APPLIED STATISTICS: USING LARGE DATABASES IN EDUCATION RESEARCH

RESCH.GE.2110
Course Syllabus – Spring 2012

Professors:
Sean P. Corcoran
665 Broadway, Suite 805 (IESP)
Phone: (212) 992-9468
Email: sean.corcoran@nyu.edu
Office hours: Wed. 3 – 4:30 or by appt.

Joel A. Middleton
246 Greene Street, 803W (PRIISM)
Phone: (212) 992-6701
Email: joel.middleton@nyu.edu
Office hours: Mon. 12:30 – 2 or by appt.

Lecture Wed. 4:55 – 7:35 p.m.
Waverly Building Room 668

Course description

This course is designed to serve as a bridge between introductory statistics/econometrics and practical work with real, large-scale databases. Although the focus is mainly on datasets relevant to education and education policy research, the skills taught in the course are broadly transferable across subject areas in social, behavioral, and health sciences.

At the conclusion of this course students will be prepared to produce descriptive statistics about a population using data collected under complex survey design and to estimate simple cross-sectional and panel regression models of the sort frequently employed in real data analysis. The emphasis throughout the course is on hands-on data preparation and modeling using the Stata statistical software package.

Prerequisites

At a minimum, one semester of introductory statistics is required. Topics covered should have included simple linear regression, hypothesis testing, and basic topics in descriptive statistics and probability. The course RESCH.GE.2001 (Statistics for the Behavioral and Social Sciences I) fulfills this requirement.

In addition, students should have either completed or be concurrently enrolled in a course on multiple linear regression or econometrics (e.g. RESCH.GE.2002, Statistics for the Behavioral and Social Sciences II, or PADM-GP.2902). Students not meeting this requirement must demonstrate satisfactory knowledge of multiple regression prior to enrolling in the course. No prior experience with Stata is assumed or required. If you have concerns about your prior preparation, please see one of us.

Books

There is no required text for this course. Lecture notes prepared by Professor Jack Buckley, *Analysis of Large-Scale Education Data with Stata* (2009) will be available on Blackboard, along with other required and recommended readings. Advanced graduates may find the following book useful:

Applied Survey Data Analysis, by Heeringa, West, and Berglund, 2010, CRC Press (HWB).

We do recommend you buy a guide to Stata for your own reference, especially if you are new to this software. There are a number of good books on this topic, all available from the Stata Press. In order of most basic to most advanced, we recommend:

(*) A Gentle Introduction to Stata, 3rd edition by Alan C. Acock, 2010 (AA).

An Introduction to Modern Econometrics Using Stata by Christopher Baum, 2006 (CB).

Microeconometrics Using Stata, revised edition by Cameron and Trivedi, 2010 (C&T).

We also highly recommend the UCLA Stata guide, which includes tutorials, references, examples, and useful links (http://www.ats.ucla.edu/stat/stata/). If you plan to create a lot of graphics in Stata, the following book is indispensible:

(*) A Visual Guide to Stata Graphics, 3rd edition by Michael N. Mitchell, 2012.

Computer labs and software

Successful completion of this course will require the use of Stata (any version after 9.0 should work, but we recommend the most recent release, 12.0). Access to Stata is possible through any of three methods: (1) campus computer labs, (2) the NYU Virtual Computer Lab, and (3) purchase.

Stata is installed in labs at the ITS Washington Place Academic Technology Center and the NYU Data Service Studio (DSS) on the 6th floor of Bobst. The DSS offers free on-site help with Stata and other software. Contact them or stop by for more information, or to make an appointment. They also offer occasional tutorials on Stata (click here for more information).

NYU recently began a Virtual Computer Lab which provides access to university-licensed software from anywhere with an NYU student login. You can access this site through NYUHome, or here.

