Home Work On Data Analysis 1

Md Shamsuzzaman

10.11.2017

1 Executive Summary

The study has been carried out on auto data set from ISLR package which has 392 observations along with 9 variables. Among the variables; displacement, cylinders, mpg, horsepower, weight, acceleration,origin, name and year, name is the qualitative variable and others are numeric. In this study, a multiple linear regression model will be applied to predict the model for mpg data. At the very beginning, the data has been checked on the assumptions; normality, constant variance and independency. It has been found that these three assumptions are met for the auto data set. The p-values of acceleration, cylinders and horse power (less than 0.05) from the multiple linear regression indicates (rejection of null hypothesis) that these three data are not important while the adjusted R-square value is (0.8184). However, it also has been found that the mpg model, reduced model, excluding name variable provides the lowest AIC value. This model also provides same adjusted R-square value like full-model. So, the name variable excluded model is the good model. It is worth to mention that the ANOVA test has revealed the independency in effects except acceleration. 81.82% response variability has been explained by the variables except name in this model so that is why is has been treated as good model.

2 Introduction

This data set (auto) has been taken from the StatLib library which is cared at Carnegie Mellon University. The main objective of this study is, to observe whether this data can be used to predict the miles per gallon (mpg) of the car. At the very beginning, the data has been checked based on the assumptions; normality, constant variance and independency. Different visualising plots like scattered plot, QQ plot etc. have been used to check the linearity justification for this data set. We have the following hypothesis for the model and ANOVA.

```
The hypothesis for Model.
Ho: Coefficient of covariate is equal to zero
Ha: Coefficient of covariate is not zero
The hypothesis for ANOVA.
Ho: Coefficient of covariate ha the same effect with other variables or levels
Ha: Coefficient of covariate has different effects from other variables or levels
Finally we have the model searching formula
>step(mpgmodel,data= Auto,direction="backward")
```

The lowest AIC value is the best indicator to find good model.

3 Data Collection

Except the variable ?name?, rest of the variables are numerical which were collected by StatLib library. The original data contained 408 observations but 16 observations with missing values were removed.

4 Data Analysis and Summary

5 Analysis

Analysis of the auto data set has been carried out through the linearity justification along with other tests.

5.1 Linearity justification of the data set

Scattered plot is one of the most important visualisation through which the linearity among the variables of auto data set can be carried out.

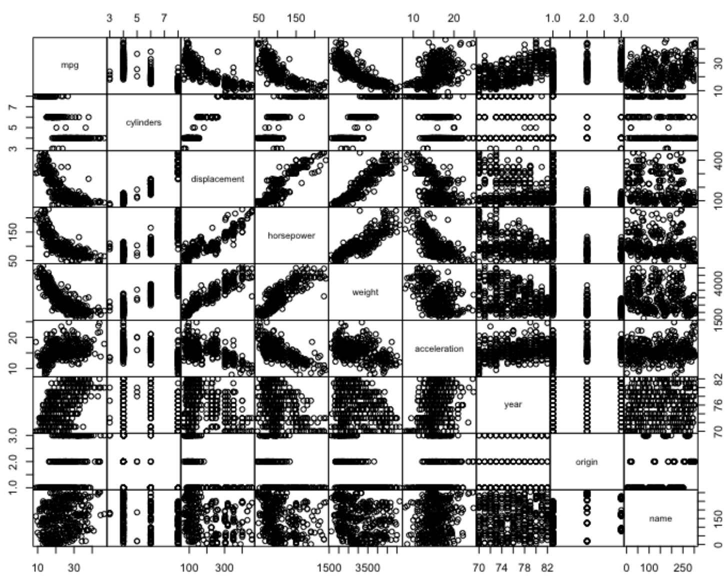

Fig 1. Scatter plot matrix

According to the scattered plot visualisation (figure 1)we can say that the auto dada set shows linearity character due to linearity found in mpg, horsepower, weight, displacement, and accelerate.

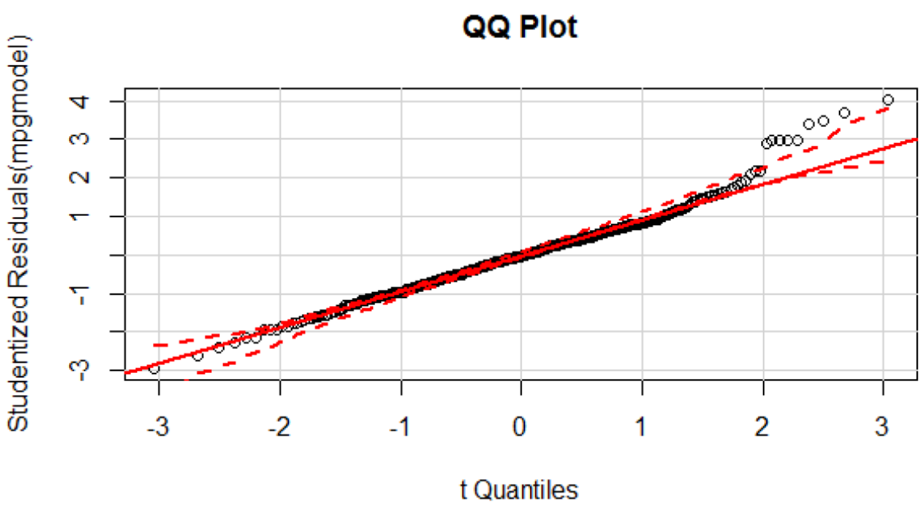

Fig 2.

QQ plot of auto data set

QQ plot (figure 2) of the auto data set indicates the linearity along with some outliers.

