Regression Models

July 25, 2015

The aim of the project is to explore the Motor Trend Car data set. Motor Trend Car data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973-74 models).

Exploratoy Data Analysis

Using the summary function, the range of values is explored for variables in the dataset. Also, a correlation chart and matrix is evaluated. From the correlation matrix (Figure 2), it is noticed that -

a. MPG has a high negative correlation with cyl, disp, hp and wt

b. MPG has a strong correlation with drat, qsec, vs and am

```
summary(mtcars)
```

```
##       mpg             cyl             disp             hp       
##  Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0  
##  1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5  
##  Median :19.20   Median :6.000   Median :196.3   Median :123.0  
##  Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7  
##  3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0  
##  Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0  
##       drat             wt             qsec             vs        
##  Min.   :2.760   Min.   :1.513   Min.   :14.50   Min.   :0.0000  
##  1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000  
##  Median :3.695   Median :3.325   Median :17.71   Median :0.0000  
##  Mean   :3.597   Mean   :3.217   Mean   :17.85   Mean   :0.4375  
##  3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000  
##  Max.   :4.930   Max.   :5.424   Max.   :22.90   Max.   :1.0000  
##        am              gear            carb      
##  Min.   :0.0000   Min.   :3.000   Min.   :1.000  
##  1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000  
##  Median :0.0000   Median :4.000   Median :2.000  
##  Mean   :0.4062   Mean   :3.688   Mean   :2.812  
##  3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000  
##  Max.   :1.0000   Max.   :5.000   Max.   :8.000  
```

```
head(mtcars, 10)
```

```
##                    mpg cyl  disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8 360.0 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6 225.0 105 2.76 3.460 20.22  1  0    3    1
## Duster 360        14.3   8 360.0 245 3.21 3.570 15.84  0  0    3    4
## Merc 240D         24.4   4 146.7  62 3.69 3.190 20.00  1  0    4    2
## Merc 230          22.8   4 140.8  95 3.92 3.150 22.90  1  0    4    2
## Merc 280          19.2   6 167.6 123 3.92 3.440 18.30  1  0    4    4
```

Comparing Automatic and Manual Transmission

The two sample t-test is used to determine if Automatic or Manual Transmission is better for mpg. Under the null hypothesis, it is assumed that mean of mpg is same for both Automatic and Manual Transmission.

From the t-test it is seen that Manual Transmissions have a higher mean mpg of 24.39 compared to Automatic Transmissions mean of 17.147. This is also visible in Figure 3 - the boxplot of mpg vs am.

From the t-test, the p value is less than 0.001, therefore the Null Hypothesis is rejected. The mean of mpg for both Automatic and Manual Transmissions are different and from t-test it can be inferred that they come from two different populations.

```
t.test(mtcars$mpg~mtcars$am)
```

```
## 
##  Welch Two Sample t-test
## 
## data:  mtcars$mpg by mtcars$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
```

Building regression Models

A full regression model is built to predict mpg with all the other predictor variables. From the model -

a. The overall model has a high adjusted R-squared of 0.8066. The model can explain 80.6% of the variance in mpg.

b. Keeping all other factors constant, it can be estimated that mpg increases by 2.5 when we move from automatic to manual transmissions. Thus it again proves the point that Automatic Transmissions are better.

c. However, all the predictors do not seem to have a significant contribution in the model. It looks like certain variables can be removed from the model.

A Step wise regression model is used for better selection of variables. The new model has:

a. A better adjusted R-squared of 0.8336.
b. It only uses three predictor variables - wt, qsec and am, and all are significant.
c. From the model, it can be inferred, that mpg increases with increase in qsec and when transmission changes from automatic to manual. However, mpg decreases with increase in weight of car.

Model 2 looks better.

```
l1<-lm(mpg=~.,data=mtcars)
```

```
## Warning in lm.fit(x, y, offset = offset, singular.ok = singular.ok, ...):
## extra argument 'mpg' is disregarded.
```

```
summary(l1)
```

```
## 
## Call:
## lm(data = mtcars, mpg = ~.)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -3.4506 -1.6044 -0.1196  1.2193  4.6271 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 12.30337   18.71788   0.657   0.5181  
## cyl         -0.11144    1.04502  -0.107   0.9161  
## disp         0.01334    0.01786   0.747   0.4635  
## hp          -0.02148    0.02177  -0.987   0.3350  
## drat         0.78711    1.63537   0.481   0.6353  
## wt          -3.71530    1.89441  -1.961   0.0633 .
## qsec         0.82104    0.73084   1.123   0.2739  
## vs           0.31776    2.10451   0.151   0.8814  
## am           2.52023    2.05665   1.225   0.2340  
## gear         0.65541    1.49326   0.439   0.6652  
## carb        -0.19942    0.82875  -0.241   0.8122  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 2.65 on 21 degrees of freedom
## Multiple R-squared:  0.869,  Adjusted R-squared:  0.8066 
## F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```