Students can purchase a one-year or perpetual Stata license for their own computer. Student proces are available through Stata's GradPlan. At last check, a six-month license for the "Intercooled" (IC) version of Stata was $65; a one-year license was $98; and a perpetual (forever) license was $179. Do not purchase the "Small" version of Stata as this is insufficient for the datasets we will examine. Stata IC is installed in the labs and should suffice for most purposes. Power users who expect to use very large databases in the future may wish to buy the (more expensive) Stata SE or MP.

Please bring some form of data storage (e.g. a flash drive) to class each week. A Dropbox account is another alternative for storing data and working files.

Course requirements

Your grade for this course will be based on six (6) practical problem sets that will require the use of Stata and real datasets to complete. Each problem set is weighted equally (16.6% each) and the dates of assignment and completion are listed in the course outline below.

Unless prior arrangements have been made with the professors, problem sets submitted past the original due date will be penalized at the rate of 10 percentage points per week (approximately one complete letter grade). In addition, each student must hand in his or her own work for each problem set. Collaborative work will not be accepted.

Other class information

1. Blackboard: All materials pertaining to this course (lecture notes, readings, problem sets, data) will be made available via Blackboard, accessible through NYUHome. Enrollment in the course should automatically give you access to the Blackboard site. Check frequently for new material and announcements. Lecture notes will generally be posted 1-2 days in advance of class. However, occasional (hopefully rare) delays can be expected

2. Lab etiquette: The class is held in a computer lab. To help promote a productive learning environment, please keep all other internet activities (e.g. email) to a bare minimum. Please do not use Facebook, instant messaging, or other such services while in the lab, and do not use class time to work on your problem sets (unless we formally give you class time to do so)

3. Academic integrity: NYU Steinhardt policies on academic integrity will be *strictly enforced* in this class. You can find the school's official statement on academic integrity here. You are encouraged to study and work together on problem sets, but all submitted work must be that of the individual student.

4. Withdrawal: If you wish to withdraw from the course, please do so formally with the University Registrar. If you withdraw without authorization, you are at risk for receiving a failing grade for the course. *February 12 is the last day for graduate and undergraduate students to withdraw without receiving a "W" on their transcripts.*

5. Accommodations: Any student requiring an accommodation due to a chronic psychological, visual, mobility and/or learning disability, or who is Deaf or Hard of Hearing, should register with and consult with the Moses Center for Students with Disabilities at 212-998-4980, 726 Broadway, 2nd floor (www.nyu.edu/csd). Of course, we are happy to provide any and all accommodations recommended by the Moses Center.

CLASS SCHEDULE

January 25	1. Introduction to "large scale" datasets and Stata	*PS #1 assigned*
February 1	2. Stata – advanced commands and graphs	*PS #1 due* *PS #2 assigned*
February 8	3. Extracting and preparing longitudinal survey databases	
February 15	4. Sampling and sampling distributions	*PS #2 due*
February 22	5. Complex survey designs	*PS #3 assigned*
February 29	6. Applications of multiple linear regression (I)	
March 7	7. Applications of multiple linear regression (II)	*PS #3 due*
March 14	**NO CLASS—SPRING BREAK**	
March 21	8. Guest speakers: Research Alliance for New York City Schools: large-scale databases in New York City	*PS #4 assigned*
March 28	9. Panel data methods (I)	
April 4	10. Panel data methods (II)	*PS #4 due*
April 11	11. Panel data methods (III)	*PS #5 assigned*
April 18	12. Replication exercise	
April 25	13. Generalized linear models (I)	*PS #5 due* *PS #6 assigned*
May 2	14. Generalized linear models (II)	

Problem set #6 due on or before Wednesday May 9th, 6:00 p.m.

