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

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

ICPSR Summer Program
July 19-August 13, 2010

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 and particularly using a set of special commands for post-estimation interpretation. The course assumes familiarity with the linear regression model. Familiarity with Stata is not assumed.

Lectures: 3pm-5pm
Office Hours: 1pm-3pm or by appointment (Newberry House)
Computing Labs: Teaching assistants will be available in the Newberry Labs each day. Specific times TBA

Texts
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.


Articles illustrating the application of the methods discussed will be posted on the class Wiki.
**Course Outline:** The content of the workshop will vary depending on the background of class members. You will get the most out of the lectures if you try to read the material before the class in which it is discussed.

<table>
<thead>
<tr>
<th>Day</th>
<th>Topic</th>
<th>Reading</th>
<th>Due</th>
</tr>
</thead>
<tbody>
<tr>
<td>M, 7-19</td>
<td>Registration</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T, 7-20</td>
<td>Overview; Short Intro to Stata</td>
<td>RMStata Ch. 1, 2; RMSage Ch. 1</td>
<td></td>
</tr>
<tr>
<td>W, 7-21</td>
<td>Review of linear regression; linear and nonlinear models; interpreting the LRM</td>
<td>RMStata Ch. 3.1; RMSage Ch. 2</td>
<td></td>
</tr>
<tr>
<td>R, 7-22</td>
<td>Identification; Maximum Likelihood Estimation; binary outcomes; LPM</td>
<td>RMStata Ch. 3.5, 4; RMSage Ch. 3</td>
<td></td>
</tr>
<tr>
<td>F, 7-23</td>
<td>Latent variable models for binary outcomes; Identification of Pr(y=1); nonlinear probability models</td>
<td></td>
<td>LRM</td>
</tr>
<tr>
<td>M, 7-26</td>
<td>Using Pr(y=1) to interpret the BRM; plots; discrete change</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T, 7-27</td>
<td>Odds ratios; articles using the BRM</td>
<td>Example Articles</td>
<td></td>
</tr>
<tr>
<td>W, 7-28</td>
<td>Catch-up; Hypothesis testing; Wald tests; LR tests</td>
<td>RMStata Ch. 3.3-3.4; RMSage Ch. 4</td>
<td>BRM</td>
</tr>
<tr>
<td>R, 7-29</td>
<td>Confidence intervals; residuals and outliers; BIC and AIC; measures of fit; delta method, bootstrap</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F, 7-30</td>
<td>Catch-up; RHS issues; functional forms; Lowess smoothing</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M, 8-2</td>
<td>Interactions; group differences</td>
<td>RMStata Ch. 5; RMSage Ch. 5</td>
<td>T&amp;F</td>
</tr>
<tr>
<td>T, 8-3</td>
<td>Ordinal variables; a latent variable model</td>
<td></td>
<td></td>
</tr>
<tr>
<td>W, 8-4</td>
<td>Estimation; latent variable interpretations; Pr(y=k)</td>
<td>RMStata Ch. 5; RMSage Ch. 5</td>
<td></td>
</tr>
<tr>
<td>R, 8-5</td>
<td>Odds ratios; parallel regression assumption and proportional odds; articles using ordinal models</td>
<td>Example Articles</td>
<td></td>
</tr>
<tr>
<td>F, 8-6</td>
<td>Multinomial logit as a set of BLMs; Tests for the MNLM</td>
<td>RMStata Ch. 6; RMSage Ch. 6</td>
<td>ORM</td>
</tr>
<tr>
<td>M, 8-9</td>
<td>Interpretation; discrete change; odds ratio plots</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T, 8-10</td>
<td>Class examples of odds ratio plots; articles using MNLM</td>
<td>Example Articles</td>
<td></td>
</tr>
<tr>
<td>W, 8-11</td>
<td>Count models; adding heterogeneity; PRM</td>
<td>RMStata Ch. 7; RMSage Ch. 8</td>
<td>MNLM</td>
</tr>
<tr>
<td>R, 8-12</td>
<td>Predicted probabilities in count models, NBRM, unobserved heterogeneity; Zero-modified and zero-inflated models</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F, 8-13</td>
<td>Hurdle models; articles using count models; review</td>
<td>Example Articles</td>
<td>COUNT</td>
</tr>
</tbody>
</table>
Computing: This course focuses on using Stata 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.

Getting Started using Stata: A document titled “Getting Started using Stata” is available for download from my website: http://www.taitmedina.com/teaching. If you have never used or are not comfortable with 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 installed. To check if SPost is installed, type `help prchange`. If a help window pops up, then SPost is installed. If not, then 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 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 data can be found at the end of the Lab Guide.

Course Materials: Copies of the syllabus, assignments, lecture notes, and Lab Guide are available on my website: http://www.taitmedina.com/teaching. Datasets and sample do-files are available via the internet from Stata. In Stata, type `findit icpsr`. You will then see the package `icpsrmedina`. Click the link to download. These files are also available in the class folder, Z:\medina. Remember that do and data files downloaded via Stata are saved to your “working directory.”

Questions and getting help: Teaching assistants are 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 “ICPSRCDA: ” 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.

Assignments: Assignments should be placed in my mailbox in Newberry House by 3pm or brought to class on the due date. Because of the condensed nature of this course, late assignments will not be accepted. For each assignment, 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 stapled together.
   c. Your Stata log stapled together as a separate document.

2) **Answers:** Label your answers with the question number; you do not need to type the question itself. Include the Stata output that corresponds to what you are reporting. The Stata output must be in a fixed font such as Courier New (9pt font works well) and must not wrap. Indicate the numbers from the output used in you answer; you can make them bold or underlined or simply highlight them with a highlighter. For example:

```stata
. 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.

3) **Tables:** Tables should correspond to those in the Sage text or from a major journal in your field. Here are some things to remember: line up the decimal points and be consistent with the number of decimal digits.

4) **Do file:** Use comments in your do file to indicate which commands correspond to which questions in the assignment. The 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.

5) **Stata log:** The log must be in fixed font and output must not wrap.

6) **Significant findings:** In all regression models, you must include at least one continuous independent variable and one binary independent variable, both of which must be statistically significant.
a. **TIP:** If you have trouble finding significant effects, you can use a “go to” model provided at the bottom of each codebook. Points will not be deducted for using a “go to” model as long as this is indicated at the top of the assignment sheet.

7) *Working Together:* You are encouraged to discuss your work with other students, but you must complete all assignments on your own.

8) *Example Assignment:* I have provided an example of how to complete the first assignment (LRM). Please see icpsrcda01a-mock.docx and icpsrcda01a-mock.do for how I want do-files and answers formatted for all assignments. These files are available for download via Stata and are also available in the class folder, Z:\medina.