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

ICPSR Summer Program
July 22 - Aug 16, 2019

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

Course website: http://www.shawnasmith.net/icpsrcda/

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

Teaching Assistants: Muna Adem, Indiana University
munaadem@iu.edu

Kelly Hunter, Duke University
kelly.hunter@duke.edu

Lectures: 1:05pm-3pm

Office Hours: Monday, Thursday & Sunday 5-7PM (State Street Espresso Royale)
or by appointment (Institute for Social Research office; book at
http://icpsrcda.youcanbook.me)

Course overview:
Many variables of interest to social, political and behavioral scientists are non-continuous, either by nature or by measurement. Outcomes like vote choice, social class, condom use, and/or number of Instagram followers necessarily violate key assumptions of the linear regression framework and require other model estimation strategies. These types of variables are common in both inferential statistics as well as predictive analytics (e.g., election forecasting, machine learning algorithms). Although advances in statistical software have made estimation of models for these variable types trivial, identification assumptions of the different models have substantive repercussions for investigators considering different data generating processes. Further, 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 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, as well as a brief overview of the **method of maximum likelihood estimation** (MLE). From there, we will 'derive' the **binary logit and probit models** for use with binary outcomes, and also introduce a variety of post-estimation tools for interpreting nonlinear models.

Testing and fit for non-linear models will also be considered. We will then extend the binary logit and probit models and methods of interpretation to ordinal outcomes using the **ordered logit and probit** models, and nominal outcomes with the **multinomial logit**. 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 purposes, I will use Stata 15 in course lectures for model estimation and interpretation. Some support and code will also be provided for running these models and commands in R.
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. This coursepack 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.
- This coursepack is available to purchase at the Institute for Social Research. For participants that prefer electronic versions, component parts are also available on the course website. Cost TBC, but typically around $40.

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

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

For the Stata devotees: Long, J. Scott & Jeremy Freese. 2014. *Regression Models for Categorical Dependent Variables Using Stata.* 3rd Edition. College Station, TX: Stata Press. *Hereafter:* **L&F**

Or if you like R: I'm still searching for my favorite here, but a couple of good ones are:
- Monogan, James E. III. 2015. *Political Analysis Using R.* New York, NY: Springer. *Hereafter:* **Monogan**.
- Fox, John & Sanford Weisberg. 2010. *An R Companion to Applied Regression.* Thousand Oaks, CA: Sage. *Hereafter:* **F&W**.

Course Outline

N.B.: The exact content of the course will vary depending on the background and 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 (Stata); **F&W** Ch. 1-2 or **Monogan** Ch. 1-2 (R)	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; Odds 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; difference at means vs. mean of difference; partial change/margins		
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, in-sample prediction (ROC/C-statistic), AIC, BIC		BRM2
W2: R	BRM redux: Group differences & interactions		
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 (to TAs by 10am)

Computing
You are strongly encouraged to provide your own computer for this course.

Course lectures will use Stata for model estimation and interpretation; course support will be provided for both Stata and R. Demonstrations will use Stata version 15, but Stata versions 12 or higher will suffice. While Stata—and most popular statistical software packages—includes native estimation (and even post-estimation) commands for categorical models, we will also use a set of ado files written for Stata by Scott Long & Jeremy Freese that facilitate the (at times complicated) interpretation of categorical models within Stata. This suite of commands is called SPost. These post-estimation commands can also be emulated in R although this will require more investigation on the part of the student. A variety of packages now exist in R relevant to the course that we are happy to provide guidance on when possible. *If you are taking this course for credit,* **_you will need to complete assignments using either Stata and SPost 13 commands or R with appropriate post-estimation commands_**.

- **Getting Access to Stata/R:**
  - *Stata:* Access to Stata MP is available through both the ICSPR Summer Program VDE and (for short-term access) through a special Summer Program access code.
  - *R:* R is free. You can download it at https://www.r-project.org/
    - **R Studio** is a free program that greatly upgrades R's user-interface and can be downloaded at https://www.rstudio.com/

- **Getting Started using Stata:** New to Stata? No worries—this course will catch you up quickly. However, I strongly suggest working through the "Getting Started using Stata" document available on my website (http://www.shawnasmith.net/icpsrcda/) prior to Day 1 of class. Feel free to get in touch with either TA or me if you have questions.
  - *New to R?* As R has a steeper learning curve, I would not recommend attempting to learn R solely for the purposes of this course. However, I am happy to recommend resources for those of you so inclined:
    - The two textbooks recommended above provide good introductions for 'getting started' with R, as well as lots of *in situ* code.
    - Mike Marin (UBC Public Health) has a great series of videos introducing R online at http://www.statslectures.com/index.php/r-stats-videos-tutorials/getting-started-with-r.
    - John Fox offers a great lecture series for learning R in the evenings during the first two weeks of Second Session. I strongly recommend this for both new and beginner R users.

