ST 307 Spring 2018

Activity 10

This week we will be working on the data set in the file SAT.csv, which can be downloaded from Moodle. For each high school in New York City, the average score for each of the three SAT subjects, reading, mathematics and writing, are collected, along with the number of students taking SAT in 2010. More information can be found from link. (https://catalog.data.gov/dataset/sat-college-board-2010-school-level-results-5c6d6).

Missing data: Records with 5 or fewer students are suppressed (marked 's').

To do:

1. Create a library called Topic10.
2. Read in the data with all the 6 variables, then drop the variable DBN. Here are two points you need to pay attention to:
    a. Some of the school names contain a comma in between. This will cause some trouble since the delimiter in this data set is also comma, and SAS will break those school names apart. To solve this issue, add a DSD option on the INFILE statement.
    b. By default, SAS will treat a period (.) as missing numeric value, but not a character 's'. To make this happen, try adding a MISSING statement in the DATA step.
3. Setting aside the number of students in each high school, perform a naive correlation analysis among the average score of three subjects.
    a. Have SAS produce a scatter plot matrix with histograms along the diagonal.
    b. Answer in a comment: which two of them have the strongest correlation? What is their correlation coefficient?
4. Taking into account the number of students this time, perform a weighted correlation analysis among those three variables. So each data point will be treated as if they appear the number of times equal to the number of students taking the exam in that high school. A WEIGHT statement will help you with it.
    a. Have SAS produce a scatter plot matrix with histograms along the diagonal.
    b. Answer in a comment: which two of them have the strongest correlation? What is their correlation coefficient?
5. Find a 95% confidence interval for the correlation between the school-average reading and writing scores, under Fisher's z transformation. (A weighted test will be preferred.)
6. Fit a simple linear model with the school-average reading score as predictor and writing score as response (ignore the number of students).
    a. Find the 95% confidence interval for the slope of this model. Interpret you result.
    b. Have SAS produce the diagnostic plots and comment on the assumptions of this model.