SYS-601 Homework Cover Sheet

Date: HW #:

Author:

Collaborators:

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:
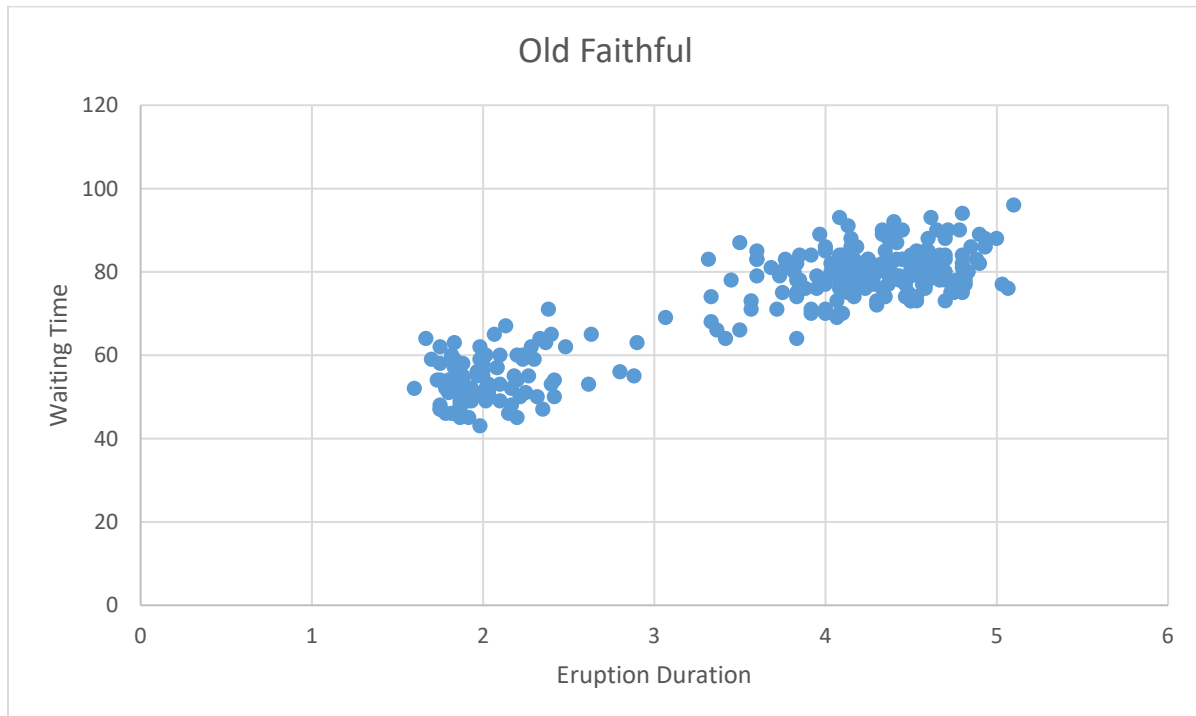
(b) Compute the Pearson correlation coefficient r.

Solution:

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

(c) Assuming the regression model ŷ = β_0 + β_1x, 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:

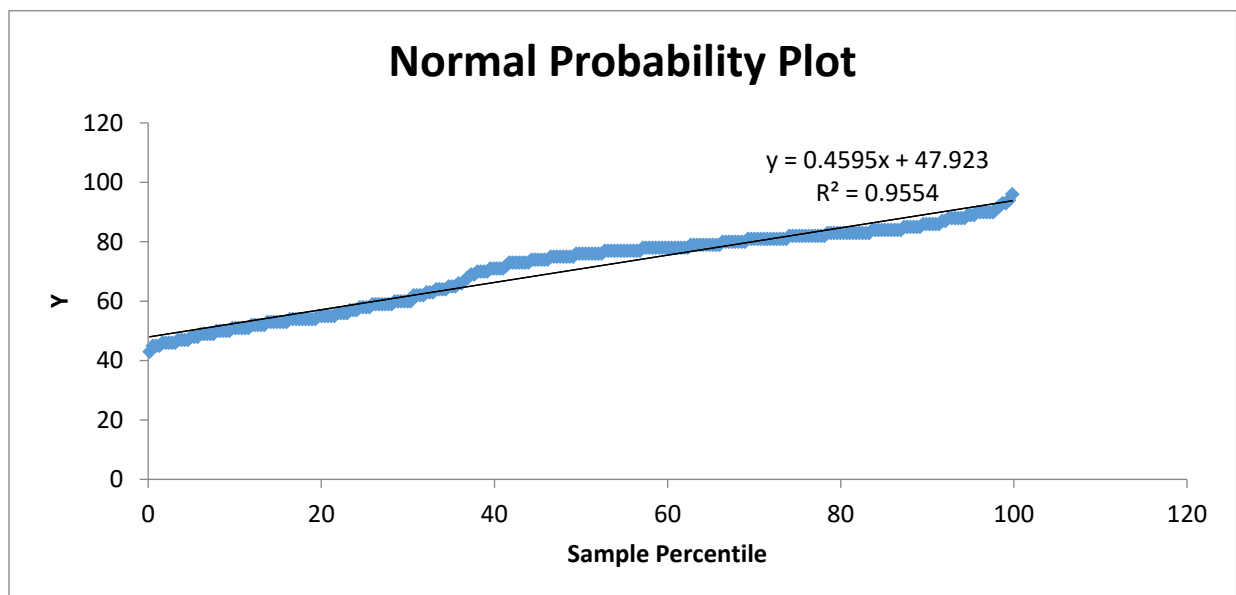B_0 = 10.72

   (iii) Coefficient of determination r^2

Solution:

$r^2 = (\,r\,)^2$

∴ r^2 = 0.811

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

Solution:

Normal Probability Plot

y = 0.4595x + 47.923
R^2 = 0.9554

(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:

$\hat{y} = \beta 0 + \beta 1 x$
$= 33.47 + 10.72(4)$
≈ 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

```
R R Data Miner - [Rattle (gasvapor.rattle)]                                    —    □    ×

Project  Tools  Settings  Help                        (i)  Rattle Version 5.1.0 togaware.com

    Execute    New    Open    Save    Report   Export    Stop    Quit    Connect R

Data  Explore  Test  Transform  Cluster  Associate  Model  Evaluate  Log

Type: ○ Tree  ○ Forest  ○ Boost  ○ SVM  ◉ Linear  ○ Neural Net  ○ Survival  ○ All

◉ Numeric  ○ Generalized  ○ Poisson  ○ Logistic  ○ Probit  ○ Multinomial      Model Builder:  lm

  Plot

Summary of the Linear Regression model (built using lm):

Call:
lm(formula = Hydrocarbons ~ ., data = crs$dataset[crs$train,
    c(crs$input, crs$target)])

Residuals:
    Min      1Q  Median      3Q     Max
-7.7050 -1.6932  0.3375  1.5199  7.0403

Coefficients:
                  Estimate Std. Error t value     Pr(>|t|)
(Intercept)        0.51082    1.43325   0.356      0.72266
Gasoline.Temp      0.19003    0.05675   3.348      0.00134 **
Gasoline.Pressure  4.58815    0.74274   6.177 0.0000000435 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.851 on 67 degrees of freedom
Multiple R-squared:  0.8822,    Adjusted R-squared:  0.8787
F-statistic: 250.9 on 2 and 67 DF,  p-value: < 2.2e-16

==== ANOVA ====|

Analysis of Variance Table
```

Image 2 Regression Modelling (Part 1)

R Data Miner - [Rattle (gasvapor.rattle)]

Project Tools Settings Help

Rattle Version 5.1.0 togaware.com

| Execute | New | Open | Save | Report | Export | Stop | Quit | Connect R |

Data Explore Test Transform Cluster Associate Model Evaluate Log

Type: ○ Tree ○ Forest ○ Boost ○ SVM ● Linear ○ Neural Net ○ Survival ○ All

● Numeric ○ Generalized ○ Poisson ○ Logistic ○ Probit ○ Multinomial   Model Builder: lm

Plot

```
Gasoline.Temp       0.19003    0.05675   3.348      0.00134 **
Gasoline.Pressure   4.58815    0.74274   6.177 0.0000000435 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.851 on 67 degrees of freedom
Multiple R-squared:  0.8822,    Adjusted R-squared:  0.8787
F-statistic: 250.9 on 2 and 67 DF,  p-value: < 2.2e-16

==== ANOVA ====

Analysis of Variance Table

Response: Hydrocarbons
                  Df Sum Sq Mean Sq F value       Pr(>F)
Gasoline.Temp      1 3769.7  3769.7  463.68    < 2.2e-16 ***
Gasoline.Pressure  1  310.2   310.2   38.16 0.0000000435 ***
Residuals         67  544.7     8.1
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
[1] "\n"
Time taken: 0.05 secs

Rattle timestamp: 2018-04-16 19:38:24 rhbar
==============================================================================
```

