SC709: QUANTITATIVE DATA MANAGEMENT

Instructor: Natasha Sarkisian
Email: natasha@sarkisian.net
Phone: (617) 755-3178
Office: McGuinn 417

Class time: Wednesdays 4:30-7 PM
Class location: O’Neill 245
Office hours: By appointment
Webpage: http://www.sarkisian.net/sc709/

COURSE DESCRIPTION
The goals of this course are to develop the skills necessary to prepare complex datasets for analysis using an efficient, straightforward, error-free, and well documented data management process. This course is designed for graduate students with a prior background in statistics at the level of SC703: Multivariate Statistics (or its equivalent). This means that students should have considerable experience with at least one statistical software package (e.g., SPSS, SAS, Stata, etc.).

We will be using Stata for all the analyses throughout the course. No previous Stata experience is necessary: I will provide an introduction to Stata in the beginning of the course and guide you throughout the course. For your assignment, you can use Stata on Citrix: see http://apps.bc.edu.

The major topics of the course will include reading datasets into Stata, combining and reshaping datasets, saving and storing them, cleaning, recoding, and labeling the data, dealing with skip patterns and missing data, organizing the workflow, and basic programming in Stata.

COURSE POLICIES
For each data management topic covered in the course, I will give a lecture involving a demonstration in Stata, and you will get a chance to follow along and do all the necessary steps in Stata on your laptop. After that, you will work on applying these data management steps to a different dataset or do exercises, either individually or in groups.

The course is based on an interactive relationship between the instructor and students, as well as on collaboration among the students. You are strongly encouraged to ask questions and discuss the material in class. I also encourage collaboration among the students; therefore, many of you will work in groups to prepare presentations on specific datasets that will be delivered in class later in the semester. You will also check each other’s work and provide feedback on a preliminary draft of your final data management assignment.

I also would like to stress that you are always welcome to come and see me with any additional questions. Email is the best way to get in touch with me to get a quick question answered or to set up an appointment to discuss something at length. You are also welcome to call me either in my office or at home; however, be prepared to leave your name and number if I am not available to pick up the phone. Also, please check your email regularly: I will let you know by email when course notes are posted on the website and send other announcements from time to time.

Finally, a note on feedback. I would like to know how I could make this course experience as useful and interesting as possible. Therefore, every class in the end of class I will ask you to submit a sheet of paper (or send an email) with the date and at least one sentence of reaction to that class meeting, indicating what you learned, or something you liked or did not like, found clear or too simplistic, or found confusing and in need of further (or better) explanation. You may also submit comments on the course in general.
COURSE REQUIREMENTS AND GRADING
Two required books will be available at the Boston College bookstore:
Other required readings will be available on electronic reserve in the library: http://www.bc.edu/reserves

The main assignment for this course will be to conduct all the necessary data management steps for a research project using a large, complex dataset. You can select a dataset yourself, or I can provide one. This data management assignment includes the following steps.

Proposal. The proposal will involve selecting a dataset, identifying a research question, and selecting the relevant variables from the codebook. Please consult me early on if you need help selecting a dataset. (10% of your grade.)

First Draft. For the first draft of your assignment, you will submit all your data management do files, log files, and the research log (make sure to include comments to document all the decisions) to one of your peers for review (please send a copy to me as well). If at all possible, also share the data with your peer. (10% of your grade.)

Peer Feedback. You will thoroughly review and provide comments on the do files and log files submitted by one of your peers. You will get a chance to discuss the project with that peer twice during class: a couple of days after you receive the draft (to ask any clarifying questions), and soon after you provide your comments (to discuss the feedback). (10% of your grade.)

Final Draft. The final draft will incorporate peer feedback and include all your do files, log files, the research log, and (if possible) datasets. (40% of your grade.)

In addition to the main data management project, you will prepare written answers (20% of your grade) and deliver an in-class presentation (10% of your grade) based on a set of questions for a specific dataset assigned to you, either individually or in a group. I will distribute the questions and assign datasets early in the course based on your interests.

<table>
<thead>
<tr>
<th>Assignment</th>
<th>Contribution to your grade</th>
<th>Due date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposal for data management project</td>
<td>10%</td>
<td>September 25</td>
</tr>
<tr>
<td>Written responses to dataset questions</td>
<td>20%</td>
<td>October 30</td>
</tr>
<tr>
<td>First draft of data management project</td>
<td>10%</td>
<td>November 17</td>
</tr>
<tr>
<td>Peer feedback on data management</td>
<td>10%</td>
<td>December 1</td>
</tr>
<tr>
<td>Dataset presentations</td>
<td>10%</td>
<td>November 13-Decemeber 11</td>
</tr>
<tr>
<td>Final draft of data management project</td>
<td>40%</td>
<td>December 13</td>
</tr>
</tbody>
</table>

All assignments will be submitted electronically. Small files can be sent by email; any large files should be submitted using MyFiles or another file sharing website.

The letter grades will be determined as follows:

<table>
<thead>
<tr>
<th>Score</th>
<th>Grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>93-100</td>
<td>A</td>
</tr>
<tr>
<td>90-92</td>
<td>A-</td>
</tr>
<tr>
<td>87-89</td>
<td>B+</td>
</tr>
<tr>
<td>83-86</td>
<td>B</td>
</tr>
<tr>
<td>80-82</td>
<td>B-</td>
</tr>
<tr>
<td>60-79</td>
<td>C</td>
</tr>
<tr>
<td>0-59</td>
<td>F</td>
</tr>
</tbody>
</table>
**COURSE OUTLINE**

**September 4. Introduction to Data Management using Stata.**

**September 11. Reading and Writing Datasets.**

**September 18. Data Cleaning.**

**September 25. Creating, Changing, and Labeling Variables and Values.**

***Proposal for data management project is due September 25***

**October 2. Planning the Data Management Workflow.**

**October 9. Automating Your Work.**

**October 16. Combining and Reshaping Data Files.**

**October 23. Processing Observations Across Subgroups.**

**October 30. Dealing with Missing Data.**

***Responses to dataset questions are due October 30***

**November 6. Multiple Imputation for Missing Data.**

**November 13. Do File Programming.**
Chapters 4, 6, 7 from Baum, Christopher F. 2009. An Introduction to Stata Programming. Stata Press.

--*Dataset Presentations, Part 1*

***Drafts of data management assignments are due for peer feedback on November 17***
November 20. Do File Programming (continued).
Chapters 8, 9, 10 from Baum, Christopher F. 2009. An Introduction to Stata Programming. Stata Press.
--Dataset Presentations, Part 2

November 27: No class, Thanksgiving break

***Peer feedback on drafts is due on December 1***

December 4 Programming Using Ado Files.
Chapters 11 and 12 from Baum, Christopher F. 2009. An Introduction to Stata Programming. Stata Press.
--Dataset Presentations, Part 3

December 11 Programming Using Ado Files (continued).
--Dataset Presentations, Part 4

***Final draft of data management assignment due December 13 at 5PM***