ST505/ST697R: Fall 2012 Homework 4. Due Friday Oct. 12

First three go back to concepts from earlier in past week. Problems 1, 2 and 3 are very brief. The fifth problem involves model assessment (some of this done Friday, Oct. 5; more on Tuesday the 9th). SAS and R code for diagnostics will be posted (and are in notes).

1. 2.21

2. 2.41

3. Show algebraically that $SSE = \sum_{i=1}^{n} e_i^2 = \sum_i (Y_i - \bar{Y})^2 - b_1^2 \sum_i (X_i - \bar{X})^2$. Hint: First write $e_i = Y_i - \hat{Y}_i = Y_i - \bar{Y} - b_1(X_i - \bar{X})$ (why is this okay?) and then expand to get $\sum_i e_i^2 = \sum_i (Y_i - \bar{Y})^2 - 2b_1 \sum_i (Y_i - \bar{Y})(X_i - \bar{X}) + b_1^2 \sum_i (X_i - \bar{X})^2$ and simplify.

   Implication of this result: This can be used to show the expression for $\hat{\sigma}^2$ on page 41 of notes (I'm not asking you to do that). It also shows that $SSE = SSTO - b_1^2 \sum_i (X_i - \bar{X})^2$ and so $SSR = b_1^2 \sum_i (X_i - \bar{X})^2$. Use this to show that $r^2 = R^2$, where $R^2 = SSR/SSTO$ and r is the sample (Pearson) correlation (this should be relatively quick given the above).

4. Write the residual e_i as $\epsilon_i + something$ where the something has expected value 0. This motivates treating the e_i as approximately ϵ_i for diagnostics. (Explained in class).

5. For this use the nitrogen balance-intake data, used for an example in class. The data is posted. We start using the simple linear regression model for balance on intake. Use R or SAS to do the following:

   (a) Obtain plots of the residual versus intake and predicted value and a plot of the absolute residual versus intake.

   (b) Find a normal probability plot for the residuals, a histogram (along with a smooth curve overlayed) and tests for normality.

   (c) Use the above plots and output to evaluate the assumptions that i) that a linear regression model is adequate ii) that the errors have constant variance and iii) that the errors are normally distributed. Summarize your conclusions explaining carefully how you are using the plots and tests. For i) and ii) just use the plots, for iii) use appropriate plots and the tests for normality to back up your conclusions.

   - Comment also on whether normality of the errors is important here or not. Will it matter and if so, for what.

   The next two parts used two approximate tests of constant variance arising from different models on the variance.

   (d) Assuming the model (3.10) in the book is correct, use the regression of $log(e_i^2)$ on ni to get a t-test of $\gamma_1 = 0$ and hence of constant variance. This is an alternative to the Breusch-Pagan test of constant variance using the same model. We'll use this alternative since it is easily implemented. Does model (3.10) seem reasonable from a plot of $log(e_i^2)$ on ni?

   (e) Now use the plot of $log(|e_i|)$ versus $log(\hat{Y}_i)$ to see if the model $\sigma_i = \theta_1 \mu_i^{\theta_2}$ is reasonable (if the model is reasonable then the fit of $log(|e_i|)$ versus $log(\hat{Y}_i)$ should look roughly linear). Then assuming the model is at least reasonable use the regression of $log(|e_i|)$ on $log(\hat{Y}_i)$ to test for the assumption of constant variance; i.e., test that θ_2 is 0.

   State your conclusions regarding the constant variance assumption from these last two parts.

Other suggested problems (not to be handed in, but worth looking over at some point:)
2.22, 2.40 and 2.36. 2.36s is worded a bit badly in my view. With random X's one can use both a regression and a correlation model. What they are really trying to ask here is "would it make sense to estimate a correlation here?" (or correspondingly, make use of R^2). 1.43 and 3.25 (connected).