Quantitative Analysis SOCI6112, POLS6340, SOWK6112
Course Outline

Fall/Winter 2017-18

Instructor: Michael Ornstein, ornstein@yorku.ca,
Office hour in TEL5057, Thursdays 1-2pm and by appointment

Description

This course teaches the skills to write a journal article, thesis or dissertation based on a survey or other quantitative social data. The main focus is regression models, which are the standard analytical tool of contemporary quantitative social research, a prerequisite for reading and evaluating published research and the basis of advanced models. Quantitative analysis is approached as a craft that combines a knowledge of statistics, an understanding of social science data and their limitations, a feeling for the translation of theoretical questions into testable models and the ability to interpret and write about analytical results.

The fall term begins with a review of elementary statistics, combined with basic data management and statistical computing. The main part of the course begins with a detailed discussion of ordinary least squares regression and then considers models for categorical outcomes, such as electoral party support, and for "counts". The last part of the course introduces more advanced topics. The emphasis is on the analysis of survey data. The teaching incorporates many examples, especially from the study of inequality and political attitudes.

Skills in data management and statistical computing and integrated throughout the course, as is consideration of the quality of social data, including non-response and measurement error.

The view of data analysis informing this course understands regression as a very nice way to describe cross-sectional, observational data. Often it is possible to make causal inferences. Without longitudinal or experimental data, however, reaching conclusions about cause and effect are often difficult. Regression is not a mathematical trick to overcome fundamental weaknesses in theory or data. Still, careful analysis of somewhat flawed data beats ignoring evidence and often it is important to describe a social phenomenon even when it is impossible to properly understand its causes.

While many social phenomena cannot be fully understood without quantitative data, quantitative analysis is not inherently superior to qualitative methods. Many fundamental concerns about empirical research apply equally to quantitative and qualitative approaches.

Lectures and Texts

This is a lecture course. Slides providing extensive explanations for each topic, as well as the datasets, command files and output for examples will be available on Moodle prior to each lecture. The course texts provide helpful explanations as well as a systematic reference, but the lectures do not follow them. Students are expected to read actively to solidify their understanding of the lecture material and to practice statistical techniques with their own data.

The course text is Kohler and Kreuter's *Data Analysis Using Stata* (3rd ed. College Station, TX: Stata Press, 2012). It provides good, but somewhat limited treatment of statistical topics in the context of a very effective guide to Stata. Stata has great help screens and thousands of pages of excellent statistical and computational documentation. For a more detailed and very clear explanation of regression read the supplementary text by Rachel Gordon, *Regression Analysis for the Social Sciences* (*2nd ed.*, New York, Routledge, 2015), which has been ordered.

Students without a working knowledge of elementary statistics would profit from reading a basic text. There are many good ones tailored to different disciplines and it is not necessary to have the latest edition. One very nice introduction to elementary statistics is David Freedman, Robert Pisani and Roger Purves, *Statistics*, (4th ed., New York: Norton, 2007). A light hearted and effective guide is Roberta Garner's *The Joy of Statistics* (2nd ed., Toronto: U of T Pr., 2010).

Data for Your Assignments and Paper

The best way to learn quantitative methods is by analyzing data in your own research area. In the past, some students begin the course knowing what data they want to use, while others have to locate some interesting data. It is important to do this early in the course, because effective data analysis requires a familiarity with a dataset that takes time to develop. Of course, you can get advice.

Ontario's ODESI archive (available from the York library site - just search for ODESI and log in with Passport York) and many other websites offer a huge variety of free, suitable datasets. The only restrictions are that your dataset should not be too small and there should include at least one reasonably continuous "outcome" measure. Your dataset and the topic of your paper must be approved by the instructor.

Software

The course uses Stata, a clean, elegant and powerful environment for social data analysis. You are advised to purchase it, for $89US for a download of the IC version, one-year license. The "perpetual" license at $198US is a very good buy if you intend to use Stata for further research or coursework. A perpetual license means you have no annual fee though eventually you may want to buy a newer version. Unless there is a new feature that you need, you can go for 3-4 years before an upgrade.

IC and SE are statistical except that SE allows you to use datasets with a very large number of variables (>2000). You can upgrade from IC to SE at any point by paying the difference in price.

Order Stata at http://www.stata.com/coursegp between August 7, 2017 and April 30, 2018. Select the package you want, enter your address information, and specify the GradPlan ID MO6112 in the GradPlan ID field of the End-User Information tab during the checkout process.

In USD prices are: Stata/IC software: $89US/one year, $198/perpetual
Stata/SE software: $235/one year, $395/perpetual

Software is delivered electronically. Download instructions and license information are sent after orders have been processed, typically within one day of order receipt.

