Linear Regression for MPG Prediction

# Objective:

The data concerns city-cycle fuel consumption in miles per gallon to be predicted in terms of 3 multivalued discrete and 5 continuous attributes having 398 instances.

# Data Description:

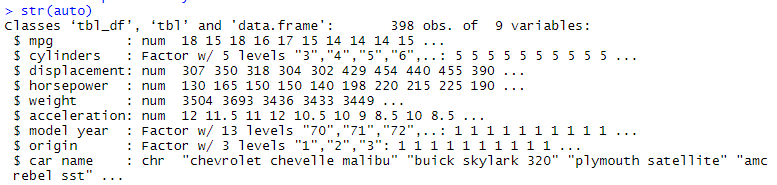
|  |  |
| --- | --- |
| **Variable Name** | **Variable Type** |
| Mpg | Continuous |
| Cylinders | Multi-valued discrete |
| Displacement | Continuous |
| Horsepower | Continuous |
| Weight | Continuous |
| Acceleration | Continuous |
| model year | Multi-valued discrete |
| Origin | Multi-valued discrete |
| car name | String (unique for each instance) |

# Approach:

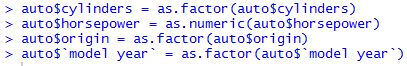
1. EDA – Missing values were treated, variables were changed to factor and numeric, Histogram & boxplot were plotted.
2. Model fitting was done, Adjusted R2 = **86.35%.**
3. Outliers were treated using **Cook’s distance**.
   1. Cook’s distance is used in Regression Analysis to find influential outliers in a set of predictor variables. In other words, it’s a way to identify points that negatively affect your regression model. The measurement is a combination of each observation’s leverage and residual values; the higher the leverage and residuals, the higher the Cook’s distance. (<https://www.statisticshowto.datasciencecentral.com/cooks-distance/>)
   2. Here, we investigate any point over 4/n, where n is the number of observations. Those points were dropped
4. Model fitting was done again, and assumptions were verified
   1. 3 out of the 5 underlying assumptions: Homoscedasticity, Multicollinearity and Auto-correlation were being violated
5. Using Box-cox transformation, Response variable was transformed, since value of Lambda was obtained close to 0, **Log Transformation** was done on y.
6. Outliers were then treated again, based on the new model to satisfy the assumptions and final Adjusted R2 = **90.13%;** **MAPE = 2.5%**

# Codes and Output:

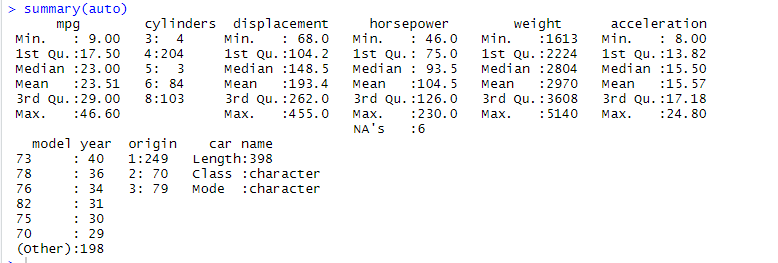
* 1. Data structure



* 1. Changing the variable formats:



* 1. Data summary:

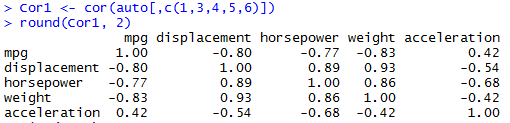


* 1. Treating Missing Values:

Replacing missing values in Horsepower with **Mean**

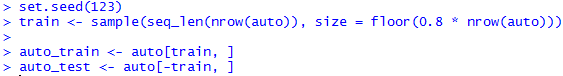
> auto$horsepower[is.na(auto$horsepower)] <- mean(auto$horsepower, na.rm = TRUE)

* 1. Correlation between the continuous variables:



Thus, we can see the independent variable mpg is highly correlated to displacement, horsepower and weight.