COURSE OUTLINE

(*) = required reading, all others are recommended
(S) = sample papers using the databases introduced in class

Week 1: Introduction to "large scale" datasets and Stata

- What are "large scale" datasets? Examples
- Microdata issues: data security, privacy, confidentiality, and FERPA
- Electronic access to large scale data—query tools, table and model servers
- Obtaining longitudinal survey data
- Dataset documentation and codebooks; electronic codebooks
- Introduction to Stata
    - Basic operations, reading/writing data files, importing and converting files, "do" files
    - Stata syntax
    - Labeling variables and values
    - Creating and re-coding variables
    - Summarizing data: descriptive statistics and tables
    - Correlation, t-tests, and simple linear regression
    - Help files and user-written commands (.ado)

Readings:
(*) Buckley, chapters 1-3.

(*) Schneider et al. (2007) *Estimating Causal Effects Using Experimental and Observational Designs*, chapter 1, "Introduction," and skim chapter 4, "Analysis of Large-Scale Datasets: Examples of NSF-Supported Research"

(AA) chapters 1-5, 7-8; (CB) chapters 1-2 and Appendix A; (C&T) chapters 1-2.

Databases:
Common Core of Data, or CCD (http://nces.ed.gov/ccd/aboutCCD.asp)

(S) Clotfelter, Charles T. (2001). "Are Whites Still Fleeing? Racial Patterns and Enrollment Shifts in Urban Public Schools, 1987-1996." *Journal of Policy Analysis and Management* 20, no. 2: 199-221. http://www3.interscience.wiley.com/cgi-bin/jhome/34787. [CCD]

Week 2: Stata: advanced commands and graphs

- Combining and merging datasets
- Re-shaping datasets
- Processing observations within subgroups (e.g. egen)
- Stata programming: local and global macros, loops
- Univariate graphs (e.g. boxplot, bar chart, histogram, kernel density estimation)
- Bivariate graphs (e.g. scatterplot, violin, sunflower, line fit, lowess)

Readings:
(*) Buckley, chapters 3-4

(AA) chapters 5-6; (CB) chapter 3 and Appendix B; (C&T) chapters 1-2

Databases:
Digest of Education Statistics (http://nces.ed.gov/Programs/digest/)

(S) Card, David and Alan Krueger. 1992. "Does School Quality Matter? Returns to Education and the Characteristics of Public Schools in the United States," *Journal of Political Economy,* 100(1): 1—40. [Digest]

Week 3: Extracting and preparing longitudinal databases

- Obtaining longitudinal data from NCES and other sources (e.g. NCHS, IPUMS, ICPSR)
- Dataset documentation and codebooks; electronic codebooks

Databases:
ECLS-K K-8 Manual – Part 1 Chapter 1 – 2
(http://nces.ed.gov/ecls/data/ECLSK_K8_Manual_part1.pdf)

ECLS-K K-8 Manual – Part 2 Chapters 7, 10
(http://nces.ed.gov/ecls/data/ECLSK_K8_Manual_part2.pdf)

(S) Fryer, Roland G. and Steven D. Levitt. (2004). "Understanding the Black-White Test Score Gap in the First Two Years of School," *Review of Economics and Statistics,* 86(2): 447-464. [ECLS-K]

Week 4: Sampling and sampling designs

- Simple random sampling
- Stratification
- Cluster sampling
- Unequal probability sampling
- Weights

Readings:
(*) Buckley, chapter 5.

(HWB) Chapter 2 – on Blackboard

Databases:
National Household Education Survey (NHES) (http://nces.ed.gov/nhes/index.asp)

(S) Howell, Jessica S. and Michelle E. Sheran (2008). "Homeschooling in the United States: Revelation or Revolution?" Working Paper, CSU Sacramento. [NHES and GSS]

(S) Carver, P.R., and Iruka, I.U. (2006). *After-School Programs and Activities: 2005* (NCES 2006-076). U.S. Department of Education. Washington, DC: National Center for Education Statistics

Week 5: Complex survey designs

- Methods of variance estimation
- Consequences of ignoring design
- Stata's `svy` command
- Examples: NHES and PISA public use files

Readings:
(*) Kreuter, F. and R. Valliant. 2007. "A Survey on Survey Statistics: What is Done and can be Done in Stata." *Stata Journal,* 7(1): 1-21

(HWB) Chapter 3 – on Blackboard

(C&T) chapter 3, section 7.

