Statistics 512: Homework#1
Due January 24, 2014 BEFORE CLASS

A reminder – Please do not hand in any unlabeled or unedited SAS output. Include in your write-up only those results that are necessary to present a complete solution. In particular, questions must be answered in order (including graphs), and all graphs must be fully labeled (main title should include question number, and all axes should be labeled). Don't forget to put all necessary information (see course policies) on the first page. Include the SAS input for all questions at the very end of your homework. You will often be asked to continue problems on successive homework assignments. So save all your SAS code.

1. A regression analysis relating test scores (Y) to training hours (X) produced the following fitted question: $\hat{y} = 25 - 0.5x$.

   (a) What is the fitted value of the response variable corresponding to $x = 7$?

   (b) What is the residual corresponding to the data point with $x = 3$ and $y = 30$?

   (c) If x increases 3 units, how does $\hat{y}$ change?

   (d) An additional test score is to obtained for a new observation at $x = 6$. Would the test score for the new observation necessarily be 22? Explain.

   (e) The error sums of squares (SSE) for this model was found to be 7. If there were $n = 16$ observations, provide the best estimate for σ^2.

   (f) Rewrite the regression equation in terms of x^* where x^* is training time measured in seconds. Show that your answer makes sense, i.e., gives the same predictions as the original equation (an example is sufficient).

2. Explain the difference between the following two equations:

$$\begin{aligned} \hat{Y} &= b_0 + b_1 X \\ Y &= \beta_0 + \beta_1 X + \epsilon. \end{aligned}$$

3. For this problem, use the "grade point average" data described in KNNL Problem #1.19. The data are on the disk that accompanies the text and can also be found on the class web site (`CH01PR19.DAT`). Make sure you understand which column is X and which is Y and read in the data accordingly. See Topic 1 or `nknw060.sas` for an example of how to read in a data file.

   (a) Plot the data using `proc gplot`. Include a smoothed function on the plot by preceding the `plot` statement with "`SYMBOL1 i = smNN`" where `NN` is a number between 1 and 99. Note that larger numbers cause greater smoothing. Make sure to indicate the smoothing number in the title of the plot. Is the relationship approximately linear?

   (b) Run a linear regression to predict GPA based on the entrance exam. Give the complete ANOVA table for this regression.

   (c) Give a point estimate and a 95% confidence interval for the slope and intercept and interpret each of these in words. (*Point estimate* is another word for parameter estimate.)

   (d) Would it be reasonable to consider inference on the intercept for this problem? Please provide justification for your answer.

4. For this problem, use the "plastic hardness" data described in the text with problem 1.22 on page 36. (`CH01PR22.DAT`) Make sure you understand which column is X and which is Y and read in the data accordingly.

   (a) Plot the data using `PROC GPLOT`. Include a regression line on the plot (`i = rl`). Is the relationship approximately linear?

   (b) Run the linear regression to predict hardness from time. Give

      i. the linear model used in this problem

      ii. the estimated regression equation.

   (c) Describe the results of the significance test for the slope. Give the hypotheses being tested, the test statistic with degrees of freedom, the p-value, and your conclusion in sentence form.

   (d) Explain why or why not inference on the intercept is reasonable (i.e. of interest) in this problem.

5. An investigative study collected 40 observations from the Wabash river at random locations near Lafayette. Each observation consisted of a measure of water pH (X) and fish count (Y). The researchers are interested in how the acidity of the water affects the number of fish. Complete the following ANOVA table for the regression analysis (the p-value need not be exact). State the null and alternative hypotheses for the F-test as well as your conclusion in sentence form.

Source	degrees of freedom	Sum of Squares	Mean Square	F-value	Pr $> F$
Model	_____	55.30	_____	_____	_____
Error	_____	_____	_____		
Corrected Total	_____	60.00			