* 1. Splitting of train data and test data into 80:20 ratio:



# Data Visualization:

Boxplot helps identify the outliers present in the data and Histogram helps visualize the distribution.

* 1. Boxplot:

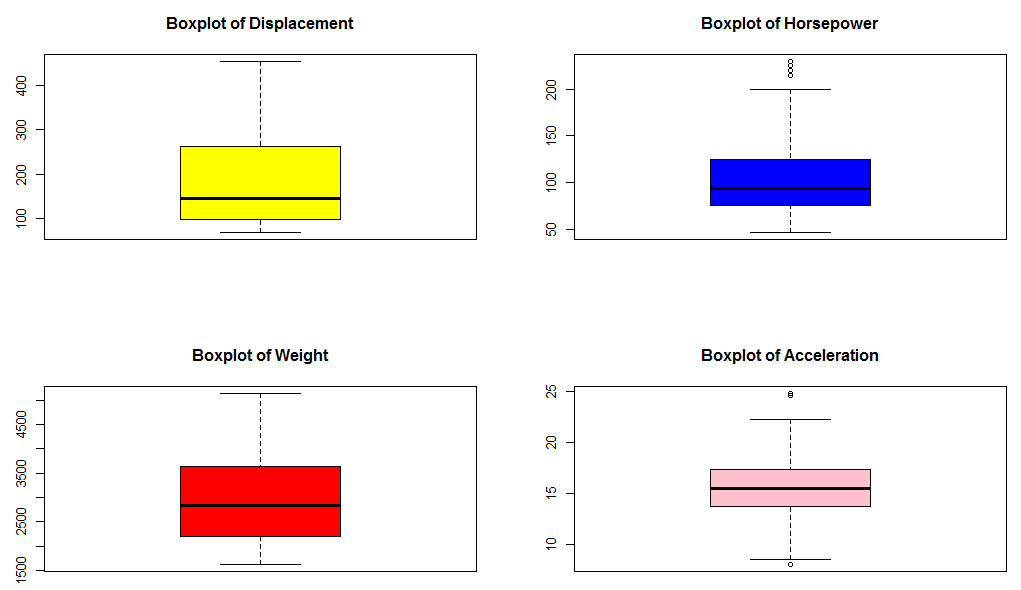
> par(mfrow = c(2,2))

> boxplot(auto\_train$displacement, col = "Yellow", main = "Boxplot of Displacement")

> boxplot(auto\_train$horsepower, col = "Blue", main = "Boxplot of Horsepower")

> boxplot(auto\_train$weight, col = "Red", main = "Boxplot of Weight")

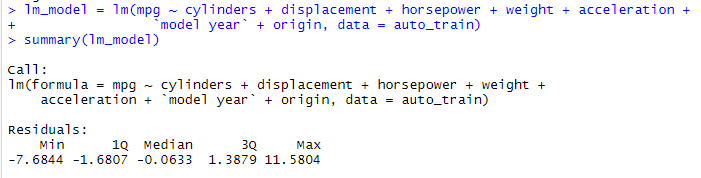
> boxplot(auto\_train$acceleration, col = "Pink", main = "Boxplot of Acceleration")

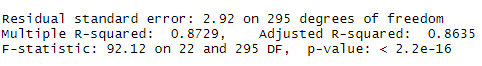


We can see that there are outliers present in Horsepower and acceleration variable.

