E10.2110  
Applied Statistics: Using Large Databases in Education Research  
Spring, 2010  
Thursday, 2:00-4:45 LC19

Jack Buckley  
Associate Professor of Applied Statistics  
NYU Steinhardt

1. Contact Information:  
Postal: 726 Broadway, 7th Floor, Room 752  
Electronic: spb5@nyu.edu or jack.buckley@gmail.com  
Phone: 992-7676  
Office Hours: Thursdays 12:00-2:00 and by appointment

2. Overview:  
This course is designed to serve as a bridge between more theoretical coursework in applied statistics (introductory course in statistics and linear models) and practical work with real, large-scale data. Although the focus is mainly on datasets relevant to education and educational policy research, the skills taught in the course are broadly transferrable 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.

3. Prerequisites:  
At minimum, **one semester of introductory statistics** is required. Topics covered should have included simple linear regression ($Y = a + bx + e$), hypothesis testing (e.g. z-tests, t-tests), 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, P11.2902). Students not meeting this requirement must demonstrate satisfactory knowledge of multiple linear regression methods prior to enrolling in this course.

No prior experience with Stata is required.

4. Course Text:  
The only required text for this course is:

The text is available in the bookstore and online at http://www.stata.com/bookstore/mus.html.

The textbook is a useful reference for using Stata to estimate models of the sort discussed in the course. It should be noted, however, that the text covers a great deal of advanced material beyond the scope of this course as well. In some ways the book can be considered a concise manual for the use of Stata in public policy (and related) empirical work and will serve as a useful reference for future study and for real, practical analysis.

Additional course materials will be made available on Blackboard or via links to BOBCAT or other websites as required.

5. Software:
Successful completion of this course will require the use of Stata software (any version after 7.0 should be sufficient; the latest release is 11.0). Stata is available on the lab computers at the ITS Washington Place Academic Technology Center, the Third Avenue NYU HotSpot (see http://www.nyu.edu/its/map/), and at the Data Service Studio on the 6th Floor of Bobst (http://library.nyu.edu/dataservice/). The Data Service Studio also offers short courses and on-site help with Stata (among other packages).

In addition, students can purchase either a one-year or perpetual license for Stata to work with on their own computer. Pricing is about $98 for one-year, $179 forever (do not purchase the “small” Stata version for $49 as this is insufficient to handle some datasets we will examine). For more information see (http://www.stata.com/order/new/edu/gradplans/gp-campus.html) and also the Politics department website (http://politics.as.nyu.edu/object/politics.stata) as this department is the on-campus representative for Stata’s GradPlan. Note also that many of the homework assignments will require an Internet connection.

6. Course Requirements:
The course will be graded on the basis of five (5) practical problem sets that will require the use of Stata and real datasets to complete. Each problem set is weighted equally (20% of the final course grade) and the dates of assignment and completion are listed in the detailed syllabus 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 and students are reminded that university policy concerning academic integrity applied to this course and will be enforced.

7. Accommodations:
NYU is committed to facilitating equal access for students with disabilities, including hearing and visual impairments, mobility impairments, learning disabilities and attention
deficit disorders, chronic illnesses, and psychological impairments. Per NYU policy, students seeking accommodation are required to contact the Moses Center on 240 Greene Street, 2nd Floor, 212-998-4980.

8. Schedule

Note: all readings assignments are due upon the day listed.

1 (Jan 21): Introduction to education data in the US and beyond
   - Data sources:
     - NCES collections
     - CCD/PSS/IPEDS/Census of governments
     - NAEP and other national assessments
     - High school longitudinal studies
     - Early childhood longitudinals
     - NPSAS/BPS/B&B
     - SASS/TFS
     - NHES
     Other Federal data sources relevant to education/child development
     - Census/ACS
     - Bureau of Labor Statistics
     - NICHD/NCHS (e.g. NHANES)
     International Data
     - TIMSS/PIRLS
     - PISA
     State and municipal administrative data
     - Public use vs. restricted
     - Privacy, confidentiality and FERPA
     - Downloading and converting files

   Reading:
   - Familiarize yourself with the NCES website (http://nces.ed.gov/)

2 (Jan 28): Online Data Analysis and Table Generation
   - Model and table servers in general
   - CCD Build-A-Table
   - NAEP Data Explorer
   - DAS and related model servers (postsecondary data, ECLS, SASS, etc)
   - Limits of online data analysis in education policy research

   Reading:
   - Review specific model/table servers previewed in previous week

3 (Feb 4): Introduction to Stata
   - Basic operations (reading/writing data files, importing and converting files)
   - “do” files and log files in Stata
   - Summarizing data/simple descriptives
- Simple linear regression (one independent variable)
- Extending Stata via user-written commands
- Stata help system and web resources
- Merging data
- Examples using the National Household Education Surveys (NHES) series of cross-sectional household surveys

Reading:
- C&T Chapter 1 and 2
- Also, briefly familiarize yourself with the NHES by exploring the study website at http://nces.ed.gov/nhes/index.asp.

