Homework 9

9.1 Old Faithful

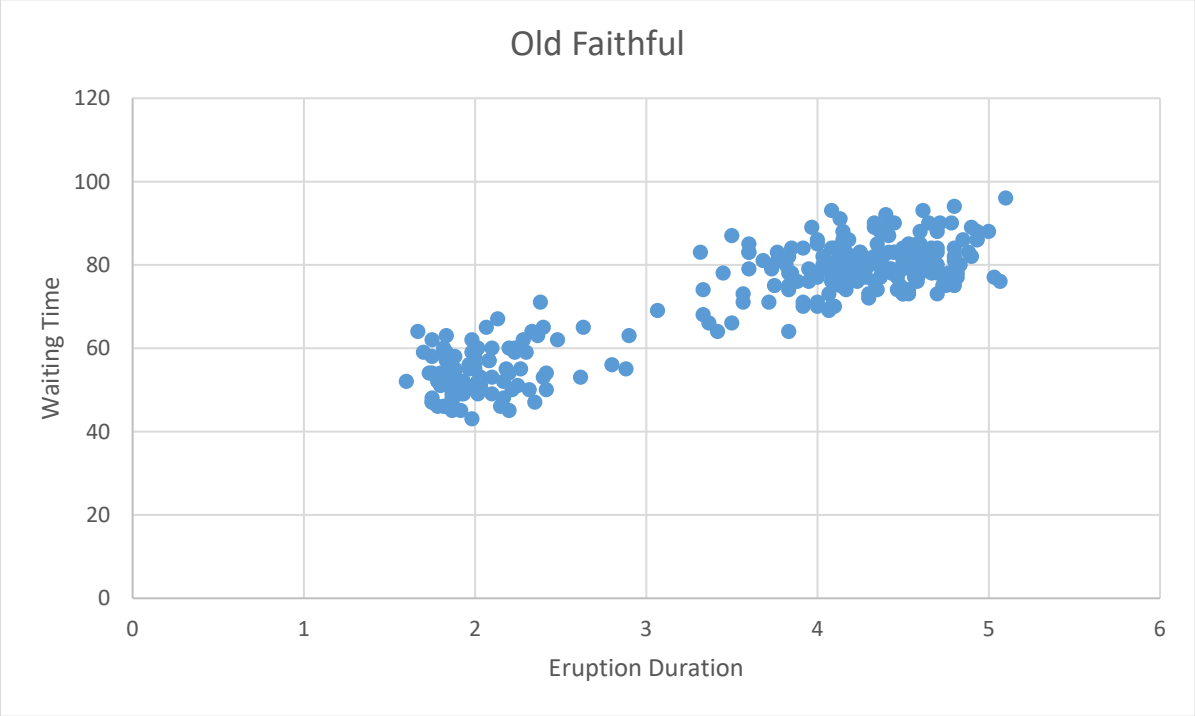
The attached file faithful.csv contains 272 observations of Old Faithful geyser eruptions. The data included:

- The duration of 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) Create a scatter plot with waiting time on the y-axis and duration on the x-axis.

Solution:



(b) Compute the Pearson correlation coefficient r.

Solution:

Array 1 = Eruptions
Array 2 = Waiting
Correlation coefficient = 0.901

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

Assuming eruptions as output and waiting as input.

(i) B_0 coefficient

Solution:

$B_0 = 33.47$

(ii) B_1 coefficient

Solution:

