Regression Models Course Project

Mahesh

12 September 2016

Summary

Motor Trend, a magazine about the automobile industry, is interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). The questions of interest are:

1. "Is an automatic or manual transmission better for MPG"
2. "Quantify the MPG difference between automatic and manual transmissions"

DataSet

mtcars

```
library(ggplot2)
data(mtcars)
```

Inspect the data

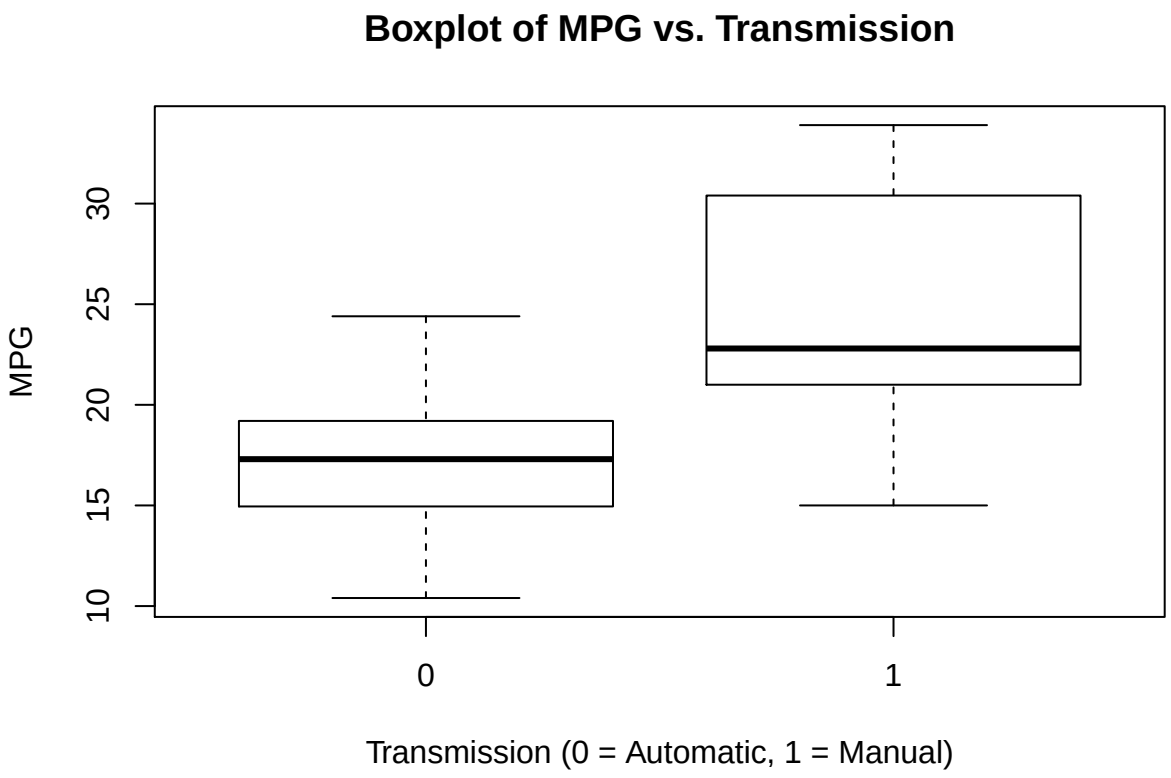
```
head(mtcars)
```

```
##                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

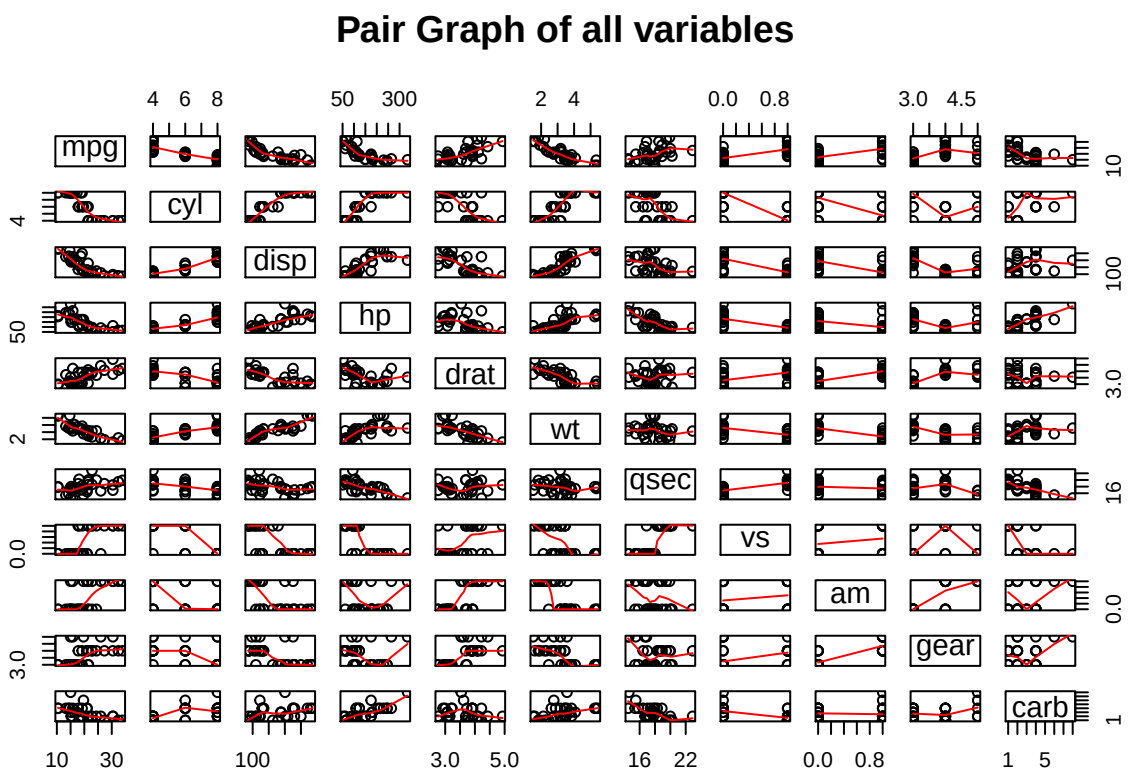
```
str(mtcars)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

```
boxplot(mpg ~ am, data = mtcars, xlab="Transmission (0 = Automatic, 1 = Manual)", ylab="MPG",main="Boxpl
```




```
pairs(mtcars, panel=panel.smooth, main="Pair Graph of all variables")
```



