APPLIED STATISTICS: USING LARGE DATABASES IN EDUCATION RESEARCH

APSTA.GE.2110
Course Syllabus – Spring 2014

Professors:
Sean P. Corcoran
665 Broadway, Suite 805 (IESP)
Phone: (212) 992-9468
Email: sean.corcoran@nyu.edu
Office hours: Mon. 2:30 – 4:30 or by appt.

Kathleen Ziol-Guest
726 Broadway, Room 524 (Institute for Globalization and Education)
Phone: (212) 998-5478
Email: kmz204@nyu.edu
Office hours: Wed. 10-11:30 or by appt.

Lecture Wed. 4:55 – 7:35 p.m.
Tisch Hall Room LC19

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. Emphasis throughout the course is on hands-on data preparation, workflow, and modeling using the Stata statistical software package.

Course objectives

Upon completion of this course, students will be able to:

- Identify, acquire, and prepare a large-scale database for use in a research project
- Understand and apply the necessary steps in planning a research project with large data
- Understand and apply principles of dataset preparation and workflow, including cleaning, documentation, automation, and replication
- Create a codebook and other data documentation appropriate for a research project
- Understand statistical sampling distributions and the implications of complex survey designs for statistical inference
- Produce descriptive statistics using data collected under a complex survey design
- Estimate simple cross-sectional and panel regression models of the sort frequently used in analyses of large-scale databases
- Replicate the empirical analysis of an existing piece of published research

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, as does CORE.GP.1011 (Statistical Methods for Public, Nonprofit, and Health Management).

In addition, students should have either completed or be concurrently enrolled in a course on multiple linear regression or econometrics, such as RESCH.GE.2002 (Statistics for the Behavioral and Social Sciences II) or PADM.GP.2902 (Multiple Regression and Introduction to Econometrics). No prior experience with Stata is assumed or required. If you have concerns about your prior preparation, please see one of us.

Books

The following book by J. Scott Long is required:

(*) The Workflow of Data Analysis Using Stata, by J. Scott Long, 2009, Stata Press.

Many of the practical topics we will cover in class come from this book. If you are new to Stata, we recommend you buy a guide to Stata for your own reference. There are a number of good books on this topic, all available from the Stata Press. From most basic to most advanced, we recommend:

(*) Getting Started with Stata for Windows, 2013. (*free*) Also: Mac and Unix versions.

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

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

Microeconometrics Using Stata, revised edition by Cameron and Trivedi, 2010.

We also recommend the UCLA Stata guide, which includes tutorials, references, examples, and useful links (http://www.ats.ucla.edu/stat/stata/). We will post other useful Stata references on the class website. For creating graphs in Stata, the following book is indispensible:

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

Advanced students may find the following books on survey methodology useful. We will make some use of both:

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

Survey Methodology, 2nd edition by Robert M. Groves et al., 2009, John Wiley & Sons.

Computer lab and software

Successful completion of this course will require the use of Stata (any version 11.0 or later should work, but we recommend the most recent release, 13.0). Access to Stata is possible through any of three methods: (1) the Virtual Computer Lab, (2) the (real) computer labs, and (3) purchase.

(1) NYU operates a service called the Virtual Computer Lab (VCL) which provides access to university-licensed software from anywhere with an NYU student login. You can access the VCL through NYUHome or: https://vcl.nyu.edu/vpn/index.html. Currently, version 13 of Stata SE is accessible through the VCL. Please note that students have experienced problems with the VCL in the past (e.g. downtime, slow connections). Use at your own risk.

(2) As a student you have access to campus computer labs with your ID. (Click here for a list of campus labs that offer Stata). Lab attendants are not typically experts in Stata, but they can answer system-level questions about opening files, saving, printing, etc. NYU Data Services, located on the 5th floor of Bobst, offers consulting to students who need assistance with statistical software. Contact them for more information, or to make an appointment. Data Services offers occasional tutorials on Stata, SPSS, and other software.

(3) You may be interested in buying Stata for your own computer. Stata version 13 can be purchased at a discounted student rate (the "Campus GradPlan"). "Small" Stata is the least expensive ($35 for six months or $49 for a year), but is limited in the size of datasets it can manage. "Intercooled" Stata is the next level up ($69 for six months or $98 for a year); it can accommodate most projects, but for very large databases a more expensive version may be needed (e.g., SE or MP, available only in labs). For most purposes, you will notice few differences between versions 11-13. However, be aware that minor differences do exist.

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. NYU Classes: All materials pertaining to this course (lecture notes, readings, problem sets, data) will be made available via NYU Classes. Enrollment in the course should automatically give you access to the class site. Check in frequently for new materials and announcements. Lecture notes and other relevant materials will generally be posted in advance of class. However, occasional (hopefully rare) delays are to 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).