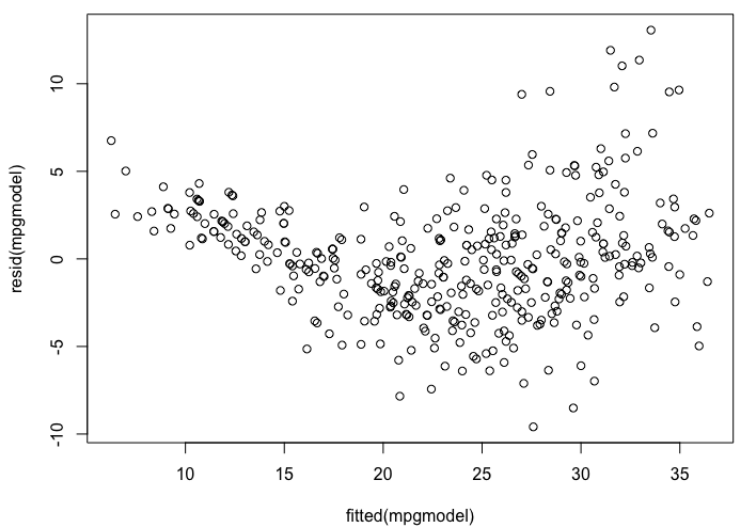

Fig 3. Residual plot of auto data set

According to the residual plot (figure 3), we can see the no systematic pattern with a constant variance. So, here in this data set, the constant variance is met.

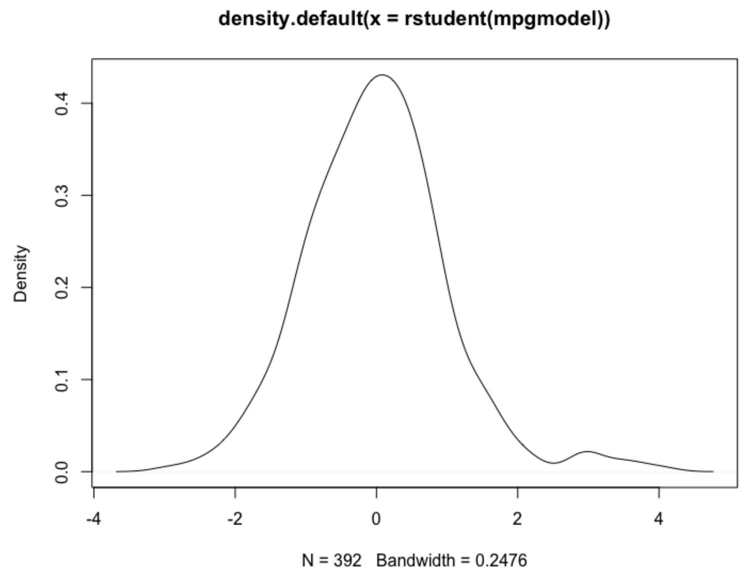

Fig 4. Density plot for auto data set

The density plot (figure 4) shows the presence of normal distribution in the auto data set.

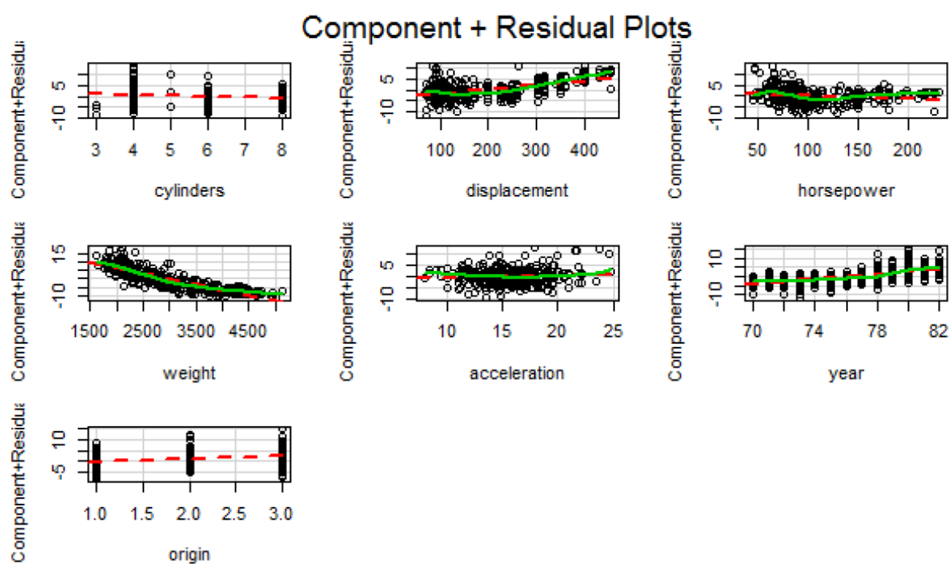

Fig 5. Partial residual plots of information

This partial residual plots present component and residual plots for linear and generalised linear models. This visualisation of the auto data shows that there are systematic patterns in displacement and weight data which might (linear pattern) also reveals the information regarding the missing of constant variance of these two variables. However, for getting the better understanding regarding the constant variance met, we can carry out the var.test and nvc test.

