press.

Zorn, "Generalized Estimating Equation Models for Correlated Data: A Review With Applications," *American Journal of Political Science, 45(2):* April, 2001.

SYLLABUS

Students are expected to read over the assigned material prior to the lecture on the topic. A second reading after the lecture will probably be a good idea. Both schedule and readings are subject to change depending on the pace of the class and the interests of students.

Week	Topic	Assigned Readings
Part 1. Contingency Tables		
1	Distributions and inference for categorical data	Agresti, pp. 1-16 (Ch. 1)
2	Describing contingency tables	Agresti, pp. 21-44, 49-54
Part 2. Regression Models for Categorical Responses		
3	Generalized linear models	Agresti, pp. 65-90 (Ch. 3)
4	Logistic regression models	Agresti, 99-108, 115-121, 144-147, 152-157.
5	Multicategory logit models	Agresti, 173-194
6	Log-linear models	Agresti, 204-223
Part 3. Limited Dependent Variables: Partial Observability		
7	Censored and sample-selection models I	Breen, pp. 1-17, 23-48
8	Censored and sample-selection models II	Breen, pp. 49-72
Part 4. Models of Choice		
9	Discrete choice logit models	Train, pp. 1-29, 34-50, 64-66, 71-74
10	Discrete choice probit models; mixed logit	Train, pp. 97-110; 134-139, 141-144, 145-150
Part 5. Two-Sided Models of Choice		
11	Bayesian inference methods	Train, pp. 282-305, 313-314
12*	Two-sided models	Logan 1996; Logan et al., 2008; Boyd, et al. 2013; Lin and Lundquist 2013
Part 6. Clustered Response and Random Effects Models		
13	Models for matched pairs	Agresti, pp. 244-258, line 8
14	Clustered responses	Agresti, Ch. 9
15	Random effects: generalized linear mixed models	Agresti, pp. 297-316

* Week 12 conflicts with Thanksgiving. We will try to work around that.