Convert the variables of interest into factors for analysis

```
mtcars$cyl <- factor(mtcars$cyl)
mtcars$vs <- factor(mtcars$vs)
mtcars$am <- factor(mtcars$am)
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)
```

mean of manual and auto mpg

```
# Automatic mpg mean
round(mean(mtcars[mtcars$am==0, "mpg"]),2)
```

```
## [1] 17.15
```

```
# Manual mpg mean
round(mean(mtcars[mtcars$am==1, "mpg"]),2)
```

```
## [1] 24.39
```

Regression Model

t test

```
t.test(mpg ~ am, data = mtcars)
```

```
##
##  Welch Two Sample t-test
##
## data:  mpg by am
## t = -3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.280194  -3.209684
## sample estimates:
## mean in group 0 mean in group 1
##        17.14737        24.39231
```

p-value is 0.001374 and we can reject null hypothesis. mean difference is 7

```
par(mfrow=c(2,2))
fullmodel <- lm(mpg ~ ., data=mtcars)
summary(fullmodel)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

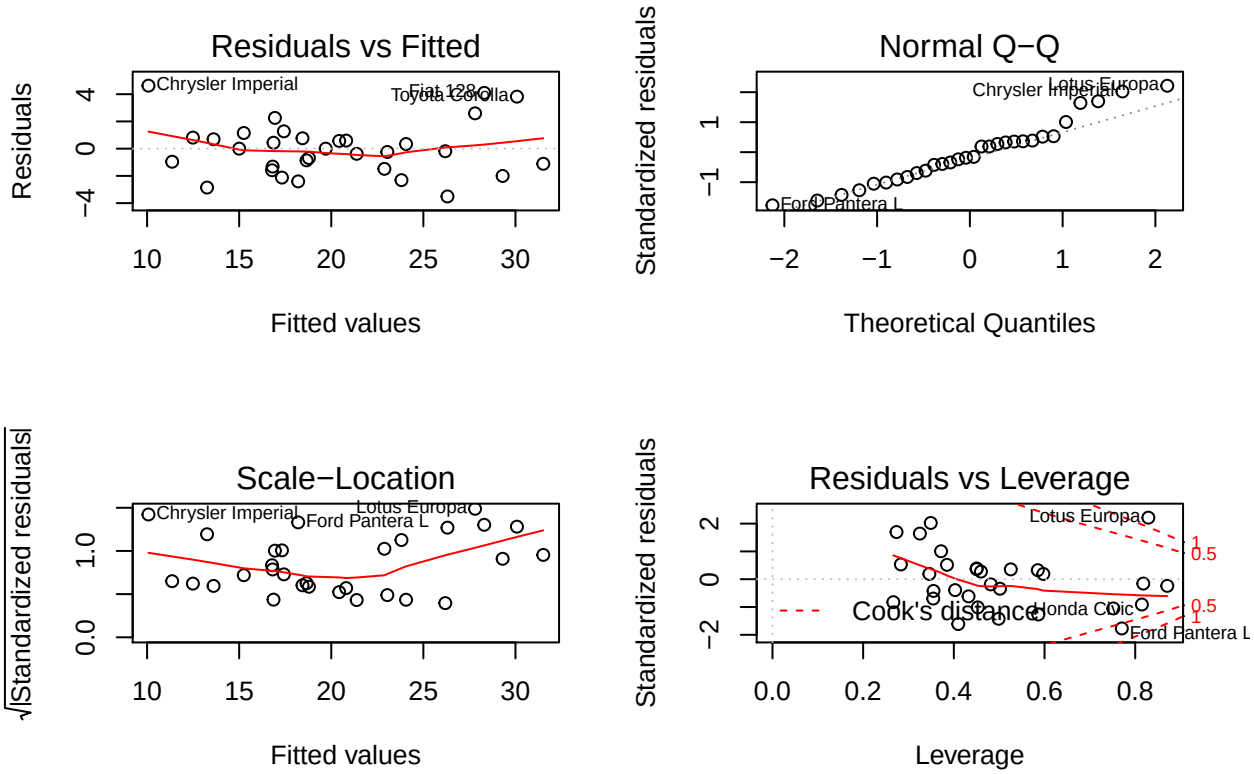
```
## -3.5087 -1.3584 -0.0948  0.7745  4.6251
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 23.87913   20.06582   1.190   0.2525
## cyl6        -2.64870    3.04089  -0.871   0.3975
## cyl8        -0.33616    7.15954  -0.047   0.9632
## disp         0.03555    0.03190   1.114   0.2827
## hp          -0.07051    0.03943  -1.788   0.0939 .
## drat         1.18283    2.48348   0.476   0.6407
## wt          -4.52978    2.53875  -1.784   0.0946 .
## qsec         0.36784    0.93540   0.393   0.6997
## vs1          1.93085    2.87126   0.672   0.5115
## am1          1.21212    3.21355   0.377   0.7113
## gear4        1.11435    3.79952   0.293   0.7733
## gear5        2.52840    3.73636   0.677   0.5089
## carb2       -0.97935    2.31797  -0.423   0.6787
## carb3        2.99964    4.29355   0.699   0.4955
## carb4        1.09142    4.44962   0.245   0.8096
## carb6        4.47757    6.38406   0.701   0.4938
## carb8        7.25041    8.36057   0.867   0.3995
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.833 on 15 degrees of freedom
## Multiple R-squared:  0.8931, Adjusted R-squared:  0.779
## F-statistic:  7.83 on 16 and 15 DF,  p-value: 0.000124
```