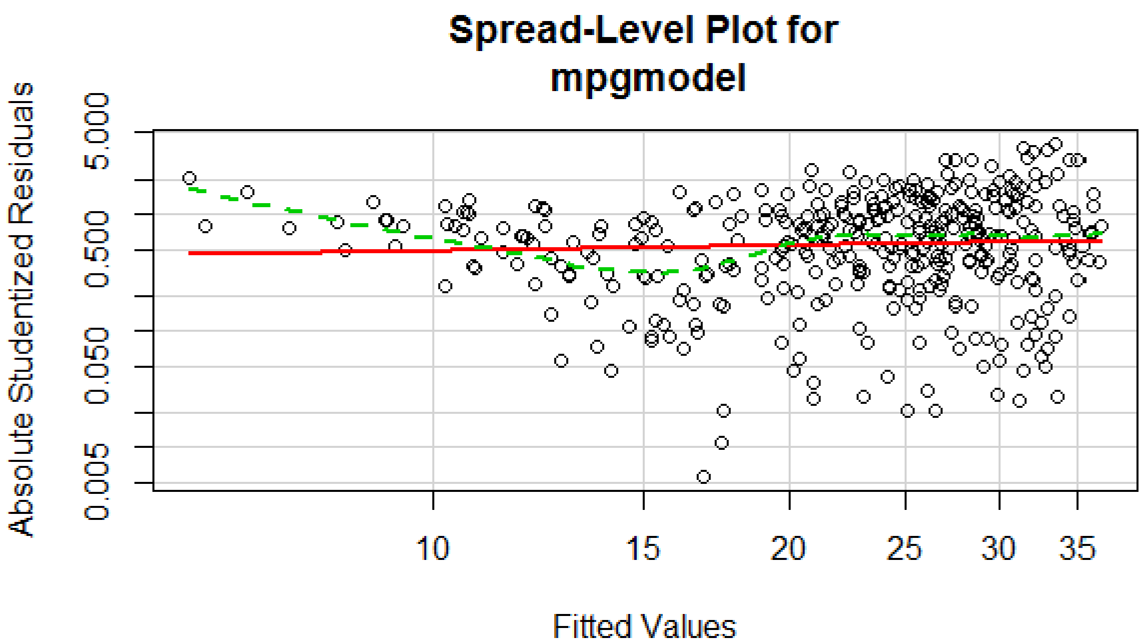

Fig 6. Spread-Level plot for mpgmodel

Spread level plot has been used here which examines the possible dependence of spread on level, or an extension of these plots to the studentized residuals from linear models. Here in this plot (figure 6), we can see that the data are distributed around up and down of the fitted values which indicates the constant variance has been met.

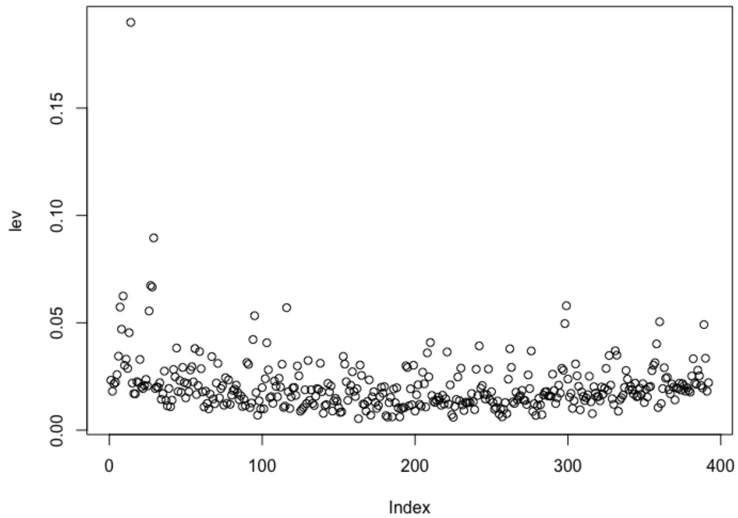

Fig 7. Leverage plot on each observations

From the Leverage plot (figure 7), we can see on point is far away than the rest of the points. This point can be assumed as a outlier or potential influential effect containing data.

6 Different models

```
mpgmodel=lm(mpg~cylinders+displacement+horsepower+weight+acceleration+year+origin)

> summary(mpgmodel)

Call:
lm(formula = mpg ~ cylinders + displacement + horsepower + weight +
    acceleration + year + origin)

Residuals:
    Min      1Q  Median      3Q     Max
-9.5903 -2.1565 -0.1169  1.8690 13.0604

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -17.218435   4.644294  -3.707  0.00024 ***
cylinders     -0.493376   0.323282  -1.526  0.12780
```

```
displacement   0.019896   0.007515   2.647  0.00844 **
horsepower    -0.016951   0.013787  -1.230  0.21963
weight        -0.006474   0.000652  -9.929  < 2e-16 ***
acceleration   0.080576   0.098845   0.815  0.41548
year           0.750773   0.050973  14.729  < 2e-16 ***
origin         1.426141   0.278136   5.127 4.67e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.328 on 384 degrees of freedom
Multiple R-squared:  0.8215,Adjusted R-squared:  0.8182
F-statistic: 252.4 on 7 and 384 DF,  p-value: < 2.2e-16
```

From this multiple linear regression, we can get the value of median is near about zero (0) which indicates the normality of the data set (auto). We have already have been noticed this normality of this data set through other tests such as density plot, QQ plot at the beginning of analysis. The individual p-values of displacement (0.00844), weight (< 2e-16), year (< 2e-16) and origin (4.67e-07) are less than predefined p-values (0.005). So, these covariates have significant and important effect. Additionally, the adjusted R-squared value (81.82 %) indicates this model is a good model and about 82 % of the response variability can be explained by this model. So, we can conclude that this model reveals a good relationship between the predictors and the response. Now, we can perform ANOVA to check the independency of covariates.

```
> anova(lm(mpg~cylinders+displacement+horsepower+weight+acceleration+year+origin))
Analysis of Variance Table

Response: mpg
              Df  Sum Sq Mean Sq   F value    Pr(>F)
cylinders      1 14403.1 14403.1 1300.6838 < 2.2e-16 ***
displacement   1  1073.3  1073.3   96.9293 < 2.2e-16 ***
horsepower     1   403.4   403.4   36.4301 3.731e-09 ***
weight         1   975.7   975.7   88.1137 < 2.2e-16 ***
acceleration   1     1.0     1.0    0.0872    0.7679
year           1  2419.1  2419.1  218.4609 < 2.2e-16 ***
origin         1   291.1   291.1   26.2912 4.666e-07 ***
Residuals    384  4252.2    11.1
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the carried out ANOVA test, we can see that the all of the p-values (< 2.2e-16,< 2.2e-16, 3.731e-09, < 2.2e-16,< 2.2e-16,4.666e-07) except acceleration (0.7679) are less than

0.05. So, it defines the independency regarding the different effects on the response. As a result, we can also make the conclusion that the independency of covariates has been met. Finally, we can say that all the covariates except acceleration have significant effect on the response. Now, var.test and ncvtest can be performed to check the assumption of constant variance.

```
Here, the hypotheses are following below for var.test and ncvtest.
Null Hypothesis = Variance is non-constant.
Alternative = Variance is constant.

> w<-lm(mpg~.-name,data= Auto)
> z<-lm(mpg~cylinders,data= Auto)
> var.test(w,z)

F test to compare two variances

data:  w and z
F = 0.45865, num df = 384, denom df = 390, p-value = 4.02e-14
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.3757091 0.5600003
sample estimates:
ratio of variances
         0.4586549

> ncvTest(mpgmodel)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 30.89489      Df = 1       p = 2.723876e-08
```

The results got from var.test and nvctest, we can see the p-values (4.02e-14, 2.723876e-08 respectively) are less than 0.05. So, the null hypothesis is rejected and alternative is accepted. It means, the assumption of constant variance has been met. At this point, we can make the conclusion that the results of ANOVA and multiple linear regression are accepted and reliable because the assumptions of linearity, constant variance and independency have been met. Now, we can find if there is other model with less cobariates that can fit this prediction of mpg.

