

**E10.2110 APPLIED STATISTICS:
USING LARGE DATABASES IN EDUCATION RESEARCH
NEW YORK UNIVERSITY
SPRING 2011**

Professor: Sean P. Corcoran Phone: (212) 992-9468
665 Broadway, Suite 805 FAX: (212) 995-4564
Institute for Education and Social Policy Email: sean.corcoran@nyu.edu

Lecture: Wednesdays, 3:30 – 6:10 p.m. (194 Mercer Room 304)
Office hours: Thursdays, 2:00 – 4:00 p.m., or by appointment

Teaching assistant: Gundula Loffler (gundula.loffler@nyu.edu)

COURSE DESCRIPTION This course is designed to serve as a bridge between more theoretical coursework in applied statistics (introductory courses in statistics and 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 longitudinal regression models of the sort frequently employed in real applied data analysis. The emphasis throughout the course is on hands-on data preparation and modeling using the Stata statistical software package.

PREREQUISITES At 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 E10.2001, for example, fulfills this requirement.

In addition, students should have either completed or be concurrently enrolled in an additional semester of linear regression (e.g. E10.2002 or P11.2902). Students not meeting this requirement must demonstrate satisfactory knowledge of multiple linear regression methods prior to enrolling in the course. No prior experience with Stata is assumed or required.

TEXTBOOK There is no required text for this course. Lecture notes from Professor Jack Buckley, *Analysis of Large-Scale Education Data with Stata* (2009) will be available on Blackboard, along with other required and recommended readings. I do recommend you buy a guidebook 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](http://www.stata.com). In order of most basic to most advanced:

Data Management Using Stata: A Practical Handbook by Michael N. Mitchell, 2010 (MM).

Data Analysis Using Stata, 2nd edition by Ulrich Kohler and Frauke Kreuter, 2010 (K&K).

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 by A.C. Cameron and P.K. Trivedi, 2009 (C&T).

I also highly recommend the UCLA Stata guide, which includes tutorials, references, examples, and useful links (<http://www.ats.ucla.edu/stat/stata/>). Michael Mitchell keeps a blog of “Stata tidbits of the week” (<http://www.michaelnormanmitchell.com/>). If you plan to do a lot of graphics in Stata, the following book is immensely useful:

A Visual Guide to Stata Graphics, 2nd edition by Michael N. Mitchell, 2008.

**COMPUTER LAB
AND SOFTWARE**

Successful completion of this course will require the use of Stata software (any version after 7.0 should be sufficient, but I recommend using the most recent release, 11.0). Stata is available on lab computers at the ITS Washington Place Academic Technology Center, the Third Avenue NYU Hotspot, and at the NYU Data Service Studio on the 6th floor of Bobst Library (www.nyu.edu/its/statistics/).

As a student you should have access to the computer labs with your NYU ID. Lab attendants are not typically experts in Stata, but they can answer system level questions about accessing the program, opening, saving, printing files, etc. The Data Service Studio offers free on-site help with Stata and other software packages. 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).

Students can purchase a one-year or perpetual license for Stata to work with their own computer. At last check, the price was \$65 for a six-month license, \$98 for a one-year license, and \$179 for a perpetual (forever) license. Do not purchase the “small” Stata version for \$29-\$49 as this is insufficient for the datasets we will examine. To order Stata/IC 11.0, see: <http://www.stata.com/order/new/edu/gradplans/gp-campus.html>

**COURSE
REQUIREMENTS**

Your grade for this course will be determined based on five (5) practical problem sets that will require the use of Stata and real datasets to complete. Each problem set is weighted equally (20% each) and the dates of assignment and completion are listed in the course outline below.

Unless prior arrangements have been made with the instructor, 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.

BLACKBOARD

All materials pertaining to this course (lecture notes, readings, problem sets, data) will be made available via Blackboard, which can be accessed through NYUHome. Enrollment in the course should automatically give you access to the class Blackboard site. Check in with Blackboard frequently for new announcements.

