Stevens Institute of Technology

SYS-601 Homework #9

Due Apr. 16 2018

Submit the following using the online submission system: 1) Completed assignment cover sheet, 2) Written responses in PDF format, 3) All saved models (e.g. .xlsx or .py files).

9.1 Old Faithful [10 points]

The attached file `faithful.csv` contains 272 observations of Old Faithful geyser eruptions. The data include:

- The duration of an eruption (minutes)
- The waiting time between eruptions (minutes)

This problem seeks to build a regression model to predict the waiting time between eruptions (y) as a function of the duration of an eruption (x).

(a) 2 pts Create a scatter plot with waiting time on the y-axis and duration on the x-axis.

(b) 1 pt Compute the Pearson correlation coefficient r.

(c) 3 pts Assuming the regression model $\hat{y} = \beta_0 + \beta_1 x$, compute the following:

(i) β_0 coefficient

(ii) β_1 coefficient

(iii) Coefficient of determination r^2

(d) 1 pts Overlay a line plot of the regression model on the scatter plot from (a).

(e) 2 pt Analyze the health of the residuals. Do the residuals appear to be:

(i) Homoscedastic?

(ii) Independent?

(iii) Normally distributed?

(f) 1 pt Predict the waiting time until the next eruption after one 4.0 minutes in duration.

9.2 Hydrocarbon Emissions [10 points]

(Problem based on material from Dr. Roy Welsch)

Refueling automobiles can create hydrocarbon emissions as the dispensed gasoline displaces vapor-rich air in the fuel tank. The attached file `gasvapor.csv` contains 100 samples gathered during experimental refueling trials. Columns measure the following factors:

- x_1: Temperature of gasoline in tank (°F)
- x_2: Vapor pressure of gasoline in tank (psi)
- x_3: Temperature of dispensed gasoline (°F)
- x_4: Vapor pressure of dispensed gasoline (psi)
- y: Mass of hydrocarbons emitted during refueling (g)

The goal of this problem is to develop a predictive model to anticipate the mass of hydrocarbons emitted during refueling.

Randomly select 70 samples as your *training set* to build the regression model. Use the remaining 30 samples as your *testing set* to evaluate accuracy of predictions. This approach helps avoid over-fitting the regression model to the particular samples selected.

Using at **least 2 iterations** of a stepwise procedure (i.e. forward selection or backward elimination), develop a multiple regression model to predict hydrocarbon emissions using the *training* data set. Consider exploring interaction factors and mathematical transformations as desired. **Report results of each** incremental regression analysis including:

(a) 2 PTS Regression equation including coefficient values.

(b) 2 PTS Coefficient t statistics and p-values.

(c) 2 PTS Coefficient of multiple determination R^2.

(d) 2 PTS Root mean square error (RMSE) of the *training* data set.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{70} (y_i - \hat{y}_i)^2}{70}}$$

(e) 2 PTS Root mean square error (RMSE) of the *testing* data set.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{30} (y_i - \hat{y}_i)^2}{30}}$$

Clearly mark your **final** model believed to make the best predictions. Anyone with a unique model yielding better RMSE results than the solution from a *validation set* of 25 additional samples will receive a 2 bonus points added to their assignment grade.