LAB ASSIGNMENT 3

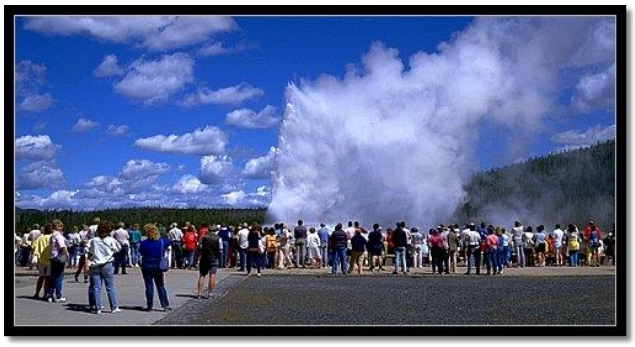
SIMPLE LINEAR REGRESSION

In the simple linear regression model, the mean of a response variable is a linear function of an explanatory variable. The model and associated inferential tools are valid as long as the assumptions of independence, normality, and constant variance are not violated.

In this assignment, you will use linear regression tools in SPSS to predict the time between eruptions from the duration and estimated height of the previous eruption for the Old Faithful Geyser in Yellowstone National Park, Wyoming. Two simple linear regression models will be examined and compared.

Old Faithful

Old Faithful Geyser, which throws about ten thousand gallons of water and steam up to one hundred seventy feet in the air, derives its name from the regularity of its eruptions. Old Faithful is not, in fact, the largest geyser in the park, but it has become a popular destination because it erupts more frequently than any of the other big geysers.



Every geyser consists of a reservoir for water storage, an opening through which steam and hot water can be released, a heat source, and underground channels to replenish the water supply after an eruption. The huge amounts of water which erupt from geysers, such as Old Faithful, deplete water from the geyser's two reservoirs. Since a long eruption depletes more water than a short eruption, it takes more time to refill the reservoir after a longer eruption. In other words, a longer time interval is needed until the next eruption. Thus, there is a relationship between the duration of an eruption and the waiting time until its next eruption. In the lab assignment, you will explore the relationship and you will use the relationship to predict times of eruption for a geyser. Accurate predictions of eruption times allow the Park Service to inform visitors of the approximate time of the next eruption.

The data from July 1995 eruptions are available in the SPSS file *lab3.sav* located on the *STAT 337* Laboratories web site at http://www.stat.ualberta.ca/statslabs/index.htm (click *Stat 337* link, and *Data* for *Lab 3*). The data are not to be printed in your submission. Note some values are missing in the data file.

Variable Name	**Description of Variable**
DAY	Day of July 1995 on which data were collected,
DURATION	Length of the previous eruption in minutes,
INTERVAL	Length of time (in minutes) from the end of the previous eruption to the beginning of the next eruption (waiting time until next eruption),
HEIGHT	Estimated height of previous eruption (in feet).

1. In this part, you will use the *Line Chart* feature in the *Graphs* menu to examine the pattern in the mean duration and the mean height of Old Faithful's eruptions over the 31-day period.

   (a) Obtain a line chart of the mean eruption height over the 31-day period. Paste the chart into your report. Is there any trend? What is the typical mean height? How does the mean height vary over the period? What factor might distort the height measurements?

   (b) Obtain a line chart of mean duration over the 31-day period. Paste the chart into your report. Is there any trend? What is a typical duration of an eruption?

2. Obtain a scatterplot of the interval between eruptions (or waiting time) and the duration of the previous eruption.

   (a) Paste the scatterplot with the title and the names of the axes (*Duration, Interval*) into your report.

   (b) Use the scatterplot to describe the distributions of interval between eruptions and duration of eruption. Specifically, answer the following four questions: What is the shape of each distribution, its center, and variability? How long do Old Faithful's eruptions last? How do intervals between eruptions vary? What does the distribution of the interval between eruptions tell you about the regularity of Old Faithful's eruptions?

   (c) Describe the overall pattern of the relationship involving the interval between eruptions and the duration of eruptions. Is the relationship reasonably strong or quite weak? Is it linear? Is the association positive, negative, or neither? Are there any outliers in the plot?

3. Now you will use the *Correlate* feature in the *Analyze* menu to quantify the strength of the relationship involving the interval between eruptions and the duration of eruptions.

   (a) Calculate the Pearson's correlation coefficient (Pearson) involving the interval between eruptions and the duration of eruptions.

   (b) Do the sign and magnitude of the coefficient confirm the conclusions you reached in part (c) of Question 2? Explain briefly.

4. Imagine that you have just arrived at Yellowstone National Park and just missed the most recent eruption of Old Faithful Geyser. How long would you have to wait until the next eruption of Old Faithful? In this part, you will use simple linear regression to make predictions about waiting time until the next eruption given the duration of the previous eruption.