# Model Fitting

5.1 Fitting a Linear Regression Model



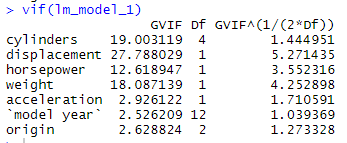


**Adjusted R2 value = 86.35%**

5.2 Treating Outliers Using Cook’s distance:

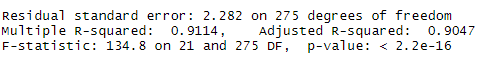


5.3 Checking VIF value:

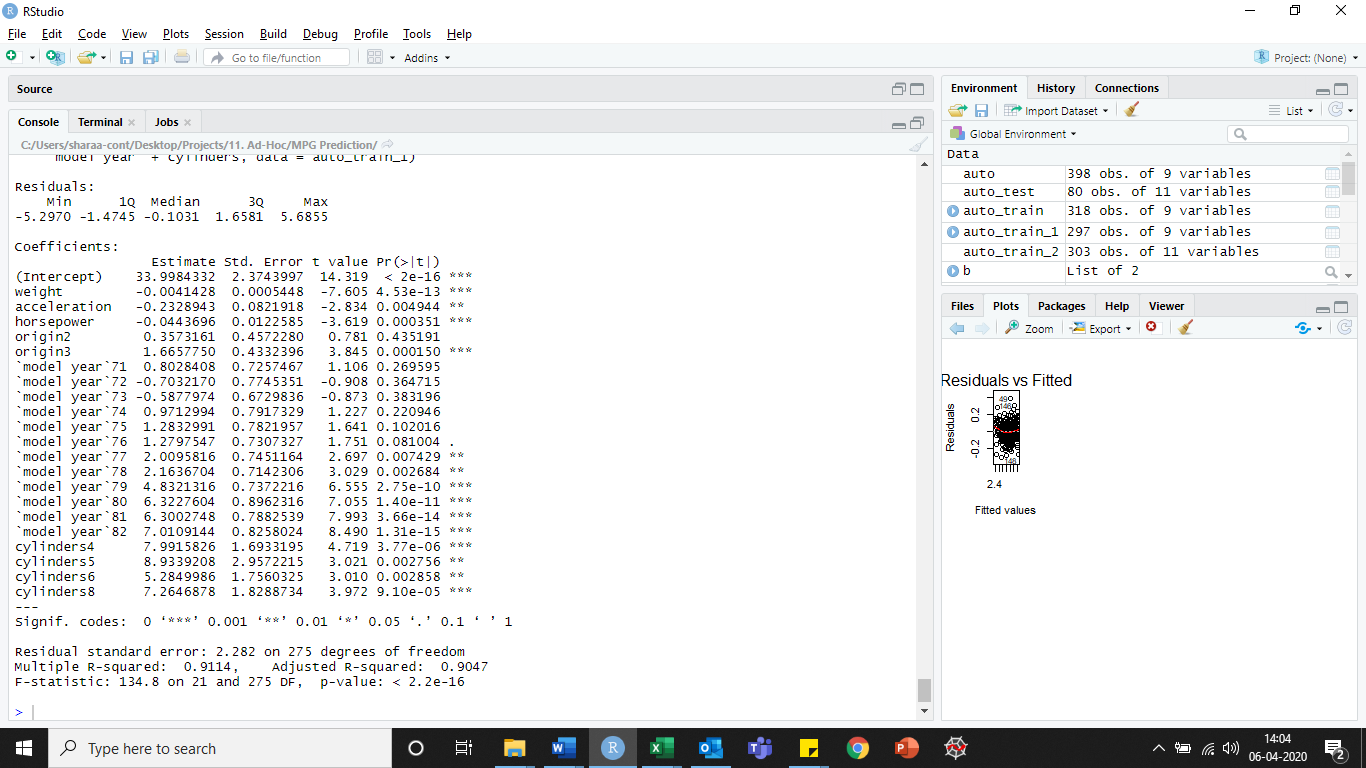


Removing ‘Displacement’ variable since the VIF value > 5

5.4 Model accuracy after removing Outliers & Displacement



> summary(lm\_model)



