Revised: March 2016

Statistical Analysis of Social Data, II

Sociology 401-2

Spring 2016

Professor: Quincy Thomas Stewart

Teaching Assistant
Jess Meyer
jessmeyer@u.northwestern.edu

COURSE DESCRIPTION: Social scientists use quantitative methods to explore and test hypotheses, describe patterns in survey and census data, analyze experimental findings, and dynamically model social relations among individuals and groups. This course is third part of the quantitative methods sequence for graduate students in sociology. In this course, we will focus on regression-like methods for categorical outcomes, notably binary outcomes, ordered outcomes, nominal outcomes, as well as, hopefully, event and count outcomes. Additionally, we will cover many of the practical issues in performing a statistical analysis of secondary data where the outcome of interest is a categorical variable. Our close consideration of modeling strategies for categorical outcomes and related practical issues is designed to help cultivate an understanding of the fundamental principles of statistical inference, data analysis, and modeling that extend beyond any specific set of techniques (i.e., a foundation). In building this foundation, our intent is to advance your capacity to be an independent social scientist who can learn new methods, perform analyses, and communicate results with others in the scientific community.

CLASS STRUCTURE/INSTRUCTIONAL TECHNIQUE: Although the traditional method of teaching statistical methods is lectures and/or proofs, we will use an alternative teaching style for this course. Specifically, we will employ a collaborative design characterized by mini-lectures (~1 hour) each course, problem based discussion (~1 hour), application demonstrations, as well as independent research, problem solving and application outside of class. This collaborative pedagogy is designed to facilitate a hands-on understanding of the respective methods, and impart and reinforce the skill of independently learning/applying advanced quantitative methods. To this end, students can expect to research methods, the related benefits and drawbacks, and apply, interpret and present them to others in the class.

COURSE GOALS: The major goals of the course for students are:

1. Advance your understanding of the basic logic, flexibility and elegance of maximum likelihood estimation.
2. Increase familiarity with a variety of models that are often used in contemporary social science.

3. Proficiency in regression models for categorical data, inclusive of their interpretation and presentation in data analysis—competence to perform an analysis that would appear in a sociology/social science journal.
4. Expertise in using statistical software—specifically, Stata—to analyze quantitative data with categorical outcomes to answer social research questions.
5. Impart and reinforce capacity to independently learn, apply, interpret, and present new quantitative methods.
6. Appreciation of both the work and fun of applied data analysis.

Cautionary Point: The sociology statistics sequence, for which this is the final course, is not intended to impart the requisite tools for a social science research career that is based primarily on quantitative analysis. For that, you will need additional training while in graduate school, more first-hand experience with the craft of data analysis, and a commitment to staying fresh with training over your career. Methodological competence, much less expertise in a particular method, is an ongoing project for all those who are continually engaged in research.

Methodological Admonition: For students who do not plan to perform quantitative research after the conclusion of this statistics sequence, the idea of becoming a literate consumer of quantitative research instead of producer may be appealing. "Please, I just want to be able to read quantitative articles and feel like I understand what's going on," may be the stylistic refrain. Of course, the cultivation of such a "consumption" capacity is all to the good, and we will take up the evaluation of quantitative evidence. But, software mastery aside, this instructor has never been sure if there really exists a competence base for quantitative-research-consumption that is separate from the competence base for quantitative-research-production. A neat parallel used by Jeremy Freese is foreign language. *Can you imagine a course of language instruction that is directed only to being able to read the language and not to being able to write it?* Analogously, our focus will be on both the consumption <u>and</u> production (i.e., reading and writing) of quantitative methods with categorical outcomes.

PREREQUISITES: This course builds upon Sociology 400 and 401-1. Thus, students are assumed to have a working knowledge of elementary statistics and multiple linear regression, as well as a basic knowledge of Stata statistical software Students who did not take Sociology 400, 401-1, and/or are unfamiliar with Stata will need to (independently) do extra work at the start of the quarter to assure they can keep pace with the course. Although other statistical packages may be used instead of Stata, the course staff will not assist with the use of other software packages and students in the course will be examined on their Stata proficiency.

COURSE REQUIREMENTS: The requirements include: class attendance and participation (i.e., challenges) (10-20%), homework (50%), and a final program/exam (30-40%).

Attendance and participation: Class attendance, a very easy variable to measure, is based on your regular presence in class. Attendance is mandatory. Students will be penalized 4% of their final grade for each unexcused absence. I will excuse absences for sickness, religious holidays, in-patient hospital admissions, job interviews, presentations at regional/national conferences, and military service. Please do not bring me documentation for any other reason.

Class participation will be measured through ongoing feedback and participation in the in-class discussions/problems during the course. In regards to feedback, you are required to submit one comment, question, or suggestion regarding the course every week of the quarter. A single sentence is sufficient, although longer comments/questions are welcome. Questions-comments-suggestions must be submitted by email to q-stewart@northwestern.edu and jessmeyer@u.northwestern.edu by the end of Friday each week. We will post these questions for discussion online and at the beginning of class. Students who fail submit feedback each week will be penalized 1 point per missed feedback—that means a 10 point reduction in your final grade if you do not send any feedback.

Participation will be assessed using extra research questions/tasks called *challenges*. A challenge, in this context, will involve you finding the answer to a specific problem, turning in your solution, and presenting/explaining/teaching your answer to the class if you are called upon—be prepared! We will discuss the format of challenges over the first few class meetings because they will evolve as we move from the early theoretical material (weeks 1-3) to the applied methods (weeks 3+). They will, however, involve students selecting and/or being assigned a question to solve on the Monday prior to our class meeting. The points associated with challenges will constitute between 10 and 20 percent of your final grade. As you complete more challenges and/or particularly difficult challenges the percent contribution to your final grade will increase—and the contribution from the final program/exam will go down.

