Categorical Data Analysis:
Models for Binary, Ordinal, Nominal, and Count Outcomes

ICPSR Summer Program
July 16 - Aug 10, 2012

Instructor:	Shawna Smith, Indiana University sns3@indiana.edu http://www.shawnasmith.net
Teaching Assistants:	Long Doan, Indiana University longdoan@indiana.edu Rebecca Grady, Indiana University rkgrady@indiana.edu
Lectures:	3:05pm-5pm
Office Hours:	1-3pm or by appointment (Newberry House)

Course overview:
This class focuses on the basic regression models for categorical dependent variables. Although advances in software have simplified estimation of these models, model non-linearities make post-estimation interpretation difficult. The class begins by considering the general objectives for interpreting the results of any regression-type model and then considers why achieving these objectives is more difficult with nonlinear models. Basic concepts and notation are introduced through a short review of the linear regression model. Within this familiar context, the method of maximum likelihood estimation is presented. From there, we will develop the logit and probit models for binary outcomes, as well as a variety of practical methods for interpreting nonlinear models. We will then extend these models and methods of interpretation for 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 models.

Software:
A major component of the course is using Stata to estimate and interpret models. All model demonstrations will be conducted using Stata 12. While the course assumes familiarity with the linear regression model, it does not assume familiarity with Stata.

Required Text

Lecture and Lab Notes for Categorical Data Analysis. These notes contain copies of the overheads for the lectures and materials used in the computing lab. Be sure to bring these notes to all lectures and labs.

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

Long, J. Scott & Jeremy Freese. 2005. *Regression Models for Categorical Dependent Variables Using Stata*. 2nd Edition. College Station, TX: Stata Press. *Hereafter:* **L&F**

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

Course Outline
NB: 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
W1: M	Overview of class; Introduction to models	**Long** Ch. 1	
W1: T	Review of linear regression; Identification; Maximum Likelihood Estimation; Introduction to Stata	**Long** Ch. 2; **P&X** Ch. 2; **L&F** Ch. 1-2	Math Review
W1: W	Linear probability model; Identification of Pr(y=1); Two philosophies: transformational and latent variable approach for binary outcomes	**Long** Ch. 3; **P&X** Ch.1	
W1: R	Estimation of BRM; Odd ratios		
W1: F	Using Pr(y=1) to interpret the BRM (pt. 1): tables & plots; discrete change		BRM1
W2: M	Using Pr(y=1) to interpret the BRM (pt. 2): plots and discrete change; delta method; bootstrap		
W2: T	Internal measures of fit; Hypothesis testing; Wald and LR tests; Confidence intervals	**Long** Ch. 4	
W2: W	Scalar measures of fit: pseudo-R2, AIC, BIC		BRM2
W2: R	BRM redux: Complications on the RHS; group differences		
W2: F	Ordinal variables; a latent variable model	**Long** Ch. 5; **P&X** Ch. 7	T&F
W3: M	Estimation of ORM; latent variable interpretations; Pr(y=k)		
W3: T	Odds ratios; parallel regression assumption and proportional odds		
W3: W	Multinomial logit as a set of BLMs; IIA	**Long** Ch. 6; **P&X** Ch. 8	ORM
W3: R	Tests for the MNLM; Calculating predicted probabilities; Interpretation using Pr(y=k)		
W3: F	Odds ratio plots; Discrete change plots		
W4: M	Putting it all together; catch-up (as needed)		

W4: T	Counts; Poisson process; estimation of PRM; assessing fit; the big idea of heterogeneity	**Long** Ch. 8	MNLM
W4: W	Interpretation; adding unobserved heterogeneity; estimation of NBRM		
W4: R	With-zeros models; zero-modified and zero-inflated models; comparisons among count models; course wrap-up		
W4: F	No class	COUNT (in my mailbox by 10am)	

Computing
This course will use Stata for model estimation and interpretation (demonstrations will use version 12, but Stata 10 or 11 will work just fine). While Stata (and most popular statistical packages) includes native commands for estimating our models of interest, we will also use a set of ado files written for Stata by Scott Long and Jeremy Freese that facilitate the (at times complicated) interpretation of categorical models. This suite of commands is called SPost.

- **Getting Started using Stata:** A document titled “Getting Started using Stata” is available for download from my website (http://www.shawnasmith.net/teaching). If you have never used or are not comfortable using Stata, you should work through this document prior to the first day of class. Feel free to get in touch with me or either TA should you have any questions.