```
plot(fullmodel)
```

```
## Warning: not plotting observations with leverage one:
##   30, 31
```



```
## Warning: not plotting observations with leverage one:
##   30, 31
```



The above model explains 78% variance of mpg variable Residual standard error: 2.833 on 15 degrees of freedom, we need to find alternate model

```
par(mfrow=c(2,2))
altmodel <- step(lm(mpg ~ ., data=mtcars),k =3)
```

```
## Start:  AIC=93.4
## mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##
##        Df Sum of Sq    RSS    AIC
## - carb  5   13.5989 134.00 81.828
## - gear  2    3.9729 124.38 88.442
## - cyl   2   10.9314 131.33 90.184
## - am    1    1.1420 121.55 90.705
## - qsec  1    1.2413 121.64 90.732
## - drat  1    1.8208 122.22 90.884
## - vs    1    3.6299 124.03 91.354
## - disp  1    9.9672 130.37 92.948
## <none>              120.40 93.403
## - wt    1   25.5541 145.96 96.562
## - hp    1   25.6715 146.07 96.588
##
## Step:  AIC=81.83
## mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear
##
##        Df Sum of Sq    RSS    AIC
## - gear  2    5.0215 139.02 77.005
## - cyl   2   12.5642 146.57 78.696
## - disp  1    0.9934 135.00 79.064
```

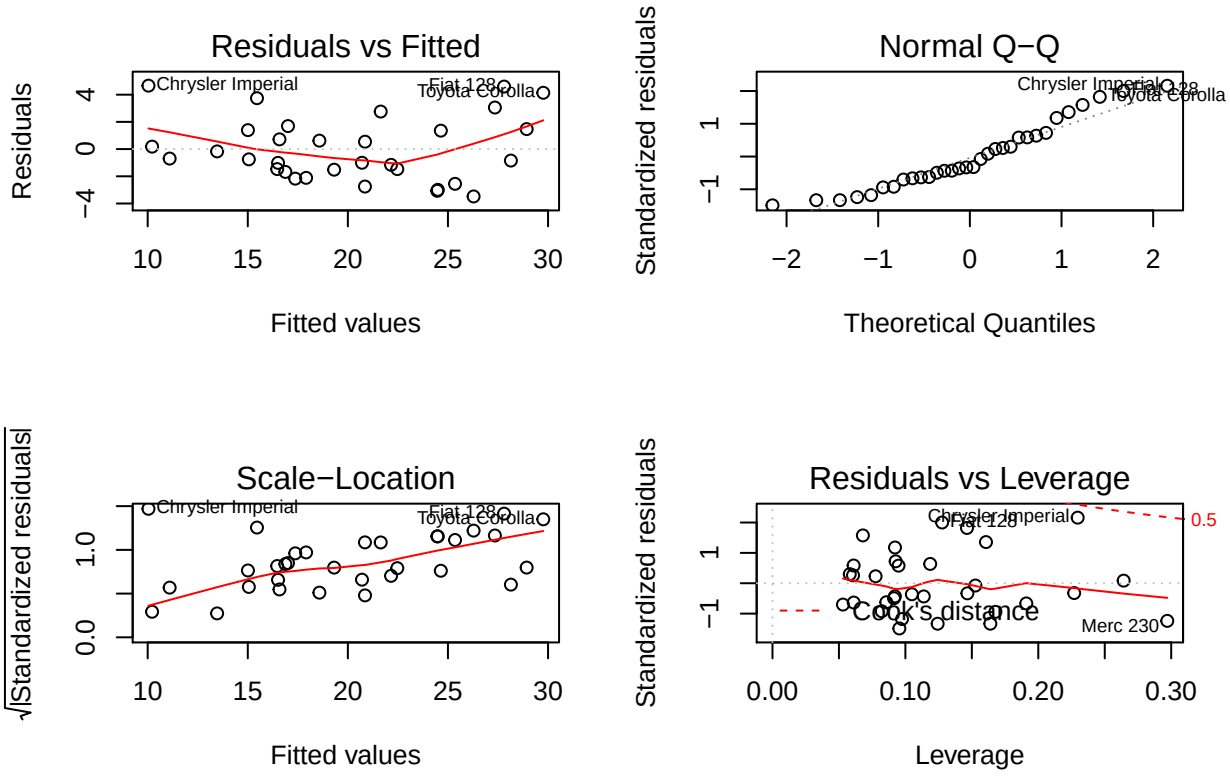
```
## - drat  1    1.1854 135.19 79.110
## - vs    1    3.6763 137.68 79.694
## - qsec  1    5.2634 139.26 80.061
## - am    1   11.9255 145.93 81.556
## <none>              134.00 81.828
## - wt    1   19.7963 153.80 83.237
## - hp    1   