Homework: I will assign homework each week to be completed by the following week. The homework is designed to have you apply the material covered in the mini-lecture to two data-sets, answer specific questions about the methods, research related topics on limitations and interpretations, and apply that new knowledge. A majority of the work will be computer based work, but there will also be a considerable portion of research on methods and a modicum of manipulation of equations by hand (i.e., using formulas).

There will be two datasets used for homework: 1) NHANES III that we will distribute, and 2) an *approved* dataset of your choice—this is part of the first assignment. The chosen dataset can be the same as another student. However, the variables you use cannot be the same (or virtually the same) as another student and the work you turn in must be your own. I encourage you to discuss your work with your fellow students and to learn from them, but you must complete your work on your own.

Output: For assignments involving the estimation and interpretation of data in Stata, as well as your final program/exam, you will turn in Stata output along with your assignment. You need only turn in output from commands involving transformation of variables and estimation and post-estimation commands for models that provide part of the answer to parts of the assignment. But:

(a) Everything that you turn in that involves data analysis, including any number in your final paper, **must be generated from a .do file** that is submitted with the exercise. The do file must be sufficient to reproduce all the submitted numbers, from a data file which may be requested by the instructor. (Stata has an interactive mode that is great for doing exploratory work quickly, but serious data analysis requires reproducibility, and that requires working from do files.)

(b) As part of the *.do* files you use for generating results for assignments, **you must use comments** (i.e., using lines preceded by * or //) that indicate what part of the output corresponds to what.

(c) You must, in some manner, emphasize/highlight numbers in your output that correspond to numbers in your assignment.

(d) **You must use a fixed-width font** (like Courier or Andale Mono) and your lines must not wrap. To have lines that do not wrap, use a sufficiently- small-but-still-readable font and/or use the set linesize command in Stata.

(e) Every binary variable that you turn in for any work in this course **must** be renamed and recoded such that the values of the variable are 0/1, and the name of the variable carries mnemonic significance consistent with 1=yes and 0=no.

In addition to individual homework, I may also assign group based homework. In the case of group based work, you will need to download a specific data set, perform an analysis covered in class, and interpret the results. The goal of group-based work is to familiarize each of you with downloading, manipulating and analyzing publicly available data. We will assign groups when there are group based homework questions.

Final Program/Exam: The final program/exam for the course takes the form of a Stata batch program, tabled results and interpretation, as well as (possibly) a related data, methodology, and results section that would mirror what we see in a published article. The program portion of this assignment will consist of a batch file where you code, notate (i.e., insert comments), and analyze a specific data set. This program must include clear documentation, be able to run without errors, and produce accurate results (i.e., statistics). You will also have to interpret the results of your program in an appended document, and (possibly) discuss the data set and variable manipulations, the respective methodology, and present the results in a format similar to that seen in a published article. The program/exam is an assessment of your ability to apply the methods learned in this class to a real world research problem.

CLASS AND OFFICE HOURS: The class meets on Wednesdays between 2:00 p.m. and 4:50 p.m. in Parkes Hall room 222. The lab for the course will meet on Thursday between 11:00 and 11:50am in the Library room B182. My office is located on the third floor of 1810 Chicago Ave in Room 322. I will be available in my sociology office on Mondays between 1:30 p.m. and 3:00 p.m. and by appointment. I prefer that you make an appointment before you come during my office hours. When you come to my office for a meeting please DO NOT wear perfume or cologne. *I will have to ask you to reschedule our meeting if you wear perfume or cologne to an office meeting*. My office phone number is 847-491-7044. My email address is q-stewart@northwestern.edu. I will be available for talking via email during my office hours if I do not have scheduled appointments.

TEXTBOOKS:
(Available at the Norris Bookstore)

Long, J. Scott and Jeremy Freese. 2014. *Regression Models for Categorical Dependent Variables Using Stata, Third Edition.* College Station, TX: Stata Press.

Sources used for additional course reading include:

Alan C. Acock. 2010. *A Gentle Introduction to Stata*. Revised 3rd Edition. College Station, Texas: Stata Press.

Long, J. Scott. 1997. *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: Sage.

SCHEDULE:

Week 1: 3/30	Introduction *Topics*: Linear Regression, Weights, Linear Probability Model
Week 2: 4/6	Maximum Likelihood Estimation *Topics*: Probability, Likelihood Function, ML Estimation
Week 3: 4/13	Maximum Likelihood Estimation & Binary Outcomes *Topics*: ML Estimation, Logistic Regression
Week 4: 4/20	Binary Outcomes *Topics*: Logistic Regression, Hypothesis Tests, Goodness of Fit
Week 5: 4/27	Ordinal Outcomes *Topics*: Ordinal Regression Model
Week 6: 5/4	Nominal Outcomes *Topics*: Multinomial Regression Model
Week 7: 5/11	Count Outcomes, I *Topics*: Poisson Regression Model
Week 8: 5/18	Count Models, II *Topics*: Negative Binomial Regression Model
Week 9: 5/25	Event History Models *Topics*: The Life Table, Cox Proportional Hazards, Exponential Hazard
Week 10: 6/1	Open/Spillover
Week 12:	**Final Program/Exam Due at 5:00pm, 6/7/15**