- **Downloading SPost:** If you will be using Stata on a personal computer, then you will need to install the current SPost suite of commands. Here's the step-by-step:
  - *Pre-reqs:* Internet access & administrative privileges
  - In Stata, type `search spost13` into the command line.
  - In the viewer window that appears, click the link for "spost13_ado from http://www.indiana.edu/~jslsoc/stata"
  - Follow directions to install

o   Double-check install by typing `help mchange` in the command line. If a help window pops up, SPost 13 has been installed correctly.

Computers in Newberry or other campus computing labs may or may not have SPost 13 commands installed. Before starting your work, check by typing `help mchange` into the command line. If a help window pops up, then SPost 13 is installed. If not, install following the steps above.

- **Setting Working Directories:** Stata has a default 'working directory' that it looks to for saving documents. If you are working on a computer that is not your own (e.g., a Summer Program or Michigan computing lab computer), this default folder is typically …/My Documents/work. However, as lab computers are frequently shared access (i.e., any other participant can log on to the same machine and access the same 'My Documents' folder), you will want to change your 'working directory' to a personal space, either on a portable drive or cloud storage space (e.g., Dropbox). We will discuss working directories on the first day of class; see the "Getting Started using Stata" document on my website for further information.

- **Lab Guide:** Lab Guides for both R and Stata are included in the back of the *Course Notes and Lab Guide* (hereafter: The Guide). These guide work through examples of all models, interpretations and commands used in the assignments, and should be your first stop before starting any class assignment. The amount of time you will need to spend with The Guide will depend on your past experience with Stata/R and your familiarity with the methods being discussed. The Guide is divided into sections that correspond to the course 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 Guide – *icpsr_scireview4* – **_cannot_** be used for assignments.

- **Datasets:** Five 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 my demonstrations will focus on Stata, and **_participants taking this course for credit will need to complete assignments using either Stata/SPost 13 or R_**, I strongly encourage and welcome participants to explore model estimation and interpretation using other software packages (e.g., Python, SAS, JMP…). Nowadays, nearly all packages come with some type of post-estimation package or commands for categorical models; however as we will discuss, differences in content and utility of the commands across packages is often nuanced, yet very important for substantive interpretation. I'll do my best to discuss some of these differences in lecture, but I also encourage participants to explore these commands and bring examples, questions, concerns or ambiguities to my attention and the attention of the class.

Course Materials
Course materials are available on the course website (http://www.shawnasmith.net/icpsrda/). The course Twitter feed http://www.twitter.com/icpsrcda/ is also used for posting relevant just-in-time course readings, updates to assignments and a biased sample of miscellany from other quant-inclined Twitter feeds.

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 announced on Day 1. 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 "ICPSRCDA19:" followed by a short description of your question or problem.

Grading
Grades are based on assignments. Final grades are 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 not to write up insignificant effects, all models must include at least one continuous independent variable (C) and one dichotomous independent variable (D) that are statistically significant at the $\alpha < 0.05$ level or better.

   o **HINT:** Having trouble finding significant predictors? Ask a 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 Stata do-file or R script.** All commands for an assignment should be included in a single do-file or script. 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 log/output file.
   o **HINT:** See course website for Stata do-file template.

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 or R output that corresponds to what you are reporting (single-spaced output in 9pt `Courier New` font prevents wrapping and other shenanigans). Highlight or indicate the specific number(s) used in your answer. See example for Stata output below:

```
. 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 points, holding all other variables constant.

> o **HINT:** Consult the mock BRM1 assignment for more *in situ* examples.

4) **Include a(n edited) log with all relevant (but no irrelevant!) analyses.** The log should be printed in a fixed font (again, I recommend single-spaced 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 log-file stapled together as a separate document.