**Case Study: Linear Regression Analysis**

After pre-processing the mileage data using exploratory data analysis (EDA) the researcher wants to understand how mileage is driven by diﬀerent aspects of automobile design and performance using some appropriate modeling technique. She wants to understand the magnitude, direction and signﬁcance of individual attributes and the overall relationship, and also wants to ensure if the model ﬁts the data well and has a good predictive strength.

Further, she wants to ensure that the model is free from all possible problems and satisﬁes all the assumptions. Also she wants to include as less as possible attributes to understand the mileage without loosing much of goodness of ﬁt, for the reasons of cost and overﬁtting.

Explain how this model can be used for mileage predictions and in what all situations the model will become obselete and what can be the possible remedies in those situations.

**Methodology:**

let y be a response variable and x1,…….xn be n features then we are interested in exploring a functional relationship of the following form:

Y = f(x1…..,xn) + e

Now for the individual regressor we fit a simple linear model like given below:

Y = Xβ + e

After fitting a model we obtain the value of the regression coefficient which describe the relationships between different regressors and the response variable. For this we set up the following hypothesis:

Ho: βj=0, i.e. jth feature is insignificant ; j = 1,…….n and

H1: βj≠0, i.e. jth feature is significant ; j = 1,…….n

**R Code and Output:**

> #SIMPLE MODEL BUILDING

> # Building simple Linear Regression of mpg on cyl

>

> model<-lm(mpg~cyl,data=mileage)

> summary(model)

Call:

lm(formula = mpg ~ cyl, data = mileage)

Residuals:

Min 1Q Median 3Q Max

-4.9814 -2.1185 0.2217 1.0717 7.5186

Coefficients:

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

(Intercept) 37.8846 2.0738 18.27 < 2e-16 \*\*\*

cyl -2.8758 0.3224 -8.92 6.11e-10 \*\*\*

---

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

Residual standard error: 3.206 on 30 degrees of freedom

Multiple R-squared: 0.7262, Adjusted R-squared: 0.7171

F-statistic: 79.56 on 1 and 30 DF, p-value: 6.113e-10

**Conclusion :** At 5% level of significance , Here p value < α then we reject null hypothesis and conclude that there is a linear relationship between feature(cyl) and the response(mpg). The value of adjusted R-squard (>.70) considered as high value is regarded as good model.

> #Building simple Linear Regression of mpg on disp

> model<-lm(mpg~disp,data = mileage)

> summary(model)

Call:

lm(formula = mpg ~ disp, data = mileage)

Residuals:

Min 1Q Median 3Q Max

-4.878 -2.181 -0.948 1.641 7.239

Coefficients:

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

(Intercept) 29.595733 1.233018 24.003 < 2e-16 \*\*\*

disp -0.041276 0.004734 -8.719 1.01e-09 \*\*\*

---

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

Residual standard error: 3.259 on 30 degrees of freedom

Multiple R-squared: 0.717, Adjusted R-squared: 0.7076

F-statistic: 76.02 on 1 and 30 DF, p-value: 1.006e-09

**Conclusion:** At 5% level of significance , Here p value < α then we reject null hypothesis and conclude that there is a linear relationship between feature(disp) and the response(mpg). The value of adjusted R-squard (>.70) considered as high value is regarded as good model.

> #Building simple Linear Regression of mpg on hp

> model<-lm(mpg~hp,data = mileage)

> summary(model)

Call:

lm(formula = mpg ~ hp, data = mileage)

Residuals:

Min 1Q Median 3Q Max

-5.043 -2.172 -1.030 1.823 7.875

Coefficients:

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

(Intercept) 31.26519 1.60093 19.53 < 2e-16 \*\*\*

hp -0.07735 0.01018 -7.60 1.78e-08 \*\*\*

---

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

Residual standard error: 3.582 on 30 degrees of freedom

Multiple R-squared: 0.6581, Adjusted R-squared: 0.6468

F-statistic: 57.76 on 1 and 30 DF, p-value: 1.784e-08

**Conclusion:** At 5% level of significance , Here p value < α then we reject null hypothesis and conclude that there is a linear relationship between feature(hp) and the response(mpg). The value of adjusted R-squard (>.70) considered as high value is regarded as good model. Here the adjusted R-squar less than 0.70 so the model is not good as previous for mpg~disp.