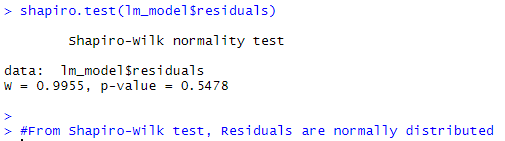
**Adjusted R2 value = 90.47%**

**Linear Regression:**

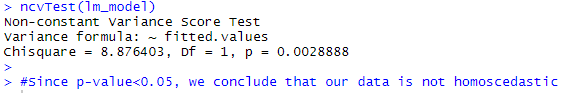
|  |  |
| --- | --- |
| **Data** | **RMSE Value** |
| **Train Data** | 2.195574 |
| **Test Data** | 3.02862 |

## Model Adequacy Checks:

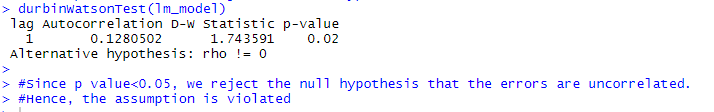
1. Normality Assumption –



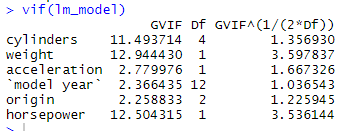
1. Homoscedasticity Assumption –



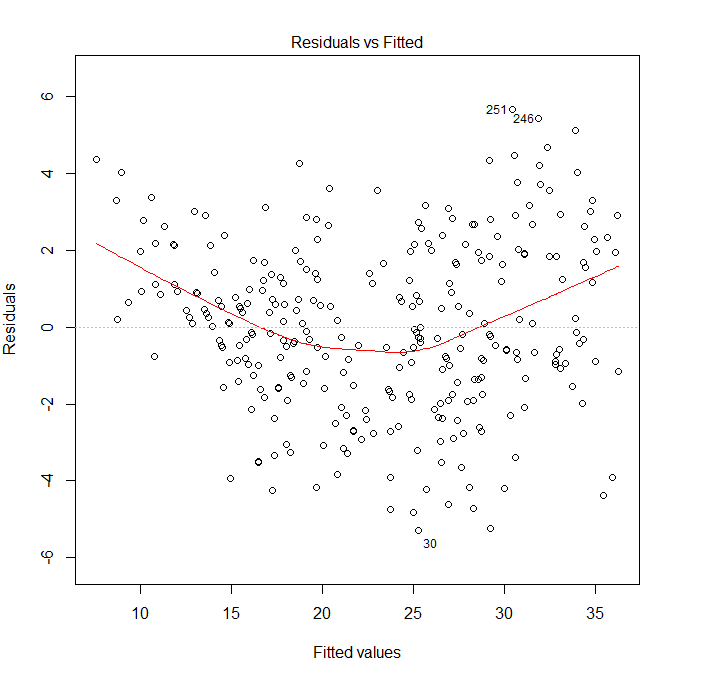
1. Auto-correlation Assumption –



1. Multicollinearity Assumption –



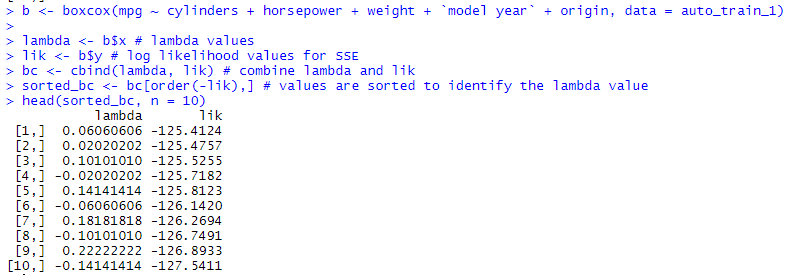
1. Linearity Assumption –



There is a non-linearity present in this model.

Since **3 out of 5 assumptions for the model were not satisfied**, we go for **Box-Cox transformation** in order to satisfy the assumptions.

