Old Faithful Geyser

1. Suppose a linear regression was performed using *n* independent pairs of observations (x_i, y_i). Denote the i^{th} residual as e_i. Show that

$$\sum_{i=1}^{n} x_i e_i = 0$$

2. Suppose a linear regression was performed using *n* independent pairs of observations (x_i, y_i). Denote the i^{th} residual as e_i. Show that

$$\sum_{i=1}^{n} \hat{y}_i e_i = 0$$

3. Consider the regression through the origin model below. Suppose the model is fit to *n* independent pairs of observations (x_i, y_i).

$$y = \beta x + \epsilon; \text{ where } \epsilon \sim N(0, \sigma^2)$$

Find the maximum likelihood estimator for the slope.

4. Most of us have grown up to think of the geyser at Yellowstone National Park (Wyoming) named Old Faithful as just that--faithful and reliable. But actually it isn't very faithful at all, with times between eruptions varying between about 45 minutes and 90 minutes.

Thousands of visitors come to see Old Faithful every year. The Park Service would like to be able to inform visitors about the expected time for the next eruption (the interval between eruptions). That way, visitors can enjoy the park's other attractions and return to see the geyser erupt.

The park service has collect data to see if one could develop a strategy for helping inform the public. The data collected are the length of previous eruption and the interval between eruptions.

The data can be found here: http://www.stat.cmu.edu/~larry/all-of-statistics/=data/faithful.dat.

Use R to do all calculations and plots

a) Construct a histogram of the interval between eruptions. What would you conclude? What does the distribution of Intervals tell us about the "faithfulness" or "reliability" of Old Faithful?

b) Consider the standard descriptive statistics for the interval data (mean, standard deviation)? Would you use these to describe the data (why or why not?)

c) Divide the data into two parts: intervals that had a previous eruption of 3 or less minutes and intervals that had a previous eruption that is larger than 3 minutes. Construct histograms of intervals between eruptions for both sets of data. Use the 68-95-99.7 empirical rule (from stat 270) to construct a rule based on the length of the previous eruption (3 minutes or less or greater than 3 minutes) to estimate the interval between eruptions.

d) Construct a scatter-plot to investigate the relationship between the interval between eruptions and the length of the previous eruption. What could you conclude? Could linear regression be used here?

e) Use R to perform a simple linear regression with the interval time (i.e., the number of minutes until the next eruption) and the length of the previous eruption. What do you conclude? Add the regression line to the scatterplot in d. (hint: the *abline* command in R will do this for you).

f) Suppose the length of the previous eruption was 2 minutes. How long would you expect to wait until the next eruption?