> #Building simple Linear Regression of mpg on drat

> model<-lm(mpg~drat,data = mileage)

> summary(model)

Call:

lm(formula = mpg ~ drat, data = mileage)

Residuals:

Min 1Q Median 3Q Max

-8.9559 -2.6680 -0.0471 2.3228 9.1441

Coefficients:

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

(Intercept) -7.272 5.494 -1.324 0.196

drat 7.590 1.508 5.034 2.12e-05 \*\*\*

---

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

Residual standard error: 4.511 on 30 degrees of freedom

Multiple R-squared: 0.4579, Adjusted R-squared: 0.4398

F-statistic: 25.34 on 1 and 30 DF, p-value: 2.117e-05

**Conclusion: :** At 5% level of significance , Here p value < α then we reject null hypothesis and conclude that there is a linear relationship between feature(drat) and the response(mpg). The value of adjusted R-squard (>.70) considered as high value is regarded as good model. Here the adjusted R-squar less than 0.70 so the model is not good as previous for mpg~disp.

> #Building simple Linear Regression of mpg on wt

> model<-lm(mpg~wt,data = mileage)

> summary(model)

Call:

lm(formula = mpg ~ wt, data = mileage)

Residuals:

Min 1Q Median 3Q Max

-9.4300 -2.0401 -0.4392 2.0033 6.4009

Coefficients:

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

(Intercept) 41.5497 2.6080 15.932 3.5e-16 \*\*\*

wt -7.0685 0.8368 -8.447 2.0e-09 \*\*\*

---

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

Residual standard error: 3.333 on 30 degrees of freedom

Multiple R-squared: 0.704, Adjusted R-squared: 0.6941

F-statistic: 71.35 on 1 and 30 DF, p-value: 1.996e-09

**Conclusion: :** At 5% level of significance , Here p value < α then we reject null hypothesis and conclude that there is a linear relationship between feature(wt) and the response(mpg). The value of adjusted R-squard (>.70) considered as high value is regarded as good model. Here the adjusted R-squar less than 0.70 so the model is not good as previous for mpg~disp.

> #Building simple Linear Regression of mpg on qsec

> model<-lm(mpg~qsec,data = mileage)

> summary(model)

Call:

lm(formula = mpg ~ qsec, data = mileage)

Residuals:

Min 1Q Median 3Q Max

-10.0519 -3.0797 -0.7772 2.3964 11.7628

Coefficients:

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

(Intercept) -9.7588 11.0197 -0.886 0.3829

qsec 1.6802 0.6179 2.719 0.0108 \*

---

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

Residual standard error: 5.487 on 30 degrees of freedom

Multiple R-squared: 0.1977, Adjusted R-squared: 0.171

F-statistic: 7.395 on 1 and 30 DF, p-value: 0.01077

**Conclusion: :** At 1% level of significance , Here p value > α then we accept null hypothesis and conclude that there is a no linear relationship between feature(qsec) and the response(mpg). The value of adjusted R-squard (>.70) considered as high value is regarded as good model. Here the adjusted R-squar less than 0.70 so the model is not good.

> #Building simple Linear Regression of mpg on vs

> model<-lm(mpg~vs,data = mileage)

> summary(model)

Call:

lm(formula = mpg ~ vs, data = mileage)

Residuals:

Min 1Q Median 3Q Max

-6.757 -3.082 -1.267 2.828 9.383

Coefficients:

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

(Intercept) 16.617 1.080 15.390 8.85e-16 \*\*\*

vs 7.940 1.632 4.864 3.42e-05 \*\*\*

---

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

Residual standard error: 4.581 on 30 degrees of freedom

Multiple R-squared: 0.4409, Adjusted R-squared: 0.4223

F-statistic: 23.66 on 1 and 30 DF, p-value: 3.416e-05

**Conclusion: :** At 5% level of significance , Here p value < α then we reject null hypothesis and conclude that there is a linear relationship between feature(vs) and the response(mpg). The value of adjusted R-squard (>.70) considered as high value is regarded as good model. Here the adjusted R-squar less than 0.70 so the model is not good as previous for mpg~cyl.

> #Building simple Linear Regression of mpg on am

> model<-lm(mpg~am,data = mileage)

> summary(model)

Call:

lm(formula = mpg ~ am, data = mileage)

Residuals:

Min 1Q Median 3Q Max

-9.3923 -3.0923 -0.2974 3.2439 9.5077

Coefficients:

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

(Intercept) 17.147 1.125 15.247 1.13e-15 \*\*\*

am 7.245 1.764 4.106 0.000285 \*\*\*

---

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

Residual standard error: 4.902 on 30 degrees of freedom

Multiple R-squared: 0.3598, Adjusted R-squared: 0.3385

F-statistic: 16.86 on 1 and 30 DF, p-value: 0.000285

**Conclusion: :** At 5% level of significance , Here p value < α then we reject null hypothesis and conclude that there is a linear relationship between feature(am) and the response(mpg). The value of adjusted R-squard (>.70) considered as high value is regarded as good model. Here the adjusted R-squar less than 0.70 so the model is not good as previous for mpg~cyl.