   (a) Define a simple linear regression model using INTERVAL as the response variable and DURATION as the explanatory variable. State the model assumptions.

   Then use the *Regression* tool in SPSS to perform the regression and use the SPSS output to answer the following questions:

   (b) What is the equation of the least-squares regression line? What is the meaning of the slope of the regression line? Insert the line into the corresponding scatterplot and paste it into your report. Does the line provide a good fit? Are there any outliers?

   (c) What percent of the variation in the time between eruptions can be explained by the duration of the previous eruption? How useful is regression to make predictions about the time interval between eruptions?

(d) Based on the estimated regression line, predict the amount of time between the current eruption and the next eruption, given that the duration of the current eruption is 2 minutes. Obtain a 95% confidence interval for the mean interval between eruptions when the duration of the current eruption is 2 minutes. What is a 95% prediction interval for the interval between eruptions when duration is 2 minutes?

(e) Is linear regression on duration of any value in explaining interval between eruptions? Define the null and alternative hypotheses about the slope of the population regression line, obtain the value of the test statistic and its P-value from the output, and give your conclusion. What is the distribution of the test statistic under the null hypothesis? What are the sums of squares of residuals for the model with the intercept only (equal means model) and the regression model specified in part (a)?

(f) Obtain the plot of standardized residuals versus standardized predicted values for the regression model and paste it into your report. Describe the pattern of the residuals. Do the residuals appear to be randomly and uniformly scattered about a horizontal line at zero? What are the consequences for violating the assumption being checked here?

(g) Does the assumption of normality for the response variable seem appropriate? Answer the question by obtaining a normal P-P plot of the standardized residuals and paste the plot into your report. Comment about the plot.

(h) Is there any evidence that the relationship between INTERVAL and DURATION may have changed through time? Answer the question by obtaining a sequence plot of residuals vs. the time order of observations (*Sequence* feature in the *Graph* menu). Paste the chart into your report. Is there any trend? Comment briefly.

5. In the previous question, you have used linear regression to predict the time interval between eruptions using duration of the last eruption as the explanatory variable. However, height of eruption may also be useful in predicting the time interval between eruptions of a geyser. In particular, the product of height and duration provides some information about the amount of water that is expelled from the geyser during eruption.

(a) Obtain a new variable which is the product of duration and height. Then obtain a scatterplot of the length of time between eruptions vs. the product of height and duration. Paste the scatterplot with a title and the names of the axes (*Interval, Duration*Height*) into your report.

(b) Describe the overall pattern of the relationship involving the time interval between eruptions and the product of height and duration. Is the relationship reasonably strong or quite weak? Is it linear? Is the association positive, negative, or neither? Are there any outliers in the plot?

(c) Fit a new regression model to predict the time interval between eruptions by using the product of height and duration as the explanatory variable and report the estimated regression line. Moreover, report the R^2 value for the new model. In order to verify the regression assumptions in this case, examine the plot of standardized residuals and normal probability plot and comment. You are not required to paste the plots into your report. Is the model better than the model discussed in Question 3? Explain briefly.

LAB 3 ASSIGNMENT: MARKING SCHEMA

Question 1 (12)

(a) Line chart of mean height: 3 points
Trend, typical mean height, mean height variation, height distortion: 4 points

(b) Line chart of mean duration: 3 points
Trend, typical duration: 2 points

Question 2 (12)

(a) Scatterplot: 3 points

(b) Shape, center, and variability of interval between eruptions: 2 points
Shape, center, and variability of duration of eruptions: 2 points
Regularity of Old Faithful's eruptions: 2 points

(c) Pattern: 2 points
Outliers: 1 point

Question 3 (5)

(a) Pearson's correlation coefficient: 3 points

(b) Comparison with scatterplot in Question 2: 2 points

Question 4 (49)

(a) Simple regression model: 2 points
Model assumptions: 2 points

(b) Equation of the least-squares line: 3 points
Meaning of the slope: 1 point
Scatterplot with the least-squares line: 4 points
Quality of the fit: 1 point
Outliers: 1 point

(c) Percent of the variation: 2 points
Utility of the linear regression: 1 point

(d) Point estimate: 2 points
95% confidence interval: 2 points
95% prediction interval: 2 points

(e) Null and alternative hypotheses: 2 points
Test statistic: 1 point
P-value: 1 point
Conclusion: 1 point
Null distribution: 2 points
Sum of squares residuals: 2 points (1 point for each model)

(f) Plot of residuals: 3 points
Pattern: 2 points
Consequences of violating the equal variance assumption: 2 points

(g) Normal probability plot: 3 points
Comments on the plot: 2 points

(h) Plot of residuals versus time: 3 points
Trend: 2 points

Question 5 (12)

(a) Scatterplot: 3 points

(b) Pattern: 2 points
Outliers: 1 point

(c) Estimated regression line: 2 points
R^2 value: 2 points
Model comparison: 2 points

TOTAL = 90