Linear Regression on the Effect of Pollution Controls

Mark Berman, Ariel Lee, Christina Li, Justin Luong, Sophia Tierney, Ryan Truong, Jenna Zarbis

3/16/2021

1 Abstract

In this report we will use general linear regression and ANOVA table evaluations to predict the amount of escaping hydrocarbons based on the tank temperature, petrol temperature, initial tank pressure, and petrol pressure. Before conducting a linear regression evaluation, we used R to conduct several tests (such as the Kolmogorov-Smirnov test for normality, Breusch-Pagan Test for constancy of variance, and ANOVA for goodness of fit check) to ensure our dataset satisfied all required conditions for the model to work before extrapolating the amount of escaping hydrocarbons using the aforementioned predictor variables. Our results yielded an R^2 value of 0.9261 indicating a strong linear regression between the predictor variables and the dependent variables. The first regression summary yielded a estimated regression function is $\hat{Y}$ = 1.01502 - 0.02861$X1$ + *0.21582*$X2$ - 4.32005$X3$ + *8.97489*$X4$. An F-Test for regression relation yielded an overall test statistic of 84.54 supporting the alternate hypothesis that not all of the βk are equal to 0. but does support at the alpha=5% level that the amount of escaping hydrocarbons is related to tank temperature, petrol temperature and petrol pressure. The Breusch-Pagan Test for constancy of error variance yielded that the error variance is not constant when we include all four predictor variables.

When graphing the interaction plots, we see that interactivity exists because the lines are not parallel. After generating different interaction models and calculating Mallow's Cp for each of them we found that using three predictor variables and excluding initial tank pressure (b3) yielded the lowest Cp value of 3.10. The new regression summary with initial tank pressure excluded yielded a regression function of Yi = 0.39742 - 0.11526 X1 + 0.27104X2 + 5.14570X3. For the new model, the Kolmogorov-Smirnov test for normality passed with a test statistic of 0.4337. The new F-test for regression relation yields a test statistic of 107.00 that indicates not all βk equal zero but does support at the alpha=5% level that the amount of escaping hydrocarbons is related to tank temperature, petrol temperature and petrol pressure. The new Breusch-Pagan Test for constancy of error variance yielded that the error variance is constant for the reduced model.

2 Introduction

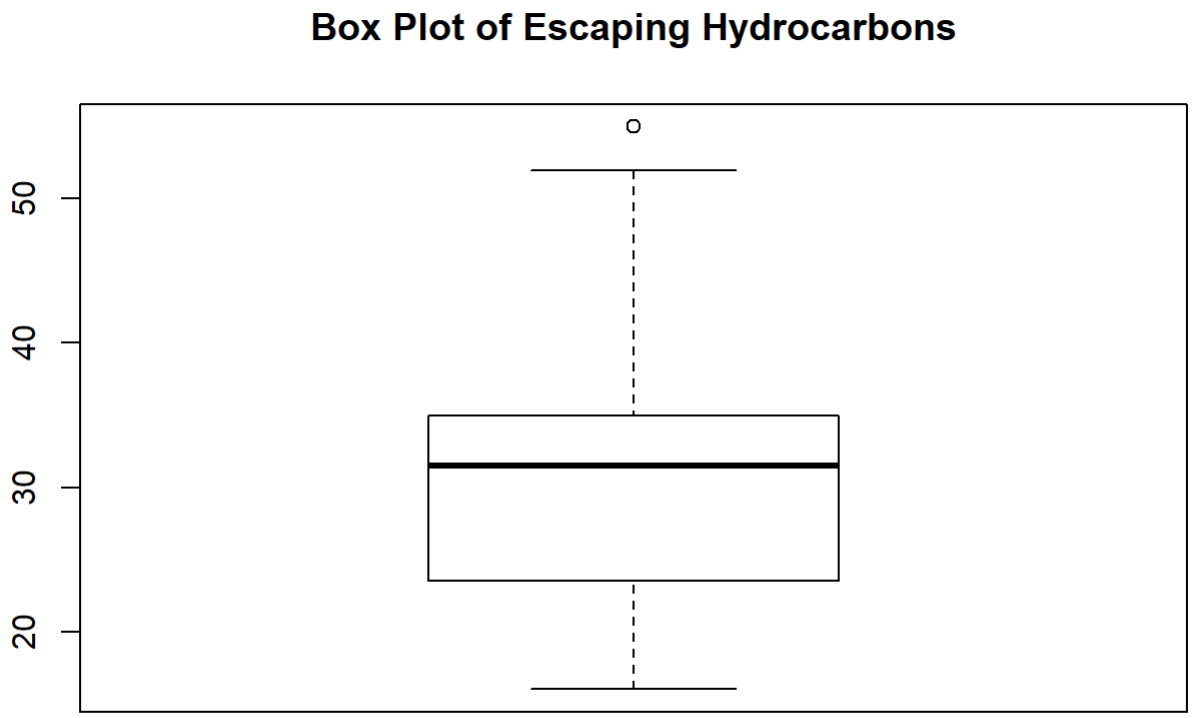
In this report, we will analyze tank temperature, petrol temperature, initial tank pressure, and petrol pressure to predict the amount of escaping hydrocarbons. These 4 predictors are combined to form a model that uses 32 rows of data across 6 columns. The purpose of the study is to determine the pollution control effectiveness when petrol is being pumped into tanks. This is important because without pollution control, small amounts of unburned fuel, either in liquid or vapor form, will escape, causing a public health concern due to unhealthy chemicals in the petrol (Hilpert, 2015). More specifically, fuel can enter groundwater or surface water via rain which is used to supply water to citizens (Hilpert, 2015). In vapor form, breathing in benzene can lead to leukemia (Hilpert, 2015). The US heavily regulates benzene under the Clean Air Act, a main substance in petrol, due to its carcinogenic impacts (Hilpert, 2015).

3 Methods

This study will be conducted via R where the data will go through a series of testing. These include the use of a General Linear Regression and ANOVA table evaluations to determine significance and level of causality versus correlation. Upon creating the regression, it was determined that two of the four predictors were not statistically significant. The multiple linear regression model was checked for assumption violations such as linearity, constancy of variance with the error term, uncorrelated errors, and normal distribution of errors. The Brush-pagan test found an inconsistency in variance. With this in mind, a high level of correlation was determined and interaction plots were created, finding issues with the beta 3 term. Given this, the model was selected using a Cp statistic with close attention to the change in R^2. The new model had 3 predictors, dropping initial tank pressure. This eliminated a statistically insignificant variable. Additionally, the linear model was transformed with a log in order to improve correlation and helped to create a constancy of variance. In order to certify these results, this study used Kolmogorov-Smirnov test to determine normality, ANOVA F-test for regression relation, and confidence intervals for all beta terms to make sure that 0 did not fall in those ranges, making the data statistically significant. A number of graphs will be presented which include boxplot, scatterplot, qqplot, interaction plots, Cp plots, R^2 vs predictors, and residual plots.

4 Data Summary and Exploratory Data Analysis

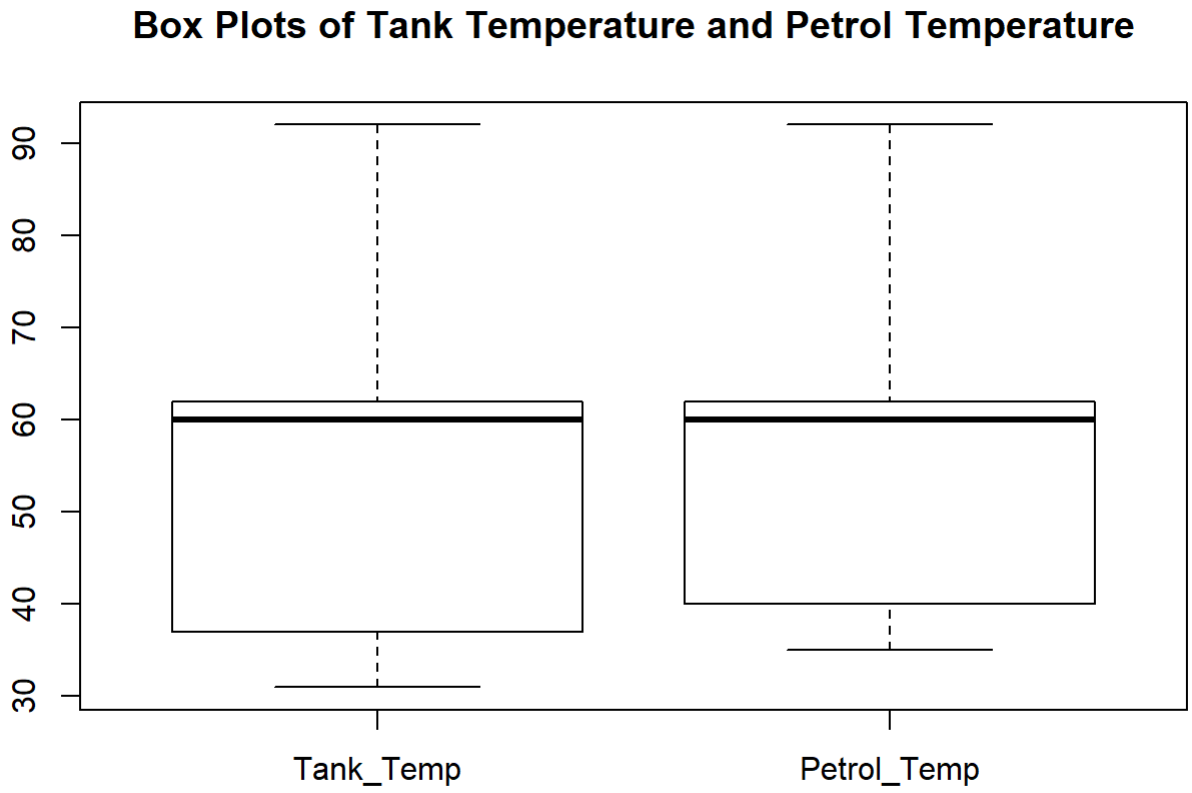
4.1 Box Plots



Five Number Summary For Escaping Hydrocarbons:

```
##    0%   25%   50%   75%  100%
## 16.00 23.75 31.50 34.50 55.00
```

```
boxplot(pollution_data[2:3], main = "Box Plots of Tank Temperature and Petrol Temperature")
```



Five Number Summary For Tank Temperature:

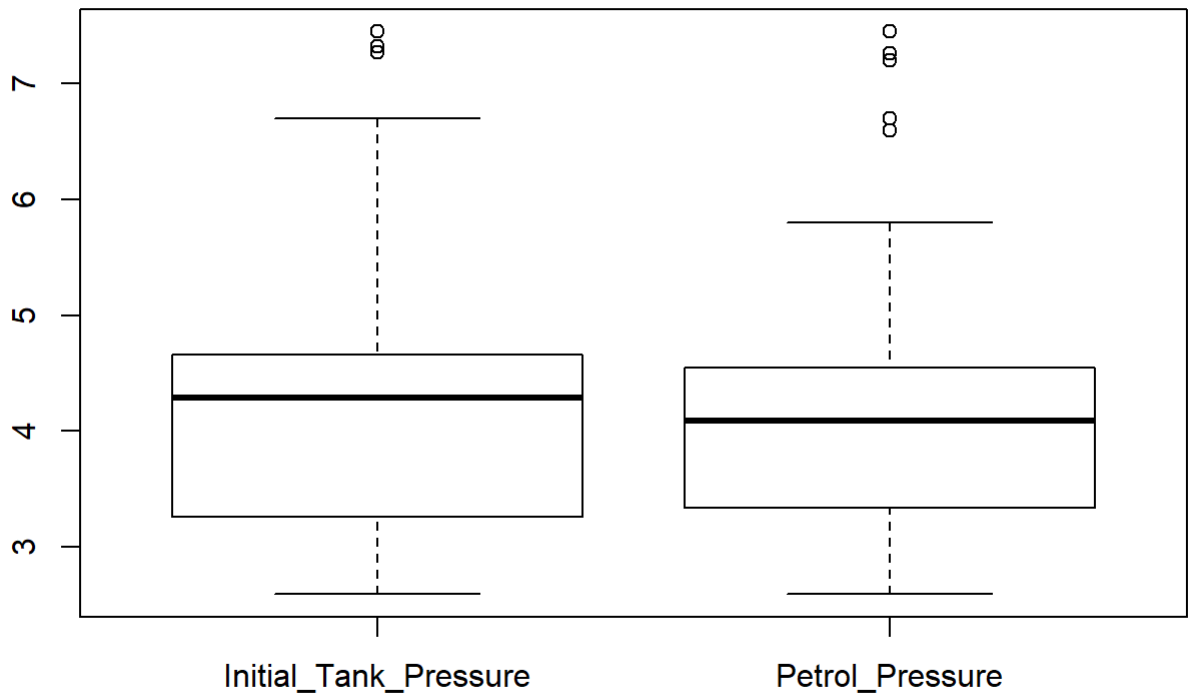
```
quantile(pollution_data$Tank_Temp)
```

```
##   0%  25%  50%  75% 100%
##   31   37   60   62   92
```