22.7935 156.79 83.855
##
## Step:  AIC=77
## mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am
##
##        Df Sum of Sq    RSS    AIC
## - cyl   2   10.4247 149.45 73.319
## - drat  1    0.9672 139.99 74.227
## - disp  1    1.5483 140.57 74.359
## - vs    1    2.1829 141.21 74.503
## - qsec  1    3.6324 142.66 74.830
## <none>              139.02 77.005
## - am    1   16.5665 155.59 77.608
## - hp    1   18.1768 157.20 77.937
## - wt    1   31.1896 170.21 80.482
##
## Step:  AIC=73.32
## mpg ~ disp + hp + drat + wt + qsec + vs + am
##
##        Df Sum of Sq    RSS    AIC
## - vs    1     0.645 150.09 70.457
## - drat  1     2.869 152.32 70.927
## - disp  1     9.111 158.56 72.212
## - qsec  1    12.573 162.02 72.904
## - hp    1    13.929 163.38 73.170
## <none>              149.45 73.319
## - am    1    20.457 169.91 74.424
## - wt    1    60.936 210.38 81.262
##
## Step:  AIC=70.46
## mpg ~ disp + hp + drat + wt + qsec + am
##
##        Df Sum of Sq    RSS    AIC
## - drat  1     3.345 153.44 68.162
## - disp  1     8.545 158.64 69.229
## - hp    1    13.285 163.38 70.171
## <none>              150.09 70.457
## - am    1    20.036 170.13 71.466
## - qsec  1    25.574 175.67 72.491
## - wt    1    67.572 217.66 79.351
##
## Step:  AIC=68.16
## mpg ~ disp + hp + wt + qsec + am
##
##        Df Sum of Sq    RSS    AIC
## - disp  1     6.629 160.07 66.515
## - hp    1    12.572 166.01 67.682
## <none>              153.44 68.162
```