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 18 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 29	**WEEK 1**: Introduction to "large scale" datasets	
February 5	**WEEK 2**: Programming in Stata	*PS #1 assigned*
February 12	**WEEK 3**: Workflow—organizing and planning a project	*PS #1 due* *PS #2 assigned*
February 19	**WEEK 4**: Accessing large scale databases	
February 26	**WEEK 5**: Workflow—data preparation and cleaning	*PS #2 due* *PS #3 assigned*
March 5	**WEEK 6**: Workflow—automation, documentation and replication	
March 12	**WEEK 7**: Workflow—descriptive analysis	*PS #3 due* *PS #4 assigned*
March 19	**NO CLASS—SPRING BREAK**	
March 26	**WEEK 8**. Guest speaker	
April 2	**WEEK 9**: Sampling and sampling distributions	*PS #4 due* *PS #5 assigned*
April 9	**WEEK 10**: Working with complex survey designs	
April 16	**WEEK 11**: Multiple regression analysis—applications	*PS #5 due*
April 23	**WEEK 12**: Methods for panel data analysis (I)	*PS #6 assigned*
April 30	**WEEK 13**: Methods for panel data analysis (II)	
May 7	**WEEK 14**: Scale development	

Problem set #6 due on or before Wednesday May 14, 6:00 p.m.

COURSE OUTLINE

(*) = required reading, all others are recommended

WEEK 1: Introduction to "large scale" datasets

(*) Buckley, chapter 1, "Introduction to Large-Scale Education Data"

National Forum on Education Statistics. 2010. *Traveling Through Time: The Forum Guide to Longitudinal Data Systems. Book One of Four: What is an LDS?* (NFES 2010–805). Washington, DC: National Center for Education Statistics. http://nces.ed.gov/pubs2010/2010805.pdf

Perez, M. and M. Socias. 2010. "Data in the Economics of Education," in Dominic J. Brewer and Patrick J. McEwan (eds.), *Economics of Education,* Amsterdam: Elsevier.

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"

WEEK 2: Programming in Stata

(*) Long, chapter 3 and Appendix A

Getting Started with Stata for Windows and/or Acock, chapters 1-4

WEEK 3: Workflow—organizing and planning a project

(*) Long, chapters 1-2

(*) Remler and Van Ryzin. 2011. *Research Methods in Practice: Methods for Description and Causation,* chapter 1, "Research in the Real World," chapter 2, "Theory and Models," and chapter 15, "How to Find, Focus, and Present Research"

WEEK 4: Accessing large scale databases

(*) Buckley, chapter 2, "Accessing Large-Scale Education Data"

(*) Remler and Van Ryzin. 2011. *Research Methods in Practice: Methods for Description and Causation,* chapter 6, "Secondary Data"

National Longitudinal Survey of Youth: Children and Young Adults. "Introduction to the Sample." https://www.nlsinfo.org/content/cohorts/nlsy79-children/intro-to-the-sample

National Longitudinal Survey of Youth: Children and Young Adults. "Using and Understanding the Data" https://www.nlsinfo.org/content/cohorts/nlsy79-children/using-and-understanding-the-data

---

WEEK 5: Workflow—data preparation and cleaning

(*) Long, chapters 5-6

Getting Started with Stata for Windows and/or Acock, chapter 3

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)

WEEK 6: Workflow—automation, documentation and replication

(*) Long, chapters 2 and 4

WEEK 7: Workflow—descriptive analysis

(*) Long, chapter 7

(*) Remler and Van Ryzin. 2011. *Research Methods in Practice: Methods for Description and Causation,* chapter 8, "Making Sense of the Numbers"

Acock, chapters 5-7

WEEK 8: Guest speaker

WEEK 9: Sampling and sampling distributions

(*) Heeringa, West, and Berglund, chapter 1, "Applied Survey Data Analysis: Overview," and chapter 2, "Getting to Know the Complex Survey Design"

(*) Remler and Van Ryzin. 2011. *Research Methods in Practice: Methods for Description and Causation,* chapter 5, "Sampling"

Groves, R.M. et al., chapter 4, "Sample Design and Sampling Error"

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

WEEK 10: Working with complex survey designs

(*) 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

(*) Buckley, chapter 5, "Analysis of Complex Survey Data"

Heeringa, West, and Berglund, chapter 3, "Foundations and Techniques for Design-Based Estimation and Inference"

Solon, G., S.J. Haider, and J. Wooldrige. 2013. "What Are We Weighting For?" NBER Working Paper No. 18859.

WEEK 11: Multiple regression analysis—applications

(*) Buckley, chapters 6-7, "Multiple Linear Regression with Stata," chapters 8-9, "Multiple Regression Pathologies"

(*) Long, chapter 7

(*) Remler and Van Ryzin. 2011. *Research Methods in Practice: Methods for Description and Causation,* chapter 9, "Making Sense of Multivariate Statistics"

Acock, chapter 8 and 10 and/or Baum chapters 4-5, 7

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

WEEK 12: Methods for Panel Data Analysis—I

(*) Buckley, chapter 10, "Introduction to Modeling Panel Data"

WEEK 13: Methods for Panel Data Analysis—II

Baum, chapter 9 (section 1) and/or Cameron and Trivedi, chapter 8.

McCaffrey, D. F., Lockwood, J. R., Mihaly, K., & Sass, T. R. 2012. "A Review of Stata Routines for Fixed Effects Estimation in Normal Linear Models." *Stata Journal*, 12(3), 406–432.

WEEK 14: Scale development

(*) Comrey, A.L. 1988. "Factor-Analytic Methods of Scale Development in Personality and Clinical Psychology," *Journal of Consulting and Clinical Psychology,* 56(5): 754-761.

Acock, chapter 12