STAT332: Generalised Linear Models

Autumn 2023

Assignment 2

Ruben Traicevski 6790021

Q1

Before we begin, we shall note the well-known following properties of the influence matrix, H.

$$H\ is\ symmetric\colon H^T = H$$

$$H\ is\ idempotent\colon H^2 = H$$

$$H1 = 1, \qquad H = 1,$$

a)

Using the linear regression model.

$$Y = X\beta + e \ \rightarrow e = y - X\beta = e = Y - \hat{Y}$$

$$\Longrightarrow \sum_{i=1}^{n} \widehat{e_i} = \widehat{e_i}^T 1 = (Y - \hat{Y})^T 1$$

From here, we will include,

$$\beta = (X^T X)^{-1} X^T Y$$

$$\Longrightarrow \hat{Y} = X\hat{\beta} = X(X^T X)^{-1} X^T Y = HY, \qquad by\ definition\ of\ H$$

$$Continuing\ on\ from\ previously,$$

$$\sum_{i=1}^{n} \widehat{e_i} = \widehat{e_i}^T 1 = (Y - \hat{Y})^T 1$$

$$(Y - \hat{Y})^T 1 = (Y - HY)^T 1 = Y^T (I - H)^T 1, \qquad where\ I\ is\ an\ identity\ matrix$$

From equation 1, we can claim,

$$\sum_{i=1}^{n} \widehat{e_i} = (Y - \hat{Y})^T 1 = (Y - HY)^T 1 = Y^T (I - H)^T 1$$

$$= Y^T (I - H)^T H = Y^T (H - H^T H)$$

$$Note, the\ last\ step\ above\ is\ made\ as\ \ the\ Tranpose\ respects\ addition, and\ I^T = I,$$
$$IH = H$$

$$Then\ as\ the\ influence\ matrix\ is\ symmetic\ amd\ idempotent,$$

$$Y^T (H - H^T H) = Y^T (H - H) = Y(0) = 0$$

$$\Longrightarrow \sum_{i=1}^{n} \widehat{e_i} = 0$$

B)

Using the definitions in A) and the following two properties

$$E\left[\hat{Y}\right] = E[HY] = HE[Y] = HX\beta + HE[e] = X\beta$$

$$E[e] = 0$$

$$Then\ we\ show\ the\ following$$

$$E[\hat{e}] = E\left[Y - \hat{Y}\right] = E[Y - HY] = E[(I - H)Y] = E[(I - H)(X\beta + e)]$$

$$Then\ by\ definition\ of\ the\ expectation\ operator$$

$$= (I - H)(X\beta + E[e])$$

$$= (\mathrm{I} - \mathrm{H})(\mathrm{X}\beta) = \mathrm{X}\beta - \mathrm{HX}\beta = \mathrm{X}\beta - \mathrm{X}\beta = 0$$

$$Where\ the\ last\ step\ was\ made\ by\ I\mathrm{X}\beta = \mathrm{X}\beta, \qquad \mathrm{X}\beta\mathrm{H} = HH = H$$

c)

$$For\ a\ model\ of\ the\ form, Y = X\beta + e, there\ is\ a\ var - covar\ matrix := \sigma^2 I$$

$$Var\left[\hat{Y}\right] = \sigma^2 H, \qquad Var[e] = \sigma^2 I$$

$$Var[\hat{e}] = Var[(I - H)Y] = Var[(I - H)(\mathrm{X}\beta + \mathrm{e})] = Var[(I - H)e], \qquad as\ H = \mathrm{X}\beta$$

$$= (I - H)^2\sigma^2 I = \sigma^2(I - H)$$

$$\Longrightarrow (I - H)^2\sigma^2 I = (I^2 - 2IH + H^2)I\sigma^2 = (I - 2H + H)I\sigma^2 = (II - HI)\sigma^2 = (I - H)\sigma^2$$

$$H, I\ is\ idempotent: H^2 = H, \qquad IA = A\ where\ A\ is\ any\ m\ x\ n\ matrix$$