```
> step(mpgmodel,data= Auto,direction="backward")
Start:  AIC=950.5
mpg ~ cylinders + displacement + horsepower + weight + acceleration +
    year + origin
```

```
               Df Sum of Sq    RSS     AIC
- acceleration  1      7.36 4259.6  949.18
- horsepower    1     16.74 4269.0  950.04
<none>                      4252.2  950.50
- cylinders     1     25.79 4278.0  950.87
- displacement  1     77.61 4329.8  955.59
- origin        1    291.13 4543.3  974.46
- weight        1   1091.63 5343.8 1038.08
- year          1   2402.25 6654.5 1124.06

Step:  AIC=949.18
mpg ~ cylinders + displacement + horsepower + weight + year +
    origin

               Df Sum of Sq    RSS     AIC
<none>                      4259.6  949.18
- cylinders     1     27.27 4286.8  949.68
- horsepower    1     53.80 4313.4  952.10
- displacement  1     73.57 4333.1  953.89
- origin        1    292.02 4551.6  973.17
- weight        1   1310.43 5570.0 1052.32
- year          1   2396.17 6655.7 1122.13

Call:
lm(formula = mpg ~ cylinders + displacement + horsepower + weight +
    year + origin)

Coefficients:
 (Intercept)     cylinders  displacement    horsepower        weight          year
  -15.563492     -0.506685      0.019269     -0.023895     -0.006218      0.747516
      origin
    1.428242
```

The value of AIC here for this model (mpg   cylinders + displacement + horsepower + weight + acceleration + year + origin) is 949.18. So, it can be checked either the reduced model is better then the full model and for this comparison of ANOVA can be carried out with assuming that both models are nested models.

```
Null Hypothesis: Reduced model is significant.
Alternative: Full model is significant.
```

```
> m1<-lm(mpg~cylinders+displacement+horsepower+weight+acceleration+year+origin)
> m2<-lm(mpg ~cylinders+displacement+horsepower+weight+year+origin)
> anova(m2,m1)
Analysis of Variance Table

Model 1: mpg ~ cylinders + displacement + horsepower + weight + year +
    origin
Model 2: mpg ~ cylinders + displacement + horsepower + weight + acceleration +
    year + origin
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    385 4259.6
2    384 4252.2  1    7.3584 0.6645 0.4155
> summary(lm(mpg~cylinders+displacement+horsepower+weight+acceleration+year+origin))

Call:
lm(formula = mpg ~ cylinders + displacement + horsepower + weight +
    acceleration + year + origin)

Residuals:
    Min      1Q  Median      3Q     Max
-9.5903 -2.1565 -0.1169  1.8690 13.0604

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -17.218435   4.644294  -3.707  0.00024 ***
cylinders     -0.493376   0.323282  -1.526  0.12780
displacement   0.019896   0.007515   2.647  0.00844 **
horsepower    -0.016951   0.013787  -1.230  0.21963
weight        -0.006474   0.000652  -9.929  < 2e-16 ***
acceleration   0.080576   0.098845   0.815  0.41548
year           0.750773   0.050973  14.729  < 2e-16 ***
origin         1.426141   0.278136   5.127 4.67e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.328 on 384 degrees of freedom
Multiple R-squared:  0.8215,Adjusted R-squared:  0.8182
F-statistic: 252.4 on 7 and 384 DF,  p-value: < 2.2e-16
```

According to the p-value (0.4155) got from the ANOVA means we fail to reject null hypothesis. So, the reduced model is important.

Now, different transformations are being used of the variables such as log X, root of X, and X2. After that, the multiple linear regression model was used like the previous complete model.

```
> loghorsepower=log(horsepower)
> logdisp=log(displacement)
> logacc=log(acceleration)
> logweight=log(weight)
> logcyl=log(cylinders)
> logmodel=lm(mpg~logcyl+logdisp+loghorsepower+logweight+logacc+year+origin)
> summary(logmodel)

Call:
lm(formula = mpg ~ logcyl + logdisp + loghorsepower + logweight +
    logacc + year + origin)

Residuals:
    Min      1Q  Median      3Q     Max
-9.6751 -1.7878 -0.0558  1.5061 12.7173

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   114.39550    9.80627  11.666  < 2e-16 ***
logcyl          1.67639    1.64739   1.018  0.30951
logdisp        -1.44495    1.49825  -0.964  0.33544
loghorsepower  -7.04654    1.55262  -4.538 7.59e-06 ***
logweight     -12.13652    2.20467  -5.505 6.77e-08 ***
logacc         -5.07430    1.59780  -3.176  0.00161 **
year            0.72585    0.04658  15.583  < 2e-16 ***
origin          0.82776    0.27792   2.978  0.00308 **
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 3.053 on 384 degrees of freedom
Multiple R-squared:  0.8497,Adjusted R-squared:  0.847
F-statistic: 310.2 on 7 and 384 DF,  p-value: < 2.2e-16
```

```
> loghorsepower=log(horsepower)
> logdisp=log(displacement)
> logacc=log(acceleration)
> logweight=log(weight)
> logcyl=log(cylinders)
```

```
> logmodel=lm(mpg~logcyl+logdisp+loghorsepower+logweight+logacc+year+origin)
> summary(logmodel)

Call:
lm(formula = mpg ~ logcyl + logdisp + loghorsepower + logweight +
    logacc + year + origin)

Residuals:
    Min      1Q  Median      3Q     Max
-9.6751 -1.7878 -0.0558  1.5061 12.7173

