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

ICPSR Summer Program
July 21 - Aug 15, 2014

Instructor: Shawna Smith, University of Michigan
shawnana@umich.edu
http://www.shawnasmith.net

Course Twitter feed: http://www.twitter.com/icpsrcda/

Teaching Assistants: Kelly Gleason, University of Wisconsin-Milwaukee
kgleason@uwm.edu
http://www.uwm.edu/~kgleason

Trent Mize, Indiana University
tdmize@indiana.edu

Lectures: 3:05pm-5pm

Office Hours: Tuesday, Thursday & Sunday 6-8PM (State Street Espresso Royale)
or by appointment (book at http://icpsrcda.youcanbook.me)

Course overview:
Many variables of interest to social, political and behavioral scientists are non-continuous, either through nature or through measurement. Outcomes like vote choice, social class, condom use, and/or number of Facebook friends necessarily violate key assumptions of the simple linear regression framework and require other model estimation strategies. Although advances in software have made estimation of these models trivial, model non-linearities make post-estimation interpretation difficult and require investigators to make choices about which aspects of the data space best represent underlying social dynamics.

The course begins by considering the general objectives for interpreting the results of any regression-type model and then considers why these objectives are more complicated within nonlinear models. Basic concepts and notation are introduced through a short review of the linear regression model, and a short overview of the method of maximum likelihood estimation. From there, we will ‘derive’ the logit and probit models for use with binary outcomes, and also introduce a variety of post-estimation tools for interpreting nonlinear models. We will then extend these models and methods of interpretation from 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 variant models.
Software:
Models for this course are presented in broad strokes; however a major component of this course is application, through model estimation, post-estimation and interpretation. For pedagogical reasons, I will use Stata 13 for model estimation and interpretation; I encourage you to do the same. N.B.: While the course assumes familiarity with the linear regression model, it does not assume familiarity with Stata.

Required Text
Lecture Notes and Lab Guide 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

(N.B.: New versions of both of Long’s books are due in 2014, so may consider borrowing)


Course Outline
N.B.: The exact content of the course will vary depending on the background & interests of participants. In other words, this schedule is subject to change.

<table>
<thead>
<tr>
<th>Day</th>
<th>Topic</th>
<th>Suggested Readings</th>
<th>Due</th>
</tr>
</thead>
<tbody>
<tr>
<td>W1:</td>
<td>M Overview of class; Introduction to models</td>
<td>Long Ch. 1</td>
<td></td>
</tr>
<tr>
<td>W1:</td>
<td>T Review of linear regression; Identification; Maximum Likelihood Estimation; Introduction to Stata</td>
<td>Long Ch. 2; P&amp;X Ch. 2; L&amp;F Ch. 1-2</td>
<td>Math Review</td>
</tr>
<tr>
<td>W1:</td>
<td>W Linear probability model; Identification of Pr(y=1); Two philosophies: transformational and latent variable approach for binary outcomes</td>
<td>Long Ch. 3; P&amp;X Ch.1</td>
<td></td>
</tr>
<tr>
<td>W1:</td>
<td>R Estimation of BRM; Odds ratios</td>
<td></td>
<td></td>
</tr>
<tr>
<td>W1:</td>
<td>F Using Pr(y=1) to interpret the BRM (pt. 1): tables &amp; plots; discrete change</td>
<td>BRM1</td>
<td></td>
</tr>
<tr>
<td>W2:</td>
<td>M Using Pr(y=1) to interpret the BRM (pt. 2): plots; difference at means vs. mean of difference; partial change/margins</td>
<td></td>
<td></td>
</tr>
<tr>
<td>W2:</td>
<td>T Internal measures of fit; Hypothesis testing; Wald and LR tests; Confidence intervals</td>
<td>Long Ch. 4</td>
<td></td>
</tr>
<tr>
<td>W2:</td>
<td>W Scalar measures of fit: pseudo-R2, AIC, BIC</td>
<td>BRM2</td>
<td></td>
</tr>
<tr>
<td>W2: R</td>
<td>BRM redux: Group differences &amp; interactions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>W2: F</td>
<td>Ordinal variables; a latent variable model</td>
<td>Long Ch. 5; P&amp;X Ch. 7</td>
<td></td>
</tr>
<tr>
<td>W3: M</td>
<td>Estimation of ORM; latent variable</td>
<td>T&amp;F</td>
<td></td>
</tr>
<tr>
<td>W3: T</td>
<td>interpretations; Pr(y=k)</td>
<td>Long Ch. 6; P&amp;X Ch. 8</td>
<td></td>
</tr>
<tr>
<td>W3: W</td>
<td>Multinominal logit as a set of BLMs; IIA</td>
<td>ORM</td>
<td></td>
</tr>
<tr>
<td>W3: R</td>
<td>Tests for the MNLM; Calculating predicted</td>
<td>Long Ch. 8</td>
<td></td>
</tr>
<tr>
<td>W3: F</td>
<td>probabilities; Interpretation using Pr(y=k)</td>
<td>MNLM</td>
<td></td>
</tr>
<tr>
<td>W4: M</td>
<td>Odds ratio plots; Discrete change plots</td>
<td></td>
<td></td>
</tr>
<tr>
<td>W4: T</td>
<td>Putting it all together; catch-up (as needed)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>W4: W</td>
<td>Counts; Poisson process; estimation of PRM; assessing fit; the big idea of heterogeneity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>W4: R</td>
<td>With-zeros models; zero-modified and zero-inflated models; comparisons among count models; course wrap-up</td>
<td></td>
<td></td>
</tr>
<tr>
<td>W4: F</td>
<td>No class</td>
<td>COUNT (to TAs by 10am)</td>
<td></td>
</tr>
</tbody>
</table>

