METHODS OF SOCIOLOGICAL RESEARCH – GRADUATE STATISTICS 2

Sociology 271C, Spring 2015
University of California, Berkeley

Instructor: David J. Harding, dharding@berkeley.edu
Office Hours: Fridays 10-noon or by appointment, 462 Barrows

GSI: Liana Prescott, lkprescott@berkeley.edu
Office Hours: TBD

Lecture: Wednesdays 9:30-noon, 402 Barrows
Lab: Thursdays 10-noon, Demography Department seminar room (2232 Piedmont Ave)

Course Website: https://bcourses.berkeley.edu/

Readings:

- Morgan and Winship, *Counterfactuals and Causal Inference: Methods and Principles for Social Research,* 2nd Edition [MW]
- Paul Allison, *Fixed Effects Regression Models* [PA]
- Various journal articles available as PDF on the course website [marked with * below]

Course Description

Sociology 271C is the second of two courses on statistical analysis of numerical data designed for sociology Ph.D. students. The course will cover regression, matching, instrumental variables, and related techniques for identifying causal effects, as well some extensions of multiple regression. Principal activities include: 1. Explore the statistical concepts and methods that sociologists most commonly use to gather and analyze quantitative evidence. 2. Use Stata (a popular computer program) to put those skills into practice. 3. Apply the skills to sociological data to gain facility and confidence in the use of these methods. *Students who have not taken Sociology 271B should consult the instructor before enrolling.*

Course Goals

After successfully completing this course, you will be able to:

(1) Understand the “counterfactual” or “potential outcomes” conceptual framework that provides the foundation for most modern applications of statistical methods to answering causal questions in the social sciences
(2) Formulate well-defined causal questions for quantitative social science research
(3) Develop and evaluate various strategies for answering causal questions using statistical methods
(4) Analyze the strengths, weaknesses, and core assumptions of applications of these statistical methods
(5) Implement in Stata basic versions of various statistical methods for answering causal questions

This course is less about statistics per se and more about the proper application of statistical methods for answering causal questions in quantitative social science research or for testing social science theories using quantitative data. Understanding the strengths, weaknesses, and assumptions of various techniques will be critical to understanding when to apply them and how to interpret their results. A central contribution of the counterfactual framework of causality is to spotlight implicit assumptions and inherent limitations in existing techniques. Consequently, many methodologists embrace the lessons of the new literature on causality as a call for analytic modesty. This may be discouraging to you at first. Yet an improved understanding of current limitations also prepares the way for novel solutions that stand on firmer ground than previous practice.

This course will emphasize practical application and intuition rather than mathematical details or statistical theory. (Students desiring more mathematically-oriented training are welcome to consult the instructor for other opportunities on campus.) We will work closely with real data throughout the semester in order to learn by doing.

Course Structure

The course is separated into five modules. For each module there will be readings, lectures, a one-hour student led class discussion of a conceptual or empirical paper that applies one or more topics covered in the module, and a homework assignment (more on class discussions and assignments is below). Each module will last 2-3 weeks. The lecture session of the final week of a module will include the one-hour class discussion followed by a one-hour introductory lecture on the next module that is intended to provide some context for the readings for the next module. Labs will be used to review material covered in readings and lectures and to learn implementation of methods in Stata. Finally, at the end of the semester, each student will write an empirical term paper on a topic of their choosing.

The five modules are the following:

1. Introduction and Causal Graphs
2. Matching
3. Regression
4. Instrumental Variables
5. Panel Data and Fixed and Random Effects

Core Elements of the Course

Lecture and Lab: Attendance and active participation in lecture and lab are essential. If you miss more than one lecture or lab session, your highest possible grade will be A-; if you miss more than two lectures or labs, your highest possible grade will be a B+; etc. If you miss more than four lectures and labs combined, you will receive a failing grade for the course. If an emergency necessitates that you miss one lab or lecture session, please contact the GSI and the instructor (in advance, if possible) so we can arrange for you to make it up. *If you don’t understand something, please ask!*

Reading before class: Each lecture period on the schedule is accompanied by a reading or readings that should be done before class that week. You may not understand all of what you read (that's why we have lectures) but you will be expected to get the "gist." See below for advice on effectively reading dense methodological pieces.

