

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

ICPSR Summer Program  
July 18 - Aug 12, 2011

*Instructor:* Tait Medina, Indiana University  
tmedina@indiana.edu  
<http://www.taitmedina.com>

*Teaching Assistant:* Shawna Smith, Indiana University  
sns3@indiana.edu

This class focuses on the basic regression models for categorical dependent variables. While advances in software have made estimating these models simple, post-estimation interpretation is difficult due to model nonlinearities. 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 review of the linear regression model. Within this familiar context, the method of maximum likelihood estimation is presented. These ideas are used to develop the logit and probit models for binary outcomes. A variety of practical methods for interpreting nonlinear models are then presented. The models and methods of interpretation for binary outcomes are extended to ordinal outcomes using the ordinal logit and probit models. The multinomial logit model for nominal outcomes is then discussed. Finally, a series of models for count data, including Poisson regression, negative binomial regression, and zero modified models are presented. A major component of the course is using Stata to estimate and interpret the models. All demonstrations will be conducted using Stata 11. While the course assumes familiarity with the linear regression model, it does not assume familiarity with Stata.

*Lectures:* 3:10pm-5pm  
*Office Hours:* 1pm to 3pm or by appointment (Newberry House)

### **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: **L**

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

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

**Course Outline:** The content of the course will vary depending on the background of class members. In other words, this schedule is subject to change.

<b>Day</b>	<b>Topic</b>	<b>Suggested Readings</b>	<b>Due</b>
W1: M	Overview of class; Introduction to models	<b>Long</b> Ch. 1	
W1: T	Review of linear regression; Identification; Maximum Likelihood Estimation; Introduction to Stata	<b>Long</b> Ch. 2; <b>P&amp;X</b> Ch. 2; <b>L&amp;F</b> 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	<b>Long</b> Ch. 3; <b>P&amp;X</b> Ch.1	
W1: R	Estimation of BRM; Odd ratios		
W1: F	Using $\Pr(y=1)$ to interpret the BRM: tables & plots; discrete change		BRM1
W2: M	Using $\Pr(y=1)$ to interpret the BRM: plots and discrete change; delta method; bootstrap		
W2: T	Internal measures of fit; Hypothesis testing; Wald and LR tests; Confidence intervals	<b>Long</b> Ch. 4	
W2: W	Scalar measures of fit: pseudo-R <sup>2</sup> , AIC, BIC		BRM2
W2: R	BRM redux: complications on the RHS; group differences		
W2: F	Ordinal variables; a latent variable model	<b>Long</b> Ch. 5; <b>P&amp;X</b> 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	<b>Long</b> Ch. 6; <b>P&amp;X</b> 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		
W4: T	Counts; Poisson process; big idea of heterogeneity; Estimation of PRM; Assessing fit	<b>Long</b> 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		
W4: F	No class		COUNT (in my mailbox by 10am)

**Computing:** This course focuses on using Stata (demonstrations use version 11, but Stata 9 or 10 will work just fine) for estimating and interpreting regression models for categorical outcomes. While Stata includes commands for estimating these models, we will use a set of ado files written by Scott Long and Jeremy Freese that make the interpretation of categorical models easier. 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. If you have never used or are not comfortable using Stata, you should work through this document prior to the first day of class.

**Downloading SPost:** The computers in Newberry lab may or may not have SPost commands installed. To check if SPost is installed, type `help prchange` into the command line. If a help window pops up, then SPost is installed. If not, type `findit spost`. A Viewer window will appear, listing links. Click on the link “spost9\_ado from <http://www.indiana.edu/~jslsoc/stata>”.

**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.” See the document “Getting Started using Stata” 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 in lab you should work through the section that corresponds to that day’s lecture. After you have worked through the appropriate section of the Guide, you should then start with the assignment. Note that the data set that is 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:\medina.

**Questions and getting help:** The teaching assistant will be available to answer your questions each day in the Newberry labs. We will decide on specific times on the first day of class. You can also meet with me during my office hours or by appointment. Many find that email works well for some questions. You are welcome to contact me by email -- please start your subject line with “ICPSRCDA11: ” 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) *Clip everything together:* Use a binder clip, with materials in the following order
  - a. The grade sheet with your name filled in. Do **not** 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.
- 2) *Do file:* Use comments to indicate which commands correspond to which questions in the assignment. These comments should be short, but clear. You do not need to hand in your do-file as it is “echoed” in your Stata log file.
- 3) *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.
- 4) *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			
Model	28.0762965	2	14.0381483	Number of obs =	408	
Residual	357.720095	405	.883259494	F( 2, 405) =	15.89	
				Prob > F =	0.0000	
				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	<b>.1083582</b>	.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.

- 5) *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.