## 5.2 Box-Cox Transformation

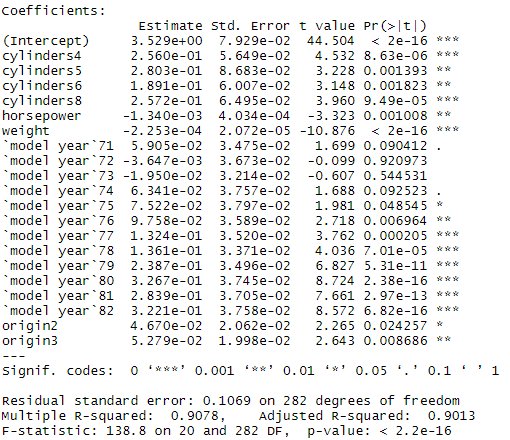


Since the lambda value obtained for Box-Cox transformation is close to zero, we go for **Log Transformation** of the Dependent Variable, mpg.

After treating the outliers using Cook’s Distance and fitting the model again, Acceleration variable is dropped since it is not significant.

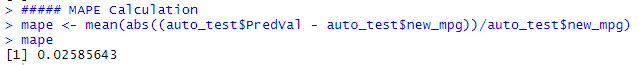
The final model is:

Model Summary –



**The Adjusted R2 is: 90.13%**

MAPE calculation:



Therefore, MAPE (Mean Absolute Percentage Error) = 2.5% which is good.

Thus, our model gives good prediction.

> rmse\_train\_lm = sqrt(mean((auto\_train\_2$PredVal - auto\_train\_2$log\_mpg)^2))

> rmse\_test\_lm = sqrt(mean((auto\_test$PredVal - auto\_test$new\_mpg)^2))

> rmse\_train\_lm

[1] 0.103158

> rmse\_test\_lm

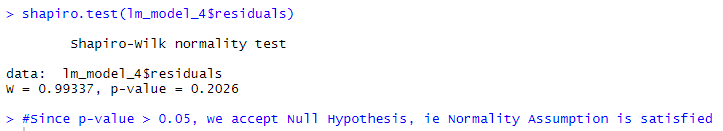
[1] 0.1010246

**Log Transformation**:

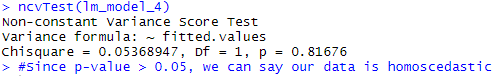
|  |  |
| --- | --- |
| **Data** | **RMSE Value** |
| **Train Data** | 0.103158 |
| **Test Data** | 0.1010246 |

## 5.2.1 Model Adequacy Checks:

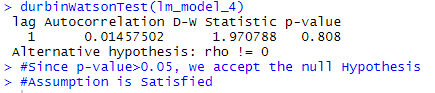
1. Normality Assumption –



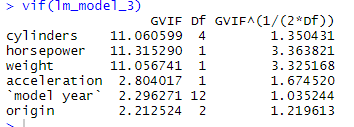
1. Homoscedasticity Assumption –

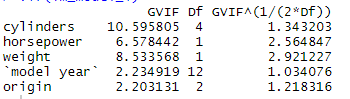


1. Auto-correlation Assumption –



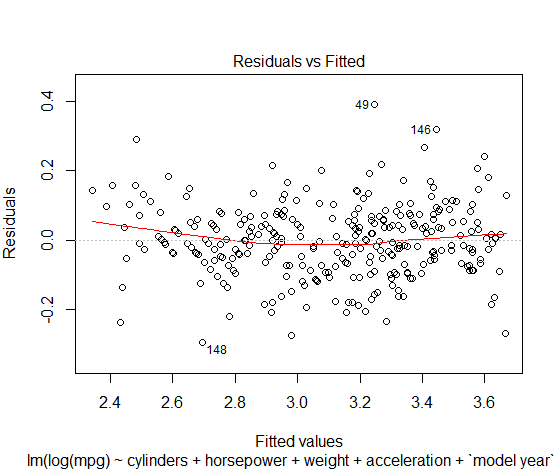
1. Multicollinearity Assumption –





Since all values are less than 5, there is no multicollinearity present in the data.

1. Linearity Assumption –



The linearity Assumption holds true.

Hence all 5 assumptions for Linear Regression Model are being Satisfied.