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   114.39550    9.80627  11.666  < 2e-16 ***
logcyl          1.67639    1.64739   1.018  0.30951
logdisp        -1.44495    1.49825  -0.964  0.33544
loghorsepower  -7.04654    1.55262  -4.538 7.59e-06 ***
logweight     -12.13652    2.20467  -5.505 6.77e-08 ***
logacc         -5.07430    1.59780  -3.176  0.00161 **
year            0.72585    0.04658  15.583  < 2e-16 ***
origin          0.82776    0.27792   2.978  0.00308 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.053 on 384 degrees of freedom
Multiple R-squared:  0.8497,Adjusted R-squared:  0.847
F-statistic: 310.2 on 7 and 384 DF,  p-value: < 2.2e-16

> sqrthp=sqrt(horsepower)
> sqrtdisp=sqrt(displacement)
> sqrtacc=sqrt(acceleration)
> sqrtweight=sqrt(weight)
> sqrtcyl=sqrt(cylinders)
> sqrtmodel=lm(mpg~sqrtcyl+sqrtdisp+sqrthp+sqrtweight+sqrtacc+year+origin)
> summary(sqrtmodel)

Call:
lm(formula = mpg ~ sqrtcyl + sqrtdisp + sqrthp + sqrtweight +
    sqrtacc + year + origin)

Residuals:
```

```
    Min      1Q  Median      3Q     Max
-9.5644 -1.9712 -0.1489  1.6737 13.0364

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.04768    6.08630   1.322   0.1869
sqrtcyl     -0.13558    1.53473  -0.088   0.9297
sqrtdisp     0.18761    0.22719   0.826   0.4094
sqrthp      -0.78036    0.30761  -2.537   0.0116 *
sqrtweight  -0.61370    0.07885  -7.783 6.63e-14 ***
sqrtacc     -0.84850    0.83330  -1.018   0.3092
year         0.73322    0.04919  14.905  < 2e-16 ***
origin       1.15363    0.28057   4.112 4.80e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.207 on 384 degrees of freedom
Multiple R-squared:  0.8342,Adjusted R-squared:  0.8312
F-statistic: 276.1 on 7 and 384 DF,  p-value: < 2.2e-16

> squarehp=(horsepower)^2
> squaredisp=(displacement)^2
> squareacc=(acceleration)^2
> squareweight=(weight)^2
> squarecyl=(cylinders)^2
> squaremodel=lm(mpg~squarecyl+squaredisp+squarehp+squareweight+squareacc+year+origin)
> summary(squaremodel)

Call:
lm(formula = mpg ~ squarecyl + squaredisp + squarehp + squareweight +
    squareacc + year + origin)

Residuals:
    Min      1Q  Median      3Q     Max
-9.6507 -2.3228 -0.1115  1.8855 12.9932

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.939e+01  4.306e+00  -6.825 3.44e-11 ***
squarecyl    -8.620e-02  2.519e-02  -3.423 0.000687 ***
squaredisp    5.954e-05  1.384e-05   4.301 2.16e-05 ***
```

```
squarehp     -4.143e-05  4.983e-05  -0.831 0.406248
squareweight -9.416e-07  8.955e-08 -10.514  < 2e-16 ***
squareacc     6.148e-03  2.685e-03   2.289 0.022594 *
year          7.636e-01  5.363e-02  14.239  < 2e-16 ***
origin        1.749e+00  2.766e-01   6.322 7.16e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.533 on 384 degrees of freedom
Multiple R-squared:  0.7988,Adjusted R-squared:  0.7951
F-statistic: 217.8 on 7 and 384 DF,  p-value: < 2.2e-16
```

According to the result, log transformation technique gives the highest improved adjusted R-squared value: 0.847. This adjusted R square value is increasing from the previous complete model. Thus, the log transformation is suggested to be done for the prediction model.

7 Conclusion

However, it also has been found that the mpg model, reduced model, excluding name variable provides the lowest AIC value. This model also provides same adjusted R-square value like full-model. So, the name variable excluded model is the good model. It is worth to mention that the ANOVA test has revealed the independency in effects except acceleration. 81.82% response variability has been explained by the variables except name in this model so that is why is has been treated as good model.

8 Appendix

8.1 R Code

```
attach(Auto)
class(Auto)
names(Auto)
str(Auto)
summary(Auto)
par(mex=0.5)
pairs(Auto, gap=0, cex.labels=0.8)
cor(Auto [,-9])
mpgmodel=lm(mpg~cylinders+displacement+horsepower+weight+acceleration+year+origin)
plot(mpgmodel)
summary(mpgmodel)
```

```
anova(lm(mpg~cylinders+displacement+horsepower+weight+acceleration+year+origin))
w<-lm(mpg~.-name,data= Auto)
z<-lm(mpg~cylinders,data= Auto)
var.test(w,z)
spreadLevelPlot(mpgmodel)
ncvTest(mpgmodel)
step(mpgmodel,data= Auto,direction="backward")
m1<-lm(mpg~cylinders+displacement+horsepower+weight+acceleration+year+origin)
m2<-lm(mpg ~cylinders+displacement+horsepower+weight+year+origin)
anova(m2,m1)
summary(lm(mpg~cylinders+displacement+horsepower+weight+acceleration+year+origin))
summary(lm(mpg~cylinders+displacement+horsepower+weight+acceleration+year))
summary(lm(mpg~cylinders+displacement+horsepower+weight))
summary(lm(mpg~cylinders+displacement+horsepower))
summary(lm(mpg~cylinders+displacement))
summary(lm(mpg~cylinders))
summary(lm(mpg~cylinders+displacement+horsepower+weight+year+origin))
plot(density(rstudent(mpgmodel)))
outlierTest(mpgmodel)
qqPlot(mpgmodel,simulate=TRUE, line="none")
qqPlot(mpgmodel,main="QQ Plot")
plot(fitted(mpgmodel),resid(mpgmodel))
crPlots(mpgmodel)
crPlots(mpgmodel,ask=FALSE)
lev=hat(model.matrix(mpgmodel))
plot(lev)
summary(lm(mpg~year))
loghorsepower=log(horsepower)
logdisp=log(displacement)
logacc=log(acceleration)
logweight=log(weight)
logcyl=log(cylinders)
logmodel=lm(mpg~logcyl+logdisp+loghorsepower+logweight+logacc+year+origin)
summary(logmodel)
sqrthp=sqrt(horsepower
sqrtdisp=sqrt(displacement)
sqrtacc=sqrt(acceleration)
sqrtweight=sqrt(weight)
sqrtcyl=sqrt(cylinders)
sqrtmodel=lm(mpg~sqrtcyl+sqrtdisp+sqrthp+sqrtweight+sqrtacc+year+origin)
summary(sqrtmodel)
```