> #Building simple Linear Regression of mpg on gear

> model<-lm(mpg~gear,data = mileage)

> summary(model)

Call:

lm(formula = mpg ~ gear, data = mileage)

Residuals:

Min 1Q Median 3Q Max

-10.240 -2.793 -0.205 2.126 12.583

Coefficients:

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

(Intercept) 5.623 4.916 1.144 0.2618

gear 3.923 1.308 2.999 0.0054 \*\*

---

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

Residual standard error: 5.374 on 30 degrees of freedom

Multiple R-squared: 0.2307, Adjusted R-squared: 0.205

F-statistic: 8.995 on 1 and 30 DF, p-value: 0.005401

**Conclusion: :** At 5% level of significance , Here p value < α then we reject null hypothesis and conclude that there is a linear relationship between feature(gear) and the response(mpg). The value of adjusted R-squard (>.70) considered as high value is regarded as good model. Here the adjusted R-squar less than 0.70 so the model is not good

> #Building simple Linear Regression of mpg on carb

> model<-lm(mpg~carb,data = mileage)

> summary(model)

Call:

lm(formula = mpg ~ carb, data = mileage)

Residuals:

Min 1Q Median 3Q Max

-6.720 -3.116 -1.400 3.196 9.542

Coefficients:

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

(Intercept) 26.7970 1.9268 13.907 1.29e-14 \*\*\*

carb -2.4387 0.6231 -3.914 0.000484 \*\*\*

---

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

Residual standard error: 4.985 on 30 degrees of freedom

Multiple R-squared: 0.338, Adjusted R-squared: 0.3159

F-statistic: 15.32 on 1 and 30 DF, p-value: 0.0004836

**Conclusion: :** At 5% level of significance , Here p value < α then we reject null hypothesis and conclude that there is a linear relationship between feature(carb) and the response(mpg). The value of adjusted R-squard (>.70) considered as high value is regarded as good model. Here the adjusted R-squar less than 0.70 so the model is not good.

For testing the significance of overall Regression:

Ho: None of the βj is different from 0, i.e overall regression is insignificant.

H1: Atleast on of the βj is not 0. i.e the overall regression is significant.

R code and output:

> #OVERALL REGRESSION

> model<-lm(mpg~cyl+disp+hp+wt+drat+qsec+vs+am+gear+carb,data = mileage)

> summary(model)

Call:

lm(formula = mpg ~ cyl + disp + hp + wt + drat + qsec + vs +

am + gear + carb, data = mileage)

Residuals:

Min 1Q Median 3Q Max

-3.5450 -1.8809 0.2004 1.2982 4.0875

Coefficients:

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

(Intercept) 13.25695 24.12477 0.550 0.5884

cyl 1.10216 1.07863 1.022 0.3185

disp -0.01420 0.01190 -1.193 0.2462

hp -0.01864 0.02531 -0.736 0.4696

wt -3.02691 1.51893 -1.993 0.0594 .

drat 1.67840 1.69780 0.989 0.3341

qsec 0.34762 0.92851 0.374 0.7119

vs 0.05561 2.28797 0.024 0.9808

am 0.16399 2.16507 0.076 0.9403

gear 1.79851 1.51164 1.190 0.2474

carb -1.37571 0.60666 -2.268 0.0340 \*

---

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

Residual standard error: 2.555 on 21 degrees of freedom

Multiple R-squared: 0.8783, Adjusted R-squared: 0.8203

F-statistic: 15.15 on 10 and 21 DF, p-value: 1.826e-07

**Conclusion :** At 5% level of significance , Here p value < α then we reject null hypothesis and conclude that there is a linear relationship between feature(cyl + disp + hp + wt + drat + qsec + vs +am + gear + carb) and the response(mpg). The value of adjusted R-squard (>.70) considered as high value is regarded as good model.

