Statistics 512: Homework 1
Divisions 1, 2, 3, and 4: Due Monday, January 26, 2015
Division 5: Due Tuesday, January 27, 2015

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.

Important Note *– Every graph or plot you create should have your name printed as a subtitle. Consequently, any graph with no name will result in a* **20% points off** *on that question. Also, please attach your code at the end; any homework with no code provided will result in a* **50% points off** *on the entire assignment, NO EXCEPTIONS.*

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 $y = 30$? Is the point above or below the line?

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

   (d) An additional test score is to be 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.

2. Explain the difference between the following two equations: Also, describe the distribution of Y_i and ϵ_i. What is the meaning of β_0 and β_1?

$$\begin{aligned} \hat{Y}_i &= b_0 + b_1 X_i \\ Y_i &= \beta_0 + \beta_1 X_i + \epsilon_i. \end{aligned}$$

3. For this problem, use the "grade point average" data described in KNNL Problem #1.19, page 35 (white text) or page 41 (blue text). 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 `knnl054.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 v = square i = smNN`" where `NN` is a number between 1 and 99. Note that larger numbers cause greater smoothing. For this homework please use the number 70. 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 94% 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 (white text) or page 42 (blue text). (`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 (`v= square 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. State the null and alternative hypotheses for the F-test as well as your conclusion in sentence form. You may use the critical F (critical t) approach or the $p-$value approach.

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