Five Number Summary For Petrol Temperature:

```
##   0%  25%  50%  75% 100%
##   35   41   60   62   92
```

```
boxplot(pollution_data[4:5], main = "Box Plots of Initial Tank Pressure and Petrol Pressure")
```



Five Number Summary For Initial Tank Pressure:

```
quantile(pollution_data$Initial_Tank_Pressure)
```

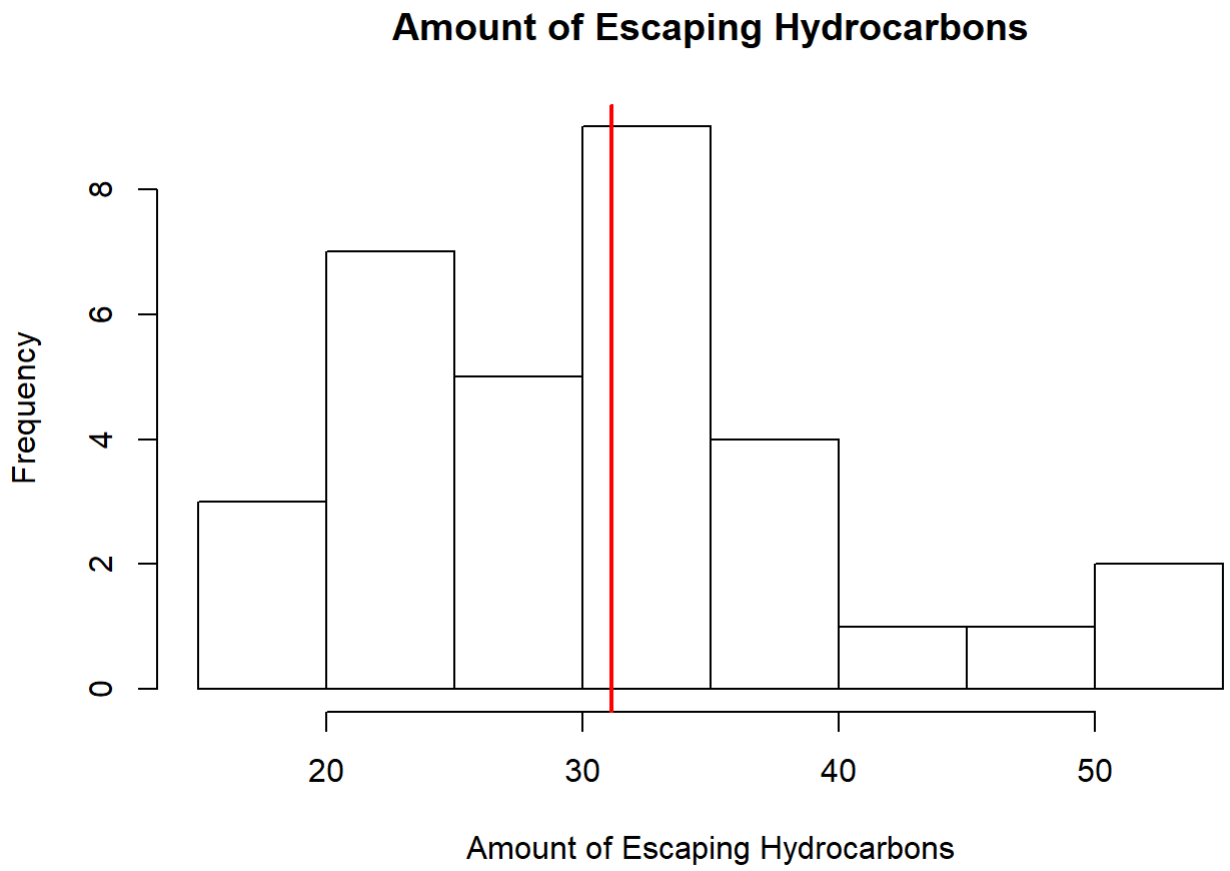
```
##    0%   25%   50%   75%  100%
## 2.590 3.290 4.285 4.630 7.450
```

Five Number Summary For Petrol Pressure:

```
quantile(pollution_data$Petrol_Pressure)
```

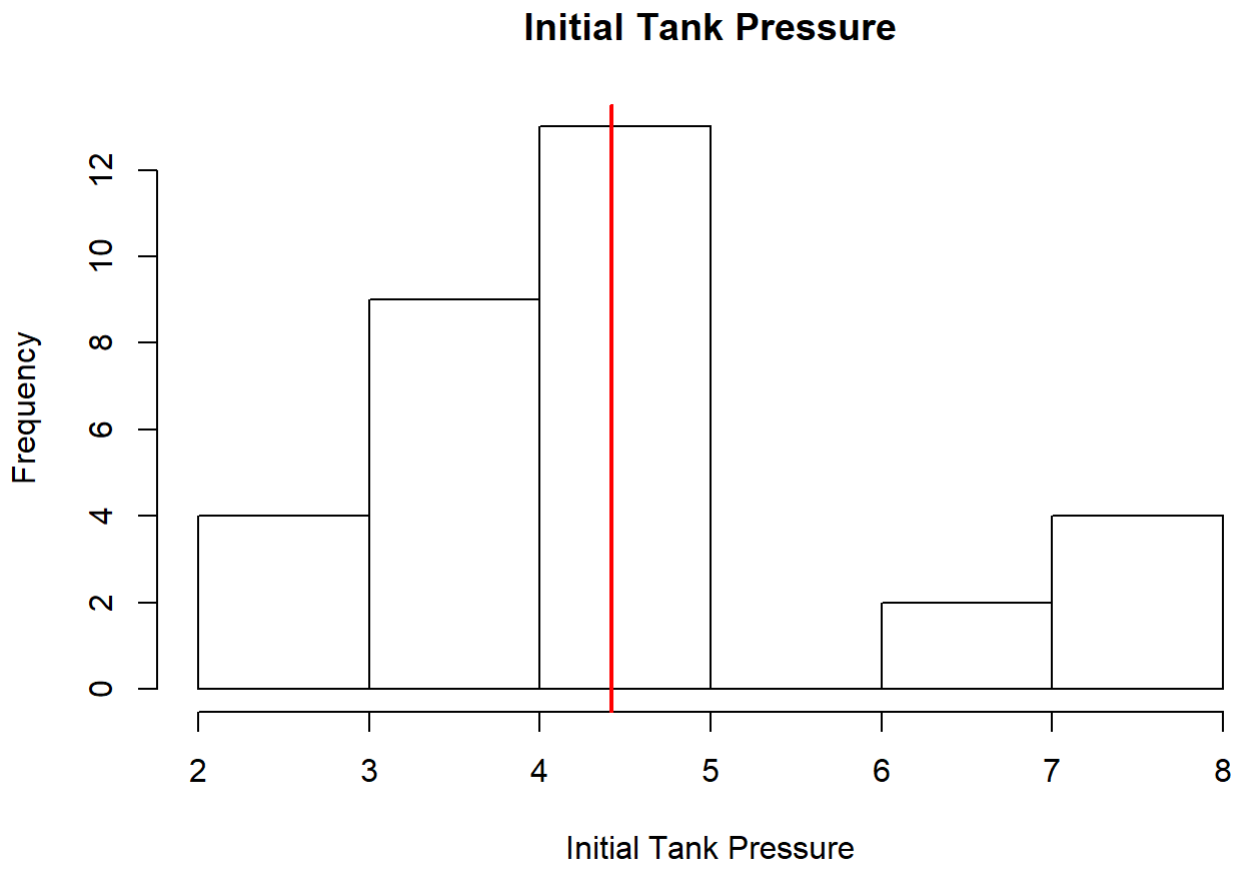
```
##     0%    25%    50%    75%   100%
## 2.5900 3.3725 4.0900 4.5400 7.4500
```

4.2 Histograms



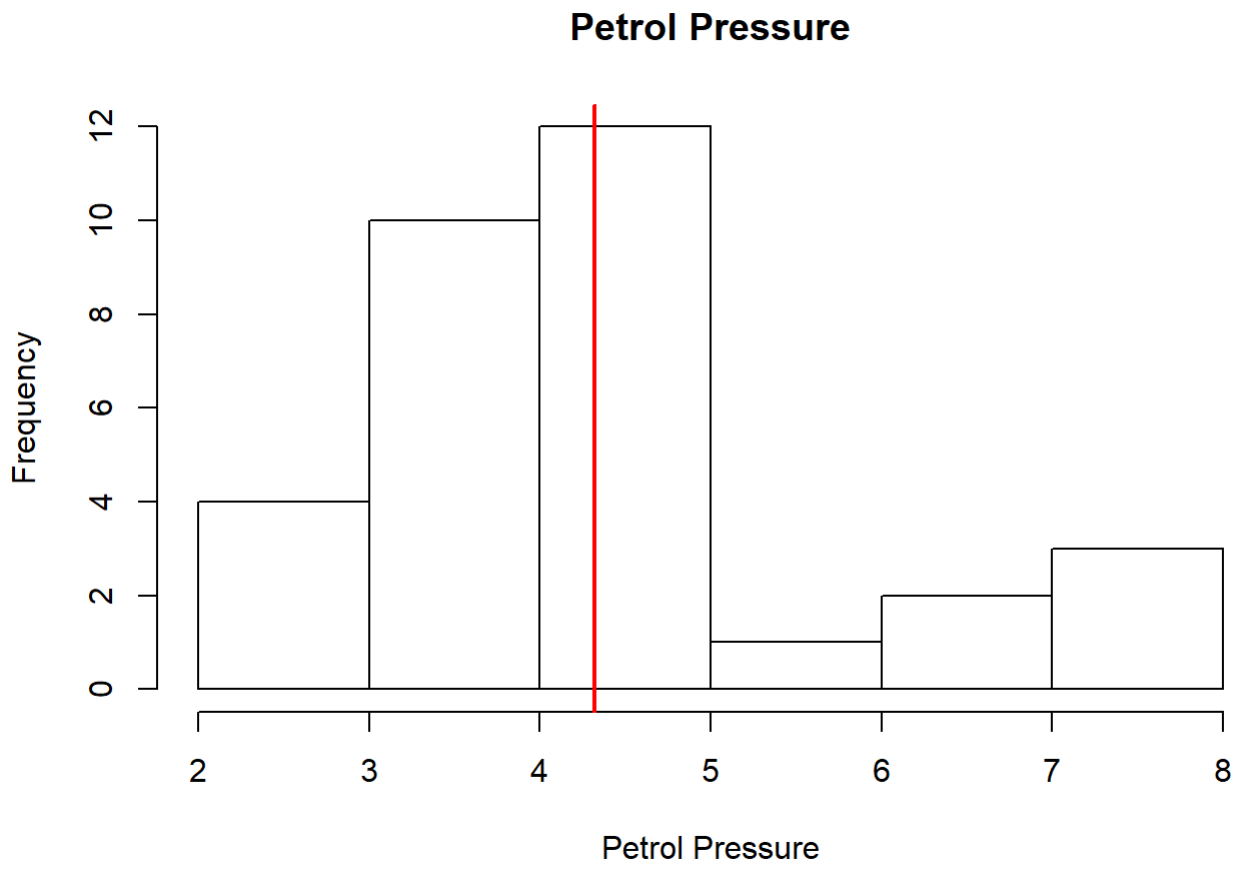
Mean for Escaping Hydrocarbons:

```
## [1] 31.125
```