```
squarehp=(horsepower)^2
squaredisp=(displacement)^2
squareacc=(acceleration)^2
squareweight=(weight)^2
squarecyl=(cylinders)^2
squaremodel=lm(mpg~squarecyl+squaredisp+squarehp+squareweight+squareacc+year+origin)
summary(squaremodel)
```

8.2 Log File

```
> install.packages("ISLR", dependencies = FALSE)
> library("ISLR", lib.loc="/Library/Frameworks/R.framework/Versions/3.3/Resources/library")
> attach(Auto)
> class(Auto)
[1] "data.frame"
> names(Auto)
[1] "mpg"          "cylinders"    "displacement" "horsepower"   "weight"
[6] "acceleration" "year"         "origin"       "name"
> str(Auto)
'data.frame': 392 obs. of  9 variables:
 $ mpg         : num  18 15 18 16 17 15 14 14 14 15 ...
 $ cylinders   : num  8 8 8 8 8 8 8 8 8 8 ...
 $ displacement: num  307 350 318 304 302 429 454 440 455 390 ...
 $ horsepower  : num  130 165 150 150 140 198 220 215 225 190 ...
 $ weight      : num  3504 3693 3436 3433 3449 ...
 $ acceleration: num  12 11.5 11 12 10.5 10 9 8.5 10 8.5 ...
 $ year        : num  70 70 70 70 70 70 70 70 70 70 ...
 $ origin      : num  1 1 1 1 1 1 1 1 1 1 ...
 $ name        : Factor w/ 304 levels "amc ambassador brougham",..: 49 36 231 14 161 141 54
> summary(Auto)
      mpg          cylinders      displacement     horsepower        weight
 Min.   : 9.00   Min.   :3.000   Min.   : 68.0   Min.   : 46.0   Min.   :1613
 1st Qu.:17.00   1st Qu.:4.000   1st Qu.:105.0   1st Qu.: 75.0   1st Qu.:2225
 Median :22.75   Median :4.000   Median :151.0   Median : 93.5   Median :2804
 Mean   :23.45   Mean   :5.472   Mean   :194.4   Mean   :104.5   Mean   :2978
 3rd Qu.:29.00   3rd Qu.:8.000   3rd Qu.:275.8   3rd Qu.:126.0   3rd Qu.:3615
 Max.   :46.60   Max.   :8.000   Max.   :455.0   Max.   :230.0   Max.   :5140

  acceleration        year           origin                      name
 Min.   : 8.00   Min.   :70.00   Min.   :1.000   amc matador       :  5
 1st Qu.:13.78   1st Qu.:73.00   1st Qu.:1.000   ford pinto        :  5
```

```
 Median :15.50   Median :76.00   Median :1.000   toyota corolla    :  5
 Mean   :15.54   Mean   :75.98   Mean   :1.577   amc gremlin       :  4
 3rd Qu.:17.02   3rd Qu.:79.00   3rd Qu.:2.000   amc hornet        :  4
 Max.   :24.80   Max.   :82.00   Max.   :3.000   chevrolet chevette:  4
                                                 (Other)           :365
> par(mex=0.5)
> pairs(Auto, gap=0, cex.labels=0.8)
> cor(Auto [,-9])
                    mpg  cylinders displacement horsepower     weight acceleration
mpg           1.0000000 -0.7776175   -0.8051269 -0.7784268 -0.8322442    0.4233285
cylinders    -0.7776175  1.0000000    0.9508233  0.8429834  0.8975273   -0.5046834
displacement -0.8051269  0.9508233    1.0000000  0.8972570  0.9329944   -0.5438005
horsepower   -0.7784268  0.8429834    0.8972570  1.0000000  0.8645377   -0.6891955
weight       -0.8322442  0.8975273    0.9329944  0.8645377  1.0000000   -0.4168392
acceleration  0.4233285 -0.5046834   -0.5438005 -0.6891955 -0.4168392    1.0000000
year          0.5805410 -0.3456474   -0.3698552 -0.4163615 -0.3091199    0.2903161
origin        0.5652088 -0.5689316   -0.6145351 -0.4551715 -0.5850054    0.2127458
                   year     origin
mpg           0.5805410  0.5652088
cylinders    -0.3456474 -0.5689316
displacement -0.3698552 -0.6145351
horsepower   -0.4163615 -0.4551715
weight       -0.3091199 -0.5850054
acceleration  0.2903161  0.2127458
year          1.0000000  0.1815277
origin        0.1815277  1.0000000
> mpgmodel=lm(mpg~cylinders+displacement+horsepower+weight+acceleration+year+origin)
> plot(mpgmodel)
Hit <Return> to see next plot: summary(mpgmodel)
Hit <Return> to see next plot: anova(lm(mpg~cylinders+displacement+horsepower+weight+accele
> w<-lm(mpg~.-name,data= Auto)
> z<-lm(mpg~cylinders,data= Auto)
> var.test(w,z)

F test to compare two variances

data:  w and z
F = 0.45865, num df = 384, denom df = 390, p-value = 4.02e-14
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.3757091 0.5600003
```

```
sample estimates:
ratio of variances
         0.4586549
> spreadLevelPlot(mpgmodel)
> ncvTest(mpgmodel)
> ncvTest(mpgmodel)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 30.89489      Df = 1       p = 2.723876e-08

> step(mpgmodel,data= Auto,direction="backward")
Start:  AIC=950.5
mpg ~ cylinders + displacement + horsepower + weight + acceleration +
    year + origin

               Df Sum of Sq    RSS     AIC
- acceleration  1      7.36 4259.6  949.18
- horsepower    1     16.74 4269.0  950.04
<none>                      4252.2  950.50
- cylinders     1     25.79 4278.0  950.87
- displacement  1     77.61 4329.8  955.59
- origin        1    291.13 4543.3  974.46
- weight        1   1091.63 5343.8 1038.08
- year          1   2402.25 6654.5 1124.06