# STEPWISE REGRESSION

## **Forward Selection**:

> # Forward Selection Regression

> modelFit\_step\_forward = lm(mpg ~ cylinders +

+ displacement +

+ horsepower +

+ weight +

+ acceleration +

+ `model year` +

+ origin, data = data)

>

> step\_forward = stepAIC(modelFit\_step\_forward, direction = "forward")

Start: AIC=860.35

mpg ~ cylinders + displacement + horsepower + weight + acceleration +

`model year` + origin

>

> summary(step\_forward)

Call:

lm(formula = mpg ~ cylinders + displacement + horsepower + weight +

acceleration + `model year` + origin, data = data)

Residuals:

Min 1Q Median 3Q Max

-8.0826 -1.6818 0.0262 1.5066 11.4175

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 30.878343 2.345504 13.165 < 2e-16 \*\*\*

cylinders4 6.839398 1.545279 4.426 1.26e-05 \*\*\*

cylinders5 6.585214 2.350852 2.801 0.005355 \*\*

cylinders6 4.420074 1.714599 2.578 0.010321 \*

cylinders8 6.580929 1.978074 3.327 0.000965 \*\*\*

displacement 0.012479 0.006782 1.840 0.066553 .

horsepower -0.035946 0.012387 -2.902 0.003927 \*\*

weight -0.005480 0.000610 -8.984 < 2e-16 \*\*\*

acceleration 0.019874 0.084457 0.235 0.814091

`model year`71 1.018528 0.806324 1.263 0.207311

`model year`72 -0.365328 0.806840 -0.453 0.650963

`model year`73 -0.468886 0.724316 -0.647 0.517801

`model year`74 1.417946 0.846511 1.675 0.094759 .

`model year`75 1.028232 0.837068 1.228 0.220077

`model year`76 1.627644 0.802868 2.027 0.043340 \*

`model year`77 3.136471 0.821780 3.817 0.000158 \*\*\*

`model year`78 3.090568 0.781433 3.955 9.15e-05 \*\*\*

`model year`79 5.034246 0.826127 6.094 2.74e-09 \*\*\*

`model year`80 9.162403 0.855399 10.711 < 2e-16 \*\*\*

`model year`81 6.727138 0.852566 7.890 3.34e-14 \*\*\*

`model year`82 7.881799 0.845452 9.323 < 2e-16 \*\*\*

origin2 1.927370 0.510788 3.773 0.000187 \*\*\*

origin3 2.349359 0.491419 4.781 2.51e-06 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.866 on 375 degrees of freedom

Multiple R-squared: 0.873, Adjusted R-squared: 0.8656

F-statistic: 117.2 on 22 and 375 DF, p-value: < 2.2e-16

> step\_forward$anova

Stepwise Model Path

Analysis of Deviance Table

Initial Model:

mpg ~ cylinders + displacement + horsepower + weight + acceleration +

`model year` + origin

Final Model:

mpg ~ cylinders + displacement + horsepower + weight + acceleration +

`model year` + origin

Step Df Deviance Resid. Df Resid. Dev AIC

1 375 3079.617 860.3513

> rmse\_train

[1] 2.651712

> rmse\_test

[1] 3.146338

**Forward Selection**:

|  |  |
| --- | --- |
| **Data** | **RMSE Value** |
| **Train Data** | 2.651712 |
| **Test Data** | 3.146338 |

## **Backward Elimination**:

> step\_backward = stepAIC(modelFit\_step\_backward, direction = "backward")