**Computing**

This course will use Stata for model estimation and interpretation. Demonstrations will use version 13, but Stata versions 11 or higher will also work. While Stata—and most popular statistical packages—includes native commands for estimating categorical models, we will also use a set of user-written ado files 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. If you are taking this course for credit, you will need to complete assignments using Stata and SPost 13 commands.

- **Getting Started using Stata:** A document entitled “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 either TA or me should you have any questions.

- **Downloading SPost:** If you will be using Stata on 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 and where you have administrative privileges, you can install the current SPost package by typing `search spost13` into the command line. A viewer window will appear, listing links. Click on the link “spost13.ado from http://www.indiana.edu/~jlsoc/stata” and follow directions to install. Computers in Newberry labs should have SPost 13 commands installed. To check, type `help mchange` into the command line. If a help window pops up, then SPost 13 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 portable drive or cloud storage space (e.g., Dropbox). 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:** The provided Lab Guide should be used to structure your working time in the computer labs. The amount of time you will need to 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_scireview4* – **cannot** be used for assignments.

• **Datasets:** Four datasets are available for you to use for assignments. Codebooks for these datasets can be found in the back of the *Course Notes and Lab Guide*.

• **Other Statistical Software:** I recognize that participants frequently use software other than Stata for their work, due to preference, availability and/or field norms. While I will focus my demonstrations on estimations from Stata, and **participants taking this course for credit will need to complete assignments using Stata/SPost 13**, I encourage and welcome you to explore model estimation and interpretation using other software packages, especially R and SAS. Do note, however, that differences in estimation presentation, algorithms and/or assumptions may differ between software packages. Some of these differences will be discussed in lecture, but feel free to bring questions, concerns or discovered ambiguities to my attention.

**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 ([http://www.shawnasmith.net/teaching](http://www.shawnasmith.net/teaching)) and (at times) through the course Twitter feed at [http://www.twitter.com/icpsrcda/](http://www.twitter.com/icpsrcda/).

**Questions & Getting Help**
The TAs and I welcome questions and feedback about this course and its materials. TAs 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 always happy to take questions by email; however to ensure a prompt response, please start your subject line with “ICPSRCDA14: ” 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 (+ liminal categories as necessary).

Assignments
Assignments are due at the beginning of class on dates listed. Due to the concentrated nature of this course, late assignments are not accepted. When handing in assignments, follow these guidelines:

1) **Interpretations should be of significant effects.** As we tend to not spend time writing up insignificant effects, all models must include at least one continuous independent variable (C) and one dichotomous independent variable (D), both of which are statistically significant at the $\alpha = 0.05$ level or better. If you have trouble finding significant predictors, ask the TA for help or use one of the suggested models at the end of each codebook in your coursepack.

2) **Commands should be documented in a single do-file.** All Stata commands for an assignment should be included in a single do-file. Use short, clear comments to indicate which commands correspond to which parts of the assignment. However, note that you do not need to hand in your do-file with your assignment as it is ‘echoed’ in your Stata log file.

3) **Answers should be labeled and organized in a Word, LaTex, etc. file.** Label your answers with the question number; (no need to type out the question itself). Include the Stata output that corresponds to what you are reporting (Stata output in 9pt Courier New font prevents wrapping and other unpleasantness). Highlight or indicate the specific number(s) used in your answer. An example can be seen below:

```stata
. regress job fem art

Source |       SS       df       MS              Number of obs =     408
-------------+--------------------------------------------------
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 points, holding all other variables constant.
Consult the mock BRM1 assignment for more in situ examples, or ask your TA for clarification.

4) **Include a Stata log with all relevant (but no irrelevant!) analyses.** The log should be printed in a fixed font (again, I recommend 9pt Courier New). It should *not* include irrelevant analyses, error messages or output that wraps or is otherwise difficult to read. Consult your TA if you have questions about your Stata log.

5) **Don’t forget your paperclip!** Assignments should be handed in at the beginning of class on their due date. Use a paperclip or binder clip to collate materials in the following order:
   a. The assignment/grade sheet with your name filled in. Please do *not* staple this sheet to the other pages.
   b. Your answers (Word, LaTeX, etc. file) stapled together.
   c. Your Stata log-file stapled together as a separate document.