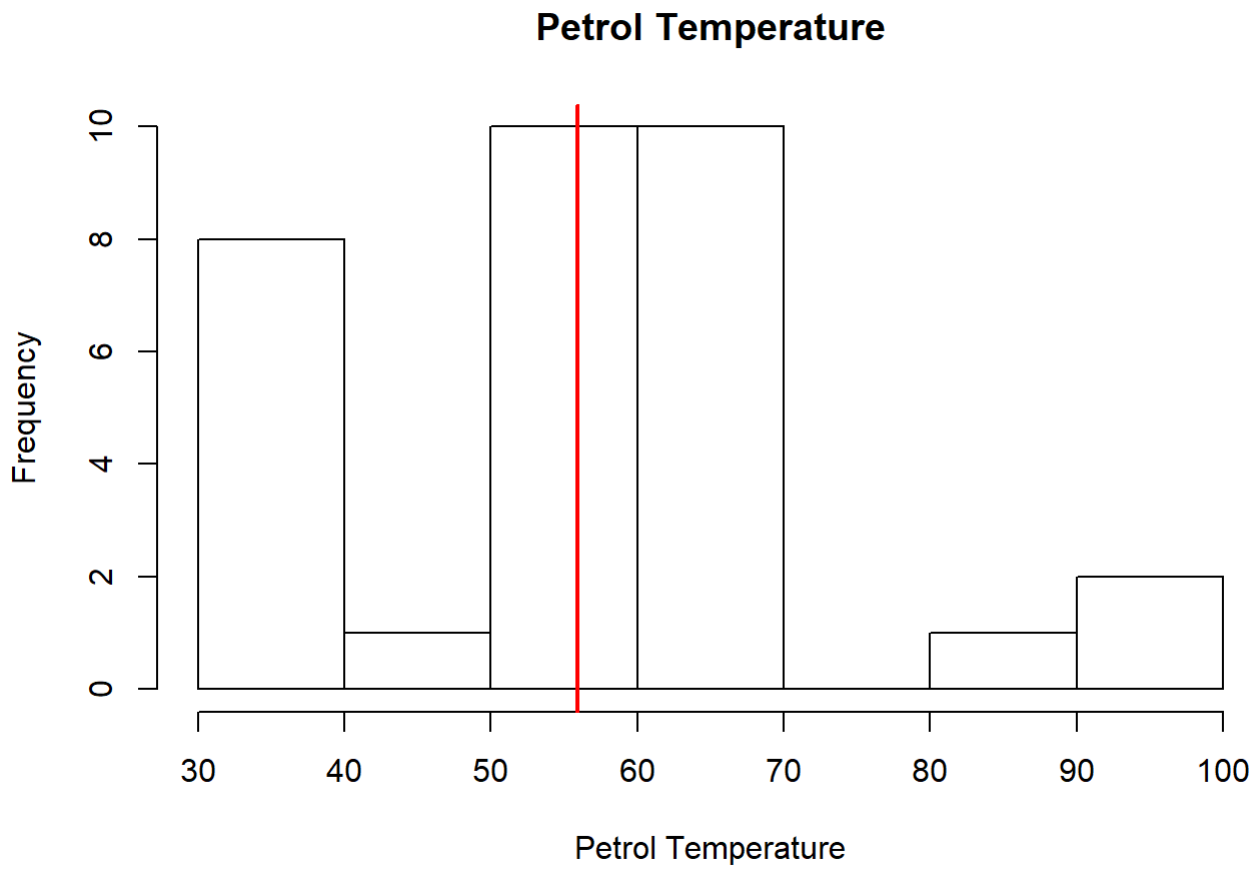
Mean for Initial Tank Pressure:

```
## [1] 4.422187
```



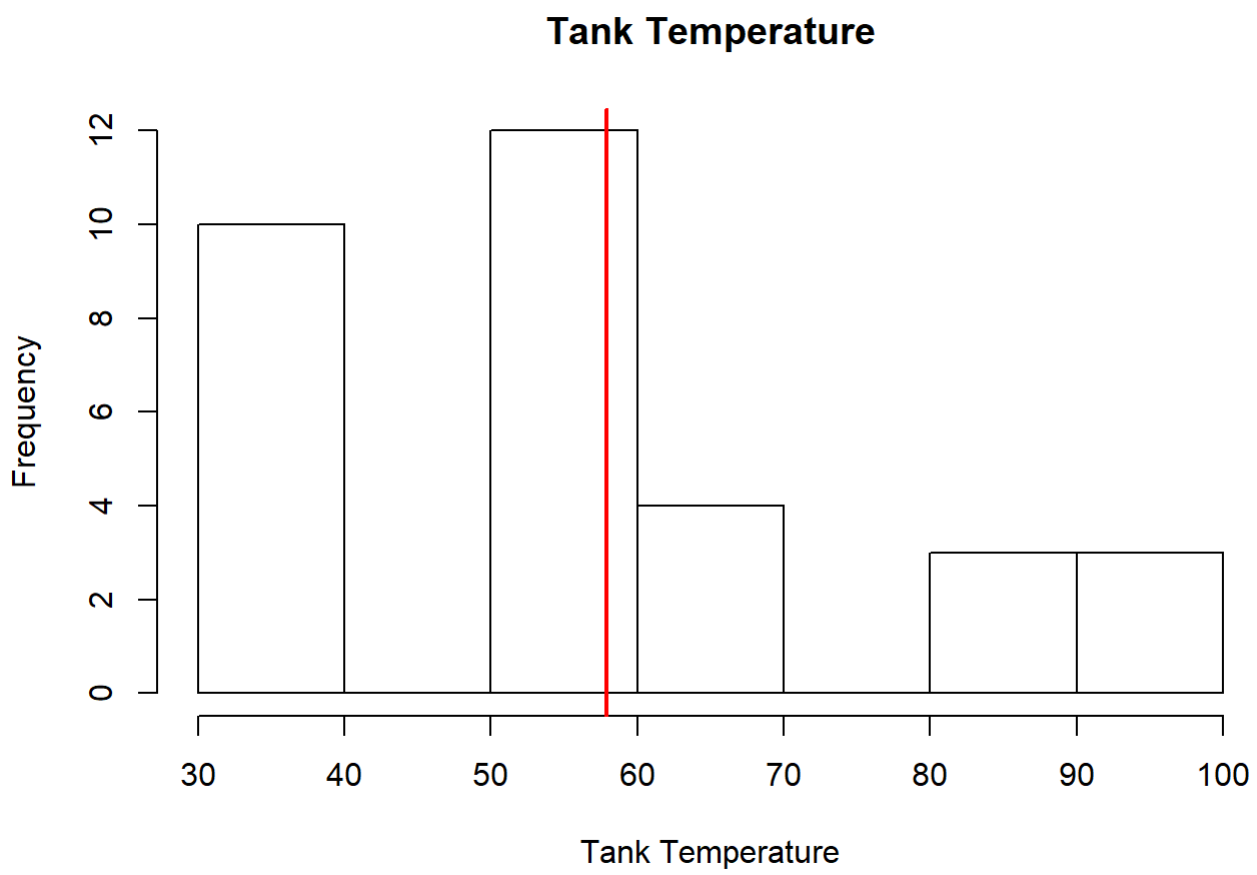
Mean for Petrol Pressure:

```
## [1] 4.32375
```



Mean for Petrol Temperature:

```
## [1] 55.90625
```



Mean for Tank Temperature

```
## [1] 57.90625
```

The data includes 32 observations consisting of 4 independent variables (tank temperature, petrol temperature, initial tank pressure, and petrol pressure) and 1 dependent variable (hydrocarbons escaping). It has a R^2 of 0.9261, which is very close to 1, showing that the predictor variables have a strong linear correlation with the dependent variable.

For escaping hydrocarbons, it has a minimum of 16, maximum of 34.5, mean of 31.125, median of 31.5, and range of 18.5. For the boxplot of escaping hydrocarbons, we can see that there is an outlier and is a little right-skewed. For tank temperature, it has a minimum of 31, maximum of 92, mean of 57.90625, median of 60, and range of 61. For the boxplot of tank temperature, there are no outliers and has a very right-skewed distribution. For petrol temperature, it has a minimum of 35, maximum of 92, mean of 55.90625, median of 60, and range of 57. For the petrol temperature boxplot, there are no outliers as well and also has a very right-skewed distribution. For initial tank pressure, it has a minimum of 2.59, maximum of 7.45, mean of 4.4221875, median of 4.285, and range of 4.86. For the initial tank pressure boxplot, there are 3 outliers and is a little right-skewed. For petrol pressure, it has a minimum of 2.59, maximum of 7.45, mean of 4.32375, median of 4.09, and range of 1.5. For the petrol pressure boxplot, there are 5 outliers and seems fairly normally distributed. We can verify the normalities with a Kolmogorov-Smirnov test.

In the escaping hydrocarbon histogram, we can see that it roughly has a normal distribution. In the initial tank pressure histogram, it is shown that there are no values between the 5 and 6 range. In the petrol pressure histogram, it shows that it is not quite normally distributed. In the petrol temperature histogram, we can see that

there are no values between the range 70 to 80. In the tank temperature histogram, there are no values between the 40 to 50 range and between the 70 to 80 range.

4.3 Correlation Matrix

```
##                       Tank_Temp Petrol_Temp Initial_Tank_Pressure
## Tank_Temp             1.0000000   0.7742909             0.9554116
## Petrol_Temp           0.7742909   1.0000000             0.7815286
## Initial_Tank_Pressure 0.9554116   0.7815286             1.0000000
## Petrol_Pressure       0.9337690   0.8374639             0.9850748
## Escaping_Hydrocarbons 0.8260665   0.9093507             0.8698845
##                       Petrol_Pressure Escaping_Hydrocarbons
## Tank_Temp                   0.9337690             0.8260665
## Petrol_Temp                 0.8374639             0.9093507
## Initial_Tank_Pressure       0.9850748             0.8698845
## Petrol_Pressure             1.0000000             0.9213333
## Escaping_Hydrocarbons       0.9213333             1.0000000
```

All of the variables that we used seem to have a strong positive correlation with each other as the correlation matrix shows no correlation less than 0.75.

5 Analysis and Interpretation

5.1 Fitting the Model

```
##
## Call:
## lm(formula = Escaping_Hydrocarbons ~ Tank_Temp + Petrol_Temp +
##     Initial_Tank_Pressure + Petrol_Pressure, data = pollution_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.586 -1.221 -0.118  1.320  5.106
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            1.01502    1.86131   0.545  0.59001
## Tank_Temp             -0.02861    0.09060  -0.316  0.75461
## Petrol_Temp            0.21582    0.06772   3.187  0.00362 **
## Initial_Tank_Pressure -4.32005    2.85097  -1.515  0.14132
## Petrol_Pressure        8.97489    2.77263   3.237  0.00319 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.73 on 27 degrees of freedom
## Multiple R-squared:  0.9261, Adjusted R-squared:  0.9151
## F-statistic: 84.54 on 4 and 27 DF,  p-value: 7.249e-15
```

From the regression summary, we see that the estimated regression function is Ŷ = 1.01502 - 0.02861$X1$ + *0.21582*$X2$ - 4.32005$X3$ + *8.97489*$X4$. The interpretation of $b0$ is that 1.01502 is the mean value that we would predict for the amount of escaping hydrocarbons if $X1$ = $X2$ = $X3$ = $X4$ = 0. It is also known as the y-intercept. The interpretation of $b1$ is that with a one unit increase in $X1$ (tank temperature), there will be a decrease of 0.02861 to the amount of escaping hydrocarbons while holding $X2$, $X3$, and $X4$ constant. The interpretation of $b2$ is that with a one unit increase in $X2$ (petrol temperature), there will be an increase of 0.21582 to the amount of escaping hydrocarbons while holding $X1$, $X3$, and $X4$ constant. The interpretation of $b3$ is that with a one unit increase in $X3$ (initial tank pressure), there will be a decrease of 4.32005 to the amount of escaping hydrocarbons while holding $X1$, $X2$, and $X4$ constant. The interpretation of $b4$ is that with a one unit increase in $X4$ (petrol pressure), there will be an increase of 8.97489 to the amount of escaping hydrocarbons while holding $X1$, $X2$, and $X3$ constant.

5.2 Transformation

Here, we apply a square root transform to the response variable: Escaping Hydrocarbons

```
##
## Call:
## lm(formula = sqrt_response ~ Tank_Temp + Petrol_Temp + Initial_Tank_Pressure +
##     Petrol_Pressure, data = transformed_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.51345 -0.13624  0.01531  0.16585  0.47079
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              2.876759   0.177078  16.246 1.84e-15 ***
## Tank_Temp               -0.000412   0.008620  -0.048  0.96223
## Petrol_Temp              0.020801   0.006443   3.229  0.00326 **
## Initial_Tank_Pressure -0.372877   0.271231  -1.375  0.18051
## Petrol_Pressure          0.729271   0.263779   2.765  0.01014 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2597 on 27 degrees of freedom
## Multiple R-squared:  0.9129, Adjusted R-squared:    0.9
## F-statistic: 70.78 on 4 and 27 DF,  p-value: 6.497e-14
```

The fitted regression function is:
$\sqrt{\hat{Y}} = 2.876759 - 0.000412X_1 + 0.020801X_2 - 0.372877X_3 + 0.729271X_4$ such that \ Intercept: $\hat{\beta_0} = 2.876759$ \ Tank temperature: $\hat{\beta_1} = -0.000412$ \ Petrol temperature: $\hat{\beta_2} = 0.020801$ \ Intial tank pressure: $\hat{\beta_3} = -0.372877$ \ Petrol pressure: $\hat{\beta_4} = 0.729271$

Breusch-Pagan Test For Constancy of Error Variance

Hypotheses: H_0: Error variance is constant vs. H_1: Error variance is not constant

Decision Rule:

If $\chi^2_{BP} \leq \chi^2_{(1-\alpha;p-1)} \Rightarrow$ fail to reject H_0

If $\chi^2_{BP} > \chi^2_{(1-\alpha;p-1)} \Rightarrow$ reject H_0, conclude H_1

```
##
##  studentized Breusch-Pagan test
##
## data:  sqrt_lm
## BP = 8.051, df = 4, p-value = 0.08973
```

```
## [1] 9.487729
```

Here we have $\chi^2_{BP} = 8.051$, and $\chi^2_{(1-\alpha;p-1)} = \chi^2_{(.95;4)} = 9.488$. Since $8.051 < 9.488 \Rightarrow$ we fail to reject H_0. Given our sample data with a transformation applied to the response variable, the evidence we have suggests the constant variance normality assumption is met.

