**STAT 440**

**Due Thursday, October 13 at 11:50pm**

# Data Report

You are allowed to use internet, and relate Materials. You are allowed to consult e-Learning modules, online notes, lecture notes, SAS documentation, and the recommended textbooks.

You must complete the exercises and turn in the SAS program file and Report just like with HW. Submissions must be uploaded to our Compass 2g site on the Exams page. No email, hardcopy, or late submissions will be accepted.

## Getting the program file ready

a. Create a folder on the hard drive with the following pathname – C:\440\midterm. Save all data files accompanying this assignment in that folder. If you cannot create the folder because you are working on a university computer and don't have permission, create the …\440\midterm folder elsewhere.
b. Assign the library reference **midterm** to the folder 'C:\440\midterm'. Use this library as your permanent library for this assignment. If you could not create the folder, assign the library reference **midterm** to your …\440\midterm folder.
   Note: If you are using a folder other than 'C:\440\midterm', you must change any pathname references in your program file to 'C:\440\midterm' before submitting your homework.

## Submitting your work to Compass 2g

You are to submit two (and only two) files for your midterm submission.

1. Your SAS program file which should be saved as **midterm_*YourNetID*.sas**. All program statements and code should be included in one program file.

2. Your Report which should be saved as **miderm_*YourNetID***, preferably as a PDF file. For this midterm, it's probably best to use ODS to send your results to a Rich Text Format (RTF) file, so that you can then mesh the output with your report dialogue.

You have an unlimited number of submissions, but only the last one will be viewed and graded. Midterm submissions must always come as a pair of files, as described above.

**Background:**
Suppose we were going to do a statistical analysis of the University of Illinois football team's record and how it impacts their AP ranking and Bowl game results. To do so, we would need a data set that's fully validated and cleaned.

**Dataset:**
You will work with a data set containing information from the 125 seasons of University of Illinois football. The raw data set **illinifb16.dat** contains data from 1892 to 2016.

| Field | Description | Notes |
|---|---|---|
| 1 | Obs | Observation number |
| 2 | Season | |
| 3 | Conf | Conference |
| 4 | W | Wins |
| 5 | L | Losses |
| 6 | T | Ties |
| 7 | Pct | Win percentage |
| 8 | SRS | Simple Rating System: A rating that takes into account average point differential and strength of schedule. Average of all teams in a season is 0. |
| 9 | SOS | Strength of Schedule: Average of all teams in a season is 0. |
| 10 | AP_pre | Rank in pre-season AP poll.<br>Possible values are 1-25 and missing if unranked. |
| 11 | AP_high | Highest rank of the team in the AP poll during that season.<br>Possible values are 1-25 and missing if unranked. |
| 12 | AP_post | Rank in final AP poll at the end of the season.<br>Possible values are 1-25 and missing if unranked. |
| 13 | ConfTitle | Did Illinois win its Conference Title: Y or N |
| 14 | Coach | Head coach (or coaches) |
| 15 | Record | Coach's record |
| 16 | Bowl | Name of post-season bowl game played in, or missing |
| 17 | BowlResult | Result of Bowl game: W or L |

**Goal:**
Adequately prepare this dataset for statistical analysis. Here is a list of items you may need to consider.

- Reading raw data files
- Creating formats and labels
- Deriving new variables via calculations or recoding
- Subsetting data
- Checking data for errors
- Validating and cleaning data

To help you along, here are some data details.

- Each record is a unique season, so each value of Season must be unique.
- The values of W, L, T, and Pct should coincide correctly. Winning percentage is equal to the number of wins (W) divided by the total number of games (W+L+T).
- No one is expected to know the proper spelling of each Head Coach's name. If there are any typos in a coach's name, each unique spelling would appear in a frequency report.
- If more than one coach is listed for a season, clean the Coach and Record variables to note which coach had more wins that season.
- If no coach is listed, it means that more than one person shared the duties of coach. Learn who they were by searching the internet and clean the Coach variable to contain the name of the coach whose last name comes first in alphabetical order. Clean the Record variable to match the W, L, and T entries.

**Midterm Report:**

A summary report that includes the following.

1. Title of the project.
2. Your name.
3. Methods section:
   - Description of the original data file including what type of input style it uses.
   - Description of the guidelines used to validate the data.
   - Description of the issues needed to be cleaned and how it will was done (though not needing to explain the programming code specifically).
   - Description of additional data preparation that you performed.
   - Description of variables to be analyzed including attributes such as name and type. You do not have to list all the variables in the original sourced file, but do mention the ones you bring to SAS for the creation of the SAS data set.
4. Results section:
   - Tables and visualizations pertaining to validation and cleaning.
   - Write-up of the results. Point out notable information from the charts and tables.
5. To verify that the cleaning was thoroughly completed, also answer these questions in your Results section:
   a. Identify which Head Coach or Coaches had the most wins in his career with the University of Illinois football team.
   b. Identify which Season(s) saw the football team with their highest ranking for the university across all seasons. Note that #1 is the highest ranking possible.
   c. Identify the number of times that Illinois won its conference title.
   d. Identify which decade had the most wins in a decade. For example, 1892-1899 will be the decade known as the 1890s; 1900-1909 is the 1900s; … 2010-2016 is the 2010s.
6. Write in complete sentences and pay attention to grammar, spelling, readability and presentation. If you include a table or chart, make sure you say something about it. If you're not discussing a result, then it doesn't belong in your report.

In terms of length, it probably shouldn't take more than 2-3 pages to explain your work on this dataset. That does not include the space occupied by tables and other output. If you have a point to make, get to it. If you find yourself writing things simply for the sake of padding the word-count, you're writing the wrong things.

**Midterm Grading:**
The grading rubric for the final project is summarized by five criteria, each worth 10 points.

- Methodology of data preparation
- Validation and cleaning
- Correctness of interpretation of results and output
- SAS file/programming (10pts, similar to HW expectation)
- Report organization and presentation (10pts, similar to HW expectation)

Maximum total points: 50