Project for SML 2013

Each team will analyze a data set of your choosing, using a regression method or methods we have discussed in this class.

The data set could be one that is of personal or academic interest to you and that fits one of the analyses studied in SML. It should be a real data set which you have not analyzed before and which has not been analyzed in a textbook (using the methods of SML, anyway).

Project Part I– Data Set Proposal/Description:

Each team should write a one- to two-page typed description of the data set you propose to study. You should include details about the response variable, about (potential) predictor variables, and about the number of observations. If there are issues such as obvious small n large p problem etc., comment on these and suggest possible remedies. Discuss the source of the data set, and whether the data come from an experimental or observational study.

In addition, please include a half-page printout of the data set (or if it is quite large, a selected part of the data set).

You should also include a general proposal for what sort of analysis you plan to do with this data set. If there are any hypotheses/research questions that are of interest from the beginning, you might mention those.

This part is due on Dec 15 (Sun) by email (pdf format preferred) and we will discuss them in ensuing classes.

Project Part II – Presentation:

The class presentation is scheduled on Dec 26 **(Thr)** and Dec 30 **(Mon)**. The presentation should be an oral presentation of your Written Report (detailed in Project Part III). The presentation time is 40 min for each team. Questions and comments will be raised/given in the presentation.

Project Part III – Written Report:

You should write a concise report summarizing your analysis. The report should be no longer than six (typed) pages, not counting any R output, graphs, etc., which you may wish to include as support or illustration for your analysis.

The style of the report is up to you, but the best reports will address many of the questions and details studied in class when we discussed the relevant type of analysis.

Some things to include (depending on the data set and choice of model) might be:

- An introduction and discussion of the data set itself
- Brief intro to the questions we would like to address and related literature
- Summary of your results (answers to the questions and implications)
- Discussion of the classifiers under your study: performance and possible explanation for the performance. Comparison of linear and nonlinear classifiers/ model-based classifiers/distance-based classifiers.
- Your overall conclusions about the data, based on your analysis

Part III of the project (the final project report) will be due ***on* Jan 6 (Mon), 2014 in class (Final)**. It will count for 30% of your final grade (5% for Part I, 15% for Part II, 10% for Part III).

Note: This description is based on the *project information* in http://www.stat.sc.edu/~hitchcock/stat704.html