5.3 Variance Inflation Factor

```
## Loading required package: carData
```

```
##             Tank_Temp           Petrol_Temp Initial_Tank_Pressure
##             12.997379              4.720998             71.301491
##       Petrol_Pressure
##             61.932647
```

```
##
## Call:
## lm(formula = Escaping_Hydrocarbons ~ ., data = new.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.7176 -1.2166 -0.2034  1.2108  5.8870
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.39742    1.85770   0.214 0.832149
## Tank_Temp       -0.11526    0.07188  -1.603 0.120080
## Petrol_Temp      0.27104    0.05838   4.643 7.36e-05 ***
## Petrol_Pressure  5.14570    1.16699   4.409 0.000139 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.792 on 28 degrees of freedom
## Multiple R-squared:  0.9198, Adjusted R-squared:  0.9112
## F-statistic:   107 on 3 and 28 DF,  p-value: 1.908e-15
```

```
##       Tank_Temp     Petrol_Temp Petrol_Pressure
##        7.820038        3.353562       10.486103
```

```
##                       Tank_Temp Petrol_Temp Petrol_Pressure
## Tank_Temp             1.0000000   0.7742909       0.9337690
## Petrol_Temp           0.7742909   1.0000000       0.8374639
## Petrol_Pressure       0.9337690   0.8374639       1.0000000
## Escaping_Hydrocarbons 0.8260665   0.9093507       0.9213333
##                       Escaping_Hydrocarbons
## Tank_Temp                         0.8260665
## Petrol_Temp                       0.9093507
## Petrol_Pressure                   0.9213333
## Escaping_Hydrocarbons             1.0000000
```

```
##
## Call:
## lm(formula = Escaping_Hydrocarbons ~ ., data = new.data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.1742 -1.1371  0.3782  1.6279  8.5390
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.22987    2.37558   0.097  0.92358
## Tank_Temp    0.14625    0.05196   2.815  0.00868 **
## Petrol_Temp  0.40114    0.06443   6.226 8.59e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.572 on 29 degrees of freedom
## Multiple R-squared:  0.8641, Adjusted R-squared:  0.8547
## F-statistic: 92.17 on 2 and 29 DF,  p-value: 2.713e-13
```

```
##   Tank_Temp Petrol_Temp
##    2.497044    2.497044
```

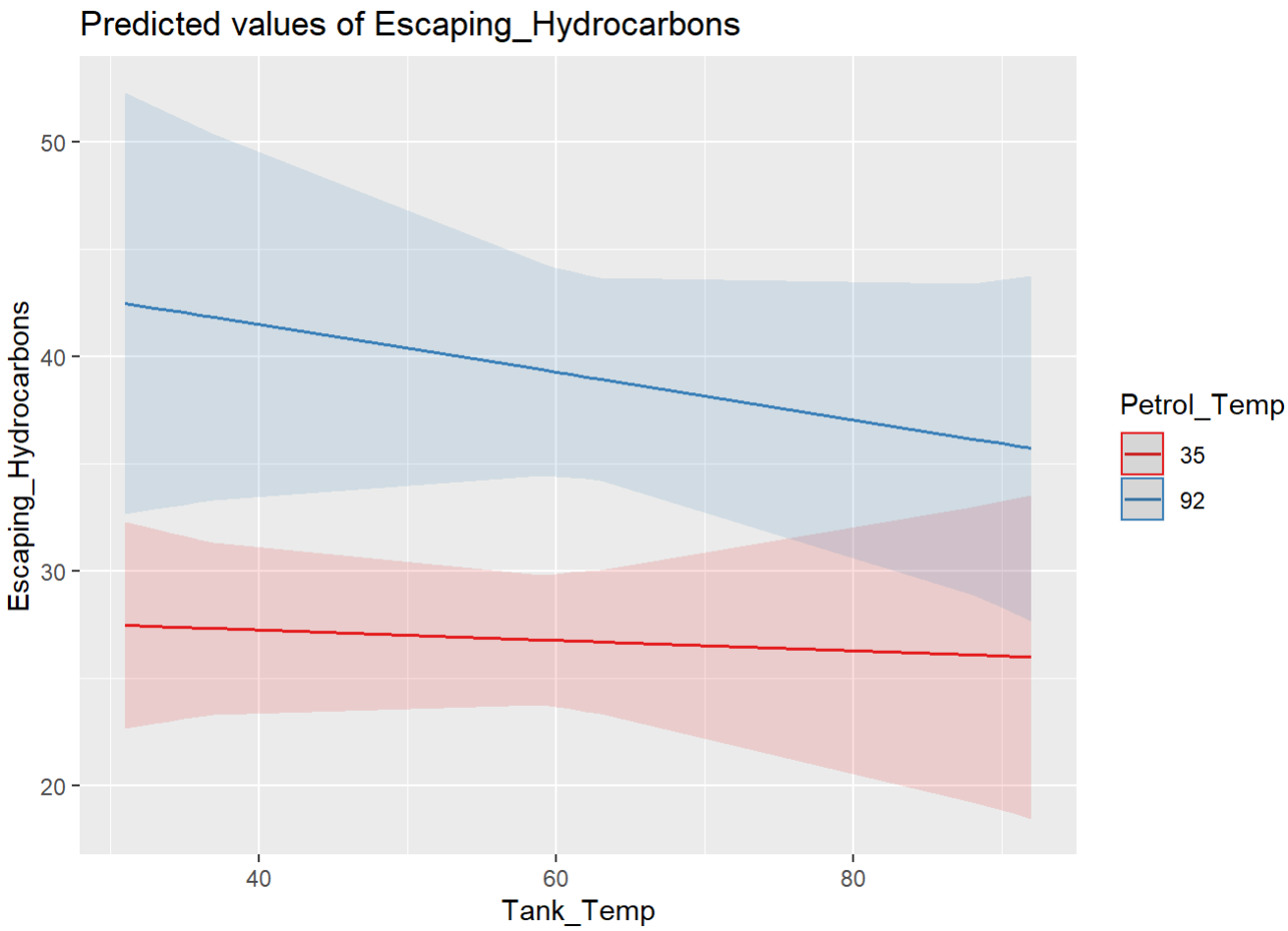
```
##                       Tank_Temp Petrol_Temp Escaping_Hydrocarbons
## Tank_Temp             1.0000000   0.7742909             0.8260665
## Petrol_Temp           0.7742909   1.0000000             0.9093507
## Escaping_Hydrocarbons 0.8260665   0.9093507             1.0000000
```

```
##
## Call:
## lm(formula = Escaping_Hydrocarbons ~ ., data = data.interaction)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.4033 -1.3389  0.1262  1.5618  5.7453
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        7.47918    3.00551   2.488  0.01905 *
## Tank_Temp         -0.04645    0.07364  -0.631  0.53326
## Petrol_Temp        0.34493    0.05819   5.928 2.22e-06 ***
## interaction.x3_x4  0.33502    0.10151   3.300  0.00264 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.084 on 28 degrees of freedom
## Multiple R-squared:  0.9021, Adjusted R-squared:  0.8916
## F-statistic: 86.03 on 3 and 28 DF,  p-value: 3.055e-14
```

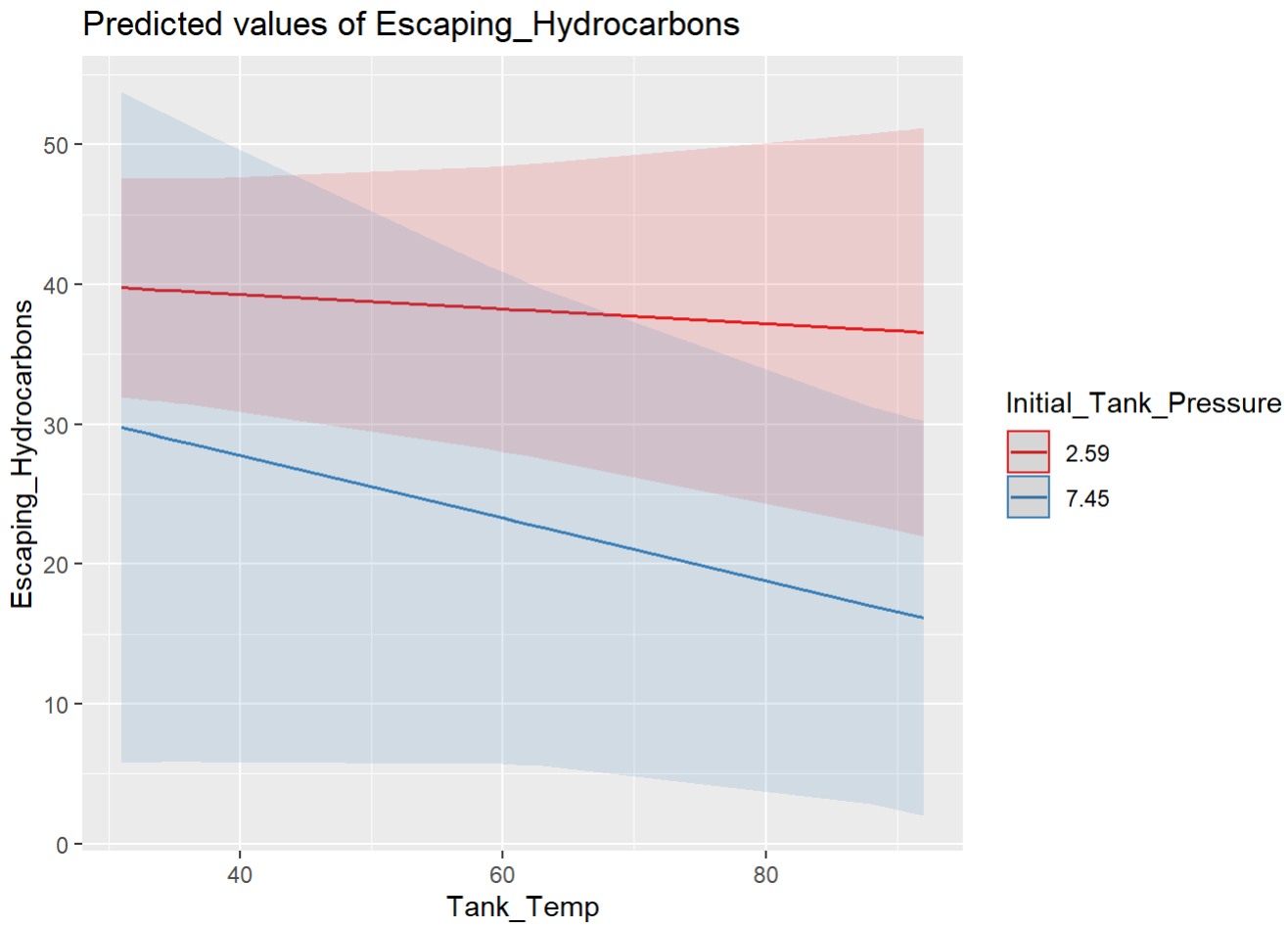
```
##                       Escaping_Hydrocarbons Tank_Temp Petrol_Temp
## Escaping_Hydrocarbons             1.0000000 0.8260665   0.9093507
## Tank_Temp                         0.8260665 1.0000000   0.7742909
## Petrol_Temp                       0.9093507 0.7742909   1.0000000
## interaction.x3_x4                 0.8818568 0.9186150   0.7844648
##                       interaction.x3_x4
## Escaping_Hydrocarbons         0.8818568
## Tank_Temp                     0.9186150
## Petrol_Temp                   0.7844648
## interaction.x3_x4             1.0000000
```

In the correlation matrix, we observed multicollinearity between each pair of predictor variables. We will use the variance inflation factor of each predictor variable to measure the correlation between each predictor against the remaining variables, a one versus all approach. This will help us see the extent of multicollinearity that we cannot see just by observing a pair of variables. We use the ‘car’ package and use the VIF function on our full regression model. The numbers that are shown in the output represent how much excess variation of a particular regression coefficent is simply due to multicollinearity.

