Stat 430, Problem Set 6, Due Friday April 24, 2009

For this assignment, provide the SAS program code used as well as the *edited* SAS output you produced to answer the questions. You may annotate your SAS output, in handwritten form if you like, but verbally explain how your output answers the questions asked, and *please* do not hand in data or printed output which is not specifically requested and does not figure in your answers to questions.

(I). The dataset bass contains data from a study of Mercury contamination in fish that live in Floridian lakes.

(a). Make a scatterplot of average mercury contamination (AvgMercury) as a function of alkalinity.

(b). Use the log transform on the response, and fit a regression line. Prepare a scatterplot with regression line, and a residual plot as well.

(c). Construct the Cook's distance measure from the data. Remove the outliers that you identify from the residual plot. Do the outliers have the largest Cook's distances?

(d). After removing cases 36 and 52, you will see that there is evidence of another outlier. Keep going until you have deleted everything outlier-like (i.e. cases 36, 52, 40, 3, and 38). Fit a regression model to the remaining points and identify the changes in the regression coefficient estimates and in `adj R-sq` between your model with all the outliers and your model with none of them.

(e). Construct the 95% prediction interval at the value of the predictor near the outlier, using SAS to generate the intervals before and after removing the outlier. Does the outlier appear to have a great effect on this interval? Based on your result, decide whether or not you would consider the outlier an influential case.

(f). Do a scatterplot of the response versus the log(predictor). Why would this produce prediction intervals that would be difficult to trust? [Note: you may be able to solve this one by observation, without running the regression program again].

(g). Based on the results from your analysis, would you hypothesize that acid rain (which decreases alkalinity) is likely to improve or make worse the average levels of mercury contamination in fish? Would your conclusion from this analysis alone be sufficient to have NOAA sending trucks to dump calcium chloride into Florida lakes? Explain briefly.

(II). (a) Create (*and save*!) by pseudo-random Monte Carlo simulation a large (n=1000) SAS dataset with the columns Y, X, Z defined as follows: $X \sim$ Uniform$[0, 5]$, $Z \sim$ Binom$(1, 0.5)$ are independent random variables in each row, and if V denotes another independent random variable with t_3 distribution, then

$$Y \;=\; 1.5 + 2 * X - 0.4 * X^2 + 7 * Z - .1 * X * Z + 2 * V$$

Hint: if you multiply c times a Uniform$[0, 1]$ random variable, you get a Uniform$[0, c]$ random variable; and the easiest way to generate a t_3 random variable is to generate four independent $N(0, 1)$ random variables W_1, W_2, W_2, W_4 using the `SAS` function `RANNOR` and then define $V = W_1 * \sqrt{3}/\sqrt{W_2^2 + W_3^2 + W_4^2}$.

(b) Fit a simple linear regression model of Y on X to your dataset. Examine a residuals plot or (use `SAS`-generated prediction intervals and/or studentized residuals or other statistical tools in `SAS` to show how you would be guided in this setting to augment the model by including a quadratic (X^2) term in the model and also a term involving Z.

(c) Now fit the multiple regression model with Y modelled in terms of X, X^2, Z. Note that in this problem, we know in advance what the correct model should be. The objective is to show which tools get us to build the correct model. What tools would you use to examine whether a fourth predictor variable $X * Z$ will actually improve the fit of the model.

(d) For the multiple regression model based on the correct model (Y regressed on $X, X^2, Z, X * Z$), do the residuals look patternless (plottedd both against X and against the predictor $\hat{Y}$? Plot a histogram of the residuals from the final fitted model and examine them for normality. (Use histograms with over-plotted normal densities with same mean and variance, or QQplot.) Do the residuals look normal ?

(III). The dataset `cathedrals` contains a list of the heights and lengths of a selection of medieval English cathedrals. Of these, the Romanesque cathedrals are indexed by `style=0` and the Gothic by `style=1`. Find the best model you can to describe height in terms of style and length. (Choose a reasonable criterion for this !) You may want to consider the following:

- transformations do not seem to be useful,

- a quadratic term may be useful,
- all interactions between the categorical (style) and measurement (length) variable should be considered, initially,
- there may be outliers, but try to find an objective basis on which to decide whether there are any and if so, which observations might reasonably be excluded from the main analysis.

Consider all possible predictor combinations when removing redundant predictors, but a close look at the scatterplot should help. You may use PROC REG along with selection options, but make sure to include some justifications of why all of the model terms you include are (sufficiently) significant to the final results, and how you know that the terms you excluded are not. Also, make sure to hand in at least some displays (eg, scatterplots and/or residuals plots) with points for the Romanesque and Gothic cathedrals plotted with different plotting symbols.