$B_0 = 10.72$

(iii) Coefficient of determination r^2

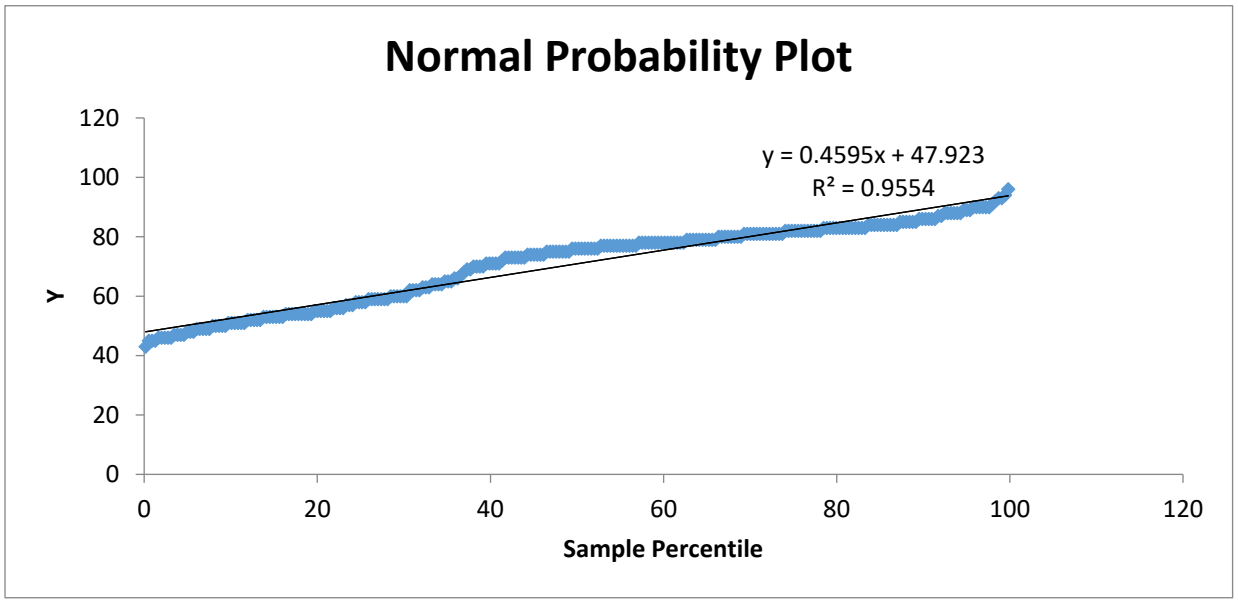
Solution:

$r^2 = (r)^2$

$\therefore r^2 = 0.811$

(d) Overlay a line plot of the regression model on the scatter plot form (a).

Solution:



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

(i) Homoscedastic?

(ii) Independent?

(iii) Normally distributed?

Solution:

The patter of errors seems to be around the 0 line. Therefore, residuals seem to be homoscedastic

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

Solution:

$\hat{y} = \beta 0 + \beta 1x$

$= 33.47 + 10.72(4)$

≈ 76.35 minutes

9.2 Hydrocarbon Emissions

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 overfitting 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) Regression equation including coefficient values.

Solution:

$$\hat{y} = 0.51082 + 0.19003x_3 + 4.58815x_4$$

(b) Coefficient t statistics and p-values.

Solution:

Model	*t Stat*	*P-value*
Intercept	0.356	0.72266
Gasoline Temp (x_3)	3.348	0.00134
Gasoline Pressure (x_4)	6.177	4.35×10^{-8}

(c) Coefficient of multiple determination R^2.

Solution:

Regression Statistics	
Multiple R Square	0.8822
Adjusted R Square	0.8787

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

$$RMSE = \sqrt{\frac{\sum_{i=1}^{70}(y_1 - \hat{y}_1)^2}{70}}$$

Solution:

I have taken help of R for this solution. Please refer the additional excel sheet.

RMSE ≈ 2.79

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

$$RMSE = \sqrt{\frac{\sum_{i=1}^{30}(y_1 - \hat{y}_1)^2}{30}}$$

Solution:

I have taken help of R for this solution. Please refer the additional excel sheet.

RMSE ≈ 3.37

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 sample will receive a 2 bonus points added to the assignment grade.

Following images are taken from Rattle (Rstudio)