Q2 a)

```
 data_import <- read_excel("C:/Users/New User/Desktop/UNIVERSITY
RESOURCES/STATISTICS/STAT332/assignment 2/data import.xlsx")

riverModel <- lm(formula = Nitrogen ~ Agr + Forest + Rsdntial + ComIndl, data = data_import)

summary(riverModel)

boxplot(data_import$Rsdntial, #suspected outliers

     ylab = "Rsdntial")

boxplot.stats(data_import$Rsdntial)$out

boxplot(data_import$ComIndl, #suspected outliers

     ylab = "ComIndl")

boxplot.stats(data_import$ComIndl)$out

pairs(~Nitrogen + Agr +  Forest + Rsdntial + ComIndl, data = data_import)

plot(lm(formula = Nitrogen ~ Agr + Forest + Rsdntial + ComIndl, data = data_import))
```

Nitrogen
Agr
Forest
Rsdntial
ComIndl
0 20 40
0 10 25
0.8 1.8
0 30
30 70
0 20
0.0 2.5
0.8 1.6
30 60 90
0.0 1.5 3.0

ComIndl
0.0 1.0 2.0 3.0

Rsdntial
0 5 15 25

Q2 a)

The pair-wise scatter plot found above is indicating that Nitrogen concentrations between 'Agr' and 'Forest' has a (+ve, -ve) collinear relationship, whereas between 'Rsdntial' and 'comIndl', there is no indication of a significant relationship with other variables. This is expected as Nitrogen concentrations in soil is vital for plant / crop. The variables 'Rsdntial' and 'comIndl' are seen to have no relationship with Nitrogen concentration either, meaning it is advisable to omit them from the model, or the current form of the model is not appropriate.

Bivariate outliers were detected and can clearly be seen with the box plots produced above in the cases of 'Rsdntial' and 'comIndl'.

We can also reinforce the notion of collinearity present by presenting that the VIF is found to be 3.44 >1 and <5, meaning slight to moderate collinearity.

Q2 b)

riverModel <- lm(formula = Nitrogen ~ Agr + Forest + Rsdntial + ComIndl, data = data_import)

summary(riverModel)

β_0	β_1	β_2	β_3	β_4
1.722214	0.005809	-0.012968	-0.007227	0.305028
Pr(>|t|)=0.1832	Pr(>|t|)=0.7046	Pr(>|t|)=0.3667	Pr(>|t|)=0.8337	Pr(>|t|)=0.0823

RSE	R-Squared	F-Statistic	P-Value
0.2649	0.7094	9.154	0.0005963

From the output, it is immediately seen that all predictor variables were found to have no statistically significant relationship with the response variable as every Pr(>|t|) > 0.05 significance level. However, the p-value for the linear regression is 0.000596 < 0.05 meaning we can reject the null hypothesis, meaning we are having sufficient evidence to conclude a $\beta \neq 0$

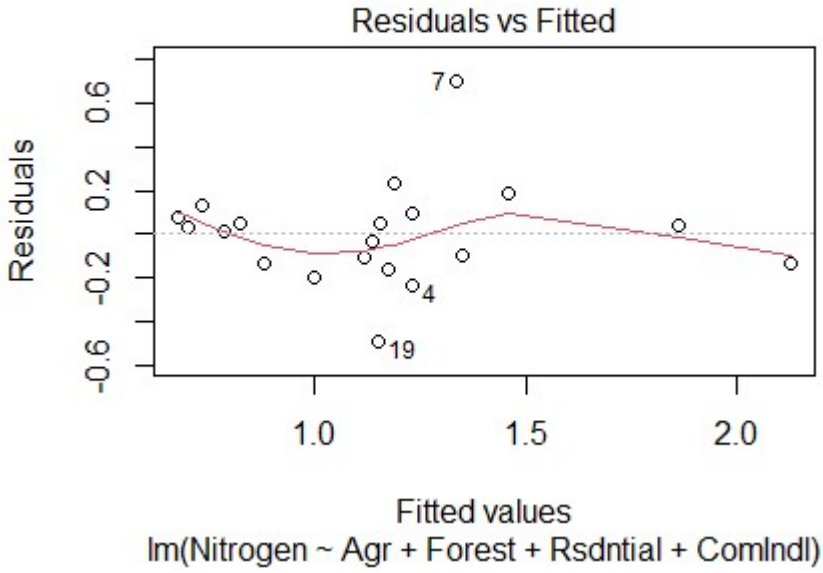
The F-Statistic yielded 9.154 which > f crit value of 3.056.. at a 0.05 significance level. Thus, we can reject the null hypothesis and conclude all the regression coefficients are $\neq 0$.

