ST505/697R: Fall 2012 Homework 2. Due: Wednesday, Sept. 26

Same general remarks apply as on homework 1 regarding presentation, etc. *Homework with scratching out will not receive credit.*

- For each question that requires computing use R and SAS to provide as much of the answer as possible. Isolate your code and output. You can edit the output so it only has the key parts needed to answer the problem. However, do not just say see output or write answers on the output. Clearly provide the answer to each question, which could involve cutting some of the output into your solution.

Do not put this off until next week to start! Be sure to look the questions over beforehand and ask questions as needed in class and office hours.

1. This question is all related to the data in Problem 1.22. Use R or SAS to obtain results.

   (a) Problem 1.22

   (b) Problem 1.26

   (c) Identify the estimated standard error for each of the estimated intercept and slope and get a 95% confidence interval for each of β_0 and β_1. You can get the confidence intervals directly from R or SAS but also write down how they are obtained from the point estimate, standard error and an appropriate t-value, showing what numerical values are involved.

   (d) Problem 2.7, parts a) and b) only!

   (e) Problem 2.16

   (f) Get a plot of the confidence intervals for $E(Y)$ and prediction intervals, plotted as a function of X.

   (g) Problem 4.9

2. Often you may be faced with working with someone else's summary of an analysis rather than the full data. Below is analysis of 2005 Dallas marathon results regressing finish time for the full marathon (Y = full) versus the time at the half (X = half), assuming the simple linear regression model (with constant variance) is appropriate. This shows partial results, here from SAS, but the labeling is like that we've seen in R.

```
                    Dependent Variable: full
                      Analysis of Variance
                              Sum of          Mean
Source               DF      Squares        Square  F Value  Pr > F
Model                 1   1968.59650    1968.59650  31213.4  <.0001
Error              3164    199.55035       0.06307
```

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.07471	0.02517	2.97	
half	1	2.10983	0.01194	......	......

Covariance of Estimates

Variable	Intercept	half
Intercept	0.0006336004	-0.000295834
half	-0.000295834	0.0001426116

(a) Identify the estimates of β_0,β_1, σ^2 and σ.

(b) Compute a 95% confidence intervals for β_1. (Most of what you need you can get from the output, but you'll need to get a Z table value to go into the confidence interval).

(c) Obtain the t-statistic for testing $H_0 : \beta_1 = 0$. Represent the P-values graphically and explain how you interpret the P-value as a probability. Use the standard normal tables (since there are 3164 degrees of freedom associated with the error term) to bound the P-value for the test.

(d) Find a 95% confidence interval for the expected finish time for those who runs the first half in 1 hour,40 minutes (half = 1.667); then get a 95% prediction interval for the finish time for an individual who ran the first half in 1 hour,40 minutes. (To do these you will need to utilize the variance/covariance information on the output,under the heading of "Covariance of Estimates" as demonstrated in class.)

3. Consider the pH example in class where the coordinators of the study were calibrating the labs. Consider two designs, each with $n = 12$.

   Design 1: The one used in the real data: Here the X values are:
   7.67,6.31,6.14,7.07,6.39,5.95,6.53,6.55,5.34, 5.74,4.94, 7.07 .

   Design 2: Another design with $n = 12$ but using X values 5.3,5.3,5.3,5.3,5.3,6,6,7,7,7,7,7.

   Suppose $\beta_0 = 0$, $\beta_1 = 1$ and $\sigma = .15$.

   For each of the above design use the simulation program to find the exact standard deviation (also referred to as the true standard error) of the estimated slope and intercept and to demonstrate the sampling distribution of the estimators of the intercept, slope and error variance. Use 5000 simulations. Show the summary statistics and histograms for each of these three estimators. Based on what you see here, which of the two designs would you prefer?