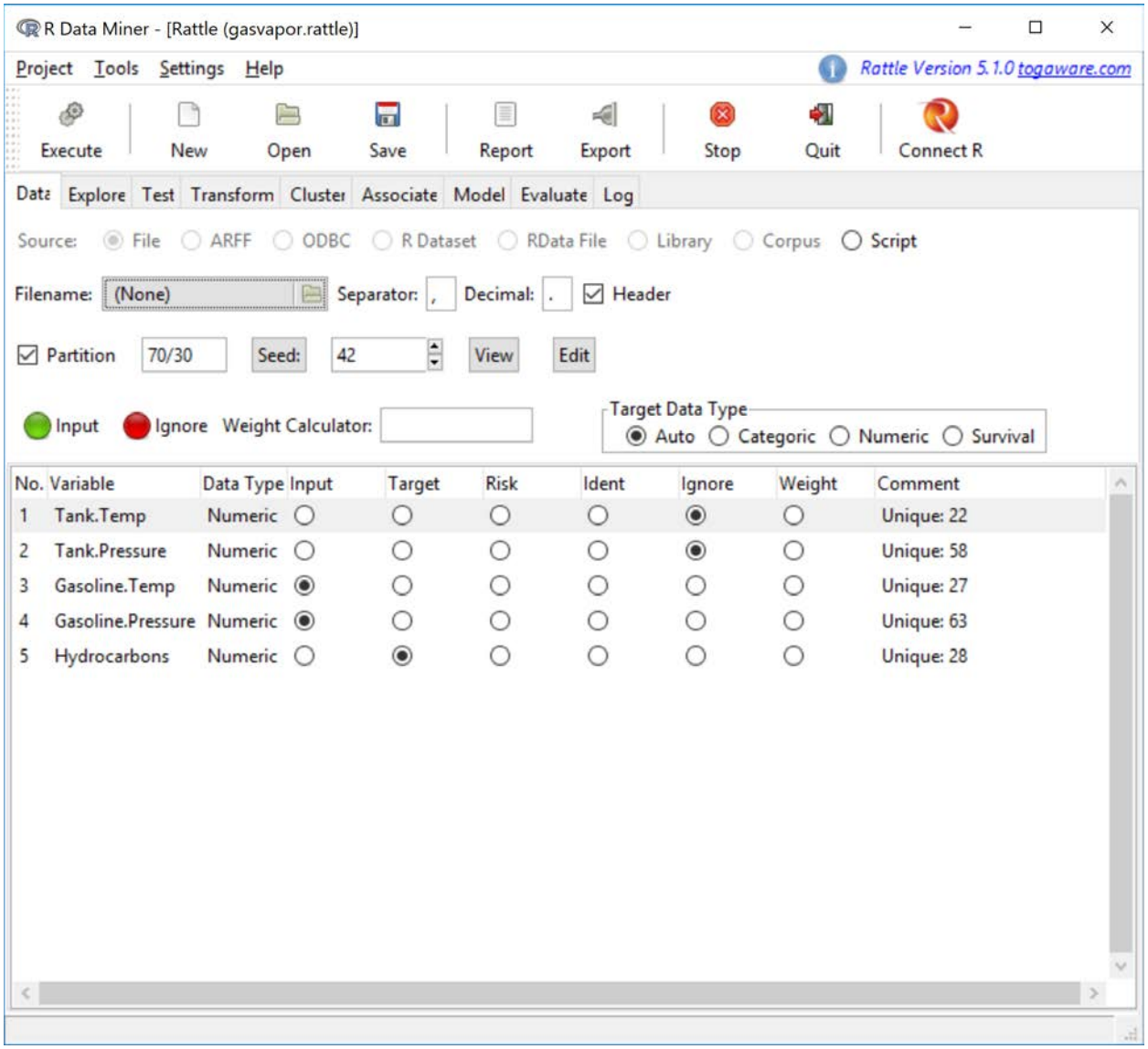


Image 1 Data Setup after iterations