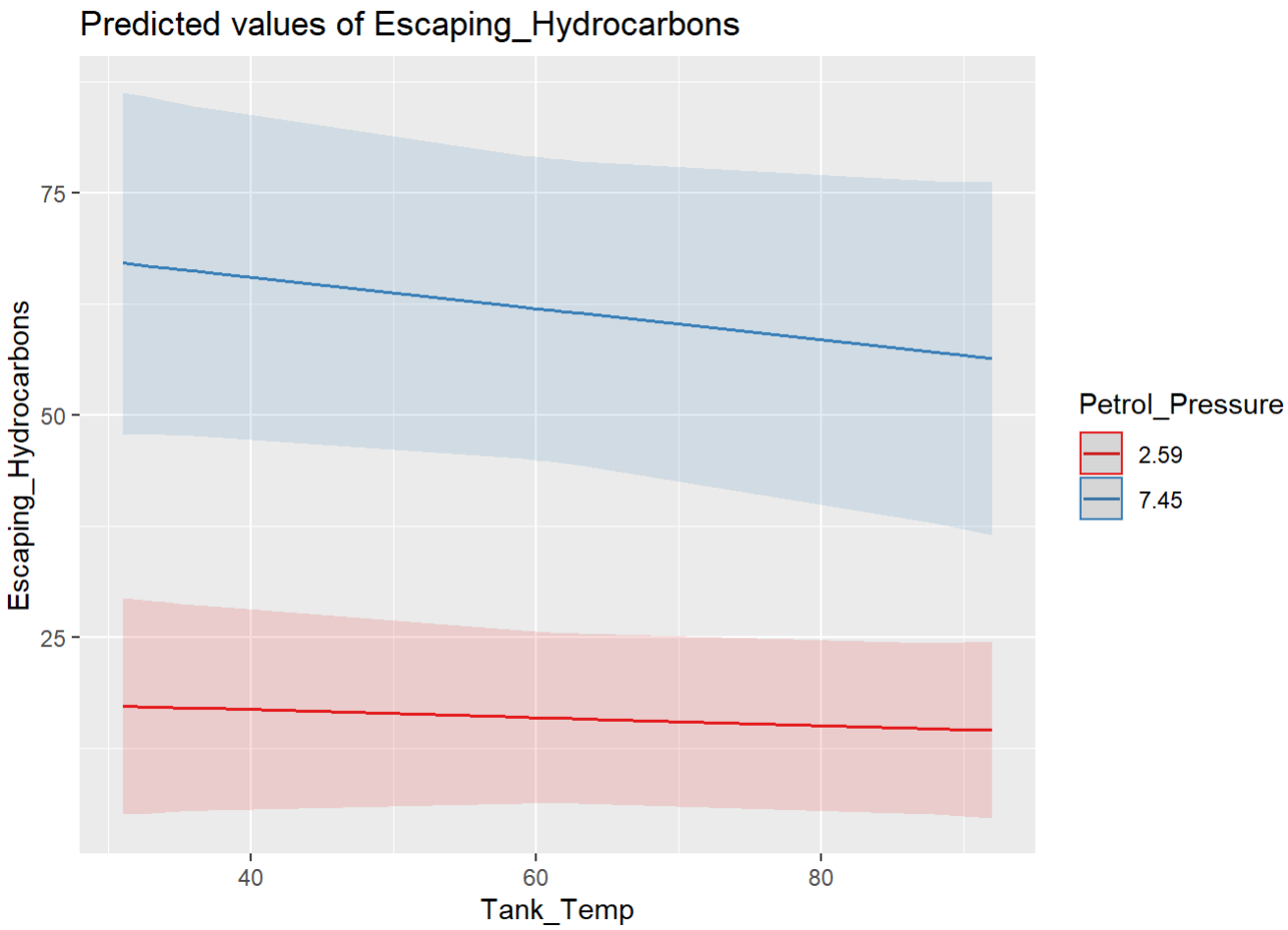
5.4 Interaction Effects



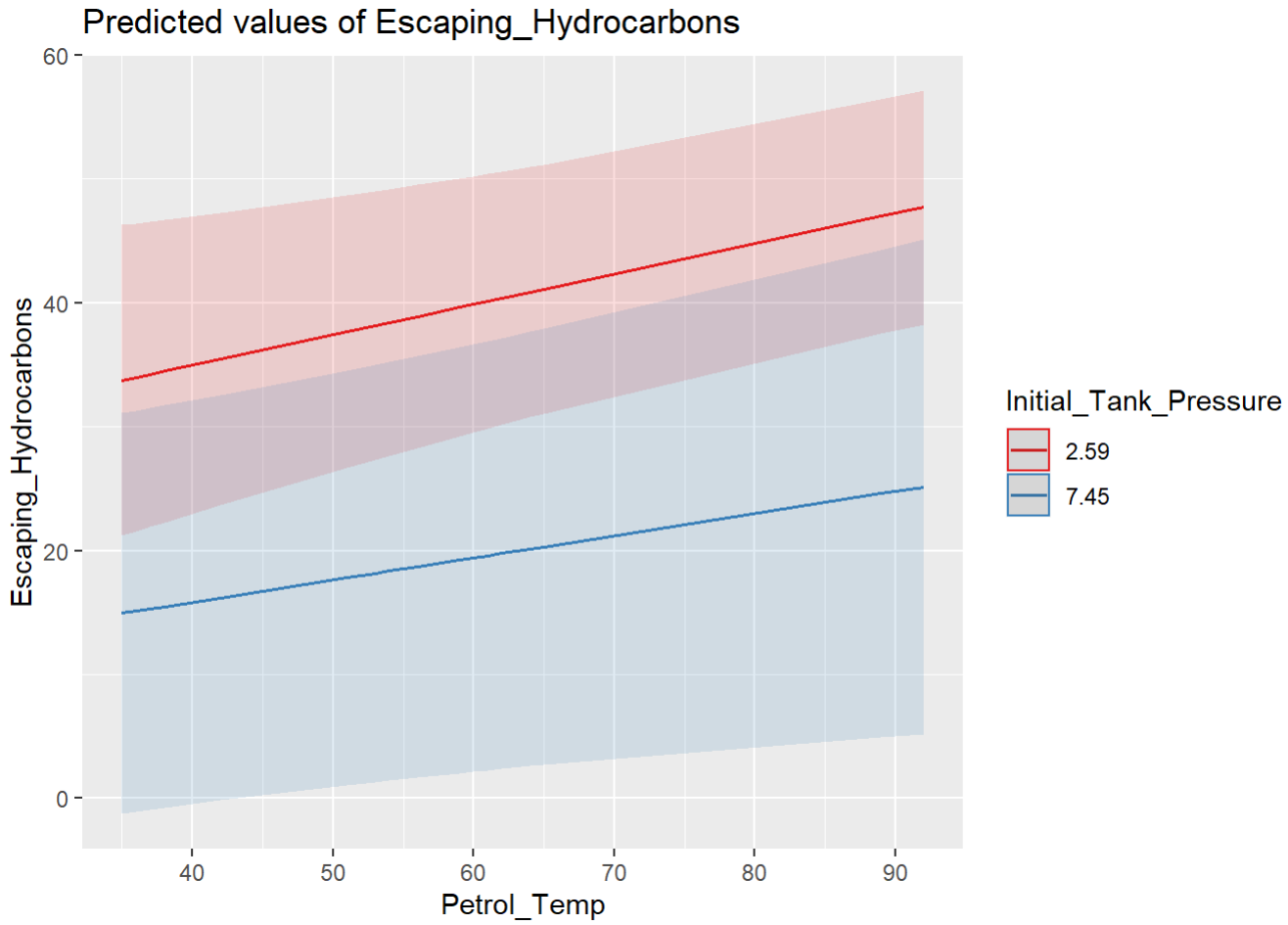
From the interaction plot above, we see that interactivity exists as the lines are not parallel.



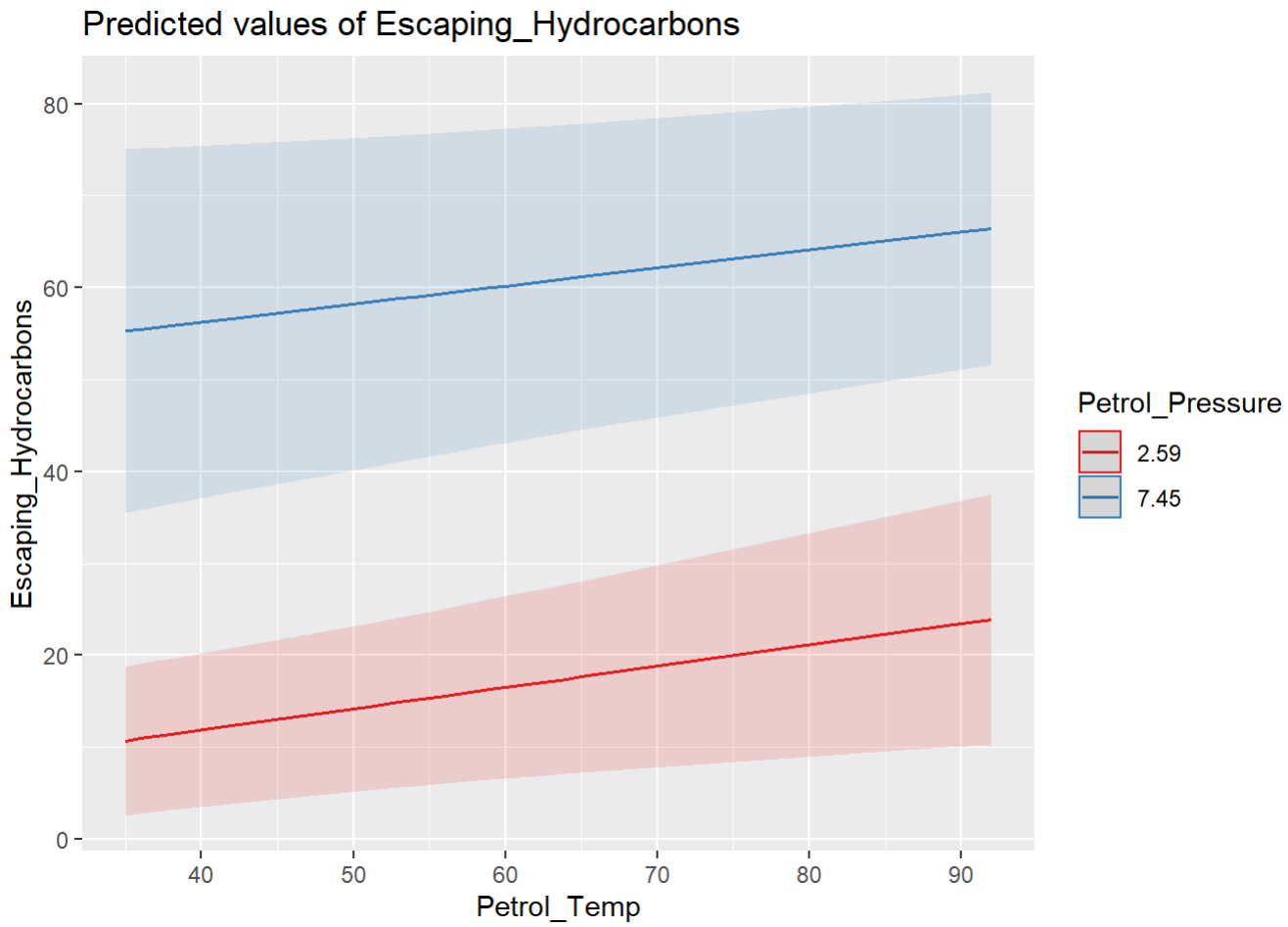
From the interaction plot above, we see that interactivity exists as the lines are not parallel.



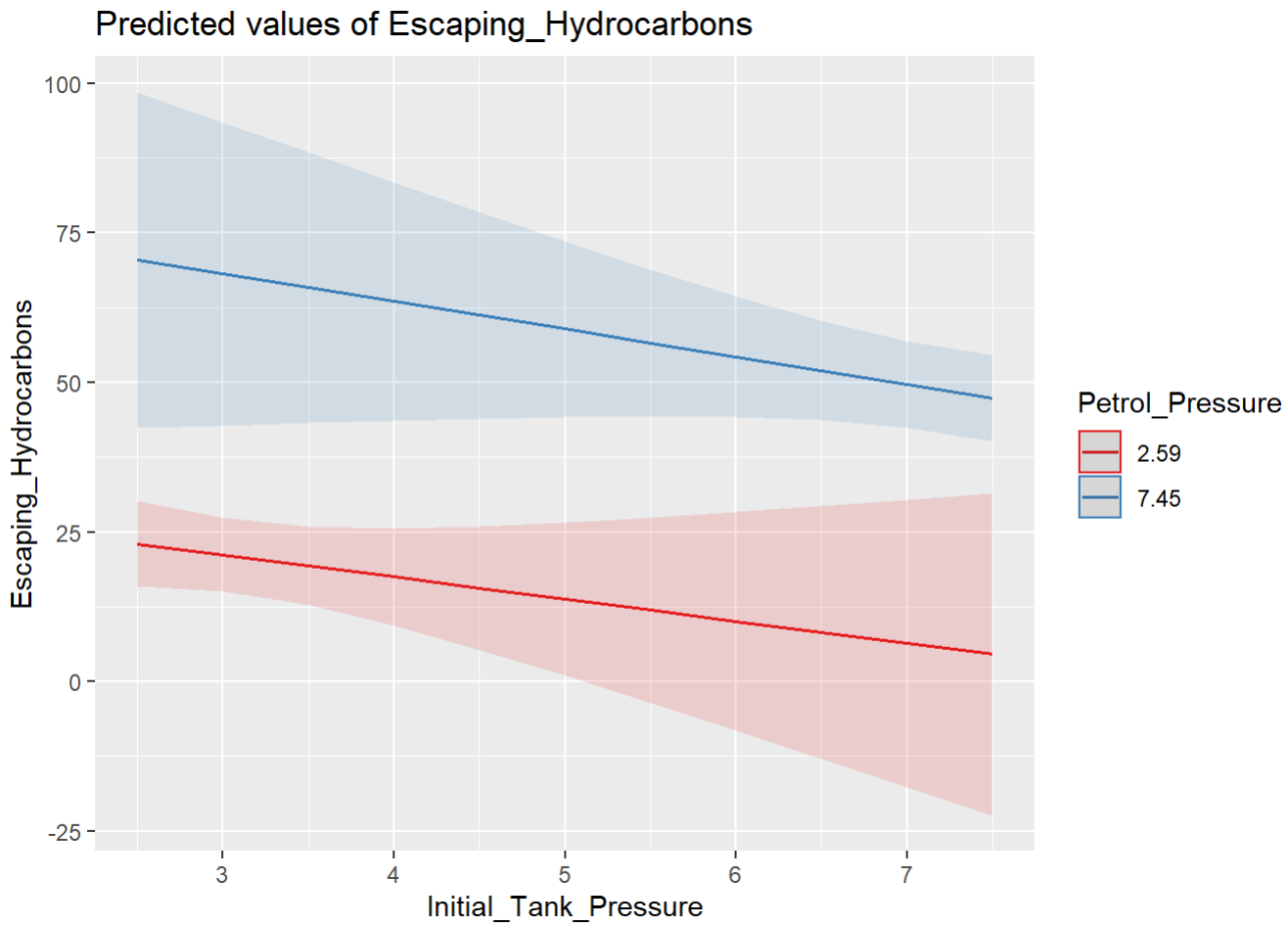
From the interaction plot above, we see that interactivity exists as the lines are not parallel.