> #Multicollinearity

> y<-mileage[,2]

> y<-as.matrix(y)

> x<-mileage[,3:12]

> x<-cbind(rep(1,dim(mileage)[1]),as.matrix(x))

> #install.packages("MASS")

> library(MASS)

> (solve(t(X) %\*% X)) %\*% t(X) %\*% y

[,1]

19.34429256

disp -0.01923223

hp -0.03122932

drat 2.71497521

> #detecting Multicollinearity for mileage model

> #install.packages("Hmisc")

> library(Hmisc)

> rcorr(cbind(mileage[,3],mileage[,7],mileage[,10]))

[,1] [,2] [,3]

[1,] 1.00 0.80 -0.52

[2,] 0.80 1.00 -0.73

[3,] -0.52 -0.73 1.00

n= 32

P

[,1] [,2] [,3]

[1,] 0.0000 0.0022

[2,] 0.0000 0.0000

[3,] 0.0022 0.0000

> cor(cbind(mileage[,3],mileage[,7],mileage[,10]))

[,1] [,2] [,3]

[1,] 1.0000000 0.8046384 -0.5226070

[2,] 0.8046384 1.0000000 -0.7340749

[3,] -0.5226070 -0.7340749 1.0000000

>

> # Detecting and removing multicollinearity for mileage model

> #install.packages("usdm")

> library(usdm)# threshold value = 10

> vif(mileage[,3:12])

Variables VIF

1 cyl 17.622965

2 disp 10.288521

3 hp 12.155298

4 drat 3.952818

5 wt 5.607788

6 qsec 10.416979

7 vs 6.315315

8 am 5.542901

9 gear 5.907280

10 carb 3.608441

>

> vif(mileage[,c(4:12)])#dropping cyl

Variables VIF

1 disp 9.243212

2 hp 12.059406

3 drat 3.650140

4 wt 4.658785

5 qsec 8.876366

6 vs 6.192887

7 am 5.503960

8 gear 4.924971

9 carb 3.482181

>

> vif(mileage[,c(4,6:12)])#dropping hp

Variables VIF

1 disp 4.790650

2 drat 3.638500

3 wt 4.513636

4 qsec 5.558485

5 vs 5.216484

6 am 5.492216

7 gear 4.896204

8 carb 2.964957

>

> vif(mileage[,c(4,6,8:12)])#dropping wt

Variables VIF

1 disp 4.782963

2 drat 3.341471

3 qsec 5.538684

4 vs 5.176907

5 am 4.133144

6 gear 4.891865

7 carb 2.633130

> #PARSIMONIOUS MODELLING OR MODEL SELECTION

> #1 forward selection

> #first make a model with just intercept

> model<-lm(mpg~1,mileage)

> # Now add variable one by one

> summary(step(model,direction = "forward",trace = 2,scope = ~cyl+disp+hp+wt+drat+qsec+vs+am+gear+carb))