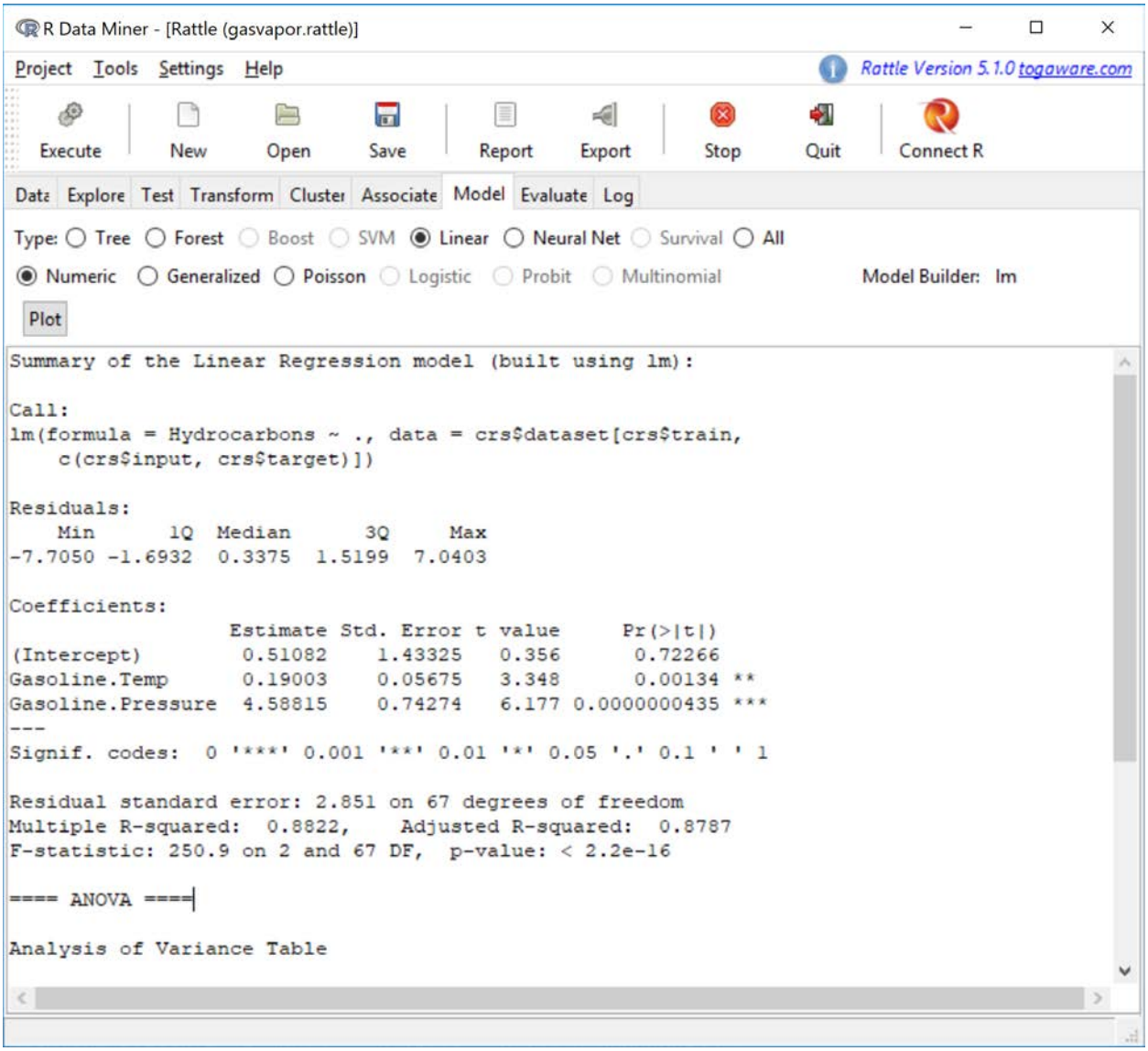


Image 2 Regression Modelling (Part 1)