From the interaction plot above, we see that interactivity does not exist as the lines are parallel.



From the interaction plot above, we see that interactivity does not exist as the lines are parallel.

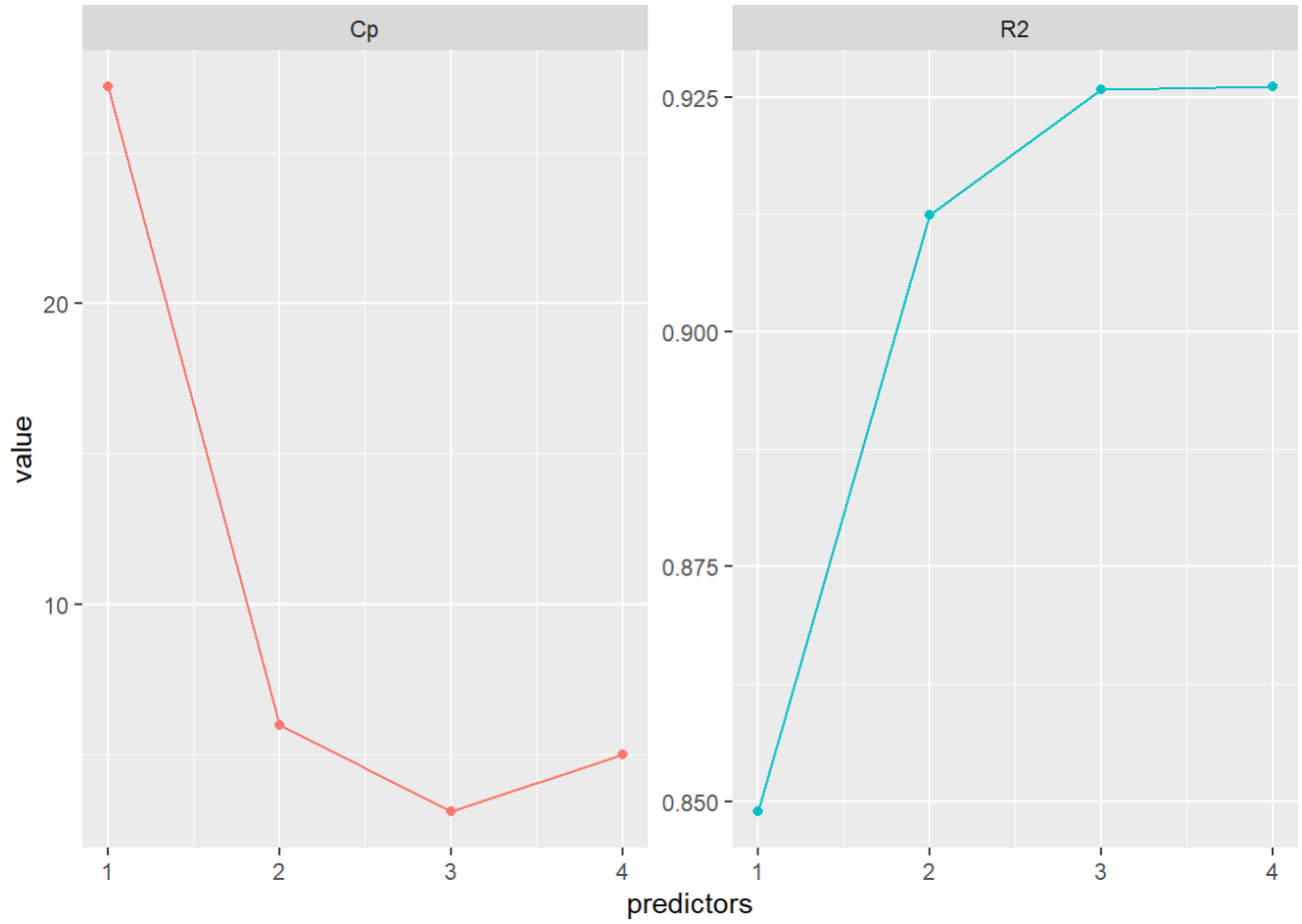


From the interaction plot above, we see that interactivity does not exist as the lines are parallel.

We exclude the initial tank pressure variable from the interaction model and add the $\beta_1\beta_2$ and $\beta_1\beta_4$ interaction terms.

```
##
## Call:
## lm(formula = Escaping_Hydrocarbons ~ Tank_Temp + Petrol_Temp +
##     Petrol_Pressure + Tank_Temp * Petrol_Temp + Tank_Temp * Petrol_Pressure,
##     data = pollution_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.2190 -1.3549 -0.4469  1.3297  5.6834
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -8.7406659  6.9488274  -1.258   0.2196
## Tank_Temp                  -0.0356932  0.0987886  -0.361   0.7208
## Petrol_Temp                 0.1900967  0.2105071   0.903   0.3748
## Petrol_Pressure             9.3486772  3.7942578   2.464   0.0207 *
## Tank_Temp:Petrol_Temp       0.0007721  0.0028675   0.269   0.7898
## Tank_Temp:Petrol_Pressure -0.0428968  0.0412871  -1.039   0.3084
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.79 on 26 degrees of freedom
## Multiple R-squared:  0.9257, Adjusted R-squared:  0.9114
## F-statistic: 64.74 on 5 and 26 DF,  p-value: 7.705e-14
```

5.5 Model Selection



From the plot of the Mallow's Cp values and amount of predictor variables, we see that having 3 predictor variables leads to the smallest Cp value.

From the plot of the R^2 and amount of predictor variables, we see that having 3 predictor variables is ideal as that is when the curve starts to flatten out.

Thus, we will select a model with 3 predictor variables.

Hypothesis Test for β_3

We conduct a hypothesis test to formally validate that we should not include $\beta_3 X_3$: inital tank pressure in our model using an alpha level of 0.05.

Hypotheses:

$H_0 : \beta_3 = 0$ vs. $H_1 : \beta_3 \neq 0$

Decision Rule:

If $t^{*} t_{(1 - , n - p)} $ fail to reject H_0

If $t^{*} > t_{(1 - , n - p)} $ reject H_0, conclude H_1

Our test statistic is: $t^* = \frac{b_3}{s(b_3)}$

```
## Initial_Tank_Pressure
##             1.515293
```

```
## [1] 0.025
```

```
## [1] 27
```

```
## [1] 2.051831
```

Here we have $|t^*| = 1.5153$, and $t_{(1-\frac{\alpha}{2},n-p)} = t_{(.975,27)} = 2.0518$. Since $1.5153 < 2.0518 \Rightarrow$ we fail to reject H_0. We have sufficient evidence that suggests we should not include the predictor initial tank pressure variable in our model.

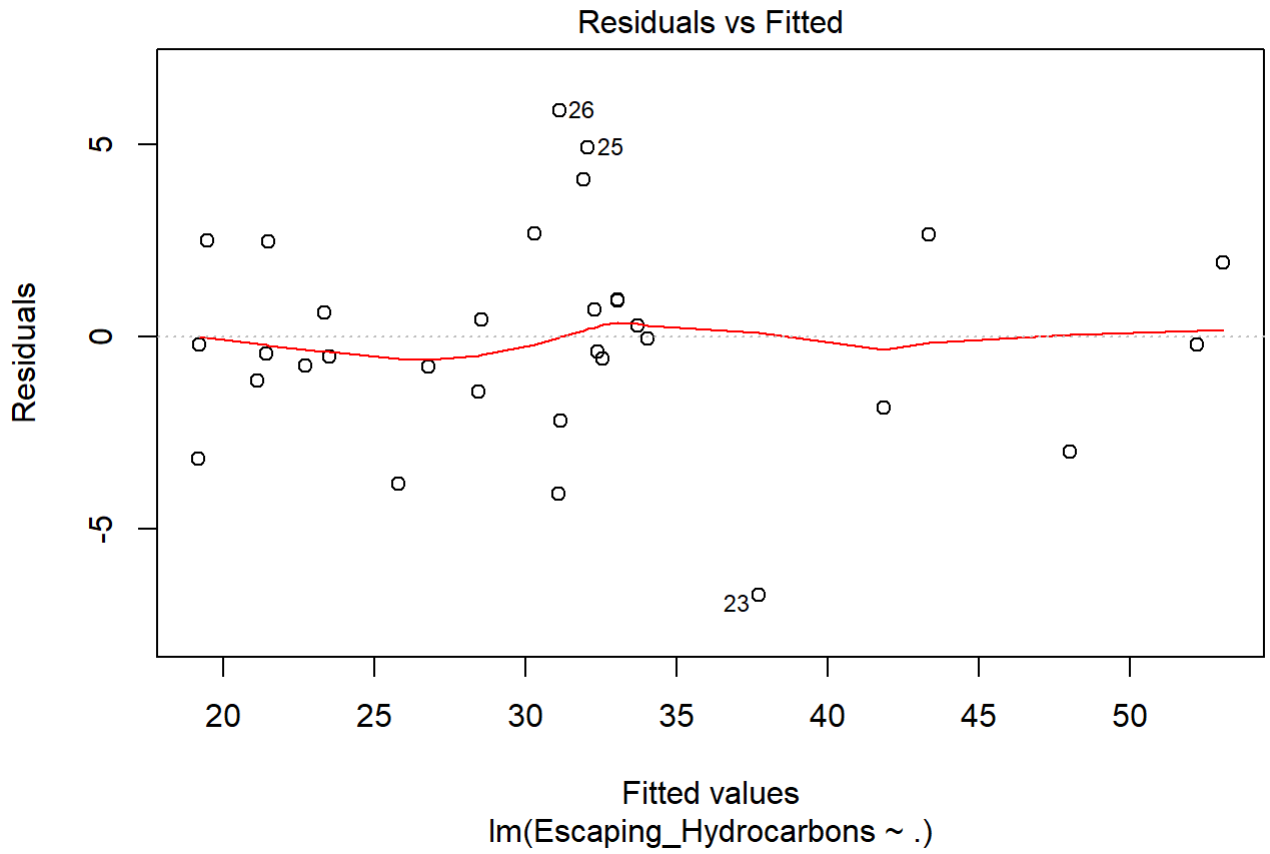
The model we will choose is $Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon_i$ where $\epsilon_i \sim N(0, \sigma^2)$ and