**MISC.
POLICIES**

- 1) NYU and Steinhardt policies toward academic integrity will be *strictly enforced* in this class. You can find the school's official statement on academic integrity here: http://steinhardt.nyu.edu/policies/academic_integrity. All work submitted must be that of the individual student and must be work original to this course.
 - 2) Please make an effort to be on time (I will do the same) and please turn off your cell phone—and other digital distractions—while in class.
 - 3) The class is being held in a computer lab. To help promote a productive learning environment, please keep all other internet-related 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 to do so).
 - 4) Please see me immediately if you have any conflicts with scheduled assignments and/or exams, or if you anticipate being absent due to religious observances.
 - 5) If you wish to withdraw from this course, please do so formally with the University Registrar. If you withdraw without authorization, you are at risk for receiving an “F” for the course. *February 13 is the last day for graduate and undergraduate students to withdraw without receiving a “W” on their transcripts.*
 - 6) Any student attending NYU who needs an accommodation due to a chronic psychological, visual, mobility and/or learning disability, or is Deaf or Hard of Hearing, should register with the Moses Center for Students with Disabilities at 212-998-4980, 726 Broadway, 2nd floor (www.nyu.edu/csd).
-

CLASS SCHEDULE

Wednesday January 26	1. Introduction: education data sources and online data analysis	
February 2	2. Introduction to Stata (I) – Basics	<i>Problem set #1 assigned</i>
February 9	3. Introduction to Stata (II) – Advanced commands and graphics	<i>Problem set #1 due</i>
February 16	4. Issues of survey design	<i>Problem set #2 assigned</i>
February 23	5. Applications of multiple linear regression (I) – Basics	<i>Problem set #2 due</i>
March 2	6. Applications of multiple linear regression (II) – Non i.i.d errors	
March 9	7. Program evaluation and causal inference (I)	<i>Problem set #3 assigned</i>
March 16	NO CLASS—SPRING BREAK	
March 23	Guest speaker from NYU’s Research Alliance for New York City Schools – large-scale databases in NYC	<i>Problem set #3 due</i>
March 30	8. Program evaluation and causal inference (II)	
April 6	9. Panel data methods (I)	<i>Problem set #4 assigned</i>
April 13	10. Panel data methods (II)	
April 20	11. Panel data methods (III)	<i>Problem set #4 due</i>
April 27	12. Qualitative dependent variable models (I)	<i>Problem set #5 assigned</i>
May 4	13. Qualitative dependent variable models (II)	

Problem set #5 due on final exam date, Wednesday May 11th, by 6:00 p.m.

COURSE OUTLINE

1. Introduction: education data sources and online data analysis

- a. What are “large scale” datasets?
- b. Types and examples of large-scale databases used in education research
- c. Public vs. restricted-use data
- d. Privacy, confidentiality, and FERPA
- e. Electronic access to large scale data—query tools, table and model servers
- f. Downloading and converting files
- g. Dataset documentation and codebooks
- h. NCES electronic codebooks
- i. Limitations of online data analysis

Reading:

- Buckley, chapters 1-2
- 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”

References:

- Dataset descriptions on NCES website (<http://nces.ed.gov>)
-

2. Introduction to Stata (I) – Basics

- a. Basic operations: reading/writing data files, importing and converting files, “do” files
- b. Stata syntax
- c. Data cleaning
- d. Labeling variables and values
- e. Creating and re-coding variables
- f. Summarizing data: descriptive statistics and tables
- g. Correlation, t-tests, and simple linear regression
- h. Help files and user-written commands (.ado)
- i. Examples: National Household Education Surveys (NHES)

Reading:

- Buckley, chapter 3

References:

- (MM) 1-5 and Appendix A, (K&K) 1-5, 10 (AA) 1-5, 7-8, (CB) 1-2 and Appendix A, (C&T) 1-2
 - NHES dataset description on NCES website (<http://nces.ed.gov/nhes/index.asp>)
-

3. Introduction to Stata (II) – Advanced commands and graphics

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

Reading:

- Buckley, chapter 4

References:

- (MM) 6-9, (K&K) 6-7, 11 (AA) 5-6, (CB) 3 and Appendix B, (C&T) 1-2
-

4. Issues of survey design

- a. Weights and design variables
- b. Consequences of ignoring design for descriptive statistics
- c. Computing descriptive statistics accounting for design and nonresponse
- d. Methods of variance estimation
- e. Examples: NHES and PISA public use files