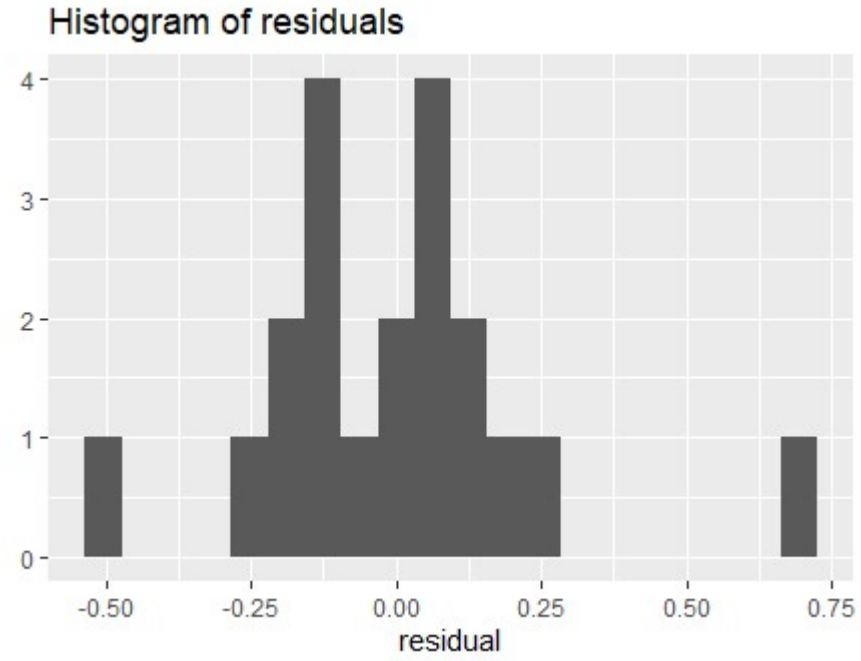
The goodness of fit measure, R-Squared, was given to be 70.94% which indicates the model is moderately suited to explain the variation of the response variable. I.e, it is able to explain 70% of the variation.

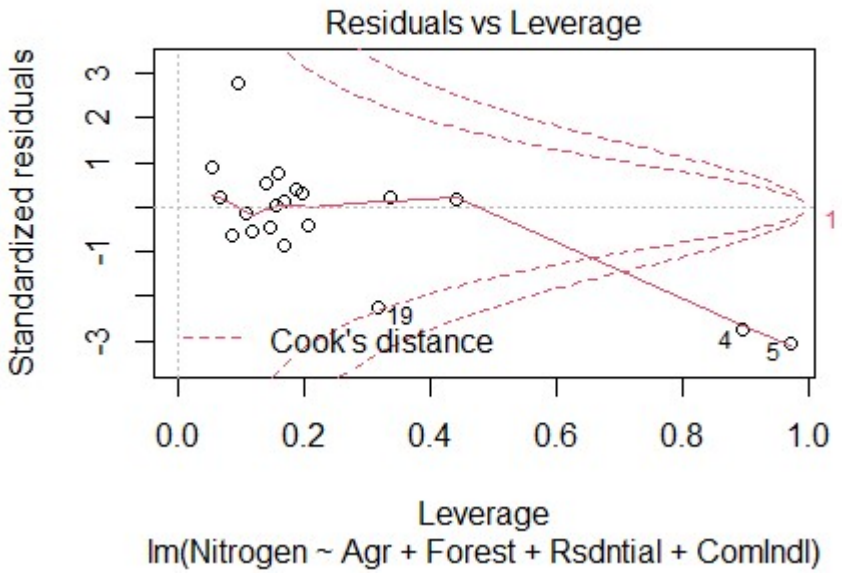
The RMSE was found to be 0.2294 which is low, indicating the current model is able to fit the dataset.