Start: AIC=860.35

mpg ~ cylinders + displacement + horsepower + weight + acceleration +

`model year` + origin

Df Sum of Sq RSS AIC

- acceleration 1 0.45 3080.1 858.41

<none> 3079.6 860.35

- displacement 1 27.80 3107.4 861.93

- horsepower 1 69.16 3148.8 867.19

- origin 2 209.08 3288.7 882.49

- cylinders 4 450.49 3530.1 906.69

- weight 1 662.87 3742.5 935.94

- `model year` 12 2968.63 6048.2 1104.99

Step: AIC=858.41

mpg ~ cylinders + displacement + horsepower + weight + `model year` +

origin

Df Sum of Sq RSS AIC

<none> 3080.1 858.41

- displacement 1 27.35 3107.4 859.93

- horsepower 1 111.18 3191.3 870.52

- origin 2 209.03 3289.1 880.54

- cylinders 4 455.03 3535.1 905.25

- weight 1 824.72 3904.8 950.84

- `model year` 12 2971.56 6051.6 1103.21

> summary(step\_backward)

Call:

lm(formula = mpg ~ cylinders + displacement + horsepower + weight +

`model year` + origin, data = data)

Residuals:

Min 1Q Median 3Q Max

-8.111 -1.715 -0.007 1.478 11.482

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 31.1736365 1.9790949 15.751 < 2e-16 \*\*\*

cylinders4 6.8914564 1.5274395 4.512 8.61e-06 \*\*\*

cylinders5 6.6385952 2.3369398 2.841 0.004746 \*\*

cylinders6 4.4671498 1.7007479 2.627 0.008977 \*\*

cylinders8 6.6163074 1.9698733 3.359 0.000863 \*\*\*

displacement 0.0122661 0.0067129 1.827 0.068454 .

horsepower -0.0375943 0.0102046 -3.684 0.000263 \*\*\*

weight -0.0054135 0.0005395 -10.034 < 2e-16 \*\*\*

`model year`71 0.9991438 0.8010973 1.247 0.213094

`model year`72 -0.3735228 0.8050753 -0.464 0.642945

`model year`73 -0.4790886 0.7221081 -0.663 0.507444

`model year`74 1.3979274 0.8411671 1.662 0.097368 .

`model year`75 1.0040198 0.8296762 1.210 0.226987

`model year`76 1.6068480 0.7969862 2.016 0.044494 \*

`model year`77 3.1176769 0.8168615 3.817 0.000158 \*\*\*

`model year`78 3.0733250 0.7770119 3.955 9.14e-05 \*\*\*

`model year`79 5.0152520 0.8211408 6.108 2.52e-09 \*\*\*

`model year`80 9.1470189 0.8518248 10.738 < 2e-16 \*\*\*

`model year`81 6.7012512 0.8443757 7.936 2.42e-14 \*\*\*

`model year`82 7.8649956 0.8413727 9.348 < 2e-16 \*\*\*

origin2 1.9259304 0.5101090 3.776 0.000186 \*\*\*

origin3 2.3496576 0.4908000 4.787 2.43e-06 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.862 on 376 degrees of freedom

Multiple R-squared: 0.873, Adjusted R-squared: 0.8659

F-statistic: 123.1 on 21 and 376 DF, p-value: < 2.2e-16

> step\_backward$anova

Stepwise Model Path

Analysis of Deviance Table

Initial Model:

mpg ~ cylinders + displacement + horsepower + weight + acceleration +

`model year` + origin

Final Model:

mpg ~ cylinders + displacement + horsepower + weight + `model year` +

origin

Step Df Deviance Resid. Df Resid. Dev AIC

1 375 3079.617 860.3513

2 - acceleration 1 0.454751 376 3080.072 858.4100

> rmse\_train

[1] 2.650192

> rmse\_test

[1] 3.150992

**Backward Elimination**:

|  |  |
| --- | --- |
| **Data** | **RMSE Value** |
| **Train Data** | 2.650192 |
| **Test Data** | 3.150992 |

## **Stepwise Regression:**

> step = stepAIC(modelFit\_step, direction = "both")

