APPLIED STATISTICS: USING LARGE DATABASES IN EDUCATION RESEARCH

APSTA.GE.2110
Course Syllabus

Professor:
Sean P. Corcoran
665 Broadway, Suite 805 (IESP)
Phone: (212) 992-9468
Email: sean.corcoran@nyu.edu

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 APSTA.GE.2001 (Statistics for the Behavioral and Social Sciences I) fulfills this requirement, as does Wagner's CORE.GP.1011 (Statistical Methods for Public, Nonprofit, and Health Management).

It is recommended, but not required, that students complete (or be concurrently enrolled in) a course on multiple linear regression or econometrics, such as APSTA.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 me.

Books

The following book is required and has been ordered by the NYU Bookstore:

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

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

(*) Getting Started with Stata for Windows, 2015. (*free*) Also: Mac and Unix versions.

(*) A Gentle Introduction to Stata, 5th edition by Alan C. Acock, 2016.

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

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

I also recommend the UCLA Stata guide, which includes tutorials, references, examples, and useful links (http://www.ats.ucla.edu/stat/stata/). The Stata YouTube site is also very informative. I 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.

Later in the semester, advanced students may find the following books on survey methodology useful. They are not required, but I 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 12.0 or later should work, but I recommend the most recent release, 14.0). 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/Citrix/VirtualComputerLabWeb/ Currently, version 14 of Stata SE is accessible through the VCL. Please note that students have experienced intermittent 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 14 can be purchased at a discounted student rate. "Small" Stata is the least expensive ($38 for six months or $54 for a year), but is limited in the size of datasets it can manage. I don't recommend Small Stata for this course. "Intercooled" Stata is the next level up ($75 for six months or $125 for a year; $198 for a perpetual license); it can accommodate most projects, but for *very* large databases a more expensive version may be needed (e.g., SE or MP, which are available in the NYU labs). For most purposes, you will notice few differences between versions 12-14. 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 or NYU Box account is another alternative for storing data and working files.

Course requirements

Your grade for this course will be based exclusively on **10** problem sets that require the use of Stata and real datasets to complete. Each problem set is weighted equally (10% each) and the dates of assignment and submission are listed in the course outline below. I will assign 11 problem sets over the course of the semester, but will only count 10 of these. (I will drop your lowest score).

Late assignments will not be accepted after problem set solutions have been posted. Each student must hand in his or her own work for each problem set. While I encourage you to work together, duplicate work will not be accepted.

Please submit your completed problem set as a PDF document via email. Use your last name and problem set number as the filename (e.g., *Smith Problem Set 2.pdf*). Doing so will allow us to grade your assignment quickly and return it to you electronically.

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 XX 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

WEEK 1: Introduction to "large" datasets

WEEK 2: Programming in Stata	*PS1 assigned*
WEEK 3: Workflow—organizing and planning a project	*PS1 due* *PS2 assigned*
WEEK 4: Accessing relevant databases	*PS2 due* *PS3 assigned*
WEEK 5: Workflow—data preparation and cleaning	*PS3 due* *PS4 assigned*
WEEK 6: Workflow—automation, documentation and replication	*PS4 due* *PS5 assigned*
WEEK 7: Workflow—descriptive and regression analysis	*PS5 due* *PS6 assigned*
NO CLASS—SPRING BREAK	
WEEK 8. Guest speaker	*PS6 due*
WEEK 9: Sampling and sampling distributions	*PS7 assigned*
WEEK 10: Working with data from complex survey designs	*PS7 due* *PS8 assigned*
WEEK 11: Working with data from complex survey designs: applications	*PS8 due* *PS9 assigned*
WEEK 12: Methods for panel data analysis (I)	*PS9 due* *PS10 assigned*
WEEK 13: Methods for panel data analysis (II)	*PS10 due* *PS11 assigned*
WEEK 14: Stata graphs and visualizations	
NO CLASS – FINALS WEEK	*PS11 due*

COURSE OUTLINE

(*) = required reading, all others are recommended

WEEK 1: Introduction to "large" datasets

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

(*) Pirog, M. A. 2014. "Data Will Drive Innovation in Public Policy and Management Research in the Next Decade." *Journal of Policy Analysis and Management*, 33(2), 537–543.

(*) Cook, T. D. 2014. "'Big Data' in Research on Social Policy." *Journal of Policy Analysis and Management*, 33(2), 544–547.

Figlio, D. N., Karbownik, K., & Salvanes, K. G. 2015. "Education Research and Administrative Data." *National Bureau of Economic Research Working Paper* No. 21592.

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

Schneider, B., Saw, G., & Broda, M. (2016). "A Future for the National Education Longitudinal Program." *AERA Open*, 2(2).

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 relevant databases

(*) Buckley lecture notes, 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"

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.

Lovenheim and Turner – Appendix A "Description of Datasets Commonly Used in the Economics of Education"

WEEK 5: Workflow—data preparation and cleaning

(*) Long, chapters 5-6

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

WEEK 6: Workflow—automation, documentation and replication

(*) Long, chapters 2 and 4

WEEK 7: Workflow—descriptive and regression analysis

(*) Buckley lecture notes, 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 8, "Making Sense of the Numbers," chapter 9, "Making Sense of Multivariate Statistics"

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

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

Williams, R. 2012. "Using the Margins Command to Estimate and Interpret Adjusted Predictions and Marginal Effects." *The Stata Journal*, 12(2), 308–331.

Stata Manuals Ch. 25. "Working with Categorical Data and Factor Variables." http://www.stata.com/manuals13/u25.pdf

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 and 11: 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 12: Methods for Panel Data Analysis—I

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

Thompson, M. E. 2015. "Using Longitudinal Complex Survey Data." *Annual Review of Statistics and Its Application*, 2(1), 305–320.

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: Stata Graphs and Visualizations

Mitchell, *Visual Guide to Stata Graphics*

OTHER PAPERS WHICH MAY BE OF INTEREST:

Glaeser, E. L., Kominers, S. D., Luca, M., & Naik, N. 2015. "Big Data and Big Cities: The Promises and Limitations of Improved Measures of Urban Life." *National Bureau of Economic Research Working Paper* No. 21778.

Conaway, C., Keesler, V., & Schwartz, N. 2015. "What Research Do State Education Agencies Really Need? The Promise and Limitations of State Longitudinal Data Systems." *Educational Evaluation and Policy Analysis*, 37(1 suppl), 16S–28S.

Dynarski, S. M., Hemelt, S. W., & Hyman, J. M. 2015. "The Missing Manual: Using National Student Clearinghouse Data to Track Postsecondary Outcomes." *Educational Evaluation and Policy Analysis*, 37(1 suppl), 53S–79S.

Hahs-Vaughn, D. L. 2005. "A Primer for Using and Understanding Weights with National Datasets." *The Journal of Experimental Education*, 73(3), 221–248. Retrieved from http://www.tandfonline.com/doi/abs/10.3200/JEXE.73.3.221-248