

# CATEGORICAL DATA ANALYSIS AND SELECTED TOPICS

Sociology 230:605

Course Syllabus

Fall 2003

Class time: W 10:00-12:00

Office hour: W 1:30-2:30 or by appointment

Professor Lingxin Hao

526 Mergenthaler

509 Mergenthaler

516-4022; hao@jhu.edu

---

## Course Objectives

The course introduces the main tools of categorical data analysis and a selected topic on multi-level modeling for continuous and categorical data. The focus is on applications of the techniques and interpretations of results. The first two thirds of the course introduce generalized linear models for dichotomous and multiple-category dependent variables, including linear probability, logit, probit, multinomial logit, Poisson, and negative binomial models. The last one third of the course introduces multilevel analysis for continuous and categorical data. Applications of the various techniques will use the statistical packages STATA, SAS and HLM.

## Prerequisite

Sociology 604 (regression analysis) or equivalent. This means that students should have considerable experience with ordinary-least-squares (OLS) regression.

## Grading:

40% Homework (seven assignments, each of which may include two parts—paper-and-pencil exercise and a computer application using STATA, SAS or HLM; the due date is one week after the distribution of the assignment)

30% Midterm exam (take-home 11/5-12)

30% Final exam or a research paper (take-home 12/3-10; If the paper option is chosen, a 12-15 page research report of a mini project applying one of the techniques learned in this course using your own data or a data set provided)

Incompletes will only be given under very unusual circumstances.

## Lab

STATA, SAS and HLM are available on the PC's in the Sociology Computer Lab (534 Mergenthaler). Enrolled students will have their accounts in the lab. Data sets to be used will be placed in the server Coleman R:\Cat03\cat03.dta.

## Required Readings

Agresti, Alan. 1996. *An Introduction to Categorical Data Analysis*. New York: Wiley.

Raudenbush, Stephen W. and Anthony S. Bryk. 2002. (2<sup>nd</sup> ed.) *Hierarchical Linear Models: Applications and Data Analysis Methods*. Thousand Oaks: Sage Publications.

## Recommended Readings

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

Hosmer, D.W. and Lemeshow, Stanley. 1989. *Applied Logistic Regression*. New York: Wiley.

Greene, William H. 1993. *Econometric Analysis*. New York: Macmillan Publishing Company.

Singer, Judith D. 1998. "Using SAS PROC MIXED to Fit Multilevel Models, Hierarchical Models, and Individual Growth Models." *Journal of Educational and Behavioral Statistics* 24:323-355.

Raftery, A.E. 1995. "Bayesian Model Selection in Social Research." In Peter V. Marsden (ed.) *Sociological Methodology* pp. 111-164.

## Empirical Papers

Brooks, Clem and Jeff Manza. 1997. "The social and ideological bases of middle-class political realignment in the United States." *American Sociological Review* 62:191-208. (Logistic regression)

Browne, Irene. 1997. "The black-white gap in labor force participation among women." *American Sociological Review* 62:236-252. (Logistic regression)

Minkoff, Debra C. 1997. "The Sequencing of Social Movements." *American Sociological Review* 62:779-799. (Count data)

Biblarz, Timothy J. and Adrian E. Raftery. 1993. "The Effects of Family Disruption on Social Mobility." *American Sociological Review* 58:97-109. (Log-linear model)

Hao, Lingxin and Melissa Bonstead-Bruns. 1998. "Parent-Child Difference in Educational Expectations and Academic Achievement of Immigrant and Native Students." *Sociology of Education* 71:175-198. (Hierarchical linear models)

## Course Schedule

Week	Topics & Reading Assignments
Week 1	<p>1. Review of the linear regression model and the normal distribution. Introduction to maximum likelihood (ML) and the OLS-ML connection. Introduction to the binomial, Poisson, negative binomial, and multinomial distributions and their ML estimators.</p> <p>Readings: Agresti Ch. 1; Long, Ch. 2</p>
Week 2	<p>2. Two-way and three-way contingency table analysis</p> <ul style="list-style-type: none"><li>(a) probability structure</li><li>(b) odds ratio</li><li>(c) independence, association, and conditional association</li></ul> <p>Readings: Agresti Ch. 2-3</p>
Week 3-4	<p>3. Generalized linear models—models for categorical response data</p> <ul style="list-style-type: none"><li>(a) connections between linear regression models and generalized linear models (linear probability, logit and probit and their connections)</li><li>(b) binary data</li><li>(c) count data: Poisson distribution and the negative binomial distribution; model formulation; ML estimation; interpretation</li><li>(d) model inference and model checking: the likelihood ratio and the Wald test</li></ul> <p>Readings: Agresti Ch. 4; Long Ch. 3.1-3.6, 4.1-4.6; Greene Ch. 21.1-21.3 Empirical paper for discussion: Minkoff</p>
Week 5-6	<p>4. Logit models</p> <ul style="list-style-type: none"><li>(a) logit models for dichotomous dependent variables</li><li>(b) linear probability models and the connection to logit models</li><li>(c) motivating the logit model using the binomial distribution</li><li>(d) odds ratio interpretation</li><li>(e) ML estimation of the models</li><li>(f) predicted probabilities for the logit models; Pseudo R-squares; Hosmer-Lemeshow goodness-of-fit test.; likelihood ratio tests of nested specification; diagnostic methods for binary models</li></ul> <p>Readings: Agresti Ch. 5; Long Ch. 3.7-3.8; Hosmer Ch. 3 Empirical papers for discussion: Brooks and Manza; Browne</p>
Week 7-8	<p>5. Multinomial logit models</p> <ul style="list-style-type: none"><li>(a) Pairs of logit models constitute multinomial logit model</li><li>(b) Interpretation in terms of log odds, odds ratio, and change in probabilities</li><li>(c) Hypotheses testing, model goodness of fit, and model checking</li><li>(d) Testing the assumption of independence of irrelevant alternative (IIA)</li><li>(e) Compare with conditional logit models</li></ul> <p>Readings: Agresti Ch. 8; Long Ch. 6</p>
Week 9	<p>6. Data and variable issues of applying logit and multinomial logit models</p> <ul style="list-style-type: none"><li>(a) data and variable construction</li><li>(b) model fitting and comparing coefficients between models</li><li>(c) comparing goodness of fit among non-nested models</li></ul> <p>Readings: Agresti Ch. 7; Hosmer Chs. 4, 5; Greene Ch. 9.1-9.4; Raftery 1995</p>

Week	Topics & Reading Assignments
11/5-12	Midterm take-home exam
Week 10-12	<p>8. Multi-level analysis</p> <p>(a) hierarchical linear models (HLM) for multi-level units</p> <p>(b) hierarchical linear models (HLM) for panel data</p> <p>Readings: Raudenbush &amp; Bryk Ch. 1-6; Singer 1998  Empirical paper for discussion: Hao and Bonstead-Bruns</p>
Weeks 13	<p>9. Multi-level analysis (continued)</p> <p>(a) hierarchical generalized linear model (HGLM) for dichotomous data</p> <p>(b) hierarchical generalized linear models (HGLM) for multicategory data</p> <p>Readings: Raudenbush &amp; Bryk Ch.10</p>
12/3-10	Final take-home exam or research paper