Tae Coding
Introduction to Data Science: CS61
Summer 2018
Class Exercise#6

Date Given: June 28, 2018 Due Date:

==

Old Faithful is a geyser located in Yellowstone National Park in Wyoming, US. It is a highly predictable geothermal feature and has erupted every 44 to 125 minutes since 2000.

The 'oldfaithful.csv' file contains waiting time between eruptions and the duration of the eruption. This file contains 272 observations on 2 variables.

Variable Name	Type	Semantics
Time Eruption	Numeric	Eruption duration time in mins
Time Waiting	Numeric	Waiting time between eruption in mins

Build your regression model.

- Use 'Time Eruption' as the predictor variable
- Use 'Time Waiting' as the response variable

Use Python/Scikit-Learn package for this homework assignment.

1. Compute the regression equation and the R-Square metrics of your regression model.

2. Split the dataset into training and testing set with 70/30 ratio randomly. Build a regression model using training data set. Compute the predicted value of 'Time Waiting' variable of training and testing data set using the model built. The Root Mean Square Error (RMSE) and the Mean Square Error (MSE) metrics are defined as follows.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}[f(x_i) - y_i)]^2} \qquad MSE = \frac{1}{N}\sum_{i=1}^{N}[f(x_i) - y_i)]^2$$

Here $f(x_i)$ is the computed value and y_i is the true (observed) value.

Compute the training error (RMSE or MSE) and the testing error (RMSE or MSE). Which one is greater – RMSE(Training) or RMSE (Testing)?