Q2 c)

```
plot(lm(formula = Nitrogen ~ Agr + Forest + Rsdntial + ComIndl, data = data_import))
residualsNitr <- residuals(riverModel)
p1 <- qplot(residualsNitr,
        geom = "histogram",
        bins = 20) +
 labs(title = "Histogram of residuals",
     x = "residual")
p1
```







Q2 d)

```
#setup influence matrix H = X(transpose(X)*X)inv * transpose(X)
dmatrix <- model.matrix(riverModel) #The design matrix X
## calculating hat matrix, H
hmatrix <- dmatrix%*%(solve(t(dmatrix)%*%dmatrix))%*%t(dmatrix)
## Then to obtain our influence scores, use diag function for diagonal elements
inflscores <- diag(hmatrix)
## Cooks distance calculation
cd <- cooks.distance(riverModel) ##Using inbuilt function
# using equation with 4 parameters
yhat = predict(riverModel)
yobs = data_import$Nitrogen
yhatvar = var(yhat)
cd2 = (((yobs - yhat)^2)/((4+1)*yhatvar^2))*(inflscores/((1-inflscores)^2))
cd2 # 2nd method of cooks distance calculation
# Using D > F(0.5)(p,n-p-1) criteria
fstat <- df(0.5, df1=4,df2=20-4-1)
cd > fstat # obs 4, 5 are influential
```

cd2 > fstat # obs 4, 5 are influential

Locate influential observations using influence scores

inflscores > 2*4/20 # obs 3 and 4, 5 are influential

```
> cd2 = (((yobs - yhat)^2)/((4+1)*yhatvar^2))*(inflscores/((1-inflscores)^2))
> cd2 # 2nd method of cooks distance calculation
           1            2            3            4            5            6            7
1.885178e-03 2.909821e-02 2.160834e-02 5.071483e+01 2.509982e+02 3.488965e-02 6.303298e-01
           8            9           10           11           12           13           14
8.781966e-02 2.527104e-02 2.095362e-03 2.802097e-02 2.739354e-02 2.252445e-03 2.279117e-04
          15           16           17           18           19           20
2.054156e-02 3.669021e-02 1.095332e-01 1.704625e-02 1.796151e+00 3.653899e-02
```

```
> cd
           1            2            3            4            5            6            7            8
4.913989e-04 7.584870e-03 5.632525e-03 1.321955e+01 6.542632e+01 9.094491e-03 1.643046e-01 2.289146e-02
           9           10           11           12           13           14           15           16
6.587261e-03 5.461864e-04 7.304071e-03 7.140521e-03 5.871323e-04 5.940847e-05 5.354455e-03 9.563833e-03
          17           18           19           20
2.855140e-02 4.443351e-03 4.681928e-01 9.524415e-03
```

```
> # Using D > F(0.5)(p,n-p-1) criteria
> fstat <- df(0.5, df1=4,df2=20-4-1)
> cd > fstat # obs 4, 5 are influential
    1     2     3     4     5     6     7     8     9    10    11    12    13    14    15    16
FALSE FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
   17    18    19    20
FALSE FALSE FALSE FALSE
> cd2 > fstat # obs 4, 5 are influential
    1     2     3     4     5     6     7     8     9    10    11    12    13    14    15    16
FALSE FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
   17    18    19    20
FALSE FALSE  TRUE FALSE
>
> # Locate influential observations using influence scores
> inflscores > 2*4/20 # obs 3 and 4, 5 are influential
    1     2     3     4     5     6     7     8     9    10    11    12    13    14    15    16
FALSE FALSE  TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
   17    18    19    20
FALSE FALSE FALSE FALSE
```

```
> inflscores
          1           2           3           4           5           6           7           8           9
0.10599034 0.08326469 0.44166295 0.89678338 0.97208197 0.05306997 0.09505121 0.15875826 0.14358055
         10          11          12          13          14          15          16          17          18
0.06411348 0.18575308 0.11518135 0.16711604 0.15546492 0.19637030 0.13841724 0.16837203 0.33677914
         19          20
0.31543831 0.20675079
```