Hahs-Vaughn, D.L. 2006. "Weighting Omissions and Best Practices When Using Large-Scale Data in Educational Research," *Association for Institutional Research*, Professional Files Online No. 101.

Databases:
Programme for International Student Assessment (PISA) (http://nces.ed.gov/surveys/pisa/)

(S) Fuchs, Thomas, and Ludger Wößmann. (2007). "What Accounts for International Differences in Student Performance? A Re-examination using PISA data." *Empirical Economics* 32, no. 2: 433-464

Week 6: Applications of multiple linear regression (I)

- Review of multiple linear regression – theory
- Multiple regression in Stata
- Reading and interpreting results
- Presenting regression results (e.g. outreg)
- Hypothesis testing
- Functional form
- Dummy regressors and interaction terms
- Weighting

Reading:
(*) Buckley, chapters 6-7.

(*) (CB) chapters 4-5, 7.

(AA) chapter 8, 10; (C&T) chapter 3.

Studenmund, chapters 1-7

UCLA webbook *Regression with Stata* (http://www.ats.ucla.edu/stat/stata/webboks/reg/)

Week 7: Applications of multiple linear regression (II)

- The linear probability model
- Autocorrelation and heteroskedasticity
- Diagnostics and remedies
- "Robust" standard errors
- Partial multicollinearity

Databases
U.S. Census Elementary and Secondary Education Finance Data (F-33)
(http://www.census.gov/govs/school/)

(S) Corcoran, Sean P. and William N. Evans. (2010). "Income Inequality, the Median Voter, and the Support for Public Education." *NBER Working Paper* #16097.
http://papers.nber.org/papers/w16097

Week 8: Research Alliance for New York City Schools – Guest Lecture

Research Alliance for New York City Schools (http://steinhardt.nyu.edu/research_alliance/)

Marinell, William. 2011. *The Middle School Teacher Turnover Project - A Descriptive Analysis of Teacher Turnover in New York City's Middle Schools*. Research Alliance for New York City Schools Research Report.

Week 9: Panel data methods (I)

- Working with panel data in Stata – special commands
- Why use panel data?
- First difference model
- Fixed effects model(s)
- Random effects models

Readings:
(*) Buckley, chapter 10.

(*) (CB) chapter 9, section 1.

(C&T) chapter 8.

Studenmund, chapter 16.

Databases:
Texas Academic Excellence Indicator System (AEIS)
(http://ritter.tea.state.tx.us/perfreport/aeis/index.html)

(S) Ladd, Helen F. (1999). "The Dallas School Accountability and Incentive Program: An Evaluation of its Impacts on Student Outcomes." *Economics of Education Review*. 18(1): 1—16.

Week 10: Panel data methods (II)

(S) Acemoglu, Daron and Joshua Angrist. (2001). "Consequences of Employment Protection? The Case of the Americans with Disabilities Act." *Journal of Political Economy*. 109(5): 915—957.

Week 11: Panel data methods (III)

- Panel attrition
- Item missing data
- Reweighting and weight stabilization
- Clustered and hierarchical data

Readings:
(*) Buckley, chapter 10.

(CB) chapter 9; (C&T) chapter 9.

Week 12: Replication Exercise

- TBD

Week 13: Generalized linear models (I)

- Logit / logistic regression
- Probit models
- Marginal effects and predictions
- Extensions

Week 14: Generalized linear models (II)

- Multinomial logistic model
- Ordered logit model
- Count data models (overview)

Readings:
(*) Buckley, chapter 11.

(*) (CB) chapter 10.

(AA) chapter 11; (C&T) chapter 14.

Studenmund, chapter 13.

Week 14: Qualitative dependent variables (II) – Other models

Readings:
(CB) chapter 10; (C&T) chapter 15, 17.