Stata can be used without charge via York's WebFAS remote computing environment and this cloud version of Stata works identically to the version you purchase, *but* it requires an internet connection at all times and is more cumbersome because of the surrounding WebFAS cloud computing environment. For instructions, search for York WebFAS and install the (free) Citrix interface. The WebFAS version allows you to read and store command and data files (i.e. files on your own computer) and to read and write local command files.

http://www.youtube.com/user/StataCorp has an extensive and useful library of Stata YouTube videos, and there are many others at university websites and from freelancers.

Course requirements and Grades

The course requires four assignments and a paper based on a dataset reflecting your own research interest. Ideally, the multiple and/or logistic regression assignments will form the basis for your course paper.

1. Descriptive statistics assignment	5%
2. Simple regression assignment	10%
3. Multiple regression assignment	15%
4. Logistic regression assignment	10%
5. Course paper, first draft	45%
6. Course paper, second draft	15%

Also required, but not graded: course paper outline and some exercises

Michael Ornstein

Sociology 6112 Fall/Winter 2017-18

Thursdays, 2:30-5:30pm
begins Sept. 7, until Nov. 30, with no class Oct. 26;
in the winter, Jan. 4 to March 29, with no class Feb. 22.

Topics

The numbering of the topics below corresponds to weeks, approximately.

1. Introduction
    a. Goals and organization of the course
    b. Introductions and comments
    c. Regression analysis example
    d. How Stata works: three windows for the data matrix, commands and output
    e. Short student survey

    After this lecture: purchase and install Stata or learn how to access it on WebFAS

    Reading: Kohler and Kreuter (hereafter K & K): skim Ch. 1

    Instead of reading K & K in a task-focussed way, chapter by chapter, consider reading through the entire book, skipping any difficult parts and without trying to remember the computational details. Reading this way will help develop your ideas about the enterprise of data analysis.

2. Describing distributions with graphics and numbers
    a. Types of variables
    b. Statistics is all about distributions
    c. Graphics and display tables for distributions
    d. Measures for scalar distributions, including order statistics (and why no one number can characterize a distribution)
    e. Missing data and measurement data
    f. Using Stata to read data, create and transform variables, graph and summarize distributions, including barcharts, histograms and boxplots

    Reading: K & K: Ch 5 (sections 1, 2, 3, 5, 6), Ch. 6 (but skim 6.3), Ch. 7
    : Stata documentation[1] for `codebook summarize, generate, replace, recode, graph bar and histogram`

[1] *To obtain documentation for a command in Stata*: in the command line at the bottom of Stata's main screen type `help xyz` (where xyz is any command). This brings up a new window with a brief description of how to use the `xyz` command, including all its options and a few examples.

For more complete and prettier documentation click the blue highlighted name of the command in that initial help screen -just below "Title". The documentation appears in PDF format, to read or download or print, and it includes the basic information in 1, above.

Not-for-credit exercise: download the dataset and do file for this lecture; execute the commands; and change the variables and do some similar analysis

Practice datasets you can download from Moodle:

Toronto UrbanHEART
2011 Canadian Election Survey
2006 long form Census 10K sample
2012 Labour Force Survey 10K sample

3.-4. Sampling Distributions and Inferences about Means and Proportions

a. what is a random sample, and what is sampling error?
b. resampling exercise
c. the Central Limit Theorem
d. confidence intervals and one-sample significance tests for scalars
e. binary variables
f. the sampling distribution of the variance

Reading: K & K: Ch. 8
Assignment 1: univariate statistics
Now is the time to start thinking about the data for your next assignments and the course paper; a fallback is to use one of the datasets loaded on Moodle

5-6. Comparing distributions for groups

a. graphical comparisons
b. comparing two groups
c. 2-way contingency
d. comparing group means with one-way analysis of variance
e. comparing variances
f. variance-equalizing transformations

Stata documentation for `oneway` and `ttest`

7. Two helpful topics: sample design and weights; and constructing a simple scale

Often, data are often not from a simple random sample or cannot be treated as such, due to non-response. Weights are used to correct for differences between the sample and population, in order to generate unbiased estimates of population characteristics.

