Project

**General description:** The purpose of the term project is for you to gain familiarity with the methods presented in class as you apply the methods to your specific field of interest. Additionally, the project should help to solidify your understanding of such methods, and will enable you to compare/contrast key course concepts.

For your project, you are required to do a case study on your selected dataset. Your project can take one of two forms:

1. A comparative study where you compare the performance of three methods presented in this course on your dataset.
2. A complete and thorough analysis of your data, which includes application of at least three methods presented in this course to your dataset.

Prior to beginning your project, you must obtain approval for your proposed dataset from the instructor. In selecting a dataset, you should pay critical attention to issues concerning data quality, since a dataset consisting of many missing values and/or errors may leave you with little time to complete the project. The proposed dataset must have enough complexity (e.g., more than 8 independent variables and a decent number of observations) to support an interesting project. If you do not have a dataset, one can be provided by the instructor or you can browse [http://www.statsci.org/datasets.html](http://www.statsci.org/datasets.html), particularly those posted under ‘Larger or Less Processed’ such as those under the UCI Machine Learning Repository link. Please take time to come and talk with me about your project either during my scheduled office hours or by making an appointment.

There will be a few due dates (to be announced) throughout the semester so progress on your project can be monitored. Your project content will be submitted in the following phases:

**Phase 1 Content Components**

1. A simple and clear exposition of the data set you will be using and the question you are addressing. This should be written in the form of an Introduction/Background section(s).
2. A brief methods section indicating whether you will be performing a comparative study or a thorough analysis of your data.
3. A brief results section to include descriptive statistics for each variable in your dataset.

**Phase 2 Content Components**

1. Include any necessary modifications to phase 1 content components.
2. Update the methods section to include exposition of the methods you have selected.

**Phase 3 Content Components**

1. Finalize the methods section – this must include a clear discussion of the methods you used.
2. Present the results. Apply the methods discussed to your dataset and clearly present the results. You may include tables showing results and comparisons, graphs, and/or theoretical results.

3. Discussion and conclusion.

4. Computational component: source and sink files for your R code as well as the accompanying dataset must be e-mailed to the instructor so that it can be tested. Make sure your code is readable and documented, so I can do a little trouble shooting if necessary.

Note: Computer output should be incorporated into your report in the usual way, i.e., put tables in the text and use graphics, but do not paste piles of unedited computer output into your report. Tables and figures should be numbered and captioned. The quality of presentation will be taken into consideration for your final grade.