

Sociology 401: Statistical Analysis of Social Data: Regression Analysis I Winter 2011

Course meeting: Monday and Wednesday 2:00 – 3:20, Parkes 222

Lab meeting: Wednesday 3:40 – 4:30 in University Library B183 (PC Lab)

Prof: Lincoln Quillian
1812 Chicago Avenue, Room 303
Evanston IL 60208-1330
Office Phone: 847-491-7488

E-mail: l-quillian@northwestern.edu

Office hours: Monday 3:30 – 4:30

Or by appointment

TA: Diana Rodriguez-Franco
Room B183

dianarodriguezfranco2014 @u.northwestern.edu

Wednesday 4:30 – 5:45

Overview: This course is part of the quantitative methods sequence for graduate students in sociology. The main topic of the course is the theory and practice of linear regression analysis. We will cover multiple ordinary least squares regression, regression assumptions, regression diagnostics, basic path models, data transformations, weighting, causal inference, and basic logistic regression. If time permits, we may also cover weighted least squares, instrumental variables techniques, and fixed effects models.

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. 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 the general principles of statistical inference (hypothesis testing and confidence intervals), the application of inference procedures to one or two means, and basic (one-variable) regression. Those who do not meet this requirement should not take the course. Because the course will use Stata, 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 will also be repeated to some extent in sociology 401-1, especially those topics most applicable to multiple regression. Often these topics will receive a somewhat more in-depth discussion in sociology 401-1.

I prefer you do not use the same data as used in sociology 400 for your course paper in sociology 401-1. However, I may allow exceptions to this rule if the dataset you used before (e.g. the GSS) is truly one of the best sources of data for your topic and the analysis you perform is significantly different than your prior paper.

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. Readings will also be assigned from 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 by 9 PM the day before the lecture. My recommendation is that you print out these lecture notes before class and bring them to class to take supplemental notes.

Assignments: The major assignment for the course is a research paper using multiple regression methods. 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 early in the quarter. After the proposals are submitted, students are required to meet at least once with the professor to discuss their proposals during 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.

There will also be occasional homework assignments (likely six to nine) due during the quarter. Some of these will involve rough drafts or analyses that become part of your course paper. Others will be exercises to provide practice on the basic course materials.

Two or three times during the quarter we will have a short reading assignment of an article that uses statistical methods. These articles will be made available a week in advance on the course management web site. We will discuss the article and its use of statistical methods for roughly thirty minutes. Students are expected to read the article and participate in the discussion.

Feedback: 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 1st, 3rd, 5th, and 7th weeks of the quarter. If your last name begins with L-Z, the comment or question is due on the 2nd, 4th, 6th, and 8th 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 Prof. Quillian (e-mail: l-quillian@northwestern.edu) by Friday at 12 noon.

Grades: Grading for the course will be based 55% on the final paper; 35% on homework assignments; and 10% on turning in comments and other short assignments, participation, and attendance.

Cooperation on Assignments: 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.

Labs: There is a lab session that meets on Wednesday from 3:40-4:30 PM. Lab sessions will combine instruction and practice in statistical computing with discussion of homework assignments. New material may be covered in lab, especially relating to computer use and practical issues such as finding and reading in data. Lab WILL meet on the first Wednesday of the quarter, 1/5/2011.

Texts: There are four required texts. All are on order at the Norris Center bookstore. They are also available (often for lower prices) at major on-line retailers. 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 4e*. South-Western College Publishers. This is the basic textbook for the class.

Allison, Paul. 1999. *Multiple Regression: A Primer*. Pine Forge Press. Thousand Oaks, CA. Clear, non-technical explanations of multiple regression. A good supplement to Wooldridge. Available as a library e-book.

Kohler, Ulrich and Frauke Kreuter. 2009. *Data Analysis Using Stata*. Second Edition. College Station, Texas: Stata Press. Provides a good introduction and serves as a compact Stata manual.

Pampel, Fred. 2000. *Logistic Regression: A Primer*. Sage Publications. We will use this short monograph for discussion of logistic regression.

Optional texts:

These texts provide an alternative discussion of several topics that we will cover, often in somewhat greater depth. Optional books have been placed on reserve at the main library except for texts available as an e-book. E-book texts should be available using any web browser

through the library's web site.

Chatterjee, Samprit and Ali Hadi. 2006. *Regression Analysis by Example*. Fourth Edition. New York: Wiley-Interscience Publications. This is a book used for Sociology 401-1 in the past. Chatterjee and Hadi have an unusually in-depth coverage of diagnostics and solutions for multicollinearity.

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.

The following two books are excellent introductions to their respective topics, which we will touch on, but they go into much greater depth than covered in this course:

Allison, Paul. 2001. *Missing Data*. Newbury Park, CA: Sage.

Morgan, Stephen, and Christopher Winship. 2007. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. New York: Cambridge University Press.

Both the required and optional books have been placed on reserve at the main library. Because my reserve request was late, however, some books may not be available until a few weeks into the quarter.

Course Management System (CMS):

We will use the Course Management System (aka Blackboard) for many of the administrative tasks associated with this class. I will put materials for in-class exercises as well as the data used in the exercises and problems on the CMS site. Assignments will generally be under "assignments" on the site. Lecture notes and data will generally be under "course documents."

Preliminary Schedule:

Week	Probable Topic
1	Introduction, Data Sources, Reading in Data, Simple Regression and the Simple Linear Model
1/2	Multiple Regression
Jan 17	No Class: MLK Day Observance
3	Regression Inference
3/4	Dummy and Indicator Variables
4/5	Indicator Variables and Interactions

5/6	Data Transformations and Nonlinear Modeling
6/7	Outliers and Influence; Multicollinearity and Joint Hypothesis Tests
7/8	Path Analysis; Causal Inference
8/9	Instruments and/or Fixed Effects in Regression (if time permits)
9/10	Logistic Regression
10	Catch-Up Lecture or Additional Topic