Step:  AIC=949.18
mpg ~ cylinders + displacement + horsepower + weight + year +
    origin

               Df Sum of Sq    RSS     AIC
<none>                      4259.6  949.18
- cylinders     1     27.27 4286.8  949.68
- horsepower    1     53.80 4313.4  952.10
- displacement  1     73.57 4333.1  953.89
- origin        1    292.02 4551.6  973.17
- weight        1   1310.43 5570.0 1052.32
- year          1   2396.17 6655.7 1122.13

Call:
lm(formula = mpg ~ cylinders + displacement + horsepower + weight +
    year + origin)
```

```
Coefficients:
 (Intercept)     cylinders  displacement    horsepower        weight          year
  -15.563492     -0.506685      0.019269     -0.023895     -0.006218      0.747516
      origin
    1.428242

> m1<-lm(mpg~cylinders+displacement+horsepower+weight+acceleration+year+origin)
> m2<-lm(mpg ~cylinders+displacement+horsepower+weight+year+origin)
> anova(m2,m1)
Analysis of Variance Table

Model 1: mpg ~ cylinders + displacement + horsepower + weight + year +
    origin
Model 2: mpg ~ cylinders + displacement + horsepower + weight + acceleration +
    year + origin
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    385 4259.6
2    384 4252.2  1    7.3584 0.6645 0.4155

> summary(lm(mpg~cylinders+displacement+horsepower+weight+acceleration+year+origin))

Call:
lm(formula = mpg ~ cylinders + displacement + horsepower + weight +
    acceleration + year + origin)

Residuals:
    Min      1Q  Median      3Q     Max
-9.5903 -2.1565 -0.1169  1.8690 13.0604

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -17.218435   4.644294  -3.707  0.00024 ***
cylinders     -0.493376   0.323282  -1.526  0.12780
displacement   0.019896   0.007515   2.647  0.00844 **
horsepower    -0.016951   0.013787  -1.230  0.21963
weight        -0.006474   0.000652  -9.929  < 2e-16 ***
acceleration   0.080576   0.098845   0.815  0.41548
year           0.750773   0.050973  14.729  < 2e-16 ***
origin         1.426141   0.278136   5.127 4.67e-07 ***
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.328 on 384 degrees of freedom
Multiple R-squared:  0.8215,Adjusted R-squared:  0.8182
F-statistic: 252.4 on 7 and 384 DF,  p-value: < 2.2e-16

> summary(lm(mpg~cylinders+displacement+horsepower+weight+acceleration+year))

Call:
lm(formula = mpg ~ cylinders + displacement + horsepower + weight +
    acceleration + year)

Residuals:
    Min      1Q  Median      3Q     Max
-8.6927 -2.3864 -0.0801  2.0291 14.3607

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.454e+01  4.764e+00  -3.051  0.00244 **
cylinders    -3.299e-01  3.321e-01  -0.993  0.32122
displacement  7.678e-03  7.358e-03   1.044  0.29733
horsepower   -3.914e-04  1.384e-02  -0.028  0.97745
weight       -6.795e-03  6.700e-04 -10.141  < 2e-16 ***
acceleration  8.527e-02  1.020e-01   0.836  0.40383
year          7.534e-01  5.262e-02  14.318  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.435 on 385 degrees of freedom
Multiple R-squared:  0.8093,Adjusted R-squared:  0.8063
F-statistic: 272.2 on 6 and 385 DF,  p-value: < 2.2e-16

> summary(lm(mpg~cylinders+displacement+horsepower+weight))

Call:
lm(formula = mpg ~ cylinders + displacement + horsepower + weight)

Residuals:
     Min       1Q   Median       3Q      Max
-11.5248  -2.7964  -0.3568   2.2577  16.3221
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  45.7567705  1.5200437  30.102  < 2e-16 ***
cylinders    -0.3932854  0.4095522  -0.960 0.337513
displacement  0.0001389  0.0090099   0.015 0.987709
horsepower   -0.0428125  0.0128699  -3.327 0.000963 ***
weight       -0.0052772  0.0007166  -7.364 1.08e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.242 on 387 degrees of freedom
Multiple R-squared:  0.7077,Adjusted R-squared:  0.7046
F-statistic: 234.2 on 4 and 387 DF,  p-value: < 2.2e-16

> summary(lm(mpg~cylinders+displacement+horsepower))

Call:
lm(formula = mpg ~ cylinders + displacement + horsepower)

Residuals:
     Min       1Q   Median       3Q      Max
-11.7144  -3.1391  -0.3149   2.3481  16.5726

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  39.305268   1.324633  29.673  < 2e-16 ***
cylinders    -0.719431   0.434180  -1.657 0.098331 .
displacement -0.029120   0.008623  -3.377 0.000807 ***
horsepower   -0.059935   0.013498  -4.440 1.17e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.523 on 388 degrees of freedom
Multiple R-squared:  0.6667,Adjusted R-squared:  0.6641
F-statistic: 258.7 on 3 and 388 DF,  p-value: < 2.2e-16

> summary(lm(mpg~cylinders+displacement))

Call:
lm(formula = mpg ~ cylinders + displacement)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-13.2304  -3.0383  -0.5243   2.4307  18.3134

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  36.537707   1.196611  30.534  < 2e-16 ***
cylinders    -0.576348   0.443276  -1.300    0.194
displacement -0.051118   0.007226  -7.074 7.02e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.631 on 389 degrees of freedom
Multiple R-squared:  0.6498,Adjusted R-squared:  0.648
F-statistic: 360.8 on 2 and 389 DF,  p-value: < 2.2e-16

> summary(lm(mpg~cylinders))

Call:
lm(formula = mpg ~ cylinders)

Residuals:
     Min       1Q   Median       3Q      Max
-14.2413  -3.1832  -0.6332   2.5491  17.9168

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  42.9155     0.8349   51.40   <2e-16 ***
cylinders    -3.5581     0.1457  -24.43   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.914 on 390 degrees of freedom
Multiple R-squared:  0.6047,Adjusted R-squared:  0.6037
F-statistic: 596.6 on 1 and 390 DF,  p-value: < 2.2e-16