Creating scales by combining two or more survey items is a common part of data craft.

a. Types of samples: simple random samples, stratified samples, cluster samples, multistage samples
b. Effects of stratification and clusters on estimation of and confidence intervals for the population mean
c. Sample weights in Stata
d. Creating a simple additive scale

Reading: K & K: Ch. 3.3
Requirement: Course paper proposal, including the research question, a brief description of the data you intend to use and very brief reviews of 2-4 articles with similar research, ***due in two weeks***

Read about weights in the Stata documentation and about the command `alpha`

If you intend to use a scale in your assignments or paper, now is a good time to practice building a scale, and it would be a good idea to submit it for ungraded comment

8-9. Correlation and Simple Regression

a. displaying bivariate relations with scatterplots
b. Pearson's correlation
c. simple linear regression
   i. defining regression as the comparison of means
   ii. where to put the line: the OLS criterion and ANOVA for the regression
   iii. model predictions and residuals
   iv. regression diagnostics, including Anscombe's quartet
d. What if X is binary or ordinal? What if Y is binary or ordinal?

Reading: K & K, Ch. 9.1
Assignment 2: simple regression, due in class in two weeks
Due today: plan for your course paper, including reviews of 2-4 articles using the same or similar data

When you have received feedback on the paper proposal, start work!

10-13. Multiple Regression

a. The multiple regression model
b. Regression with two predictors in detail - how the formulas help understand multicolinearity
c. Causality in regression

K & K: Ch. 8, 9.2

d. Model specification
   i. non-linear effects
   ii. factors
   iii. interactions
e. Effects and effect plots

Reading: *K & K: Ch. 9.4*

Stata help on `margins and marginsplot`

Shelley Phipps, Peter Burton and Lynn Lethbridge "In and out of the Labour Market: Long-Term Income Consequences of Child-Related Interruptions to Women's Paid Work" Canadian Journal of Economics 34(2, May, 2001): 411-429

f. ANOVA for the entire model and for factors
g. Model diagnostics
    i. Multicolinearity
    ii. Non-linearity
    iii. Heteroscedasticity
h. Outliers and influential observations

Reading: K & K: Ch. 9.3

Thomas Lemieux, "The "Mincer Equation" Thirty Years after *Schooling, Experience, and Earnings*", Chapter 11 of Shoshana Grossbard, ed., *Jacob Mincer: A Pioneer or Modern Labor Economics.* Berlin: Springer, 2006: pp. 127-145. (search the author's name and the title to find the pdf download from UBC)

Also try Solomon W. Polachek (2007) "Earnings Over the Lifecycle: The Mincer Earnings Function and Its Application" IZA DP No. 3181.

i. Model selection
j. Comparing the effects of individual variables and factors
k. Separate models for different groups, and the Blinder-Oaxaca decomposition
l. Interval regression for ordinal and censored outcomes
m. Robust regression

Required, not-for-credit exercise: specifying models

14. Two Regression examples in detail
    a. Wage equation
    b. Predicting attitudes
    c. How to report the results of regression analysis in a paper, using tables and graph and in text

Reading: K & K, Ch. 9.5

Assignment 3: multiple regression, due in class in two weeks
Reading TBA: example of a regression article

15. Handling Missing Data
    a. How does missing data arise?
    b. Easy fixes: omit observations, substitute means
    c. Predictive strategies to preserve observations and improve estimates
    d. Multiple imputation, the rocket science approach

16-17. Binary Logistic Regression

a. the model, its coefficients and estimation
b. Parameters and goodness-of-fit measures
c. Diagnostics for logistic regression
d. How to report the results of logistic regression in a paper, using tables and graph and in text

Reading: K & K, Ch. 10.1-10.6

Telles, Edward and Stanley Bailey, "Understanding Latin American beliefs about racial inequality," American Journal of Sociology 118, No. 6 (May 2013): 1559-95

Stata documentation for intreg*, attending to the comparison between the metric and log interval regressions and the ordered logistic regression*

Assignment 4 logistic regression, due in class in two weeks

18. The Generalized Linear Model

a. Limits of using OLS for discrete outcomes, count data and censored data
b. Thinking about regression models in terms of a link and error
c. Error families, the canonical link, and why GLMs are characterized by their errors
d. OLS as a special case
e. Maximum likelihood estimation

19-20. Multinominal and Ordinal Logistic Regression

a. the multinomial logistic model
b. ordinal logistic regression

Reading: K & K, Ch. 10.7

21. Regression models for count data

a. The Poisson and negative binomial distributions
b. Why a log link?
c. Interpretation of the coefficients and model effects

Two Deadlines

Week 20: First Draft of Final Paper due in class (returned in next class)
Last class of the year: Second Draft of Final Paper due in class

Weeks 22-24

If there is time at the end of the course, I will do brief Introductions to some advanced methods, decided by the class. Possible topics include:

- Multiple-equation Models
- Factor Analysis
- Structural equation models

- Multi-level models