Suaray **11.21.06**

MATH 410/510: REGRESSION ANALYSIS- PRACTICE MIDTERM 2

You are allowed 1.25 hours to complete this exam. There are a total of 50 possible points. You must show all work to get credit. Print out all relevant SAS code and output. Look through the entire exam before you start. Please do not hesitate to raise your hand if you have a question.

1. a. Express the two sample problem as a specific instance of the multiple regression model. Clearly identify the design matrix, the response variables, and the parameters.
   b. Find the least squares estimates for the parameters using the usual method.

2. The Federal Trade Commission (FTC) annually ranks varieties of domestic cigarettes according to their tar, nicotine, and carbon monoxide contents. The U.S. surgeon general considers each of these three substances hazardous to a smoker's health. Past studies have shown that increases in the tar and nicotine contents of a cigarette are accompanied by an increase in the carbon monoxide emitted from the cigarette smoke. The data set (see http://www.amstat.org/publications/jse/v2n1/datasets.mcintyre.html) lists tar, nicotine, and carbon monoxide contents (in milligrams) and weight (in grams) for a simple of 25 (filter) brands tested in a recent year. Suppose we want to model carbon monoxide content, y, as a function of tar content, x_1, nicotine content, x_2, and weight , x_3.
   a. Perform a regression analysis on the data using SAS. Are any of the variables significant?
   b. Calculate the variance inflation factors by finding the R_j values. What does this indicate about multicollinearity?
   c. What is your choice of a final model? Use forward selection to assist you in making your choice. Does a log transformation increase the coefficient of determination R^2 for your model?

3. a. There are a number of ways that we can detect outliers. Give a formula (or definition) of the following statistics, and describe the rules of thumb we use for outlier detection: i. Leverage ii. Cook's D iii. Studentized residuals

   b. Draw a sample graph illustrating a violation of the following assumptions concerning the residuals α : i. $Var(\varepsilon_i)=\sigma^2$, for all i (i.e. heteroscedasticity).
   ii. $\varepsilon_i \sim N(0,\sigma^2)$ for all i.

4. Use stepwise regression to determine the best model for the asphalt data. Use an α level of SLE=SLS=0.15. Are there any potential outliers?