APPLIED STATISTICS: LARGE DATABASES IN APPLIED RESEARCH

APSTA.GE.2110 (Section 001)
Course Syllabus – Spring 2023

Professor:
Joseph Cimpian
Email: joseph.cimpian@nyu.edu
Office hours: on zoom, Tuesdays, 4-6pm NYC time – You can reserve a 20-minute slot here.

Lecture Time: Thursdays, 4:55-7:25pm
Location: 194 Mercer, Room 306B

Course Assistant:
Jo Alkhafaji-King, jk7317@nyu.edu
Office hours: Wednesdays, 4:30-5:30 EST and Fridays, 4-5pm EST via Zoom. Reserve a slot here.

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 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.

Also, you should 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 will be used extensively. Excerpts are available on NYU Brightspace. The Kindle edition is much less expensive than the hard copy:

(*) 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, 2021. (*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 (https://stats.oarc.ucla.edu/stata/modules/). The Stata YouTube site is also very informative. I will post other useful Stata references on the class website.

Later in the semester, advanced students may find the following books on survey methodology useful, though they are not required:

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 (version 14 or later should work, but I recommend purchasing the most recent release, 17, if you don't already have Stata). Access to Stata is possible through either the Virtual Computer Lab or 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/.

(2) You may be interested in buying Stata for your own computer. Stata can be purchased at a discounted student rate. "Basic Edition" Stata (Stata/BE) is available for $48 for six months, $94 for a year, or $225 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). For most purposes, you will notice few differences between versions 14-17. However, be aware that minor differences do exist.

Course requirements

Your grade for this course will be based entirely on the completion of **11** problem sets that require the use of Stata and real datasets to complete. Each problem set is weighted equally (the number of points listed at the top of each assignment varies, but that's just for points within the assignment).

If you are taking the course for:

- 4 credits, then only your **10** highest-scoring assignments count toward your grade.
- 3 credits, then only your **8** highest-scoring assignments count toward your grade.

Each assignment will be due by Thursday 5pm NYC/Eastern time, and late assignments will not be accepted! You must complete your own work for each problem set, although you may discuss problem sets with each other.

Please submit your completed problem set as a complete, error-free (or as close as possible to error-free), annotated Stata log file via NYU Brightspace. (We will discuss this format in greater detail.) Use your last name and problem set number as the filename (e.g., *Smith Problem Set 2.log*). For some assignments, you will also need to include other types of output (for example, a graphics file with a chart).

Other class information

1. NYU Brightspace: All materials for the course (lecture notes, readings, problem sets, data) will be made available via NYU Brightspace. 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 by 5pm each Thursday. You can find these materials under the tab on the left-hand menu labeled "Lecture materials by week." Assignments will be under the "Assignments" tab.

2. 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.

3. 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. Please consult NYU's academic calendar to find the last date on which students can withdraw from a course without receiving a "W" on their transcripts.

4. 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 (www.nyu.edu/csd). Of course, I'm happy to provide any accommodations recommended by the Moses Center.

CLASS SCHEDULE

January 26	**Lecture 1**: Introduction to “large” datasets	
February 2	**Lecture 2**: Programming in Stata	*PS1 assigned*
February 9	**Lecture 3**: Workflow—organizing and planning a project	*PS1 due* *PS2 assigned*
February 16	**Lecture 4**: Accessing relevant databases	*PS2 due* *PS3 assigned*
February 23	**Lecture 5**: Workflow—data preparation and cleaning	*PS3 due* *PS4 assigned*
March 2	**Lecture 6**: Workflow—automation, documentation and replication	*PS4 due* *PS5assigned*
March 9	**Lecture 7**: Workflow—descriptive and regression analysis	*PS5 due* *PS6 assigned*
March 16	**NO CLASS – Spring Break**	
March 23	**Lecture 8**: Sampling and sampling distributions	*PS6 due* *PS7 assigned*
March 30	**Lecture 9**: Working with data from complex survey designs	*PS7 due* *PS8 assigned*
April 6	**Lecture 10**: Working with data from complex survey designs: applications	*PS8 due* *PS9 assigned*
April 13	**Lecture 11**: Methods for panel data analysis (I)	*PS9 due* *PS10 assigned*
April 20	**Lecture 12**: Methods for panel data analysis (II)	*PS10 due* *PS11 assigned*
April 27	**Lecture 13:** Stata graphics	*PS11 due*
May 4	**Lecture 14: LAST CLASS – to be arranged**	

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

Stata “cheat sheets”

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

WEEK 6: Workflow—automation, documentation and replication

(*) Long, chapters 2

WEEK 7: Workflow—descriptive and regression analysis

(*) Buckley lecture notes, chapters 6-7, “Multiple Linear Regression with Stata,” chapters 8- 9, “Multiple Regression Pathologies”

(*) 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/webbooks/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: 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 9 and 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 and 12: Methods for Panel Data Analysis—I and II

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

WEEK 13: Stata Graphs and Visualizations

Mitchell, *Visual Guide to Stata Graphics*

WEEK 14: To be arranged