Revised: January 2013

Statistical Analysis of Social Data: Regression Analysis

Sociology 401
Winter 2013

Professor: Quincy Thomas Stewart

Teaching Assistant
Justin Louie
justin-louie@kellog.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. The aim of this course is to introduce students to the the theory and practice of linear regression analysis. We will cover correlation and bivariate regression, multiple ordinary least squares regression, regression assumptions, regression diagnostics, indicator variables and interactions, data transformations, issues in causal inference, fixed effects models, and, time-permitting, an brief introduction to logistic regression.

The course will also include discussion of practical issues in performing a statistical analysis of secondary data. This includes where to find existing data, reading data into statistical packages, and practical issues in computing to do statistical analysis—some of this work will be done in the context of our labs. We will use Stata for computer-based statistical analysis.

The major goals of the course are for students (1) to become proficient enough in regression methods to understand, explain, and critique its use in articles appearing in sociology journals and (2) to be able to perform a competent analysis of data that is of sufficient quality to appear as an article in a sociology or social science journal. The major assignment for the course will be for students to write a paper that is a data analysis of secondary data. The final paper should be similar to a draft of a publishable article.

Prerequisites: This course is built upon Sociology 400, and thus students are assumed to have working knowledge of elementary statistics, including descriptive statistics, and the general principles of statistical inference (hypothesis testing and confidence intervals). Those who do not meet this requirement should not take the course. I also assume that you have a basic knowledge of Stata statistical software from Sociology 400. Students who did not take Sociology 400 and who have not had prior Stata experience will need to do extra work at the start of the quarter to

learn Stata. Alternatively, other packages may be used instead of Stata by students proficient in their use, but course staff will not assist with the use of other software packages.

For students who took Sociology 400: A few topics covered in sociology 400 may be repeated to some extent in sociology 401-1, especially those topics most applicable to regression. You may use the same dataset as you used in sociology 400 for your final paper or you may choose another dataset. For some exercises you will be required to use data other than that used for your final paper in sociology 400.

For students who did not take sociology 400: I will assume you either know or can learn on your own most of what was covered by sociology 400. This should be true if you took prior statistics and have a good grasp of the material. Only a few parts of the material from sociology 400 will be reviewed in sociology 401-1.

Lectures and Readings: The main text for the course is Wooldridge's *Introductory Econometrics: A Modern Approach 4e*, and Allison's *Multiple Regression: A Primer*. Course readings corresponding to each lecture will appear at the start of the corresponding lecture. You should do the course readings corresponding to these assignments. Lecture notes are also a main source of course materials. Lecture notes will be made available as PDF files on the course management website in the "Course Document" folder and distributed in each class meeting.

COURSE REQUIREMENTS: The requirements include: class attendance and participation (10%), homework (20%), group projects (20%), and a final paper (50%).

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, and military service. Please do not bring me documentation for any other reason.

Class participation will be measured through feedback for the course. I will require that you submit one comment, question, or suggestion regarding the course every other week for the first eight weeks of the quarter. If your last name begins with A-K, the comment or question is due on the 2nd, 4th, 6th, and 8th weeks of the quarter. If your last name begins with L-Z, the comment or question is due on the 3rd, 5th, 7th and 9th weeks of the quarter. In addition, you may (but are not required to) submit a comment or question on the off weeks. A single sentence is sufficient, although longer comments-questions-suggestions are welcome. Questions-comments-suggestions must be submitted by email to q-stewart@northwestern.edu by Mondays at 12 noon.

Homework: I will assign homework each week to be completed by the following week. The homework is designed to have you apply and interpret the material covered in the course. Students are expected to do their own assignments. Discussion of the course material among students is allowed in the spirit of increasing common understanding. Each student, however, is expected to do their own homework and write their own homework assignments. A major section of verbatim text among submitted homework assignments is not acceptable. For the course paper, each student should write their own paper and perform their own statistical analysis without help from outside sources, except for the professor and the TA.

Group Projects: Students will participate in group projects over the course of the quarter. These projects are designed to increase students' understanding of the topic that was reviewed the week prior. Specifically, my aim is for students to work together to analyze a dataset using the techniques discussed in the prior class meetings. The group will present the findings from their analysis in class and discuss issues related to their analysis. The presentations will be short (i.e., 15 minutes) followed by a short class discussion (i.e., 15 minutes). These projects will be graded by peers and instructors. Members of each group will grade the contribution of each member to the project using a 10 point scale, other students will grade the content using the same scale, and the two instructors will grade the content and presentation. The weight of each grade will be as follows: group members 25%, class 25%, and instructors 50%.

Final Paper: The major assignment for the course is a research paper using multiple regression methods (and/or possibly logistic regression). The research paper is not expected to be fully polished, but it should have all the makings of a journal article, such as a clear research question, an appropriate data set, sufficient literature review, justified hypotheses, insightful data analysis, as well as proper interpretation and presentation of the major findings. Students will submit a proposal for the paper about half way through the quarter.

The course paper may build upon a prior draft of a paper developed previously, possibly in another course or as a prior B.A. or M.A. thesis. If so, the student is required to submit the prior version of the paper to the instructor together with their proposal for the paper.

The paper should be no less than 16 pages of double-spaced 12-point text in times new Roman font (not including title page, abstract, references, notes, and tables/figures). Furthermore, the text of the paper should be no longer than 20 pages (again, not including title page, abstract, references, notes, and tables/figures). The paper should include a literature review that motivates and contextualizes the main question, a data section that reviews the data set and makeup of each variable, a methods section outlining the preliminary analysis, and a regression analysis of the data as they pertain to the motivating question. The analysis should be clear, concise and address the question in a fundamental way that adds to our understanding of the topic. Students will submit a proposal for the paper in about the fifth week of the quarter.