Assignments: Assignments will require you to apply the concepts and methods covered in the corresponding module, usually through analysis of real data in Stata. Because these assignments cover multiple weeks, they will be longer and more involved than the weekly assignments in 271B, so you are advised to start working on them before the week they are due. In most of the assignments, you will be asked to apply the methods to a research question of your choosing based on a dataset of your choosing. *Note that for this semester some of the methods are designed to be used with longitudinal data (repeated observations of the same cases), so it may not be possible to use cross-sectional data from last semester (particularly module 5). If you do not plan to continue working with the dataset you used in 271B, it is highly recommended that you spend some time during the first week of the semester to choose a dataset that interests you (see below for where to find data).* The assignments are due at the beginning of lecture on the due date. You must turn in hard copies; electronic submissions will not be accepted. One grade level will be automatically deducted from assignments that are 1-2 days late, and assignments that are more than 2 days late will not be accepted. You may collaborate on these assignments with other students but each individual student must write up his/her own work in his/her own words and submit his/her own Stata output; copying is plagiarism and will be treated as such. Assignments will be graded on the following ordinal scale: 0 = not turned in; 1 = below expectations; 2 = meets expectations; 3 = exceeds expectations. A grade of 2 = meets expectations is analogous to an A grade. See below for assignment due dates.

Class Discussions: At the end of each module, we will read a paper that applies, either conceptually or empirically, the concepts and methods covered in the module and have a one-hour discussion of the paper in class. A group of 2-3 students (depending on course enrollment) will *start* that discussion. It is the group's responsibility to orient the class to the core issues in the paper, but the group is *not* expected to "teach" the paper to the class. Each student will be assigned to one such group during the term. The group will "present" the paper briefly to the rest of the class (~15 minutes, use 10-12 slides maximum) and then provide questions/topics for discussion or elaboration (including aspects of the paper that are challenging and require further discussion to fully understand). The presentation of the paper should include a statement of the research question(s), the "counterfactuals" they imply, the data, the methodology (including assumptions, strengths, and weaknesses), and the author's interpretation of the results. The goal here is to focus on the formulation of the research question and the methodology (rather than on the theoretical motivations or implication of the findings). You should meet with the professor or GSI in office hours as you prepare (possibly more than once). Your group should come to office hours with a provisional plan and specific questions.

Final Paper: Each student will write a final paper that uses statistical methods from the course to examine one or more research questions. The paper should be 12-15 pages double spaced (not counting cover page, tables/figures, or references). The final paper will be due at **5 pm on Wednesday, May 11**. It should be uploaded on the bCourses website as a Word document. A 1-2 page single-spaced *proposal* for the final paper will be due at **5 pm on Wednesday, April 20**, also uploaded to the bCourses website. The proposal should state the research question, hypotheses, motivation, data, variables, and analysis methods. Each student will provide peer feedback to two other students on their proposals the following week. More information on the final paper, proposal, and peer feedback will be provided as the deadlines approach.

Grading: Your grade for the course will be based on the Assignments (35%), Discussion Leadership (15%), the Final Paper (35%), Peer Feedback on the final paper proposal (5%), and participation in lecture/lab (10%).

Data for Assignments and Final Paper

Over the course of the semester you will work with one (or more) datasets of your choosing in assignments and a final paper. Some students will be able to build on their papers from 271B while most others will wish to start over with a new topic or dataset. You may use any data that you like, and are encouraged to consult with the instructor and/or GSI as you choose your data. Below are some online repositories where data are available:

sda.berkeley.edu/archive.htm
icpsr.umich.edu/icpsrweb/ICPSR/access/index.jsp
norc.uchicago.edu/GSS+Website/
thearda.com/Archive/browse.asp
www.census.gov
www.ropercenter.uconn.edu/

Statistical Computing

The primary software for this course will be Stata (http://ww.stata.com). Stata is flexible, relatively user-friendly, and commonly used by social scientists. It has a large and diverse user community with many user-written commands that keep Stata continually up to data with new developments. You will need to use Stata for most assignments and for your final paper. Most Stata instruction will occur in lab, though many examples will be presented in lecture. You are encouraged to seek help with Stata questions at Berkeley's D-lab (http://dlab.berkeley.edu), located on the third floor of Barrows Hall. Another good resource is UCLA's Stata website (http://www.ats.ucla.edu/stat/stata/). You are welcome to use other statistical packages in the course (such as R), but the instructor and GSI will not be able to assist you with other software packages.