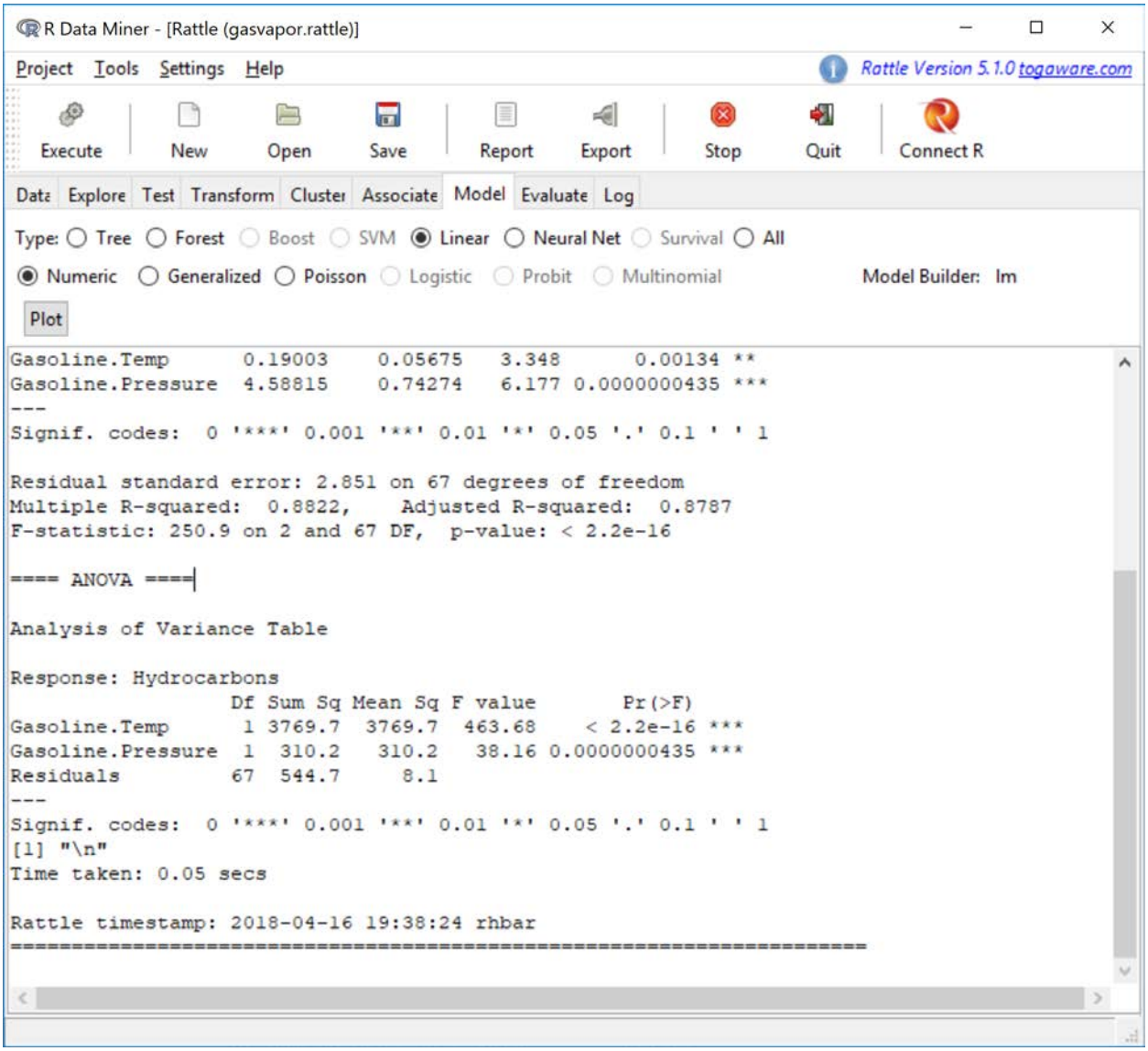


Image 3 Regression Model (Part 2)