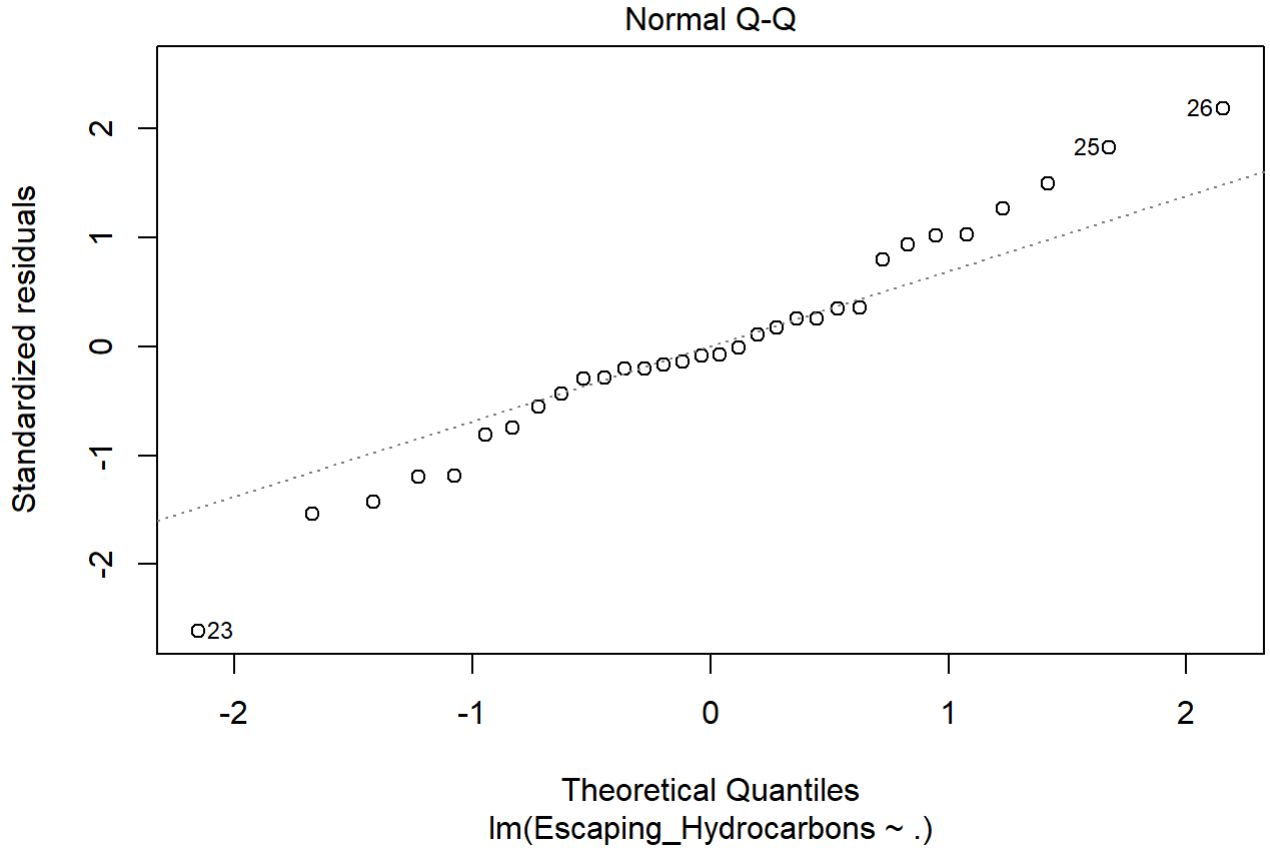
$X_{1} = $ Tank_Temp

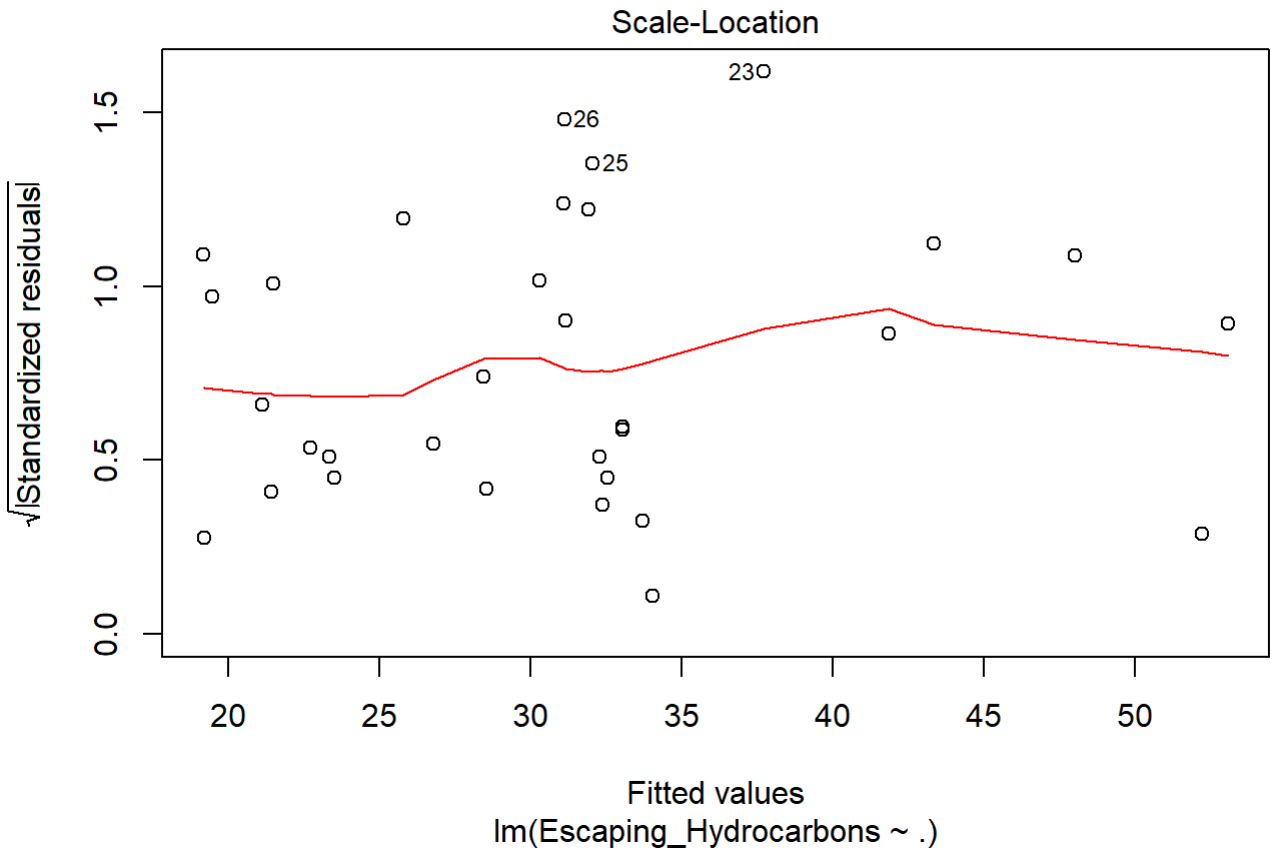
$X_{2} = $ Petrol_Temp

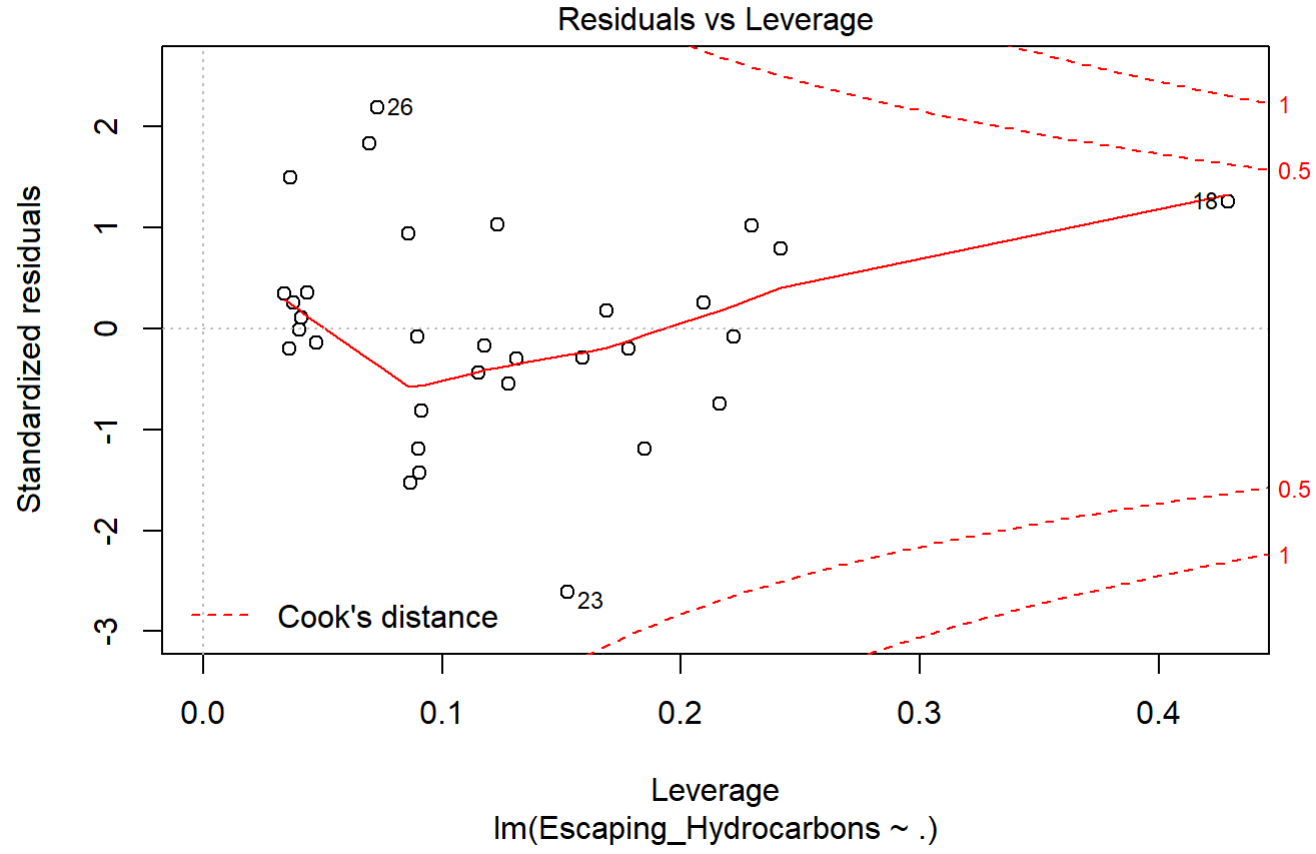
$X_{3} = $ Petrol_Pressure

5.6 Model Diagnostics


Residuals vs Fitted
Residuals
Fitted values
lm(Escaping_Hydrocarbons ~ .)
26
25
23
5
0
-5
20
25
30
35
40
45
50


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(Escaping_Hydrocarbons ~ .)
26
25
23
2
1
0
-1
-2


Scale-Location
√|Standardized residuals|
23
26
25
0.0
0.5
1.0
1.5
20
25
30
35
40
45
50
Fitted values
lm(Escaping_Hydrocarbons ~ .)


Residuals vs Leverage
Standardized residuals
26
18
23
Cook's distance
1
0.5
0.5
1
-3
-2
-1
0
1
2
0.0
0.1
0.2
0.3
0.4
Leverage
lm(Escaping_Hydrocarbons ~ .)

The Residuals vs. Fitted plot displays a fairly horizontal line and doesn't have any distinct patterns which is an indication for a linear relationship. The normal Q-Q plot shows the residuals following the Q-Q line in a normal fashion, suggesting that the normality assumption to be reasonable. The Scale-Location plot has a fairly horizontal line with spread out points which suggest our assumption of homogenous variance to be reasonable. The Residuals vs Leverage plot doesn't show any points with large Cook's distance meaning that there are no highly influential cases that will influence our model.

5.6.1 Test for Normality Assumption: KS Test

We will use a significance level of 0.05 for this KS Test.

Hypotheses: H_0: normality assumption holds vs. H_a: normality assumption does not hold

Decision Rule:

If p-value $> \alpha$: Conclude H_0

If p-value $\leq \alpha$: Conclude H_a

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  (new.lm$fitted.values - y.mean)/y.std and rnorm(32)
## D = 0.21875, p-value = 0.4337
## alternative hypothesis: two-sided
```

Conclusion: Since our p-value is greater than our alpha level of 0.05, we cannot reject the null hypothesis and thus our normality assumption holds.

5.6.2 Overall F-Test For Regression Relation

Tests whether there is a regression relation between the response variable and the set of predictor variables. We will conduct the test with a significance level of 0.05

Hypotheses: H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ vs H_a: not all β_k ($k = 1, \ldots, p-1$) equal zero

Decision Rule:

$F^* \leq F(1-\alpha; p-1, n-p)$, conclude H_0

$F^* > F(1-\alpha; p-1, n-p)$, conclude H_a

where $F^* = \frac{MSR}{MSE}, MSR = \frac{SSR}{p-1}, MSE = \frac{SSE}{n-p}$

```
## Analysis of Variance Table
##
## Response: Escaping_Hydrocarbons
##                  Df  Sum Sq Mean Sq F value    Pr(>F)
## Tank_Temp         1 1857.11 1857.11 238.157 3.201e-15 ***
## Petrol_Temp       1  494.43  494.43  63.406 1.133e-08 ***
## Petrol_Pressure   1  151.61  151.61  19.443  0.000139 ***
## Residuals        28  218.34    7.80
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$F(1-\alpha; p-1, n-p)$

```
## [1] 2.946685
```

F^*

```
## [1] 107.0019
```

Conclusion: Since 107.001915652308 > 2.94668526601727, we conclude H_a that not all β_k ($k = 1, \ldots, p-1$) equal zero. At the $\alpha = 5\%$ significance level, there is sufficient evidence that the amount of escaping hydrocarbons are related to the tank temperature, the temperature of the petrol pumped in, and the pressure of the petrol pumped in.

5.6.3 Breusch-Pagan Test For Constancy of Error Variance

We will conduct the test with a signifcance level of 0.05.

Hypotheses: H_0 = Error variance is constant vs. H_a = Error variance is not constant

Decision Rule:

$\chi^2_{BP} \leq \chi^2_{(1-\alpha;p-1)}$ = Conclude H_0

$\chi^2_{BP} > \chi^2_{(1-\alpha;p-1)}$ = Conclude H_a

χ^2_{BP}

```
##
##  studentized Breusch-Pagan test
##
## data:  new.lm
## BP = 6.6669, df = 3, p-value = 0.08331
```

$\chi^2_{(1-\alpha;p-1)}$

```
## [1] 7.814728
```

Conclusion: Since 6.6669 < 7.81472790325118, we conclude H_0 that the error variance is constant.