The general criteria for designating an observation as an outlier is any observation with a Cook's Distance > 4/n. Other criteria involves finding observations with a Cook's Distance > f statistic which was also included above. All three methods above are in accordance with each other, finding observation 4 and 5 influential, and when the criteria was set to 2*4/n, the 3rd observation was found to be influential. Their respective influence scores are tabulated below

Obs 3	Obs 4	Obs 5
0.4416	0.8967	0.9720

These findings were highlighted before in the Residuals vs Leverage plot where observation 4 and 5 were clear outliers.

Q3

a)

We want to determine whether the performance of a diesel engine depends on fuel type used. With there being 3 different fuel types, we will have following linear regression model

1. Allocate the 14 data points collected into the following 3 groups

Blended (fuel = 1)

Advanced Timing (fuel = 2)

DF-2 (fuel = 3)

We want to make fuel type a factor variable.

2. State the linear regression model

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + e_i$$

Let $x_{i,1}$ be the brake power

Let $x_{i,2}$ be a dummy variable of fuel $= 1$

Let $x_{i,3}$ be a dummy variable of fuel $= 2$

$\beta_i, e_i, be\ the\ paramters\ and\ residuals\ respectively.$

$y_i\ is\ the\ response\ (mass\ burning\ rate)$

Note: One will notice the last fuel type not mentioned. This fuel type is essentially 'sacrificed' as the base case.

3. Convert model to matrix format

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{e}$$

$$Response\ Vector: \boldsymbol{y_{14\ x\ 1}} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{14} \end{pmatrix}$$

$$Vector\ of\ Parameters: \boldsymbol{\beta_{3x\ 1}} = \begin{pmatrix} \boldsymbol{\beta_0} \\ \boldsymbol{\beta_1} \\ \boldsymbol{\beta_2} \end{pmatrix}$$

$$Design\ Matrix: \boldsymbol{X_{14\ x\ 3}} = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} \\ 1 & x_{2,1} & x_{2,2} \\ \vdots & \vdots & \vdots \\ 1 & x_{14,1} & x_{14,2} \end{pmatrix}$$

$$Vector\ of\ Residuals: \boldsymbol{e_{14\ x\ 1}} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_{14} \end{pmatrix}$$

Q3 b)

```
data_import <- read_excel("C:/Users/New User/Desktop/UNIVERSITY
RESOURCES/STATISTICS/STAT332/assignment 2/fueltype.xlsx")

y <- fueltype$BURNRATE

x1 <- fueltype$BRAKE

ft <- fueltype$FUEL

ftl <- c(0,1,2,0,1,2,0,1,2,0,1,2,0,1) ##Hard code the fuel types

#Combine data into a dataframe and apply a factor variable

obsDF <- data.frame(Y=y, X1=x1, FTL=ftl)

obsDF$FTL<-factor(obsDF$FTL, levels=0:2,labels=c("DF-2","Blended","Advanced Timing"))

lapply(obsDF[c("X1","FTL")],unique) ##So that these vectors return a list

#Creating the design matrix manually

n <- length(y)

p <- 2

dmatrix <- cbind(rep(1,n), x1, ftl)

fit <- lm(Y ~ X1 + FTL, data=obsDF)

summary(fit)

##MSE

MSE <- mean(residuals(fit)^2)

#RMSE

RMSE <- sqrt(MSE)

#Vector of Parameters Estimation

# B = (X'X)inv*X'Y

bhat <- solve(t(dmatrix)%*%dmatrix)%*%t(dmatrix)%*%y
```