Start: AIC=115.94

mpg ~ 1

Df Sum of Sq RSS AIC

+ cyl 1 817.71 308.33 76.494

+ disp 1 807.42 318.63 77.545

+ wt 1 792.72 333.33 78.989

+ hp 1 741.10 384.94 83.596

+ drat 1 515.62 610.43 98.350

+ vs 1 496.53 629.52 99.335

+ am 1 405.15 720.90 103.672

+ carb 1 380.62 745.43 104.743

+ gear 1 259.75 866.30 109.552

+ qsec 1 222.67 903.38 110.893

<none> 1126.05 115.943

Step: AIC=76.49

mpg ~ cyl

Df Sum of Sq RSS AIC

+ wt 1 75.115 233.22 69.560

+ disp 1 37.565 270.77 74.337

+ am 1 36.972 271.36 74.407

+ hp 1 24.136 284.20 75.886

+ carb 1 21.391 286.94 76.194

<none> 308.33 76.494

+ drat 1 14.664 293.67 76.935

+ qsec 1 6.972 301.36 77.762

+ gear 1 5.431 302.90 77.926

+ vs 1 2.379 305.95 78.246

Step: AIC=69.56

mpg ~ cyl + wt

Df Sum of Sq RSS AIC

+ hp 1 39.871 193.35 65.560

+ disp 1 31.207 202.01 66.963

+ carb 1 27.044 206.18 67.616

<none> 233.22 69.560

+ qsec 1 2.954 230.27 71.152

+ vs 1 2.441 230.78 71.223

+ am 1 0.823 232.40 71.447

+ drat 1 0.365 232.85 71.510

+ gear 1 0.107 233.11 71.545

Step: AIC=65.56

mpg ~ cyl + wt + hp

Df Sum of Sq RSS AIC

+ disp 1 15.5825 177.77 64.871

<none> 193.35 65.560

+ gear 1 10.8288 182.52 65.716

+ am 1 9.4826 183.87 65.951

+ drat 1 8.6054 184.74 66.103

+ qsec 1 5.2720 188.08 66.676

+ carb 1 4.6552 188.69 66.780

+ vs 1 0.5665 192.78 67.466

Step: AIC=64.87

mpg ~ cyl + wt + hp + disp

Df Sum of Sq RSS AIC

+ carb 1 17.9596 159.81 63.463

<none> 177.77 64.871

+ gear 1 2.1337 175.63 66.485

+ drat 1 2.0879 175.68 66.493

+ am 1 1.6709 176.09 66.569

+ vs 1 1.4007 176.37 66.618

+ qsec 1 0.0363 177.73 66.865

Step: AIC=63.46

mpg ~ cyl + wt + hp + disp + carb

Df Sum of Sq RSS AIC

+ gear 1 15.5498 144.26 62.187

+ drat 1 10.1281 149.68 63.368

<none> 159.81 63.463

+ am 1 5.2725 154.53 64.390

+ qsec 1 2.0107 157.80 65.058

+ vs 1 0.0002 159.81 65.463

Step: AIC=62.19

mpg ~ cyl + wt + hp + disp + carb + gear

Df Sum of Sq RSS AIC

<none> 144.26 62.187

+ drat 1 5.6776 138.58 62.903

+ qsec 1 0.5121 143.75 64.074

+ vs 1 0.3738 143.88 64.104

+ am 1 0.0470 144.21 64.177

Call:

lm(formula = mpg ~ cyl + wt + hp + disp + carb + gear, data = mileage)

Residuals:

Min 1Q Median 3Q Max

-3.8255 -1.5605 0.1185 1.3283 4.5932

Coefficients:

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

(Intercept) 29.208117 5.931472 4.924 4.53e-05 \*\*\*

cyl 0.627032 0.756542 0.829 0.4151

wt -3.194253 1.091505 -2.926 0.0072 \*\*

hp -0.020958 0.019866 -1.055 0.3015

disp -0.014778 0.009953 -1.485 0.1501

carb -1.244848 0.533848 -2.332 0.0281 \*

gear 1.777451 1.082762 1.642 0.1132

---

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

Residual standard error: 2.402 on 25 degrees of freedom

Multiple R-squared: 0.8719, Adjusted R-squared: 0.8411

F-statistic: 28.36 on 6 and 25 DF, p-value: 5.318e-10

Conclusion: **Conclusion :** At 5% level of significance , Here p value < α then we reject null hypothesis and conclude that there is a linear relationship between feature and the response. The value of adjusted R-squard (>.70) considered as high value is regarded as good model

> #2. Backward elimination

> mileage[ ,1]<-NULL

> model<-lm(mpg~.,mileage)

> # now drop variables one by one

> summary(step(model,direction = "backward",trace = 2))