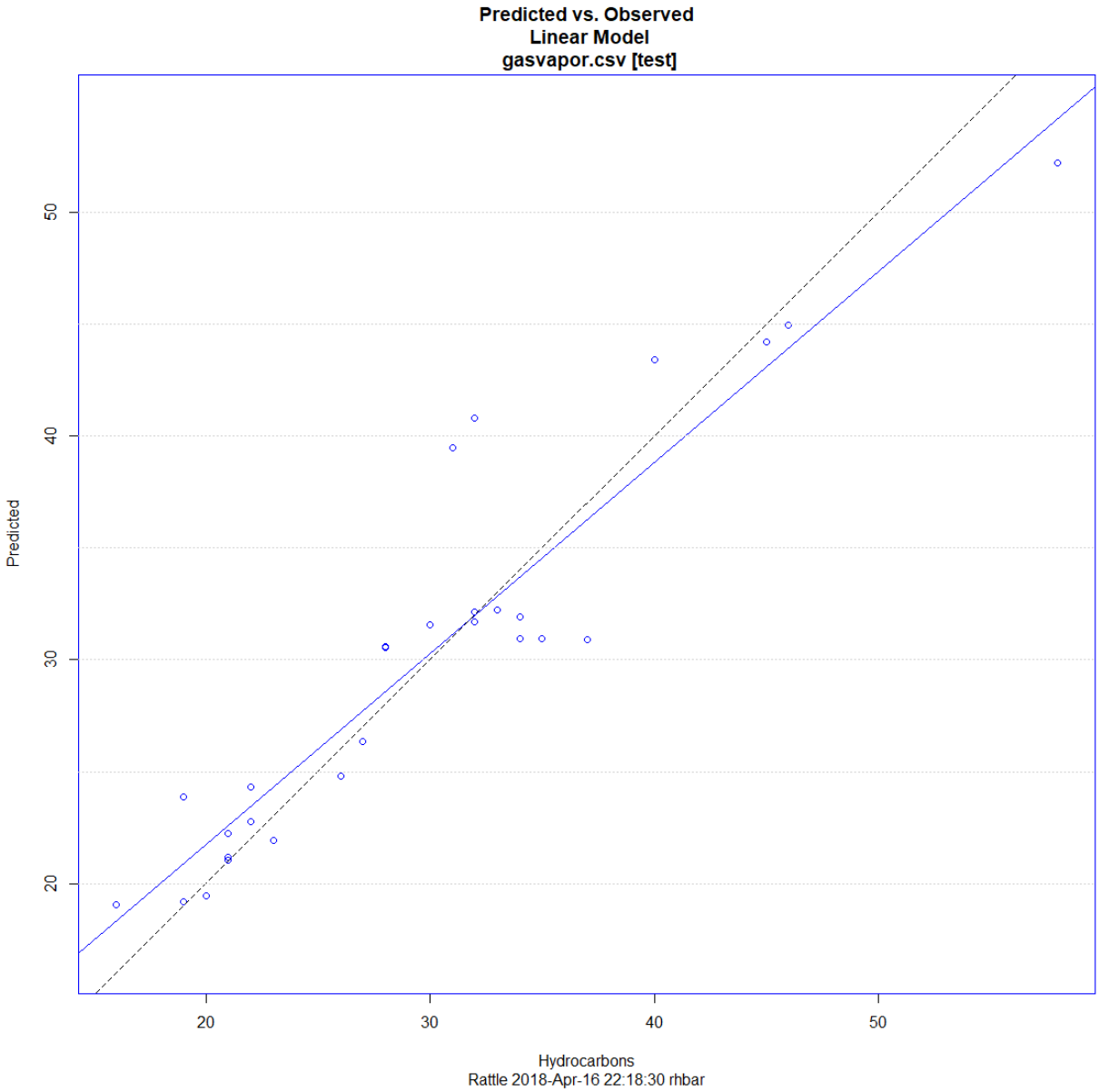
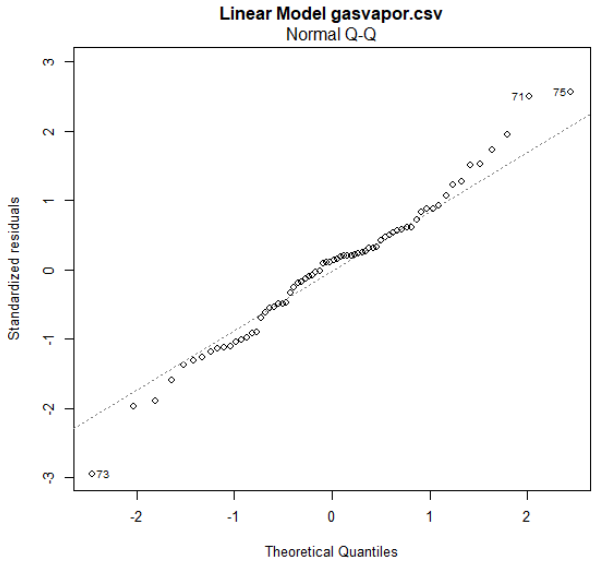


Image 4 Training vs Testing data lines – ***Dotted line = testing & Solid line = training***


Linear Model gasvapor.csv
Normal Q-Q
Standardized residuals
Theoretical Quantiles
71
75
73
3
2
1
0
-1
-2
-3
-2
-1
0
1
2