HW # 1 Assigned

4 (Feb 11): More Basic Stata Commands and Graphics
- Indicator variables and missing data
- Bootstrapping sampling distributions
- Histograms and kernel density estimation
- Local smoothing regression/local linear models
- Univariate and bivariate graphs (scatter, box, violin, sunflower, dotplots, etc).

Reading:
- C&T Chapter 2

In addition to the Stata manuals, another useful reference for Stata graphics is the UCLA ATS website: http://www.ats.ucla.edu/stat/stata/topics/graphics.htm. (the ATS website has modules and examples for many Stata functions).

HW # 1 Due

5 (Feb 18): Issues of survey design
- Weights and design variables
- Consequences of ignoring design for descriptive statistics
- Computing descriptive statistics accounting for design and nonresponse
- Different methods of variance estimation
- Examples using NHES and also introducing the Program in International Student Assessment (PISA) public use microdata file.

Reading:
- C&T Section 3.7
- Familiarize yourself with the PISA data (http://nces.ed.gov/surveys/pisa/)

HW #2 Assigned

6 (Feb 25): Multiple Linear Regression I
- Intro to multiple regression in Stata
- What about ANOVA/ANCOVA?
- Reading results
- Simple and less simple hypotheses testing
- Examples drawn from PISA and NHES

Reading:
- C&T Chapter 3
- See also the free webbook *Regression with Stata* for lots of examples: [http://www.ats.ucla.edu/stat/stata/webbooks/reg/](http://www.ats.ucla.edu/stat/stata/webbooks/reg/).

**HW# 2 Due**

7 (Mar 4): Multiple Regression, Continued
- To weight or not to weight?
- Dummy variables and simple interactions
- Transformed dependent variables
- New dataset: Early Childhood Longitudinal Study, Kindergarten Cohort (ECLS-K)

Reading:
- C&T Chapter 3, Section 5.5

**HW #3 Assigned**

8 (Mar 11): Multiple Regression Pathologies and Diagnostics
- Partial multicollinearity
- Autocorrelation
- Prais-Winsten regression

Reading:
- C&T 3.5

9 (Mar 18): Spring Recess

10 (Mar 25): Regression Pathologies Continued
- Heteroscedasticity diagnostics
- “Fixing” heteroscedasticity via GLS
- “Robust” standard errors
- Modeling heteroscedasticity via multiplicative model
- Examples drawn from various data sources, including NHANES

Reading:
- C&T Chapter 5 (skim for intuition)
- Familiarize yourself with the NHANES website [http://www.cdc.gov/nchs/nhanes.htm](http://www.cdc.gov/nchs/nhanes.htm)

**HW #3 Due**
11 (Apr 1): Introduction to Panel Data Econometrics
   - First difference model
   - Fixed effects model(s)
   - Random effects models
   - Examples drawn primarily from ECLS-K
Reading:
   - C&T Chapter 8

**HW #4 Assigned**

12 (Apr 8): Advanced Topics in Panel Data Models
   - Panel attrition
   - Item missing data
   - Reweighting and weight stabilization
   - Clustered and hierarchical data
   - Hausman-Taylor models
Reading:
   - C&T Chapter 9 (Again, don’t worry too much about the technical detail in this section. The idea is to get an intuition).

13 (Apr 15): The Linear Probability Model for Dichotomous Outcomes
   - Dichotomous dependent variables
   - Interpretation
   - Fixing heteroscedasticity

**HW #4 Due**

14 (Apr 22): More Advanced Binary Outcome Models
   - Logit/Logistic regression
   - Probit
   - Extensions of these
   - How are these models estimated?
   - Interpreting your results
Reading:
   - C&T Chapter 14

15 (Apr 29): Overview of Topics in Policy and Program Evaluation
   - Logic of causal inference
   - The intuition behind instrumental variables
   - Matching and related estimators
   - Regression discontinuity design
Readings:
   - Nichols, A. 2007. “Causal inference with observational data.” This is a pre-print of an article in the Stata Journal that gives a quick overview of all these methods and shows implementation in Stata. Available at
http://pped.org/stata/ciwod.pdf or the published version will be circulated.

HW #5 Assigned

HW #5 Due On Scheduled Exam Day