> summary(lm(mpg~cylinders+displacement+horsepower+weight+year+origin))

Call:
lm(formula = mpg ~ cylinders + displacement + horsepower + weight +
    year + origin)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-9.7604 -2.1791 -0.1535  1.8524 13.1209

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.556e+01  4.175e+00  -3.728 0.000222 ***
cylinders    -5.067e-01  3.227e-01  -1.570 0.117236
displacement  1.927e-02  7.472e-03   2.579 0.010287 *
horsepower   -2.389e-02  1.084e-02  -2.205 0.028031 *
weight       -6.218e-03  5.714e-04 -10.883  < 2e-16 ***
year          7.475e-01  5.079e-02  14.717  < 2e-16 ***
origin        1.428e+00  2.780e-01   5.138 4.43e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.326 on 385 degrees of freedom
Multiple R-squared:  0.8212,Adjusted R-squared:  0.8184
F-statistic: 294.6 on 6 and 385 DF,  p-value: < 2.2e-16

> plot(density(rstudent(mpgmodel)))
> outlierTest(mpgmodel)
> qqPlot(mpgmodel,simulate=TRUE, line="none")
> qqPlot(mpgmodel,main="QQ Plot")
> plot(fitted(mpgmodel),resid(mpgmodel))
> crPlots(mpgmodel)
> crPlots(mpgmodel,ask=FALSE)
> lev=hat(model.matrix(mpgmodel))
> plot(lev)
> summary(lm(mpg~year))

Call:
lm(formula = mpg ~ year)

Residuals:
     Min       1Q   Median       3Q      Max
-12.0212  -5.4411  -0.4412   4.9739  18.2088

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept) -70.01167    6.64516  -10.54   <2e-16 ***
year          1.23004    0.08736   14.08   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.363 on 390 degrees of freedom
Multiple R-squared:  0.337,Adjusted R-squared:  0.3353
F-statistic: 198.3 on 1 and 390 DF,  p-value: < 2.2e-16

> loghorsepower=log(horsepower)
> logdisp=log(displacement)
> logacc=log(acceleration)
> logweight=log(weight)
> logcyl=log(cylinders)
> logmodel=lm(mpg~logcyl+logdisp+loghorsepower+logweight+logacc+year+origin)
> summary(logmodel)

Call:
lm(formula = mpg ~ logcyl + logdisp + loghorsepower + logweight +
    logacc + year + origin)

Residuals:
    Min      1Q  Median      3Q     Max
-9.6751 -1.7878 -0.0558  1.5061 12.7173

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   114.39550    9.80627  11.666  < 2e-16 ***
logcyl          1.67639    1.64739   1.018  0.30951
logdisp        -1.44495    1.49825  -0.964  0.33544
loghorsepower  -7.04654    1.55262  -4.538 7.59e-06 ***
logweight     -12.13652    2.20467  -5.505 6.77e-08 ***
logacc         -5.07430    1.59780  -3.176  0.00161 **
year            0.72585    0.04658  15.583  < 2e-16 ***
origin          0.82776    0.27792   2.978  0.00308 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.053 on 384 degrees of freedom
Multiple R-squared:  0.8497,Adjusted R-squared:  0.847
F-statistic: 310.2 on 7 and 384 DF,  p-value: < 2.2e-16
```

```
> sqrthp=sqrt(horsepower)
> sqrtdisp=sqrt(displacement)
> sqrtacc=sqrt(acceleration)
> sqrtweight=sqrt(weight)
> sqrtcyl=sqrt(cylinders)
> sqrtmodel=lm(mpg~sqrtcyl+sqrtdisp+sqrthp+sqrtweight+sqrtacc+year+origin)
> summary(sqrtmodel)

Call:
lm(formula = mpg ~ sqrtcyl + sqrtdisp + sqrthp + sqrtweight +
    sqrtacc + year + origin)

Residuals:
    Min      1Q  Median      3Q     Max
-9.5644 -1.9712 -0.1489  1.6737 13.0364

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.04768    6.08630   1.322   0.1869
sqrtcyl     -0.13558    1.53473  -0.088   0.9297
sqrtdisp     0.18761    0.22719   0.826   0.4094
sqrthp      -0.78036    0.30761  -2.537   0.0116 *
sqrtweight  -0.61370    0.07885  -7.783 6.63e-14 ***
sqrtacc     -0.84850    0.83330  -1.018   0.3092
year         0.73322    0.04919  14.905  < 2e-16 ***
origin       1.15363    0.28057   4.112 4.80e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.207 on 384 degrees of freedom
Multiple R-squared:  0.8342,Adjusted R-squared:  0.8312
F-statistic: 276.1 on 7 and 384 DF,  p-value: < 2.2e-16

> squarehp=(horsepower)^2
> squaredisp=(displacement)^2
> squareacc=(acceleration)^2
> squareweight=(weight)^2
> squarecyl=(cylinders)^2
> squaremodel=lm(mpg~squarecyl+squaredisp+squarehp+squareweight+squareacc+year+origin)
> summary(squaremodel)
```

```
Call:
lm(formula = mpg ~ squarecyl + squaredisp + squarehp + squareweight +
    squareacc + year + origin)

Residuals:
    Min      1Q  Median      3Q     Max
-9.6507 -2.3228 -0.1115  1.8855 12.9932

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.939e+01  4.306e+00  -6.825 3.44e-11 ***
squarecyl    -8.620e-02  2.519e-02  -3.423 0.000687 ***
squaredisp    5.954e-05  1.384e-05   4.301 2.16e-05 ***
squarehp     -4.143e-05  4.983e-05  -0.831 0.406248
squareweight -9.416e-07  8.955e-08 -10.514  < 2e-16 ***
squareacc     6.148e-03  2.685e-03   2.289 0.022594 *
year          7.636e-01  5.363e-02  14.239  < 2e-16 ***
origin        1.749e+00  2.766e-01   6.322 7.16e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.533 on 384 degrees of freedom
Multiple R-squared:  0.7988,Adjusted R-squared:  0.7951
F-statistic: 217.8 on 7 and 384 DF,  p-value: < 2.2e-16
```