```
## - qsec  1    26.470 179.91 70.255
## - am    1    32.198 185.63 71.258
## - wt    1    69.043 222.48 77.051
##
## Step:  AIC=66.52
## mpg ~ hp + wt + qsec + am
##
##        Df Sum of Sq    RSS    AIC
## - hp    1     9.219 169.29 65.307
## <none>              160.07 66.515
## - qsec  1    20.225 180.29 67.323
## - am    1    25.993 186.06 68.331
## - wt    1    78.494 238.56 76.284
##
## Step:  AIC=65.31
## mpg ~ wt + qsec + am
##
##        Df Sum of Sq    RSS    AIC
## <none>              169.29 65.307
## - am    1    26.178 195.46 66.908
## - qsec  1   109.034 278.32 78.217
## - wt    1   183.347 352.63 85.790
```

```
summary(altmodel)
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.6178     6.9596   1.382 0.177915
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## am1           2.9358     1.4109   2.081 0.046716 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.459 on 28 degrees of freedom
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8336
## F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

```
plot(altmodel)
```

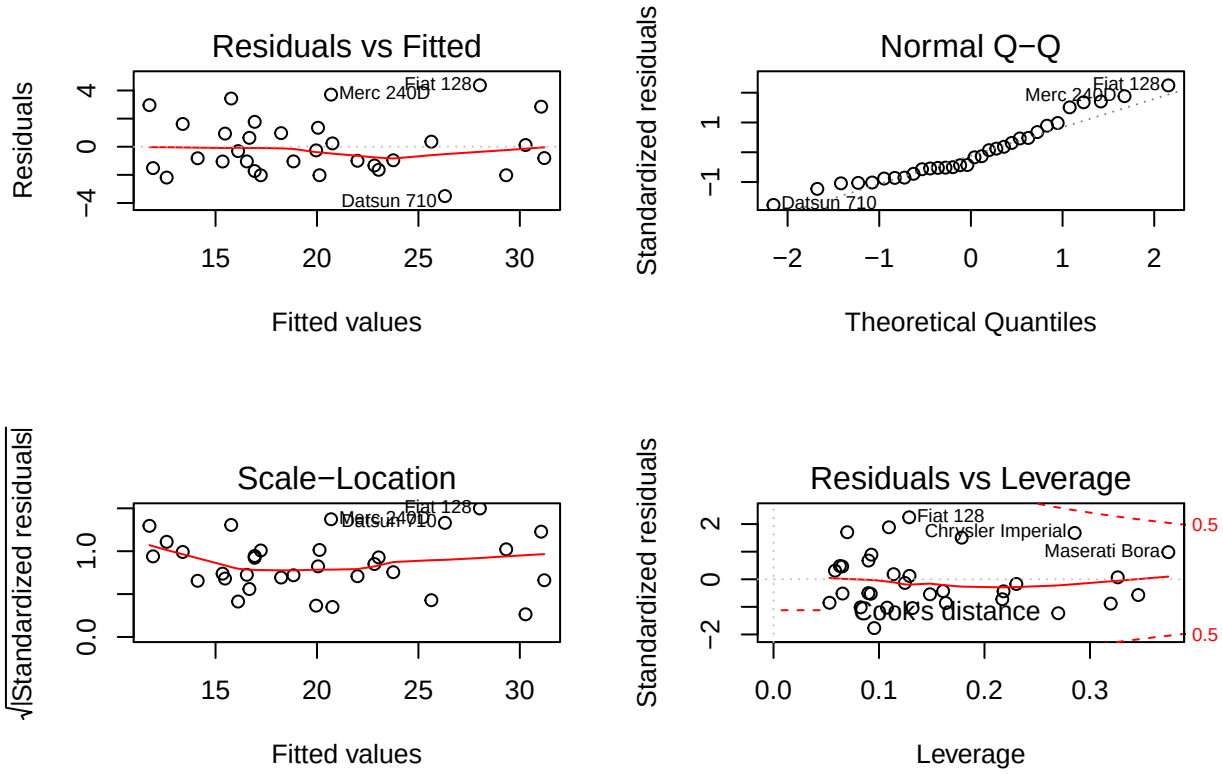


The above model explains 84% variance of mpg variable

```
par(mfrow=c(2,2))
altmodel2 <- lm(mpg ~ wt*am + qsec, data=mtcars)
summary(altmodel2)
```

```
##
## Call:
## lm(formula = mpg ~ wt * am + qsec, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5076 -1.3801 -0.5588  1.0630  4.3684
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    9.723      5.899   1.648 0.110893
## wt            -2.937      0.666  -4.409 0.000149 ***
## am1           14.079      3.435   4.099 0.000341 ***
## qsec           1.017      0.252   4.035 0.000403 ***
## wt:am1        -4.141      1.197  -3.460 0.001809 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.084 on 27 degrees of freedom
## Multiple R-squared:  0.8959, Adjusted R-squared:  0.8804
## F-statistic: 58.06 on 4 and 27 DF,  p-value: 7.168e-13
```

```
plot(altmodel2)
```

