Columbia University
MAILMAN SCHOOL
OF PUBLIC HEALTH

Principles of Large Scale Data Management
P8545.006

CLASS SESSIONS
Tuesday, 1pm-4pm, 9/3-10/15
Location: TBD

INSTRUCTOR
Matthew Neidell
212.342.4522, mn2191@columbia.edu
600 West 168th Street, 6th Floor Rm. 609
Office Hours: TBA

COURSE DESCRIPTION
Large data sets provide crucial information for monitoring the health of our nation and developing and evaluating public health policies. This class is an applied, hands-on course designed to provide an introduction to several major health data sets and guide students in managing these data. Students will hone computer and statistical skills developed in other research methods courses by analyzing these data. By accessing data that measure health variables of current importance, the class also provides a foundation for developing a variety of health policy research questions. The principal goal of the course is for students to develop the skills to identify and analyze a data set to answer a specific research or policy question. It is designed to complement skills learned in other methods courses and prepare students to perform independent research.

PREREQUISITES
P6103: Introduction to Biostatistics or equivalent.

COURSE LEARNING OBJECTIVES
Students who successfully complete this course will be able to:
- Identify potential data sets for addressing research questions
- Apply techniques for locating and obtaining appropriate data
- Organize data sets for analysis
- Construct research questions based on data availability
- Apply techniques for addressing common data issues
- Integrate techniques to conduct basic statistical analysis

ASSESSMENT AND GRADING POLICY
Student grades will be based on:
Participation..... 10%
Assignments (3)..... 15% each (total= 45%)
Final Project..... 45%

COURSE REQUIREMENTS

1. Participation (10%) is based on how well you contribute to classroom discussions. Although I have a roadmap for each class, the exact content for each class will adjust to student responses.

2. There are 3 assignments (15% each) designed to push your Stata skills while working with a new data set. The assignments are designed to be challenging, but "learning by doing" is the best way to understand how to work with data. The goal of each assignment is less about getting the correct answer but more about setting up the data for analysis. For all assignments, in addition to any statistical or graphical output, you must provide your Stata programs so I can replicate your results. Students must work in a group of 2, and must submit their assignments as a group.

3. A final project (45%) where you will replicate results from an existing (published) article and extend the analysis in at least one way. YOU MUST DISCUSS THE ARTICLE WITH ME BEFORE YOU PROCEED. You will hand-in a brief write-up of your findings, roughly 6-8 pages including tables. I will consider allowing students to perform an analysis of their own research topic if they want to pursue it, but it may be difficult to fulfill in the short time of the class. The final project is due 2 weeks after the last class (10/29). Students must again work in a group of 2.

Course materials: We will use Stata extensively in the class, so students will further their mastery of Stata in the process. No prior knowledge of Stata is required. Students can either purchase Intercooled (IC) Stata 13.0 for $69 under the 'GradPlan' for students (available at http://www.stata.com/order/new/edu/gradplans/campus-gradplan/) or use various computer labs on campus. There is no textbook or course reader; handouts will be posted on courseworks as necessary.

COURSE STRUCTURE

After a brief review of Stata and introduction to accessing data, in each of the following lectures we introduce a new data set, perform basic cleaning and processing, perform rudimentary analyses (summary statistics, regressions, graphical analysis, etc.), and introduce an advanced data analysis concept. Data sets to explore include BRFSS, NHANES, Vital Statistics natality files, and NHIS ('Google' any one of these for more information; it will be the first or second hit). Advanced concepts include creating data dictionaries, converting files to Stata format, merging data sets, collapsing data sets, imputing missing values, performing weighted analysis, and accessing geo-coded data. Additional data sets and concepts can be added depending on student interest.

MAILMAN SCHOOL POLICIES AND EXPECTATIONS

Students and faculty have a shared commitment to the School's mission, values and oath. http://mailman.columbia.edu/about-us/school-mission/

Academic Integrity

Students are required to adhere to the Mailman School Honor Code, available online at http://mailman.columbia.edu/honorcode.

Disability Access

In order to receive disability-related academic accommodations, students must first be registered with the Office of Disability Services (ODS). Students who have, or think they may have a disability are invited to contact ODS for a confidential discussion at 212.854.2388 (V) 212.854.2378 (TTY), or by email at disability@columbia.edu. If you have already registered with ODS, please speak to your instructor to ensure that s/he has been notified of your recommended accommodations by Lillian Morales (lm31@columbia.edu), the School's liaison to the Office of Disability Services.

COURSE SCHEDULE

Session 1 – Course and Stata introduction	
9/3	Learning Objectives: Students will be able to 1) describe basic concepts in Stata 2) create a do-file in Stata 3) identify techniques to access data

Session 2 – Introduction to NHANES	
9/10	Learning Objectives: Students will be able to 1) describe the major uses of NHANES 2) clean missing values 3) merge data sets •Distribute assignment 1: Analyze the association between the built environment, as proxied by living in a central city, and obesity using NHANES II Anthropometry tape

Session 3 – Introduction to BRFSS	
9/17	Learning Objectives: Students will be able to 1) describe the major uses of BRFSS 2) convert SAS data into Stata format 3) improve the presentation of data •Assignment 1 due

Session 4 – Introduction to NHIS	
9/24	Learning Objectives: Students will be able to 1) describe the major uses of NHIS 2) convert Ascii data into Stata format 3) describe processes and rationale for weighting analysis •Distribute assignment 2: Analyze trends in various health conditions using the NHIS

Session 5 – More with NHIS	
10/1	Learning Objectives: Students will be able to 1) identify conditions under which imputation is viable 2) describe techniques for accommodating missing values in data analysis •Dealing with missing values: imputation (missing at random, types of imputation) •Reading (optional): Schaefer JL (1999). "Multiple imputation: a primer." Stat Methods Med Res. 8(1): 3-15. •Assignment 2 due

Session 6 – Introduction to Vital Statistics – Birth Data	
10/8	Learning Objectives: Students will be able to 1) describe the process for collecting Vital Statistics data

	2) create a data dictionary for reading in Ascii data 3) apply advanced programming techniques appropriately and purposefully •Distribute assignment 3: Analyze tooth loss by socioeconomic status using the 2002 BRFSS

Session 7 – More with BRFSS	
10/15	Learning Objectives: Students will be able to 1) apply techniques to access non-public data 2) synthesize data merging techniques •Reading (optional): Jean Eid, Henry Overman, Diego Puga, and Matthew Turner (2008). "Fat City: Questioning the Relationship Between Urban Sprawl and Obesity in the United States." Journal of Urban Economics 63(2): 385–404. •Assignment 3 due