This paper is an assessment of your ability to apply the methods learned in this class to a real world problem that interests you. I will grade the entire paper—though I will give more weight to your ability to appropriately perform and interpret the regression analysis.

You may not use papers submitted to other courses, or submitted as B.A. or M.A. theses to satisfy this requirement. Students may submit a paper that is a more refined and substantially newer version of previous work. However, students must submit a copy of the previous paper with the proposal for their paper in the fourth week of the course.

CLASS AND OFFICE HOURS: The class meets on Wednesdays between 10:00 a.m. and12:50 p.m. in Annenberg Hall room G29. The lab for the course will meet on Thursdays between 2:00 and 2:050pm in the Library in room B182. My office is located on the first floor of 1810 Chicago Ave in Room 322. I will be available in my sociology office on Mondays and Wednesdays between 1:00 p.m. and 2:00 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 491-

7044. My email address is q-stewart@northwestern.edu. I will be available for talking via email during my office hours.

TEXTBOOKS: There are three required texts. All are on order at the Norris Center bookstore. They are also available (often for lower prices) at major on-line retailers, including rental and chapter-purchase options. These books have been placed on reserve at the main library (but may not yet be available at the very start of the quarter). The required books are:

Wooldridge, Jeffrey. 2009. *Introductory Econometrics: A Modern Approach 5e*. Cengage College Publishers.

> This is the basic textbook for the class. The book itself is expensive, but there are cheaper online options. From Cengage online, you can buy individual chapters of the book in electronic format for $12.00 per chapter. For our class you will likely need chapters 1 through 9 (I believe chapter 1 is available for free, so it is $96.00 for chapters 2-9). Go to http://www.cengagebrain.com and search for "introductory econometrics." Make sure to select the 5th edition (5e).

Allison, Paul. 1999. *Multiple Regression: A Primer*. Pine Forge Press. Thousand Oaks, CA.

> Clear, non-technical explanations of multiple regression. A addition to Wooldridge. Available as a library e-book.

A recommended (not required) text is an introduction manual to Stata. There are also many online help and reference guide for Stata which may be sufficient especially for users comfortable with command-line software. In fact, the online resources available are often better but sometimes harder to use and more fragmented:

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

> Provides an introduction to Stata and serves as a compact Stata manual. Current as of Stata 11 (which is very similar to Stata 12). You can purchase this at a reduced price ($48+shipping) through Stata gradplan from: http://www.stata.com/order/new/edu/gradplans/gp-direct.html

Optional texts:

These texts provide an alternative discussion of several topics that we will cover, often in somewhat greater depth. E-book texts are available using any web browser through the library's web site.

Agresti, Alan and Barbara Finlay. 2009. *Statistical Methods for the Social Sciences*. 4th Edition. New York: Prentice Hall.

> Our text from sociology 400. This book covers much of what we'll discuss in class. Since it is a broader statistics text, it does not engage all of the nuances of the respective methods we'll cover in this course.

Berry, William, and Stanley Feldman. 1985. *Multiple Regression in Practice*. Newbury Park, CA: Sage.

Berry and Feldman briefly discuss specification error (omitted variables), measurement errors in variables, multicollinearity, nonlinearity, heteroskedasticity, and autocorrelation. Available as a library e-book.

Hardy, Melissa A. 1993. *Regression with Dummy Variables*. Newbury Park, CA: Sage.

Discussion in some depth the use of dummy variables as independent variable in regression. Available as a library e-book.

SCHEDULE:

Week 1: 1/9	Introduction & Review *Topics*: Data Sources, Variable Distribution, Mean, Variance, Standard Deviations, Significance Tests, Correlation
Week 2: 1/16	Simple Linear Model *Topics*: Correlation to regression, graph of averages, least squares method, inferences for slope coefficients, plotting residuals, model assumptions
Week 3: 1/23	Multiple Regression *Topics*: Multiple regression models, matrix algebra, basic ideas, three-variable example, partial effects, regression through the origin, variable selection
Week 3: 1/23	Regression Inference *Topics*: Sampling distributions of OLS estimators, hypothesis tests for regression coefficients, confidence intervals, F-tests for sets of coefficients
Week 5: 2/6	Dummy and Indicator Variables *Topics*: Intercept dummy variables, dummy variables for multiple categories, ordinal dummy variables, interpreting coefficients on dummy variables
Week 6: 2/13	Interactions *Topics*: Interactions among dummy variables, different slopes, testing for group differences in regression function (*time-permitting*: binary dependent variables)
Week 7: 2/20	Data Transformations and Nonlinear Modeling. *Topics*: Fitting curves, curvilinear regression models, choosing transformations, logarithmic transformations, evaluating consequences of transformation, interpreting non-linear models

Week 8: 2/27	Specification and Data Issues *Topics*: Outliers and influence, multicollinearity, missing data, joint hypothesis
Week 9: 3/6	Extra Topics *Topics*: Fixed effects, random effects, weights
Week 10: 3/13	(***Possibly No Class; if we have class, then we'll cover topic below***) Introduction to Logistic Regression. *Topics*: Limitation of linear regression, simple logistic regression, model assumptions
Week 11:	**Final Papers Due at 10:00am, 3/18/12**