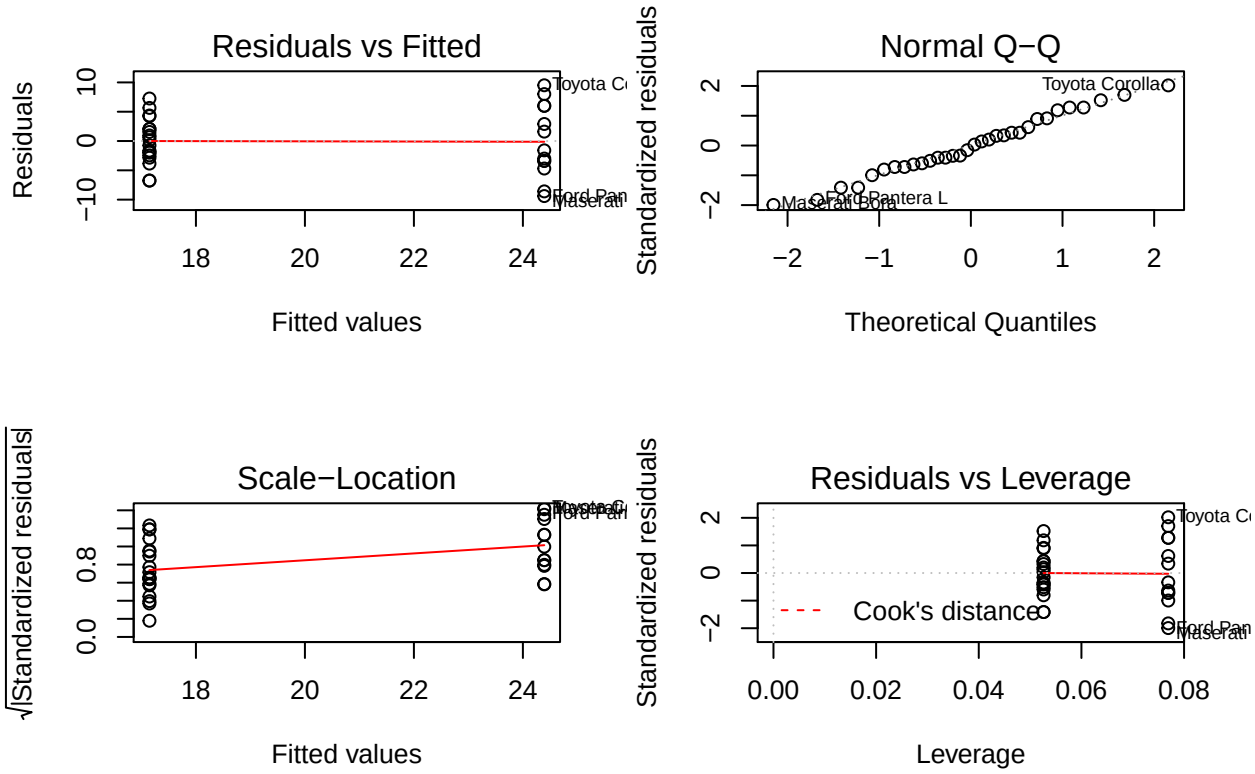


The above model explains 88% variance of mpg variable

```
par(mfrow=c(2,2))
altmodel3 <- lm(mpg ~ am, data=mtcars)
summary(altmodel3)
```

```
##
## Call:
## lm(formula = mpg ~ am, data = mtcars)
##
## 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 ***
## am1            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
```

```r
plot(altmodel3)
```



The above model explains 34% variance of mpg variable

Now we try to select the final model

```r
anova(altmodel3,altmodel,fullmodel,altmodel2)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ wt + qsec + am
## Model 3: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
## Model 4: mpg ~ wt * am + qsec
##   Res.Df    RSS  Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     28 169.29   2    551.61 34.3604 2.509e-06 ***
## 3     15 120.40  13     48.88  0.4685    0.9114
## 4     27 117.28 -12      3.13
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(altmodel2)$coef
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  9.723053  5.8990407  1.648243 0.1108925394
## wt          -2.936531  0.6660253 -4.409038 0.0001488947
## am1         14.079428  3.4352512  4.098515 0.0003408693
```

```
## qsec         1.016974  0.2520152  4.035366 0.0004030165
## wt:am1      -4.141376  1.1968119 -3.460340 0.0018085763
```

From the above "mpg ~ qsec + wt*am"" has the highest adjusted R-Squared values

above model suggest that cars with manual transmission add more mileage on decreasing weight with the equation 14.08 -2.94*wt.

```
max(cooks.distance(altmodel2))
```

```
## [1] 0.225106
```

```
par(mfrow=c(1,1))
plot(cooks.distance(altmodel2))
```