β_0	β_1 (Brake)	β_2 (Combined Fueltype)	RMSE
-8.637143	4.435238	11.485714	6.805717

Q3 C)

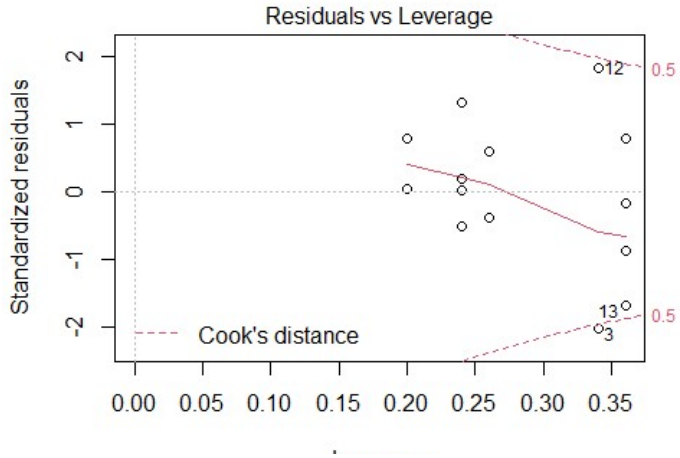
fit <- lm(Y ~ X1 + FTL, data=obsDF)

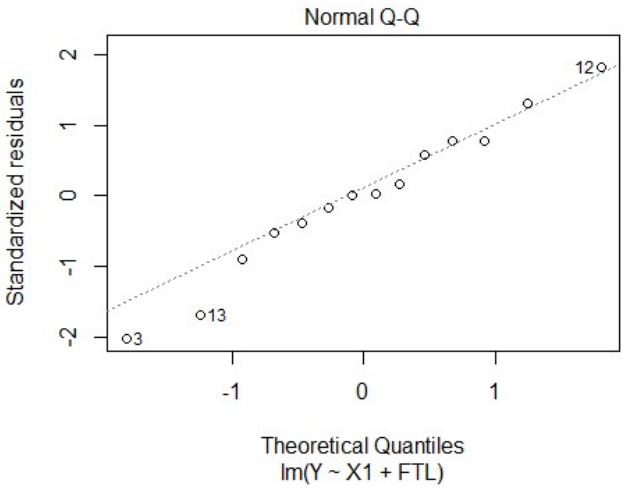
summary(fit) ### Using manually made design matrix

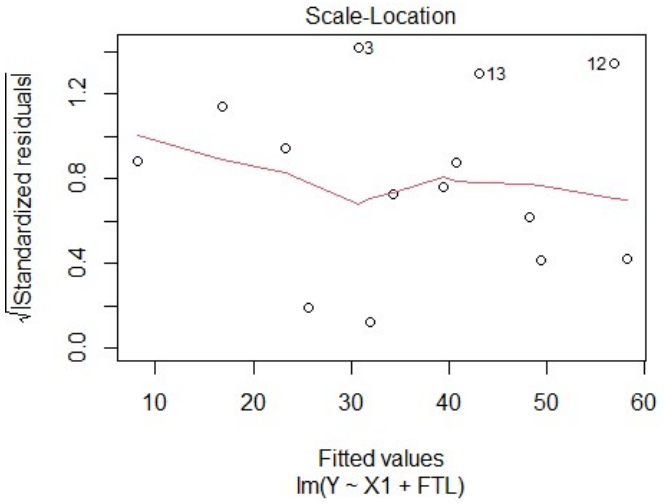
fit2 <- lm(BURNRATE ~ BRAKE + FUEL, data=fueltype)