```
l2<-step(l1)
```

```
## Start:  AIC=70.9
## mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
## 
##        Df Sum of Sq    RSS    AIC
## - cyl   1    0.0799 147.57 68.915
## - vs    1    0.1601 147.66 68.932
## - carb  1    0.4067 147.90 68.986
## - gear  1    1.3531 148.85 69.190
## - drat  1    1.6270 149.12 69.249
## - disp  1    3.9167 151.41 69.736
## - hp    1    6.8399 154.33 70.348
## - qsec  1    8.8641 156.36 70.765
## <none>              147.49 70.898
## - am    1   10.5467 158.04 71.108
## - wt    1   27.0144 174.51 74.280
```

```
## Warning in lm.fit(x, y, offset = offset, singular.ok = singular.ok, ...):
## extra argument 'mpg' is disregarded.
```

```
## 
## Step:  AIC=68.92
## mpg ~ disp + hp + drat + wt + qsec + vs + am + gear + carb
## 
##        Df Sum of Sq    RSS    AIC
## - vs    1    0.2685 147.84 66.973
## - carb  1    0.5201 148.09 67.028
## - gear  1    1.8211 149.40 67.308
## - drat  1    1.9826 149.56 67.342
## - disp  1    3.9009 151.47 67.750
## - hp    1    7.3632 154.94 68.473
## <none>              147.57 68.915
## - qsec  1   10.0933 157.67 69.032
## - am    1   11.8359 159.41 69.384
## - wt    1   27.0280 174.60 72.297
```

```
## Warning in lm.fit(x, y, offset = offset, singular.ok = singular.ok, ...):
## extra argument 'mpg' is disregarded.
```

```
##
## Step:  AIC=66.97
## mpg ~ disp + hp + drat + wt + qsec + am + gear + carb
##
##        Df Sum of Sq    RSS    AIC
## - carb  1    0.6855 148.53 65.121
## - gear  1    2.1437 149.99 65.434
## - drat  1    2.2139 150.06 65.449
## - disp  1    3.6467 151.49 65.753
## - hp    1    7.1060 154.95 66.475
## <none>              147.84 66.973
## - am    1   11.5694 159.41 67.384
## - qsec  1   15.6830 163.53 68.200
## - wt    1   27.3799 175.22 70.410
```

```
## Warning in lm.fit(x, y, offset = offset, singular.ok = singular.ok, ...):
## extra argument 'mpg' is disregarded.
```

```
##
## Step:  AIC=65.12
## mpg ~ disp + hp + drat + wt + qsec + am + gear
##
##        Df Sum of Sq    RSS    AIC
## - gear  1     1.565 150.09 63.457
## - drat  1     1.932 150.46 63.535
## <none>              148.53 65.121
## - disp  1    10.110 158.64 65.229
## - am    1    12.323 160.85 65.672
## - hp    1    14.826 163.35 66.166
## - qsec  1    26.408 174.94 68.358
## - wt    1    69.127 217.66 75.350
```

```
## Warning in lm.fit(x, y, offset = offset, singular.ok = singular.ok, ...):
## extra argument 'mpg' is disregarded.
```

```
##
## Step:  AIC=63.46
## mpg ~ disp + hp + drat + wt + qsec + am
##
##        Df Sum of Sq    RSS    AIC
## - drat  1     3.345 153.44 62.162
## - disp  1     8.545 158.64 63.229
## <none>              150.09 63.457
## - hp    1    13.285 163.38 64.171
## - am    1    20.036 170.13 65.466
## - qsec  1    25.574 175.67 66.491
## - wt    1    67.572 217.66 73.351
```

```
## Warning in lm.fit(x, y, offset = offset, singular.ok = singular.ok, ...):
## extra argument 'mpg' is disregarded.
```

```
##
## Step:  AIC=62.16
## mpg ~ disp + hp + wt + qsec + am
##
##        Df Sum of Sq    RSS    AIC
## - disp  1     6.629 160.07 61.515
## <none>              153.44 62.162
## - hp    1    12.572 166.01 62.682
## - qsec  1    26.470 179.91 65.255
## - am    1    32.198 185.63 66.258
## - wt    1    69.043 222.48 72.051
```