Image 3 Regression Model (Part 2)

SYS 601 – PROBABILITY AND STATISTICS FOR SYSTEMS ENGINEERING

Predicted vs. Observed
Linear Model
gasvapor.csv [test]

Hydrocarbons
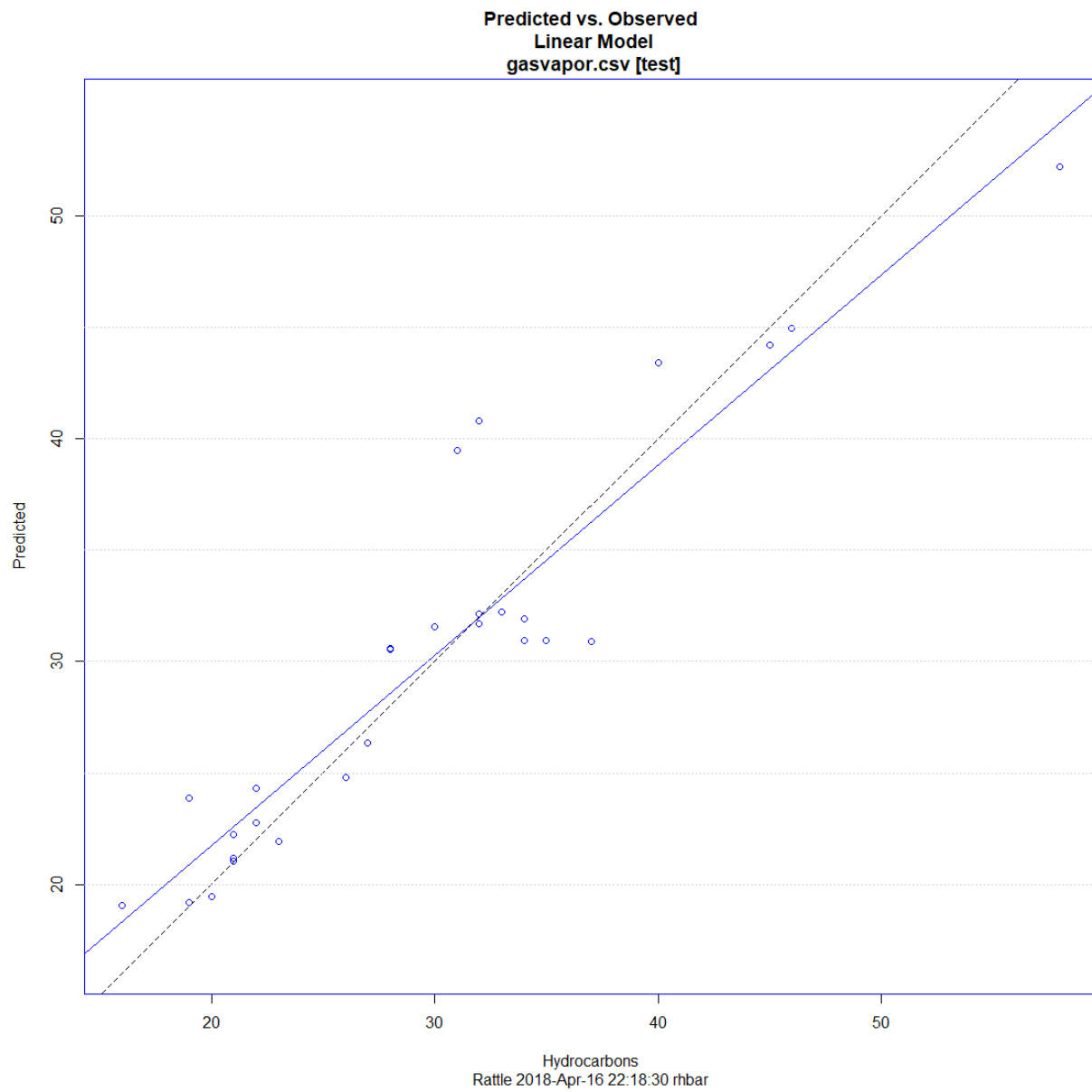Rattle 2018-Apr-16 22:18:30 rhbar

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

SYS 601 – PROBABILITY AND STATISTICS FOR SYSTEMS ENGINEERING

Linear Model gasvapor.csv
Normal Q-Q