Start: AIC=68.55

mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb

Df Sum of Sq RSS AIC

- vs 1 0.004 137.08 66.555

- am 1 0.037 137.12 66.563

- qsec 1 0.915 138.00 66.767

- hp 1 3.540 140.62 67.370

- drat 1 6.379 143.46 68.010

- cyl 1 6.815 143.90 68.107

<none> 137.08 68.555

- gear 1 9.240 146.32 68.642

- disp 1 9.289 146.37 68.653

- wt 1 25.923 163.00 72.097

- carb 1 33.567 170.65 73.564

Step: AIC=66.56

mpg ~ cyl + disp + hp + drat + wt + qsec + am + gear + carb

Df Sum of Sq RSS AIC

- am 1 0.034 137.12 64.563

- qsec 1 1.476 138.56 64.898

- hp 1 4.119 141.20 65.503

- drat 1 6.469 143.55 66.031

- cyl 1 6.904 143.99 66.128

<none> 137.08 66.555

- gear 1 10.066 147.15 66.823

- disp 1 10.513 147.60 66.920

- wt 1 26.126 163.21 70.138

- carb 1 36.051 173.13 72.027

Step: AIC=64.56

mpg ~ cyl + disp + hp + drat + wt + qsec + gear + carb

Df Sum of Sq RSS AIC

- qsec 1 1.461 138.58 62.903

- hp 1 4.159 141.28 63.520

- drat 1 6.627 143.75 64.074

- cyl 1 7.107 144.22 64.180

<none> 137.12 64.563

- disp 1 10.919 148.04 65.015

- gear 1 12.248 149.37 65.301

- wt 1 35.217 172.34 69.879

- carb 1 36.545 173.66 70.124

Step: AIC=62.9

mpg ~ cyl + disp + hp + drat + wt + gear + carb

Df Sum of Sq RSS AIC

- drat 1 5.678 144.26 62.187

- cyl 1 6.017 144.60 62.263

- hp 1 7.543 146.12 62.599

<none> 138.58 62.903

- disp 1 9.582 148.16 63.042

- gear 1 11.099 149.68 63.368

- wt 1 35.193 173.77 68.144

- carb 1 36.058 174.64 68.303

Step: AIC=62.19

mpg ~ cyl + disp + hp + wt + gear + carb

Df Sum of Sq RSS AIC

- cyl 1 3.964 148.22 61.055

- hp 1 6.422 150.68 61.581

<none> 144.26 62.187

- disp 1 12.720 156.98 62.892

- gear 1 15.550 159.81 63.463

- carb 1 31.376 175.63 66.485

- wt 1 49.418 193.68 69.614

Step: AIC=61.05

mpg ~ disp + hp + wt + gear + carb

Df Sum of Sq RSS AIC

- hp 1 3.660 151.88 59.836

- disp 1 9.479 157.70 61.038

<none> 148.22 61.055

- gear 1 12.389 160.61 61.624

- carb 1 28.017 176.24 64.595

- wt 1 45.681 193.90 67.652

Step: AIC=59.84

mpg ~ disp + wt + gear + carb

Df Sum of Sq RSS AIC

- gear 1 9.623 161.50 59.801

<none> 151.88 59.836

- wt 1 48.687 200.57 66.733

- carb 1 49.576 201.46 66.875

- disp 1 50.355 202.24 66.998

Step: AIC=59.8

mpg ~ disp + wt + carb

Df Sum of Sq RSS AIC

<none> 161.50 59.801

- carb 1 45.373 206.88 65.724

- wt 1 88.103 249.61 71.733

- disp 1 94.450 255.95 72.536

Call:

lm(formula = mpg ~ disp + wt + carb, data = mileage)

Residuals:

Min 1Q Median 3Q Max

-4.6262 -1.1745 0.0345 1.2064 4.9173

Coefficients:

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

(Intercept) 38.399799 2.067151 18.576 < 2e-16 \*\*\*

disp -0.021435 0.005297 -4.047 0.000371 \*\*\*

wt -3.553409 0.909205 -3.908 0.000537 \*\*\*

carb -0.940069 0.335176 -2.805 0.009052 \*\*

---

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

Residual standard error: 2.402 on 28 degrees of freedom

Multiple R-squared: 0.8566, Adjusted R-squared: 0.8412

F-statistic: 55.74 on 3 and 28 DF, p-value: 6.287e-12

> #3.stepwise Selection

> #starting with forward selection\

> model<-lm(mpg~1,mileage)

> summary(step(model,direction = "both",trace = 2))

Start: AIC=115.94

mpg ~ 1

Call:

lm(formula = mpg ~ 1, data = mileage)

Residuals:

Min 1Q Median 3Q Max

-9.6906 -4.6656 -0.8906 2.7094 13.8094

Coefficients:

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

(Intercept) 20.091 1.065 18.86 <2e-16 \*\*\*

---

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

Residual standard error: 6.027 on 31 degrees of freedom

> #starting with backward elimination

> summary(step(lm(mpg~.,mileage),direction = "both",trace = 2))#stepwise

Start: AIC=68.55

mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb

Df Sum of Sq RSS AIC

- vs 1 0.004 137.08 66.555

- am 1 0.037 137.12 66.563

- qsec 1 0.915 138.00 66.767

- hp 1 3.540 140.62 67.370

- drat 1 6.379 143.46 68.010

- cyl 1 6.815 143.90 68.107

<none> 137.08 68.555

- gear 1 9.240 146.32 68.642

- disp 1 9.289 146.37 68.653

- wt 1 25.923 163.00 72.097

- carb 1 33.567 170.65 73.564

Step: AIC=66.56

mpg ~ cyl + disp + hp + drat + wt + qsec + am + gear + carb

Df Sum of Sq RSS AIC

- am 1 0.034 137.12 64.563

- qsec 1 1.476 138.56 64.898

- hp 1 4.119 141.20 65.503

- drat 1 6.469 143.55 66.031

- cyl 1 6.904 143.99 66.128

<none> 137.08 66.555

- gear 1 10.066 147.15 66.823

- disp 1 10.513 147.60 66.920

+ vs 1 0.004 137.08 68.555

- wt 1 26.126 163.21 70.138

- carb 1 36.051 173.13 72.027

Step: AIC=64.56

mpg ~ cyl + disp + hp + drat + wt + qsec + gear + carb

Df Sum of Sq RSS AIC

- qsec 1 1.461 138.58 62.903

- hp 1 4.159 141.28 63.520

- drat 1 6.627 143.75 64.074

- cyl 1 7.107 144.22 64.180

<none> 137.12 64.563

- disp 1 10.919 148.04 65.015

- gear 1 12.248 149.37 65.301

+ am 1 0.034 137.08 66.555

+ vs 1 0.001 137.12 66.563

- wt 1 35.217 172.34 69.879

- carb 1 36.545 173.66 70.124

Step: AIC=62.9

mpg ~ cyl + disp + hp + drat + wt + gear + carb

Df Sum of Sq RSS AIC

- drat 1 5.678 144.26 62.187

- cyl 1 6.017 144.60 62.263

- hp 1 7.543 146.12 62.599

<none> 138.58 62.903

- disp 1 9.582 148.16 63.042

- gear 1 11.099 149.68 63.368

+ qsec 1 1.461 137.12 64.563

+ vs 1 0.575 138.00 64.770

+ am 1 0.019 138.56 64.898

- wt 1 35.193 173.77 68.144

- carb 1 36.058 174.64 68.303

Step: AIC=62.19

mpg ~ cyl + disp + hp + wt + gear + carb

Df Sum of Sq RSS AIC

- cyl 1 3.964 148.22 61.055

- hp 1 6.422 150.68 61.581

<none> 144.26 62.187

- disp 1 12.720 156.98 62.892

+ drat 1 5.678 138.58 62.903

- gear 1 15.550 159.81 63.463

+ qsec 1 0.512 143.75 64.074

+ vs 1 0.374 143.88 64.104

+ am 1 0.047 144.21 64.177

- carb 1 31.376 175.63 66.485

- wt 1 49.418 193.68 69.614

Step: AIC=61.05

mpg ~ disp + hp + wt + gear + carb

Df Sum of Sq RSS AIC

- hp 1 3.660 151.88 59.836

- disp 1 9.479 157.70 61.038

<none> 148.22 61.055

- gear 1 12.389 160.61 61.624

+ cyl 1 3.964 144.26 62.187

+ drat 1 3.625 144.60 62.263

+ am 1 0.681 147.54 62.907

+ qsec 1 0.423 147.80 62.963

+ vs 1 0.346 147.88 62.980

- carb 1 28.017 176.24 64.595

- wt 1 45.681 193.90 67.652

Step: AIC=59.84

mpg ~ disp + wt + gear + carb

Df Sum of Sq RSS AIC

- gear 1 9.623 161.50 59.801

<none> 151.88 59.836

+ hp 1 3.660 148.22 61.055

+ drat 1 3.548 148.33 61.079

+ cyl 1 1.202 150.68 61.581

+ qsec 1 0.231 151.65 61.787

+ am 1 0.165 151.72 61.801

+ vs 1 0.113 151.77 61.812

- wt 1 48.687 200.57 66.733

- carb 1 49.576 201.46 66.875

- disp 1 50.355 202.24 66.998

Step: AIC=59.8

mpg ~ disp + wt + carb

Df Sum of Sq RSS AIC

<none> 161.50 59.801

+ gear 1 9.623 151.88 59.836

+ drat 1 7.347 154.16 60.312

+ am 1 3.849 157.66 61.030

+ hp 1 0.895 160.61 61.624

+ qsec 1 0.639 160.87 61.675

+ cyl 1 0.303 161.20 61.741

+ vs 1 0.135 161.37 61.775

- carb 1 45.373 206.88 65.724

- wt 1 88.103 249.61 71.733

- disp 1 94.450 255.95 72.536

Call:

lm(formula = mpg ~ disp + wt + carb, data = mileage)

Residuals:

Min 1Q Median 3Q Max

-4.6262 -1.1745 0.0345 1.2064 4.9173

Coefficients:

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

(Intercept) 38.399799 2.067151 18.576 < 2e-16 \*\*\*

disp -0.021435 0.005297 -4.047 0.000371 \*\*\*

wt -3.553409 0.909205 -3.908 0.000537 \*\*\*

carb -0.940069 0.335176 -2.805 0.009052 \*\*

---

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

Residual standard error: 2.402 on 28 degrees of freedom

Multiple R-squared: 0.8566, Adjusted R-squared: 0.8412

F-statistic: 55.74 on 3 and 28 DF, p-value: 6.287e-12

**Conclusion:** the model looks like ggod because the value of adjusted R square is greater than 0.70.

> #VALIDATION OF ASSUMPTIONS AND RESIDUAL ANALYSIS

> # Autocorrelation

> #install.packages("lmtest")

> library(lmtest)