Reading:

- Buckley, chapter 5
- 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

References:

- (C&T) 3.7
 - 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
 - PISA dataset description on NCES website (<http://nces.ed.gov/surveys/pisa/>)
-

5. Applications of multiple linear regression (I) – Basics

- a. Introduction to multiple regression in Stata
- b. Reading and interpreting results
- c. Presenting regression results (e.g. outreg)

- d. Hypothesis testing
- e. Functional form
- f. Dummy regressors and interaction terms
- g. Weighting
- h. Examples: NHES, PISA, and/or ECLS-K

Reading:

- Buckley, chapters 6-7
- (CB) chapters 4-5, 7

References:

- (K&K) 8, (AA) 8, 10, (C&T) 3
 - Studenmund, chapters 1-7
 - UCLA webbook *Regression with Stata* (<http://www.ats.ucla.edu/stat/stata/webboks/reg/>)
 - ECLS-K dataset description on NCES website (<http://nces.ed.gov/ecls>)
-

6. Applications of multiple linear regression (II) – Non i.i.d errors and other problems

- a. Autocorrelation and heteroskedasticity
- b. Diagnostics and remedies
- c. “Robust” standard errors
- d. Partial multicollinearity
- e. Examples: ECLS-K and NHANES

Reading:

- Buckley, chapters 8-9
- (CB) chapter 6

References:

- (K&K) 8, (AA) 10, (C&T) 3.5 and 5
 - Ch. 2 of UCLA regression webbook:
(<http://www.ats.ucla.edu/stata/stata/webbooks/reg/chapter2/statareg2.htm>)
 - NHANES dataset description on CDC website (<http://www.cdc.gov/nchs/nhanes.htm>)
 - Studenmund, chapters 8-10
-

7. Program evaluation and causal inference (I)

- a. Logic of causal inference
- b. Randomized experiments
- c. Quasi- and natural experiments
- d. Internal vs. external validity

Reading:

- Buckley, chapter 12

- Schlotter, M., G. Schwerdt, and L. Woessmann, 2010. “Econometric methods for causal evaluation of education policies and practices: A non-technical guide.” *IZA Discussion Paper No. 4725*

References:

- Nichols, A. 2007, “Causal inference with observational data.” *Stata Journal* 7(4): 507-541.
 - Schneider et al. (2007) *Estimating Causal Effects Using Experimental and Observational Designs*, chapter 2, “Causality: Forming an Evidential Base,” and chapter 3, “Estimating Causal Effects Using Observational Data.”
-

8. Program evaluation and causal inference (II)

- a. Instrumental variables methods
- b. Matching estimators (overview)
- c. Regression discontinuity designs

Reading:

- Buckley, chapter 12

References:

- (CB) 8
 - Bloom, H.S. 2009, “Modern regression discontinuity analysis.” New York: MDRC
-

9. Panel data methods (I)

- a. Working with panel data in Stata – special commands
- b. Why use panel data?
- c. First difference model
- d. Fixed effects model(s)
- e. Random effects models
- f. Examples: teacher value-added and ECLS-K

Reading:

- Buckley, chapter 10
- (CB) 9.1

References:

- (C&I) 8
- Studenmund, chapter 16

10. Panel data methods (II) – Applications

- a. Examples: ECLS-K
-

11. Panel data methods (III) – Advanced topics

- a. Panel attrition
- b. Item missing data
- c. Reweighting and weight stabilization
- d. Clustered and hierarchical data

Reading:

- Buckley, chapter 10

References:

- (CB) 9, (C&T) 9
-

12. Qualitative dependent variables (I) – Binary outcome

- a. The linear probability model
- b. Logit / logistic regression
- c. Probit models
- d. Marginal effects and predictions
- e. Extensions

Reading:

- Buckley, chapter 11
- (CB) 10

References:

- (K&K) 9, (AA) 11, (C&T) 14
 - Studenmund, chapter 13
-

13. Qualitative dependent variables (II) – Other models

- a. Multinomial logistic model
- b. Ordered logit model
- c. Count data models (overview)

References:

- (CB) 10, (C&T) 15, 17