STAT 410/510 **Suaray**
Fall 2013 Class Project Instructions

Due Thursday, 12/12/13 at 12:30pm.

The following is written to provide guidelines on what is expected of you for the class project, which is 20% of your grade. Your project will include

A. A cover sheet that follows the format of the sample cover sheet.

B. An abstract, which is a one paragraph summary of the results contained in the project.

C. A table of contents, followed by a list of tables and figures.

D. The body of the paper will start with an introduction, and then give an explanation of the data set. Discussion of the analysis of the data set will follow (see below for guidelines). Note: If you are studying an actual regression technique or other theoretical consideration, you might not necessarily be following the guidelines below. This will be followed by a summary of results, and then a conclusion.

E. An appendix, in which you will give a one page example of what your data set looks like, all of your SAS code (**not** output; you will give relevant output values in D. above), pertinent graphs, and any supplementary technical discussion or proofs.

F. A bibliography, in which you will list a minimum of four references, two of which may be websites, and one of which will be the text. You will make explicit reference, including page number, to these materials in the body of the text.

By Tuesday 11/26, you will show me the work you have completed, either in office hours, or in lab on 4/18. The minimum is
1.Cover sheet.
2.Intro, and explanation of the data set.
3.Preliminary **proc reg** with output.
4.Bibliography (preliminary).

This preliminary check will comprise 3% of your class grade, so it is important that you take it seriously.

The body of the paper should be a minimum of 10 pages, and contain a maximum of 7 graphs. Please highlight areas in tables that are being emphasized.

Here is what is expected for the actual project. Feel free to go above and beyond what is mentioned here; these are minimum requirements. Report your findings in such a way that one of your classmates could take the results and easily apply them to estimating or predicting response values based on your recommendations. Essentially, you will follow the steps given in the flowchart on pg. 344 in your text. **Plagiarism will result in an F for the class.**

1. Find some data, either from the Internet or from other resources available to you (work, research, etc.) for which you desire to explore the relationship between a response (dependent) variable and several potential explanatory (independent) variables. Your data must have at least five explanatory variables (*x's*), and thirty observations (*n>30*). Determine which of the four types of data given on pages 343-346 is most applicable. Informally propose the data and what you seek to discover with the instructor.

2. Use SAS to investigate the effectiveness of the full model. Are there any variables that are not necessary? Why (give a solid explanation, mentioning theoretical considerations, and how they apply to your particular data set)? What is the individual effect of the variables you are including in the model?
a. Any evidence of multicollinearity? If so, why? Give both an intuitive as well as computational justification for your answer. Why or why not? What can you do to fix this?
b. Model specification. Is a complex (perhaps quadratic/interaction, etc.) model a possibility?
c. Conduct joint inferences (intervals and tests) to help facilitate your narrative.

3. a. Residual Diagnostics. Are there any outliers in the data set? Why do you believe that a certain case is an outlier? What happens when you fit the data without the outliers? Is there any evidence of heteroscedasticity? If so, can you remedy this?
b. Check for normality and other model assumptions.

4. Use the model selection techniques to determine which model is the best. Are your results the same using the different methods?

5. Your project will involve one of the following topics covered in the book but discussed only briefly in lecture: **1.)** Ridge regression (11.2) **2.)** Nonlinear regression (13.1-4) **3.)** Poisson Regression (14.) **4.)** Lack of fit tests (3.7, 6.8). **5.)** Topic of your choice (must be approved by professor). 510 students will provide deeper theoretical background not expected of 410 students.

6. Produce an ANOVA table for your final model, and comment on its results.

7. How confident are you of the predictive power of your regression? Why? Perform data splitting (see pages 369-375) as a way to validate your model. Use MSPR, and show how predictions for a random sample of 5 cases compare to actual values.

As discussed earlier, Stat 410 students will simply submit the project, whereas 510 students will give a 10 minute Powerpoint presentation. Sample projects will be available upon request.