Professor:
Kathleen M. Ziol-Guest, Ph.D.

Lecture Thurs. 4:55 pm -7:35 pm
Kimball Hall, Room 318E
Tisch Hall, Room LC-19
Phone: 212.998.5478
Email: ziolguest@nyu.edu
Office hours: Wed. 10:00 am – 11:00 am, or by appointment

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 U.S. datasets relevant to education and health 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 Professor Ziol-Guest.

**Course readings**

One book is required and has been ordered by the NYU Bookstore (you can also find it at Amazon, online bookstores, or Stata Press).


Another book is recommended and has also been ordered by the NYU bookstore (you can also find it at Amazon, online bookstores, or Stata Press).


Many of the practical topics we will cover in class come from these books.

If you are new to Stata, I 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. For the most basic, I recommend the following which is free.

*Getting started with Stata, 2015 for Windows* or *Mac*

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 indispensable:


**Computer lab and software**

Successful completion of this course will require the use of Stata (any version 12.0 or later should work, but we 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/vpn/index.html. Currently, version 14 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. 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. We don’t recommend Small Stata for this course. “Intercooled” Stata (IC) 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 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).

Unless prior arrangements have been made with Professor Ziol-Guest, 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. While we encourage you to work together, duplicate work will not be accepted.

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

Other class information
NYU Classes: All materials pertaining to this course (lecture notes, assignments) will be made available on NYU Classes. Enrollment in the course should automatically give you access to the site. Check 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.

Absences: Please contact Professor Ziol-Guest immediately if you have any conflicts with the scheduled assignments, or anticipate being absent for any reason.

Lab and Class 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 I formally give you class time). Further, please make every attempt to be on time.

**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.

**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.

**Accommodations:** Any student requiring an accommodation due to a chronic psychological, visual, mobility, 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](http://www.nyu.edu/csd)).
## Course Schedule: Using Large Databases in Research

<table>
<thead>
<tr>
<th>Date</th>
<th>Week</th>
<th>Topic</th>
<th>Assignments</th>
</tr>
</thead>
<tbody>
<tr>
<td>September 8</td>
<td></td>
<td>WEEK 1: Introduction to “large” datasets</td>
<td></td>
</tr>
<tr>
<td>September 15</td>
<td></td>
<td>WEEK 2: Programming in Stata</td>
<td>PS1 assigned</td>
</tr>
<tr>
<td>September 22</td>
<td></td>
<td>WEEK 3: Workflow—organizing and planning a project</td>
<td>PS1 due</td>
</tr>
<tr>
<td>September 29</td>
<td></td>
<td>WEEK 4: Accessing large datasets</td>
<td>PS2 assigned</td>
</tr>
<tr>
<td>October 6</td>
<td></td>
<td>WEEK 5: Workflow—Data preparation and cleaning</td>
<td>PS3 due</td>
</tr>
<tr>
<td>October 13</td>
<td></td>
<td>WEEK 6: Workflow—Descriptive analysis</td>
<td>PS4 due</td>
</tr>
<tr>
<td>October 20</td>
<td></td>
<td>WEEK 7: Workflow—Automatic, documentation, and replication</td>
<td>PS5 due</td>
</tr>
<tr>
<td>October 27</td>
<td></td>
<td>WEEK 8: Catch-Up and graphing extensions</td>
<td>PS6 due</td>
</tr>
<tr>
<td>November 3</td>
<td></td>
<td>WEEK 9: Sampling and sampling distributions</td>
<td>PS7 due</td>
</tr>
<tr>
<td>November 10</td>
<td></td>
<td>WEEK 10: Working with complex survey designs</td>
<td>PS8 due</td>
</tr>
<tr>
<td>November 17</td>
<td></td>
<td>WEEK 11: Multiple regression analysis</td>
<td>PS9 due</td>
</tr>
<tr>
<td>November 24</td>
<td></td>
<td>WEEK 12: No class. Thanksgiving recess!!</td>
<td></td>
</tr>
<tr>
<td>December 1</td>
<td></td>
<td>WEEK 13: Methods for panel data analysis (I)</td>
<td>PS9 due</td>
</tr>
<tr>
<td>December 8</td>
<td></td>
<td>WEEK 14: Methods for panel data analysis (II)</td>
<td>PS10 due</td>
</tr>
<tr>
<td>December 15</td>
<td></td>
<td>WEEK 15: Advanced topics</td>
<td>PS10 due</td>
</tr>
</tbody>
</table>

*Problem Set 11 will be due on or before 5 p.m. Thursday December 22.*
Course Outline: Using Large Databases in Research

(*) = required reading, all others are recommended

WEEK 1:   Introduction to “large” datasets

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


WEEK 2:   Programming in Stata

(*) Long, chapter 3 and Appendix A

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

WEEK 3:   Workflow—organizing and planning a project

(*) Long, chapters 1-2


WEEK 4:   Accessing large datasets


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](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](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/Mac and/or Acock, chapter 3

WEEK 6: 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

Stata Manuals Ch. 25. "Working with Categorical Data and Factor Variables."  

WEEK 7: Workflow—Automation, documentation, and replication

(*) Long, chapters 2 and 4

WEEK 8: Catch-up and graphing extensions

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”


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

WEEK 10: Working with complex survey designs


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

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


WEEK 11: Multiple regression analysis
(*) Buckley lecture notes, chapters 6-7, “Multiple Linear Regression with Stata,” chapters 8-9, “Multiple Regression Pathologies”

(*) Long, chapter 7


Acock, chapter 8 and 10

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


**WEEK 12:** NO CLASS.

**WEEK 13:** Methods for Panel Data Analysis (I)

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

**WEEK 14:** Methods for Panel Data Analysis (II)

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


**WEEK 15:** Advanced topics: Scale development, Multiple Imputation, HLM, Other

TBD
Descriptive statistics are limited in so much that they only allow you to make summations about the people or objects that you have actually measured. You cannot use the data you have collected to generalize to other people or objects (i.e., using data from a sample to infer the properties/parameters of a population). For example, if you tested a drug to beat cancer and it worked in your patients, you cannot claim that it would work in other cancer patients only relying on descriptive statistics (but inferential statistics would give you this opportunity).