Start: AIC=860.35

mpg ~ cylinders + displacement + horsepower + weight + acceleration +

`model year` + origin

Df Sum of Sq RSS AIC

- acceleration 1 0.45 3080.1 858.41

<none> 3079.6 860.35

- displacement 1 27.80 3107.4 861.93

- horsepower 1 69.16 3148.8 867.19

- origin 2 209.08 3288.7 882.49

- cylinders 4 450.49 3530.1 906.69

- weight 1 662.87 3742.5 935.94

- `model year` 12 2968.63 6048.2 1104.99

Step: AIC=858.41

mpg ~ cylinders + displacement + horsepower + weight + `model year` +

origin

Df Sum of Sq RSS AIC

<none> 3080.1 858.41

- displacement 1 27.35 3107.4 859.93

+ acceleration 1 0.45 3079.6 860.35

- horsepower 1 111.18 3191.3 870.52

- origin 2 209.03 3289.1 880.54

- cylinders 4 455.03 3535.1 905.25

- weight 1 824.72 3904.8 950.84

- `model year` 12 2971.56 6051.6 1103.21

> # What did we get?

> summary(step)

Call:

lm(formula = mpg ~ cylinders + displacement + horsepower + weight +

`model year` + origin, data = data)

Residuals:

Min 1Q Median 3Q Max

-8.111 -1.715 -0.007 1.478 11.482

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 31.1736365 1.9790949 15.751 < 2e-16 \*\*\*

cylinders4 6.8914564 1.5274395 4.512 8.61e-06 \*\*\*

cylinders5 6.6385952 2.3369398 2.841 0.004746 \*\*

cylinders6 4.4671498 1.7007479 2.627 0.008977 \*\*

cylinders8 6.6163074 1.9698733 3.359 0.000863 \*\*\*

displacement 0.0122661 0.0067129 1.827 0.068454 .

horsepower -0.0375943 0.0102046 -3.684 0.000263 \*\*\*

weight -0.0054135 0.0005395 -10.034 < 2e-16 \*\*\*

`model year`71 0.9991438 0.8010973 1.247 0.213094

`model year`72 -0.3735228 0.8050753 -0.464 0.642945

`model year`73 -0.4790886 0.7221081 -0.663 0.507444

`model year`74 1.3979274 0.8411671 1.662 0.097368 .

`model year`75 1.0040198 0.8296762 1.210 0.226987

`model year`76 1.6068480 0.7969862 2.016 0.044494 \*

`model year`77 3.1176769 0.8168615 3.817 0.000158 \*\*\*

`model year`78 3.0733250 0.7770119 3.955 9.14e-05 \*\*\*

`model year`79 5.0152520 0.8211408 6.108 2.52e-09 \*\*\*

`model year`80 9.1470189 0.8518248 10.738 < 2e-16 \*\*\*

`model year`81 6.7012512 0.8443757 7.936 2.42e-14 \*\*\*

`model year`82 7.8649956 0.8413727 9.348 < 2e-16 \*\*\*

origin2 1.9259304 0.5101090 3.776 0.000186 \*\*\*

origin3 2.3496576 0.4908000 4.787 2.43e-06 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.862 on 376 degrees of freedom

Multiple R-squared: 0.873, Adjusted R-squared: 0.8659

F-statistic: 123.1 on 21 and 376 DF, p-value: < 2.2e-16

> step$anova

Stepwise Model Path

Analysis of Deviance Table

Initial Model:

mpg ~ cylinders + displacement + horsepower + weight + acceleration +

`model year` + origin

Final Model:

mpg ~ cylinders + displacement + horsepower + weight + `model year` +

origin

Step Df Deviance Resid. Df Resid. Dev AIC

1 375 3079.617 860.3513

2 - acceleration 1 0.454751 376 3080.072 858.4100

> rmse\_train = sqrt(mean((pred\_train - training$mpg)^2))

> rmse\_test = sqrt(mean((pred\_test - testing$mpg)^2))

> rmse\_train

[1] 2.650192

> rmse\_test

[1] 3.150992

**Stepwise Regression**:

|  |  |
| --- | --- |
| **Data** | **RMSE Value** |
| **Train Data** | 2.650192 |
| **Test Data** | 3.150992 |