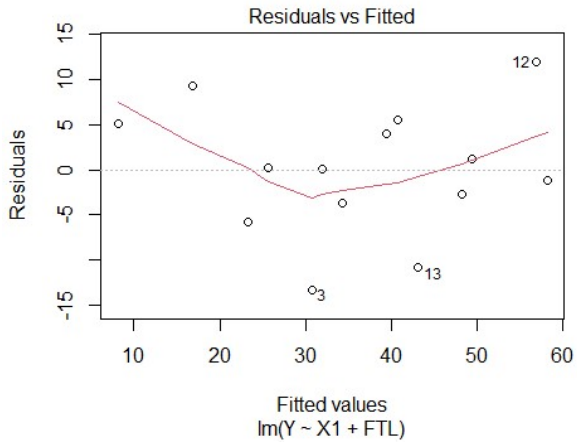
summary(fit2) ### Let R take the imported data and applying factor levels to model

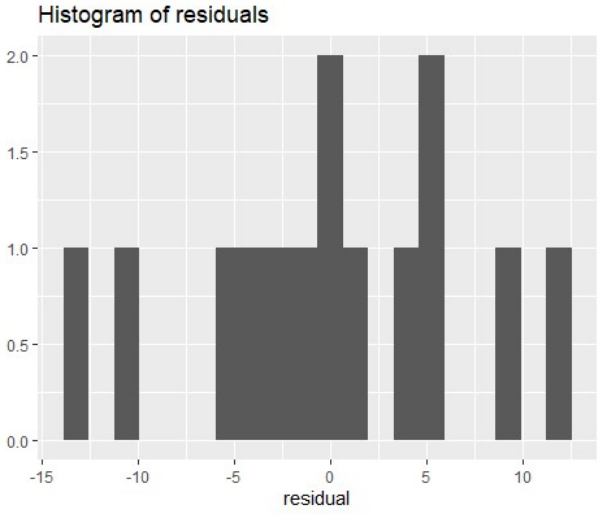
RSE	R-Squared	F-Statistic	P-Value
8.053	81.25%	14.45	0.0005759











From the evidence collected above, we can conclude that the model is an adequate choice. The R Squared was found to be ~ 81%, implying 81% of the variation of the Burn Rate can be explained by the Fuel Type and Brake Power, factored into the 3 fuel types. The p-value for the model was found to be 0.005759 < 0.05 significance level, meaning we can reject null hypothesis, concluding $\beta \neq 0$.

The RMSE was found to be 6.805717 which is low relative to the data, indicating the current model is able to fit the dataset.

The F statistic for the model is = 14.45 which is > 3.708 critical value for 3 on 10 degrees of freedom, meaning we can further support rejecting the null hypothesis and conclude $\beta \neq 0$.

As for the validity of the model in terms of OLS assumptions, the plots above highlight that the data does not violate key definitions.

Residuals vs Fitted is showing that the data is spread at random about the horizontal axis, highlighting that the error terms are i.i.d and have a constant variance. The randomness of the data also shows that E[e] = 0. The sum of the residual vector in R yielded 3.663736e-15 ~ 0.

Normal Q-Q is showing that the residuals are following a normal distribution, furthering this is the histogram of residuals showing a clustering around 0 with a distinct single peak. Taking a Shapiro Wilk test of the residuals found that the p-value = 0.9773 > 0.05 implying the distribution of data is not significantly different from a normal distribution

However, the Residuals vs Leverage is showing that multicollinearity exists as some of the data is following a vertical line. The Variance Inflation Factor was found to be 5.333857, indicating a moderate to high degree level of collinearity. This is expected if we note that from the description of fuel types, Blended is a mixture of DF-2 and coal derived, in which DF-2 also appears as another fuel type. But as there is no perfect multicollinearity, we can still use the linear regression model.