Students in 271C will have access to the Demography Computer Lab, a Linux server with a windows and terminal-based computing environment that can be accessed remotely from anywhere you have internet access. The Demography Computer Lab includes Stata and other statistical and database computing applications. This means you will not need to purchase your own copy of Stata for the course. Students are assumed to have familiarity with this system from 271B last semester. If you do not, please consult with the GSI. See also http://lab.demog.berkeley.edu/.

Keys to Success in an Introductory Statistics Course

Since this is a statistics course, it will be very different from the typical sociology course. Because most of the material is cumulative, it is absolutely essential that you keep up with the course material.

- The readings are relatively short, but they are dense and need to be read carefully. *Please do the assigned readings before lecture for the week in which they are assigned (see schedule below).* Some weeks, you will likely need to review the readings again after the lecture as well. The readings for 271C will usually be more technical than those we read in 271B. You do *not* need to understand all of the equations – any that are critical we will cover in lecture or lab. Focus on understanding the main conceptual points and the examples. You may also find it a useful strategy to read once fairly quickly to identify the main points, then re-read key sections that explain those points more carefully, and then spend time understanding the examples.

- For most students, learning statistics requires thinking through how to solve problems. Statistics cannot be learned simply by reading a book or listening to a lecture. You should not expect to fully understand the material until after you have completed the relevant assignment.
- Learning statistics is in some ways like learning a language, and it is important not to be intimidated by new terms or the use of letters (Greek letters, even) to represent quantities or concepts. It is often helpful to write in plain language the meaning of the quantities or concepts represented by a letter or symbol.
- You are strongly encouraged to do homework assignments in groups, but each student should turn in her or his own work.
- Lecture slides will be made available in advance on the bCourses website. This is so you do not have to copy formulas and diagrams during lecture, but lecture slides are not a substitute for careful note taking.
- If you find yourself falling behind, seek help *immediately* from your GSI or the professor during office hours.
- Please ask questions during lecture or during lab if you do not understand. If something is unclear to you, it is probably unclear to other students as well. Lectures and labs are planned to allow time for questions (and answers).

Weekly Schedule of Lectures/Topics, Readings, and Assignments

1	January 20	Introduction to Causal Inference	MW Ch. 1-2
2	January 27	Causal Graphs	MW Ch. 3-4
3	February 3	Endogenous Selection Discussion, Matching I	Elwert & Winship 2014*
		Assignment #1 Due February 10 in lecture	
4	February 10	Matching II	MW Ch. 5
5	February 17	Matching Discussion, Regression I	Morgan 2001*
		Assignment #2 Due February 24 in lecture	
6	February 24	Regression II	MW Ch. 6
7	March 2	Regression III (Weighting)	MW Ch. 7
8	March 9	Regression Discussion, Instrumental Variables I	Sampson et al 2006*
		Assignment #3 Due March 16 in lecture	
9	March 16	Instrumental Variables II	MW Ch. 8-9
	March 23	*No lecture or lab (spring break)*	
10	March 30	Instrumental Variables III (Regression Discontinuity)	Lee & Lemieux 2010*
11	April 6	Instrumental Variables Discussion, Panel Data and Fixed Effects I	Kirk 2009*
		Assignment #4 Due April 13 in lecture	
12	April 13	Panel Data and Fixed Effects II	PA p. 1-21
		Final Paper Proposal Due April 20, 5 pm	
13	April 20	Random Effects and Hybrid models	PA p. 21-26, MW Ch. 11
		Peer Feedback on Proposals Due April 27, 5pm	
14	April 27	Panel Data Discussion, Wrap-Up	Vaisey & Miles 2014*
		Assignment #5 Due May 4, 4 pm	
	May 11	Final Paper Due (5 pm)	

MW = Morgan and Winship book
PA = Paul Allison book
* reading available as PDF on course website