STA 302 / 1001 – Fall 2003 – Assignment 1

Due: Wednesday, October 22 at beginning of lecture. Assignments received after 1:20 will be considered late. Late assignments will be subject to a deduction of 10% of the total marks for the assignment for each day late.

Presentation of solutions is important. In particular, it is inappropriate to hand in pages of SAS output without explanation or interpretation. Marks will be deducted for poor presentation.

1. In the 2000 British Open (a golf tournament), Ernie Els had a score of 66 in the first round (low scores are good) but a score of 72 in the second round. Do you think that he choked? On the course web site, obtain the file of the 156 golfers at the British Open in 2000 and their scores on the first and second rounds.

   (a) The following golfers had the 10 lowest scores after the first round: Woods, Garcia, Flesch, Lehman, Paulson, Els, Harrington, Garbutt, Maruyuma, Dunlap. How many of these players did worse (higher score) in the second round?

   (b) The following golfers had the 10 highest scores after the first round: Ozaki, Little, Karlsson, Jacobson, DaSilva, Johnstone, Everson, Fichardt, Trevino, Gillies. Did their scores also tend to go up in the second round?

   (c) Carry out a regression of score in the second round (dependent variable) on score in the first round (independent variable). Construct a scatterplot including the fitted line.

   (d) If golfers tend to score the same in each round, the fitted line should be approximately `round2=round1`. Carry out an hypothesis test to test whether the slope is actually less than one. What does this say about golfers who do poorly or well in the first round?

   (e) Suppose a golfer in round 1 scores one standard deviation above the mean round 1 score (i.e. $X = s_X + \overline{X}$). For this golfer, how many standard deviations will the predicted score on round 2 be above its mean? (Recall: $b_1 = rs_Y/s_X$.)

2. In the pharmaceutical industry, chemical assays must be validated before use in further studies or in manufacturing. This problem involves one component of the validation process: the analysis of a method's accuracy and precision. The accuracy refers to the bias of a method, while precision refers to the variability in a method.

   To validate a method, several samples of known analyte content are prepared at several concentrations. The concentration is recorded as percentage of label strength added. The samples are then analysed, with the method reporting back an estimated content. This is recorded as percentage of label strength found. In an ideal assay, the estimated content would equal the known content.

   The questions that the chemist wishes to ask are the following:
   1. What is the method's accuracy? Is it unbiased?
   2. Is the method precise? An error standard deviation in percent recovery (defined below) $\leq 2\%$ is acceptable; over 2% is unacceptable.

Data for 10 independent tests for an assay method can be found on the course web site. Percent Recovery was found by taking the percent of label strength found, dividing it by the percent of label strength added, and multiplying by 100. Percent recovery is preferred to percent label strength found since it is interpretable without regard to the percent label strength added.

We will study the simple linear regression model with percent label strength added as the independent variable and percent recovery as the dependent variable. The closer the intercept is to 100 and that the slope is to 0, the better in terms of assay performance. It can be safely assumed that the errors are approximately normally distributed with constant variance.

(a) We will accept the method if the confidence interval for the mean predicted value is within $\pm 2\%$ of label strength over the range of the data and if the estimated standard deviation of the errors is $\leq 2\%$. Is the method acceptable under these conditions? Make sure you clearly indicate what statistical methodology you are using.

(b) We can also consider the acceptability of the method by testing hypotheses for the slope and intercept. Conduct appropriate hypothesis tests for β_0 and β_1 and draw appropriate conclusions about the method acceptability.

(c) Note that, for completeness, the approach in part (b) should also include a hypothesis test about the error standard deviation. What would be appropriate null and alternative hypotheses for this test?

3. As part of a study to investigate reproductive strategies in plants, biologists recorded the time spent at sources of pollen and the proportions of pollen removed by bumblebee queens pollinating a species of lily. The data are available on the course web site.

(a) Fit the simple linear regression of proportion of pollen removed on duration of visit. What is the fitted equation? Considering the value of r^2 and an appropriate hypothesis test for the slope, what can you say about the linear relationship?

(b) For the regression in (a), make a scatterplot including the regression line. From this plot is there any indication of a problem with the fit of the regression model?

(c) Check the model assumptions with appropriate plots. What problems are evident?

(d) Try taking natural log transformations of the indepdendent variable or the dependent variable or both and re-fit the regression line and evaluate the model assumptions. Do any of these combinations of transformations help?

(e) Try fitting the regression line only for those visit times less than 31 seconds (i.e., excluding the two longest times). Does this give a better fit? How should your description of the fitted model change when describing this line?