Stat 4197: Spring 2011
SAS as a Programming Language and a Statistical Database Management System

Prof. Robert F. Teitel; email: rteitel@gwu.edu
phone: 301.654.3684
hours: before and after class

T/A: Brandon Nedwek; email: bnedwek@gwu.edu
phone: 971.506.7739

Email Note: Please use "STAT197" at beginning of subject line if sending us email.
Replies may come from rteitel@comcast.net
Replies may come from bnedwek@gmail.com

Class: Thu : 6:10 – 8:40 Gov 103 Mostly lecture
Some demonstrations and lab time

Course content:
An alternative course in Computer Science based on the SAS system.
SAS is best known as a statistical system, but this course will not use any of the real statistical procedures. It will, however, concentrate on the use of SAS as a programming language for the manipulation of statistical data. Included will be a discussion of modern statistical database management, based on the relational model of data, and how to apply SAS to complex data collections with multiple units of analysis.

Text:
"The Little SAS Book: A Primer; Delwiche and Slaughter
Material from other SAS manuals will be distributed as necessary, or reference given to online versions.

The SAS reference manual is, as implied by its title, “reference”, not fit to be read, but to serve mostly as necessary reference material. No specific reading assignments will be made. However, you'll need it whenever you are working on an assignment.

Homework assignments:
Lecture once per week (Thursday).
Assignments distributed by email once per week, generally on Friday.
The emails are complete, in the sense that they contain both the data to be processed and the description of the required processing. Save the email as a simple text file, and use that file as the input file to the SAS system. Some data files may be distributed via BlackBoard.

At the next lecture, hand in printed versions of the LOG and the LST files produced by SAS. Some assignments will be based on previous assignments, so you should keep your previous work.

You WILL need to work on the assignments OUTSIDE scheduled Laboratory time, especially later in the semester. You MUST retain backup copies of your programs. A ZIP disk or a USB disk is strongly encouraged (instead of 3.5" disks).

You will generally work in assigned pairs on the weekly programs. The pairings will vary throughout the semester. Each pairing submits a single work product.
Learning Outcomes:
At the end of this course, students should have a good working knowledge of
1. The SAS “data step”, including
   importing data from a variety of formats;
   procedural programming; and
   generating SAS datasets.
2. Manipulating multiple datasets, including
   Aggregations; and
   Dis-aggregations.
3. SAS macro programming.

Grading:
Assignments must be done, and will be 50% of the final grade.
Two person teams will be assigned to work together on assignments; hand in a
single set of printed LST and LOG files with both names as comments.

A midterm will count 20% and a final will count 30%.
The midterm and final will be open book, open notes.

Academic Integrity Issues:
For the weekly assignments, you will generally work in pairs. You may also use
printed and electronic (internet) resources, but you must indicate that you
have done so as comments in the programs (just as you would footnote such
references in conventional papers). You may even have incidental conversations
with fellow students; helpful insights should, once again, be indicated as
comments in the programs. Communication with, or help from, persons not in this
class is not permitted. Regardless of the resources used, the work submitted
must still represent your own work product.

For midterm and the final exams, the above still hold, except that you may not
have any communication with any person, whether in this class or not. That is,
the exam work products shall be completely your own with, of course, properly
attributed researched resources.

Any work you submit will be held to these standards.
Failure to do so will be appropriately sanctioned.

Syllabus:
The following is a general outline of the subject matter.
The amount of time spent on each section, and specific assignments will depend
somewhat on class progress and interest. Let me know of specific interests.
The first three sections should occupy about half the semester; section four is
really the “meat” of the course.

Introduction: Overview of the SAS data model.
SAS concepts: dataset, value types

Section 1: Creating single SAS datasets.
This section focuses on the various methods available in SAS to create datasets
from (text) input files.

- list directed input
- fixed column input
- sequential formatted input
- absolute referenced formatted input
- relative referenced formatted input
- multiple record, single observation input
- single record, multiple observation input
Section 2: Algorithmic statements
This section focuses on the fundamental components of algorithm specification using the SAS programming constructs.

- arithmetic computation (SAS: assignment, expression, functions)
- character manipulation (SAS: conversion, functions)
- statement vs statement sequence (SAS: Do-end)
- conditional choice (SAS: If-then-else; Select-when-otherwise-end)
- count controlled iteration (SAS: Do, count controlled version)
- conditional iteration (SAS: Do-while; Do-until)
- array processing (SAS: Array)

Section 3: Multiple record types, multiple units, multiple datasets
This section introduces the relational model of data as a thought process for SAS datasets, and focuses on the creation of multiple datasets from files with various record structures.

-... from files with multiple record types
  -- (the wrong way)
  -- (the right way)
  -- (alternatives)

-... from files with records with multiple observational units types
  -- (the wrong way)
  -- (the right way)

-... from files with all kinds of records

Section 4: manipulating multiple unit-of-analysis datasets.
This section focuses on manipulating multiple dataset representing different units of analysis within a data collection.

- creating normalized datasets from existing datasets (aggregations).
- sequencing of aggregations and dis-aggregations (attributions)

SAS procedures: proc sort, proc summary.

Material on SAS macros will be woven into the sections above.