- **Downloading SPost:** If you will be using a personal computer to complete your coursework, you will need to install the current SPost suite of commands. On a computer connected to the internet & where you have administrative privileges, you can install SPost by typing `findit spost` into the command line. A viewer window will appear, listing links. Click on the link “spost9_ado from http://www.indiana.edu/~jslsoc/stata” and follow directions to install. Computers in Newberry labs may or may not have SPost commands installed. To check, type `help prchange` into the command line. If a help window pops up, then SPost is installed. If not, follow the process described above to install.

- **Working in the Newberry labs:** Once logged on to a computer in Newberry lab, you can access the “My Documents” folder. Within “My Documents” is the subfolder “work.” This subfolder is set as the default “working directory” in Stata. However, as all computers in the lab have shared access (i.e., any other participant can log on to the same machine and access the same “My Documents” folder), I suggest changing your “working directory” to a folder on your personal thumb drive or external hard drive. We will review the purpose of a “working directory” on the first day of class, as well as how to change your “working directory” as and when necessary. See the “Getting Started using Stata” document on my website for more information.

- **Lab Guide:** I have provided a Lab Guide that can be used to structure your work in labs. The amount of time you spend on the Guide will depend on your past experience with Stata and your familiarity with the methods being discussed. The Guide is divided into sections corresponding to the class lectures, and you should plan time every day to work through the section that corresponds to that day's lecture. After you have worked through the appropriate section of the Guide, you will then be prepared to start with the assignment for that section. Note that the data used in the lab guide – *icpsr_scireview3* – ***cannot*** be used for assignments.

- **Datasets:** Four datasets are available for you to use to complete the assignments. Codebooks for these datasets will be made available.

Course Materials

Copies of the course materials, including datasets, are available in the class folder, Z:\smith. Course materials will also be available on my website (www.shawnasmith.net/teaching/).

Questions & getting help

The teaching assistants and I always welcome questions and feedback about the course and course materials. The teaching assistants will be available for consultation every day in/around the Newberry labs. Specific times will be discussed on the first day and decided based on participant preferences. You can also meet with me during my office hours or by appointment.

- **Email:** We're also always happy to take questions by email, however to ensure a response (and to help me stay organized), please start your subject line with "ICPSRCDA12: " followed by a short description of your question or problem.

Grading

Grades are based on assignments. The final grade is determined by adding up the points received and dividing by the total number of possible points: 98-100% = A+; 94-97% = A; 91-93%=A-; etc. Note that if you are not taking this class for credit, we will use a simplified grading scheme for assignments: Excellent, Very Good, Good, Fair, and Poor.

Assignments

Assignments should be handed in at the beginning of class on the date they are due. Due to the concentrated nature of this class, we cannot accept late assignments. When handing in assignments, follow these guidelines:

1) *Significant findings:* All regression models must include at least one continuous independent variable and one binary independent variable, both of which must be statistically significant at the .05 level. If you have trouble finding significant effects, ask the TA for help or use one of the suggested models at the end of each codebook.

2) *Do file*: All Stata commands for an assignment should be included in a single do-file. Use comments to indicate which commands correspond to which questions in the assignment. These comments should be short, but clear. Note, however, that you do not need to hand in your do-file as it is "echoed" in your Stata log file.

3) *Answers*: Label your answers with the question number; you do not need to type the question itself, but you can if you'd like. Include the Stata output that corresponds to what

you are reporting (Stata output in 9pt Courier New font prevents wrapping). Indicate the number(s) used in your answer in the following way:

```
. regress job fem art

      Source |       SS       df       MS              Number of obs =     408
-------------+------------------------------           F(  2,   405) =   15.89
       Model |  28.0762965     2  14.0381483           Prob > F      =  0.0000
    Residual |  357.720095   405  .883259494           R-squared     =  0.0728
-------------+------------------------------           Adj R-squared =  0.0682
       Total |  385.796392   407  .947902683           Root MSE      =  .93982

------------------------------------------------------------------------------
         job |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         fem |  -.1285907   .0968463    -1.33   0.185    -.3189748    .0617935
         art |   .1083582   .0209598     5.17   0.000     .0671546    .1495618
       _cons |   2.036817   .0805349    25.29   0.000     1.878498    2.195135
------------------------------------------------------------------------------
```

- For each additional publication, the prestige of the first job is expected to increase by .11 units, holding all other variables constant.

If this is unclear, please consult the mock assignment that will be posted prior to BRM1 assignment.

4) *Stata log*: The log should be printed in a fixed font. If you do not know what a fixed font is, please ask the TA.

5) *Clip everything together & hand-in*: Use a binder clip to collate the following materials in the following order:

   a. The grade sheet with your name filled in. Please do **<u>not</u>** staple this sheet to the other pages.
   b. Your answers (Word, LaTeX, etc. file) stapled together.
   c. Your Stata log stapled together as a separate document.