5.7 Confidence Intervals

```
##                     2.5 %     97.5 %
## (Intercept)     -3.4078927 4.20274044
## Tank_Temp       -0.2625064 0.03199298
## Petrol_Temp      0.1514555 0.39062880
## Petrol_Pressure  2.7552364 7.53615830
```

Our confidence intervals suggest that petrol temperature and petrol pressure must have a positive association with escaping hydrocarbons as the confidence intervals for their slopes do not contain zero and are above zero.

To determine our regression assumption, we performed many tests. Base on our F-test for linear regression using a 0.05 level of significance, we determined that the not all Hα that not all Bk equal to zero and there is enough evidence that the escaping hydrocarbons are related to any of the measures. Another test we conducted three different times was a Breusch-Pagan test for constancy of error variance using a significance level of 0.05. With the significance level of 0.05, we do have enough evidence that suggests constant variance assumption is met. Once we conduct our confidence intervals, it suggests to us that petrol temperature and petrol pressure have a positive relationship since the slopes do not contain zero for escaping hydrocarbons. Corroborating to the positive correlation, a correlation matrix, it also shows a positive correlation. For the correlation matrix, we observed multicollinearity and used the variance to determine our regression coefficient, so remove one of the measures, initial tank pressure, to accurately determine that the other three predictors are the cause of the escaping hydrocarbons.

6 Conclusion and Discussion

The purpose of this data set is to determine how much hydrocarbons would escape depending on the different measures. Using the four different measures: tank temperature, petrol temperature, initial tank temperature, and petrol pressure to evaluate the effectiveness of pollution control. With our data consisting of six columns and thirty-two rows to see how effective petrol is being pumped. From the given data set, we used R to conduct a series of tests. For our interpretation of the model, the estimated linear regression function is $\hat{Y} = 1.01502 - 0.02861X1 + 0.21582X2 - 4.32005X3 + 8.97489X4$ which includes a mean value and depending on the X, there will be a decrease or increase of the escaping hydrocarbons while containing the other X values. After running and checking the regression summary, the two predictors were not significant and the bp test had a violation in our assumption, so we changed our model. Since there was high correlation between the variables in the correlation matrix, we created an interaction plot, but it seems the interactions were not parallel. Furthermore, we use model selection criteria test: R-squared and Mallow's Cp, choosing to drop the initial tank pressure variable because of the small marginal difference. The assumption was met when we tried a KS test for normality and a bp test for constancy of variance. The correlation matrix gives us a strong positive correlation. For multicollinearity, the initial tank pressure was the highest which we removed after completing a test statistic. Using three measures, we find a small Cp value and it is ideal to have for the curve to flatten out. Some factors that can be improved on for this data is having a larger data set with observations for a more accurate assumption and including different transformations. Lastly, all the charts and tests do match our assumptions that the amount of escaping hydrocarbons are related to our measures: tank temperature, the temperature of the petrol pumped in, and the pressure of the petrol pumped in.

7 References

Hilpert, M., Mora, B.A., Ni, J. et al. Hydrocarbon Release During Fuel Storage and Transfer at Gas Stations: Environmental and Health Effects. Curr Envir Health Rpt 2, 412–422 (2015). https://doi.org/10.1007/s40572-015-0074-8 (https://doi.org/10.1007/s40572-015-0074-8)

Helmut Spaeth, Mathematical Algorithms for Linear Regression, Academic Press, 1991,ISBN 0-12-656460-4. S Weisberg, Applied Linear Regression, Wiley, 1980, page 146.

Contributions

- Mark Berman: Introduction and Methods
- Christina Li: Conclusion and Discussion, Analysis and Interpretation

- Ariel Lee: Data Summary and Exploratory Data Analysis write up
- Justin Luong: Fit MLR model, regression summary, interpretation of regression coefficients, box plots, histograms, F-Test for regression relation, Breusch-Pagan Test, interaction plots
- Sophia Tierney: Square root transformation, regression summary for transformed MLR, BP Test for constant variance
- Ryan Truong: Model Assumptions, ANOVA, Confidence intervals, correlation matrix, model selection
- Jenna Zarbis: Abstract Write Up

Appendix

```
knitr::opts_chunk$set(fig.pos = 'H')
pollution_data <- read.table('pollutiondata.txt', col.names = c('Index', 'Tank_Temp', 'Petrol_Te
mp', 'Initial_Tank_Pressure', 'Petrol_Pressure', 'Escaping_Hydrocarbons'))
pollution_lm <- lm(Escaping_Hydrocarbons ~ Tank_Temp + Petrol_Temp + Initial_Tank_Pressure + Pet
rol_Pressure, pollution_data)
boxplot(pollution_data$Escaping_Hydrocarbons, main = "Box Plot of Escaping Hydrocarbons")
# Five Number Summary For Escaping Hydrocarbons
quantile(pollution_data$Escaping_Hydrocarbons)
boxplot(pollution_data[2:3], main = "Box Plots of Tank Temperature and Petrol Temperature")
quantile(pollution_data$Tank_Temp)
quantile(pollution_data$Petrol_Temp)
boxplot(pollution_data[4:5], main = "Box Plots of Initial Tank Pressure and Petrol Pressure")
quantile(pollution_data$Initial_Tank_Pressure)
quantile(pollution_data$Petrol_Pressure)
hist(pollution_data$Escaping_Hydrocarbons, xlab = "Amount of Escaping Hydrocarbons", main = "Amo
unt of Escaping Hydrocarbons")
abline(v=mean(pollution_data$Escaping_Hydrocarbons), col='red', lwd=2)
mean(pollution_data$Escaping_Hydrocarbons)
hist(pollution_data$Initial_Tank_Pressure, xlab = "Initial Tank Pressure", main = "Initial Tank
 Pressure")
abline(v=mean(pollution_data$Initial_Tank_Pressure), col='red', lwd=2)
mean(pollution_data$Initial_Tank_Pressure)
hist(pollution_data$Petrol_Pressure, xlab = "Petrol Pressure", main = "Petrol Pressure")
abline(v=mean(pollution_data$Petrol_Pressure), col='red', lwd=2)
mean(pollution_data$Petrol_Pressure)
hist(pollution_data$Petrol_Temp, xlab = "Petrol Temperature", main = "Petrol Temperature")
abline(v=mean(pollution_data$Petrol_Temp), col='red', lwd=2)
mean(pollution_data$Petrol_Temp)
hist(pollution_data$Tank_Temp, main = "Tank Temperature", xlab = "Tank Temperature")
abline(v=mean(pollution_data$Tank_Temp), col='red', lwd=2)
mean(pollution_data$Tank_Temp)
data <- subset(pollution_data, select=-c(Index))
cor(data)
pollution_lm <- lm(Escaping_Hydrocarbons ~ Tank_Temp + Petrol_Temp + Initial_Tank_Pressure + Pet
rol_Pressure, pollution_data)
summary(pollution_lm)
sqrt_response <- sqrt(pollution_data$Escaping_Hydrocarbons)

# save predictor variables into a new variable to make a new data frame
Tank_Temp <- pollution_data$Tank_Temp
Petrol_Temp <- pollution_data$Petrol_Temp
Initial_Tank_Pressure <- pollution_data$Initial_Tank_Pressure
Petrol_Pressure <- pollution_data$Petrol_Pressure

# New Data Frame with transformed response variable
transformed_data <- data.frame(sqrt_response, Tank_Temp, Petrol_Temp, Initial_Tank_Pressure, Pet
rol_Pressure)

# Transformed MLR
sqrt_lm <- lm(sqrt_response ~ Tank_Temp + Petrol_Temp + Initial_Tank_Pressure + Petrol_Pressure,
              transformed_data)
summary(sqrt_lm)
library(lmtest)
```

```
bptest(sqrt_lm)
qchisq(.95, df = 4)
library(car)
vif(pollution_lm)

Escaping_Hydrocarbons <- pollution_data$Escaping_Hydrocarbons
Tank_Temp <- pollution_data$Tank_Temp
Petrol_Temp <- pollution_data$Petrol_Temp
Initial_Tank_Pressure <- pollution_data$Initial_Tank_Pressure
Petrol_Pressure <- pollution_data$Petrol_Pressure

# drop Initial Tank Pressure variable first, it has the highest Variance Inflation Factor
new.data <- subset(pollution_data, select = -c(Index, Initial_Tank_Pressure))
new.lm <- lm(Escaping_Hydrocarbons ~., data = new.data)
summary(new.lm)
vif(new.lm)

corr.1 <- cor(new.data)
corr.1

new.data2 <- subset(pollution_data, select = -c(Index, Initial_Tank_Pressure, Petrol_Pressure))
new.lm2 <- lm(Escaping_Hydrocarbons ~., data = new.data2)
summary(new.lm2)
vif(new.lm2)

corr.2 <- cor(new.data2)
corr.2

## Interaction effect
interaction.x3_x4 <- pollution_data$Initial_Tank_Pressure*pollution_data$Petrol_Pressure
data.interaction <- data.frame(Escaping_Hydrocarbons, Tank_Temp, Petrol_Temp, interaction.x3_x4)
interaction.lm <- lm(Escaping_Hydrocarbons ~ ., data = data.interaction)
summary(interaction.lm)

cor(data.interaction)
## the interaction effect is significant at the alpha = .001 level
# but the correlation is still pretty high
library(sjPlot)
library(sjmisc)
# beta1 * beta2 interaction plot
b1b2_lm <- lm(Escaping_Hydrocarbons ~ Initial_Tank_Pressure + Petrol_Pressure + Tank_Temp * Petr
ol_Temp, pollution_data)
plot_model(b1b2_lm, type = "int", terms = c("Tank_Temp", "Petrol_Temp"))
# beta1 * beta3 interaction plot
b1b3_lm <- lm(Escaping_Hydrocarbons ~ Petrol_Temp + Petrol_Pressure + Tank_Temp * Initial_Tank_P
ressure, pollution_data)
plot_model(b1b3_lm, type = "int", terms = c("Tank_Temp", "Initial_Tank_Pressure"))
# beta1 * beta4 interaction plot
b1b4_lm <- lm(Escaping_Hydrocarbons ~ Petrol_Temp + Initial_Tank_Pressure + Tank_Temp * Petrol_P
ressure, pollution_data)
plot_model(b1b4_lm, type = "int", terms = c("Tank_Temp", "Petrol_Pressure"))
# beta2 * beta3 interaction plot
b2b3_lm <- lm(Escaping_Hydrocarbons ~ Tank_Temp + Petrol_Pressure + Petrol_Temp *
    Initial_Tank_Pressure, pollution_data)
```

```
plot_model(b2b3_lm, type = "int", terms = c("Petrol_Temp", "Initial_Tank_Pressure"))
# beta2 * beta4 interaction plot
b2b4_lm <- lm(Escaping_Hydrocarbons ~ Tank_Temp + Initial_Tank_Pressure + Petrol_Temp * Petrol_P
ressure, pollution_data)
plot_model(b2b4_lm, type = "int", terms = c("Petrol_Temp", "Petrol_Pressure"))
# beta3 * beta 4 interaction plot
b3b4_lm <- lm(Escaping_Hydrocarbons ~ Tank_Temp + Petrol_Temp +
    Initial_Tank_Pressure * Petrol_Pressure, pollution_data)
plot_model(b3b4_lm, type = "int", terms = c("Initial_Tank_Pressure", "Petrol_Pressure"))
interaction_lm <- lm(Escaping_Hydrocarbons ~ Tank_Temp + Petrol_Temp + Petrol_Pressure + Tank_Te
mp * Petrol_Temp + Tank_Temp * Petrol_Pressure, pollution_data)
summary(interaction_lm)
library(tidyverse)
library(leaps)
pollution_data.ms <- pollution_data[2:5]

best_subset <- regsubsets(Escaping_Hydrocarbons~ ., pollution_data.ms, nvmax = 4)
results <- summary(best_subset)
tibble(predictors = 1:4,
       R2 = results$rsq,
       Cp = results$cp) %>%
  gather(statistic, value, -predictors) %>%
  ggplot(aes(predictors, value, color = statistic)) +
  geom_line(show.legend = F) +
  geom_point(show.legend = F) +
  facet_wrap(~ statistic, scales = "free")
# Hypothesis test for beta3 = 0 or not
beta_hat.3 <- pollution_lm$coefficients[4] # beta_3 hat, estimated regression coef for intitial
 tank pressure variable
std.beta3 <- 2.85096732
t.stat <- beta_hat.3/std.beta3
abs(t.stat)
alpha <- .05
alpha/2
df <- 32 - 5
df
t.quantile <- qt(1-alpha/2, 27)
t.quantile
plot(new.lm)
set.seed(100)
y.mean <- mean(new.lm$fitted.values)
y.std <- sd(new.lm$fitted.values)
ks.test((new.lm$fitted.values - y.mean)/y.std, rnorm(32))
anova(new.lm)
qf(1-0.05, 3, 28)
# MSR/MSE
((1857.1133 + 494.4345 + 151.6120)/3) / (7.797866)
bptest(new.lm)
qchisq(1-0.05,3)
confint(new.lm)
```