> model<-lm(mpg~cyl+disp+hp+wt+drat+qsec+vs+am+gear+carb,data = mileage)

> dwtest(mpg~cyl+disp+hp+wt+drat+qsec+vs+am+gear+carb,data = mileage)

Durbin-Watson test

data: mpg ~ cyl + disp + hp + wt + drat + qsec + vs + am + gear + carb

DW = 2.1859, p-value = 0.4523

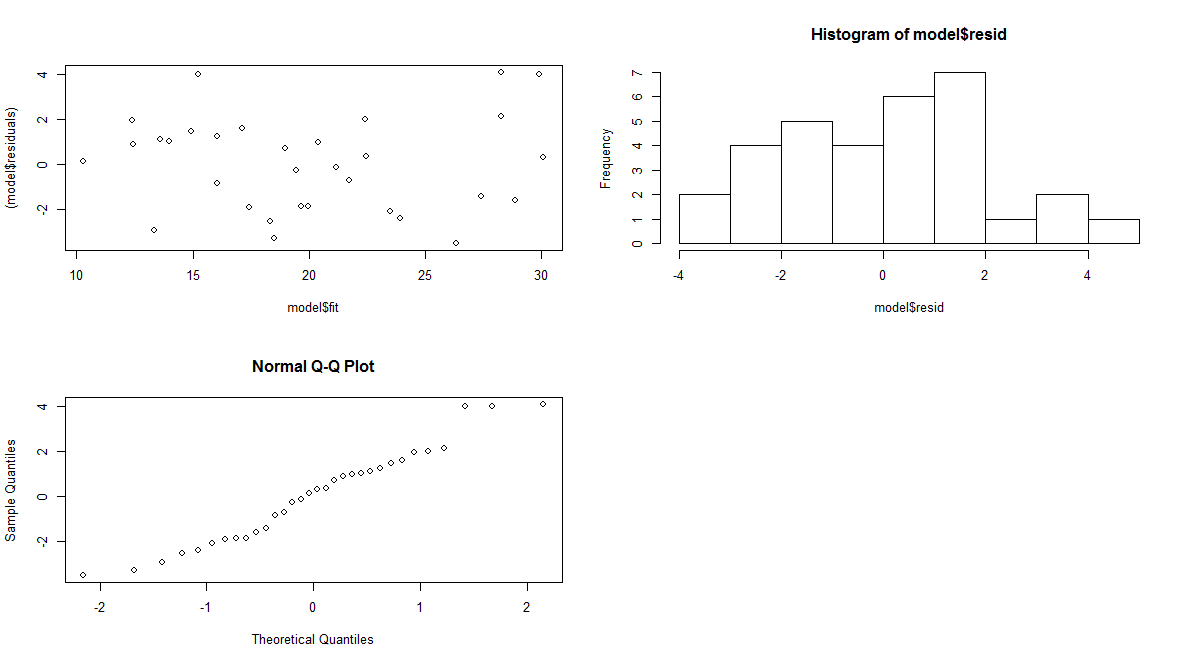
alternative hypothesis: true autocorrelation is greater than 0

**conclusion:** here the value of dw is 2.18 which is less than 2.5 , i.e there is no autocorrelation.

> # Hetroscedasticity

> plot(model$fit,(model$residuals))

|  |
| --- |
| > #Normality of errors  > #1 Histogram  > model<-lm(mpg~.,data = mileage)  > resid<-model$residuals  > hist(model$resid)  >  > #2.QQ plot  > qqnorm(model$residuals) |
|  |
| |  | | --- | |  | |



**Conclusion:** The first graph show that there is no pattern in the data i.e. all the point are randomly scattered.

There is no hetroscedasticity in the data. The 2nd graph show that data somehow looks like normal , so for large value it will tends to normal distribution.

|  |
| --- |
| > # outliers detection  > library(MASS)  > outlier.statistics<-cbind(c(1:nrow(mileage)),studres(lm(mpg~.,mileage)),hatvalues(lm(mpg~.,mileage)))  > outlier<-outlier.statistics[abs(outlier.statistics[,2])>1.5]  > print(outlier)  [1] 3.0000000 16.0000000 18.0000000 20.0000000 25.0000000 29.0000000 -1.6321062  [8] -1.5826169 1.8638391 2.0245876 1.9219444 -1.7230505 0.2200080 0.4272213  [15] 0.1764187 0.3234615 0.2640674 0.6327538  **Conclusion:** these are the outlier from the data, now the data is free from any type outlier, autocorrelation ,hetroscedasticity which help us to get valid result from the data as per our importance. |
|  |
| |  | | --- | |  | |