ST505: Fall 2012 Homework 7. Due: Wed. Nov. 21.

1. Step 2 of the IE project for ST505 students (These won't be handed in but should be completed by next Friday). See posted document labeled "IE Project Step 2".

   ST697R students: See the posted handout describing step 2 of the I.E. project. You will do the same thing but you will do this working with the SMSA data. A brief description of the data is under the link that describes potential IE projects. The data is posted.

   ST697R student should complete this and summarize your results to hand in as part of this homework. Recommended this get done by next Friday.

   NOTE: The code for getting the scatterplot matrices appears on page 103 of the notes (for both SAS and R).

2. This uses the house data which has been discussed in class; see posted command and data files also. We know want to run weighted least squares to account for the fact that the variance seems to be changing (increasing with expected price). Page 91 of the notes describes how to get estimated standard deviations and weights based on a model where the log standard deviation is linear in the log mean (or equivalently the standard deviation is proportional to the mean to a power). Create these weights and then run weighted least squares (see the esterase example on page 74 of the notes and associated posted programs). Include in your output the confidence intervals for the expected value of Y and prediction intervals which will be computed for each set of X's in the data. Comment on

   (a) How have the standard errors for coefficients changed?

   (b) is there any change in the conclusions from tests for the coefficients.

   (c) What effect does weighting have on the confidence intervals for the mean and prediction intervals? To make this comparison you'll need to run the unweighted version as I only listed a few of the intervals. You don't need to list all the intervals but you'll want more than just the first and last two to address the question.

3. Return to the patient satisfaction data from homework 6 (note that the the homework 6 solution is posted).

   (a) 6.15b, the scatterplot matrix is the matrix of plots for each pair of variables. Discussed in class.

   (b) i. 6.17 a)

      ii. 6.17 b)

      iii. Consider 6.17a). But now compute the confidence interval for the mean in the context of doing simultaneous confidence intervals the mean satisfaction at as many combinations of X_1, X_2 and X_3 as you'd like.

      iv. Consider 6.17b), suppose this was one of three prediction intervals you were going to calculate and you want them to be simultaneous 90%. Show two ways to do this and explain which is better. You don't need to calculate out; just set-up and compare the multiplies (numerically).

   (c) 7.6

4. Problem 8.21.

   In addition add: c) write a model for $E(Y)$ in terms of X_1, X_2 and X_3 and functions of them which allows for a different intercept and slope for each type of head protection.

Here are two **suggested problems**. You should look these over and know how to do the analysis (and at some point try the analysis) but I'm not asking you to write them up. These are good reinforcement of concepts illustrated in other examples from class.

1. Problem 8.37 a) and b.

   The variables in the data are *id county state land pop pop1834 pop65 phys beds crimes percenths percentb poverty unemploy income tincome region* where county and state are character variables.

   Note that we need to define Y = *crimerate=crimes/pop* and X_1 = *density=pop/land*.

   This problem also has an outlying value, one observation with a very large crime rate, that can distort the analysis. A plot of Y versus the X' and the residuals versus X's and fitted values illustrates the issue. In doing b) if you use the test option in SAS it doesn't run, saying (in the log file) that the test is inconsistent or redundant. This is an issue with numerical accuracy in some near singular matrices when it uses a matrix expression to compute the test. You can do this by fitting the full and reduced models.

2. This problem uses the setting of problem 8.15. I've posted a SAS file that reads the data from 1.20 combined with that from 8.15. The data is in a form where the first variable is service time (Y), the second variables is number of copies serviced (X_1) and the third variable is type (X_2) which is 1 if the type of copier is small and is 0 if it is large.

   (a) First do problem 8.15 a) which uses the model in (8.33). Explain what the regression model is for each copier type.

   * For the remainder of the problem we'll use a linear regression model **allowing separate coefficients for each type**. That is: if type = S then $E(Y|X_1) = \beta_{S0} + \beta_{S1}X_1$; if type = L then $E(Y|X_1) = \beta_{L0} + \beta_{L1}X_1$.
   We'll assume the variance is constant throughout for now.

   (b) Define $Z1$ = type and define $Z2 = 1-$ type (so $Z2 = 1$ when type is large and $Z2 = 0$ when type=small). Using these run **ONE!!** proc reg such that it gives you four estimated coefficients in the output and those are estimates of β_{S0}, β_{S1}, β_{L0} and β_{L1}.
   - Using the residuals from this fit run proc univariate to describe the residuals for each type. Does it appear that the errors are similarly behaved across the two types?

   (c) Regress Y on X_1, X_2 and $X_1 * X_2$ including use of the clb option in SAS or use confint in R. This is fitting the model

   $$E(Y|X_1, X_2) = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \beta_3(X_1 * X_2) \qquad (1)$$

   - Show the relationship between the β_0, β_1, β_2 and β_3 in (1) and β_{S0}, β_{S1},β_{L0} and β_{L1} defined earlier.
   - Identify how the β's in (1) relate to the differences between intercepts and slopes for the two types.
   - Use the output from the regression run to obtain i) Separate 95% confidence intervals for $\beta_{S0} - \beta_{L0}$ and $\beta_{S1} - \beta_{L1}$ and ii) separate t-tests for $H_0 : \beta_{S0} = \beta_{L0}$ and $H_0 : \beta_{S1} = \beta_{L1}$. State your conclusions.

   (d) Carry out a test that the two regression lines are equal meaning that they have both equal slopes AND equal intercepts. Do this under the assumption of equal variance throughout. Get the p-value associated with the test.