Project: Eruptions of the "Old Faithful" Geyser

(Estimation of the time interval between eruptions)

A geyser is a hot spring that occasionally becomes unstable and erupts hot water and steam into the air. The “Old Faithful” geyser at Yellowstone National Park in Wyoming is probably the most famous geyser in the world. Visitors to the park try to arrive at the geyser site to see it erupts without waiting too long; the name of this geyser comes from the fact that eruptions follow a relatively stable pattern. The National Park Service erects a sign at the geyser site predicting when the next eruption will occur. Thus, it is of interest to understand and predict the interval time until the next eruption.

A sample of inter-eruption times was taken during August 1-8. The data set is saved in the file p_ysgr.dat which is an ASCII file, and the SPSS file is p_ysgr.sav. The first column of the data set is for date, the second column is eruption duration time (in minutes) and the third column is inter-eruption time (minutes). How can we help the tourists? Try some exploratory techniques such as histogram, boxplot, scatter plot, ..., to see if there is any pattern.

A characteristic of the geyser is the duration of the previous eruption. We could think of our data as pairs of the form (duration of eruption, time until next eruption). Do you see two subgroups in this data set? Do you see that a longer duration tends to be followed by a longer time interval until the next eruption? J. S. Rinehart, in a 1969 paper in the Journal of Geophysical Research, provides a mechanism for this pattern based on the temperature level of the water at the bottom of a geyser tube at the time the water at the top reaches the boiling temperature. That a shorter eruption would be followed by a shorter inter-eruption time (and a longer eruption would be followed by a longer inter-eruption time) is also consistent with Rinehart's model, since a short eruption is characterized by having more water at the bottom of the geyser being heated short of boiling temperature, and left in the tube. This water has been heated somewhat, however, so it takes less time for the next eruption to occur. A long eruption results in the tube being emptied, so the water must be heated from a colder temperature, which takes longer.

Answer the following questions:

1. Model the relation using a simple regression modeling technique and find the **equation of the regression line**, the **coefficient of determination** for fitting this model with the data and the mean square error of the model. Is this model significant, that is, is the slope significantly different from zero? Explain why by using scatter plot, statistics in the ANOVA table and the coefficient table.

2. Use linear regression technique to predict the average waiting time until next eruption after an eruption that last 4.3 minutes. (Use a 95% confidence interval for predicting the average waiting time until next eruption.)

3. If it is 2:00 p.m. right now, and you just observed an eruption last 4.3 minutes, with 95% confidence, before what time should you suggest the visitors to come to Old Faithful in order to see the next eruption? (Find the 95% confidence interval for the prediction.)

4. What are the assumptions behind the model you used for the predictions above? Do you think these assumptions were all satisfied? Verify your answer using output from statistical software.