```
## Warning in lm.fit(x, y, offset = offset, singular.ok = singular.ok, ...):
## extra argument 'mpg' is disregarded.
```

```
##
## Step:  AIC=61.52
## mpg ~ hp + wt + qsec + am
##
##        Df Sum of Sq    RSS    AIC
## - hp    1     9.219 169.29 61.307
## <none>              160.07 61.515
## - qsec  1    20.225 180.29 63.323
## - am    1    25.993 186.06 64.331
## - wt    1    78.494 238.56 72.284
```

```
## Warning in lm.fit(x, y, offset = offset, singular.ok = singular.ok, ...):
## extra argument 'mpg' is disregarded.
```

```
##
## Step:  AIC=61.31
## mpg ~ wt + qsec + am
##
##        Df Sum of Sq    RSS    AIC
## <none>              169.29 61.307
## - am    1    26.178 195.46 63.908
## - qsec  1   109.034 278.32 75.217
## - wt    1   183.347 352.63 82.790
```

```
summary(l2)
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = mtcars, mpg = ~.)
##
## 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 ***
## am            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
```

Regression Diagnostics and Outlier Detection

Cook’s distance is found for all observations in the dataset. From the cook’s distance value, “Chrysler Imperial” and “Merc 230” are highly influential. This is also shown in Figure 5. Therefore these two values are removed from the dataset and regression model is run again. An improvement is seen in the Adjusted R-squared which increases to 0.863. Other observations from figure 5 are

1. Residual QQ plots show there is normality of the errors.
2. Residuals vs Fitted plot show constant error variance and hence no problem of Heteroskedasticity.

Also, from the Components+ Residual Plots in Figure 4, it can be observed that there is a linear trend for all predictor variables.

```
m <-mtcars
sort(cooks.distance(l2)[1:32],decreasing=TRUE)
```

```
##   Chrysler Imperial            Merc 230            Fiat 128 
##        0.3475974030        0.1620826668        0.1464019096 
##      Toyota Corolla        Lotus Europa       Toyota Corona 
##        0.1421983265        0.0879746271        0.0869042700 
##          Volvo 142E          Datsun 710    Pontiac Firebird 
##        0.0630461773        0.0584743854        0.0452965996 
##      Ford Pantera L             Valiant           Merc 240D 
##        0.0447392582        0.0375325879        0.0351911825 
##       Maserati Bora         AMC Javelin         Honda Civic 
##        0.0260597529        0.0190521001        0.0136145780 
##   Hornet Sportabout           Mazda RX4       Porsche 914-2 
##        0.0133030623        0.0091619033        0.0087942478 
##  Cadillac Fleetwood         Merc 450SLC           Merc 280C 
##        0.0076442583        0.0068973733        0.0065776629 
##       Mazda RX4 Wag        Ferrari Dino          Merc 450SE 
##        0.0061442653        0.0060670208        0.0055484603 
##    Dodge Challenger          Duster 360           Fiat X1-9 
##        0.0047376388        0.0047336096        0.0038600038 
##          Merc 450SL      Hornet 4 Drive            Merc 280 
##        0.0013871851        0.0011099473        0.0010966079 
## Lincoln Continental          Camaro Z28 
##        0.0006541961        0.0002520759
```

```
cooksd <- cooks.distance(l2)
m<-cbind(m,cooksd)
a1<-subset(m,m$cooksd>(0.1465))
m<-m[!(m$disp%in% a1$disp),]
l2<-lm(mpg~wt+am+qsec,data=m)
summary(l2)
```

```
##
## Call:
## lm(formula = mpg ~ wt + am + qsec, data = m)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8287 -1.3586 -0.1393  1.3444  4.2737
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.1780     7.2690   1.400 0.173278
## wt           -4.5365     0.7304  -6.211 1.43e-06 ***
## am            2.3060     1.3425   1.718 0.097729 .
## qsec          1.3160     0.3090   4.259 0.000237 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.269 on 26 degrees of freedom
## Multiple R-squared:  0.8772, Adjusted R-squared:  0.863
## F-statistic: 61.89 on 3 and 26 DF,  p-value: 5.704e-12
```

Conclusion

In this project, we are able to make a good model which quantifies the relationship of mpg with other variables. From the model. it is seen that -

a. Manual Transmissions are better for mpg.

b. Light cars are better for mpg.

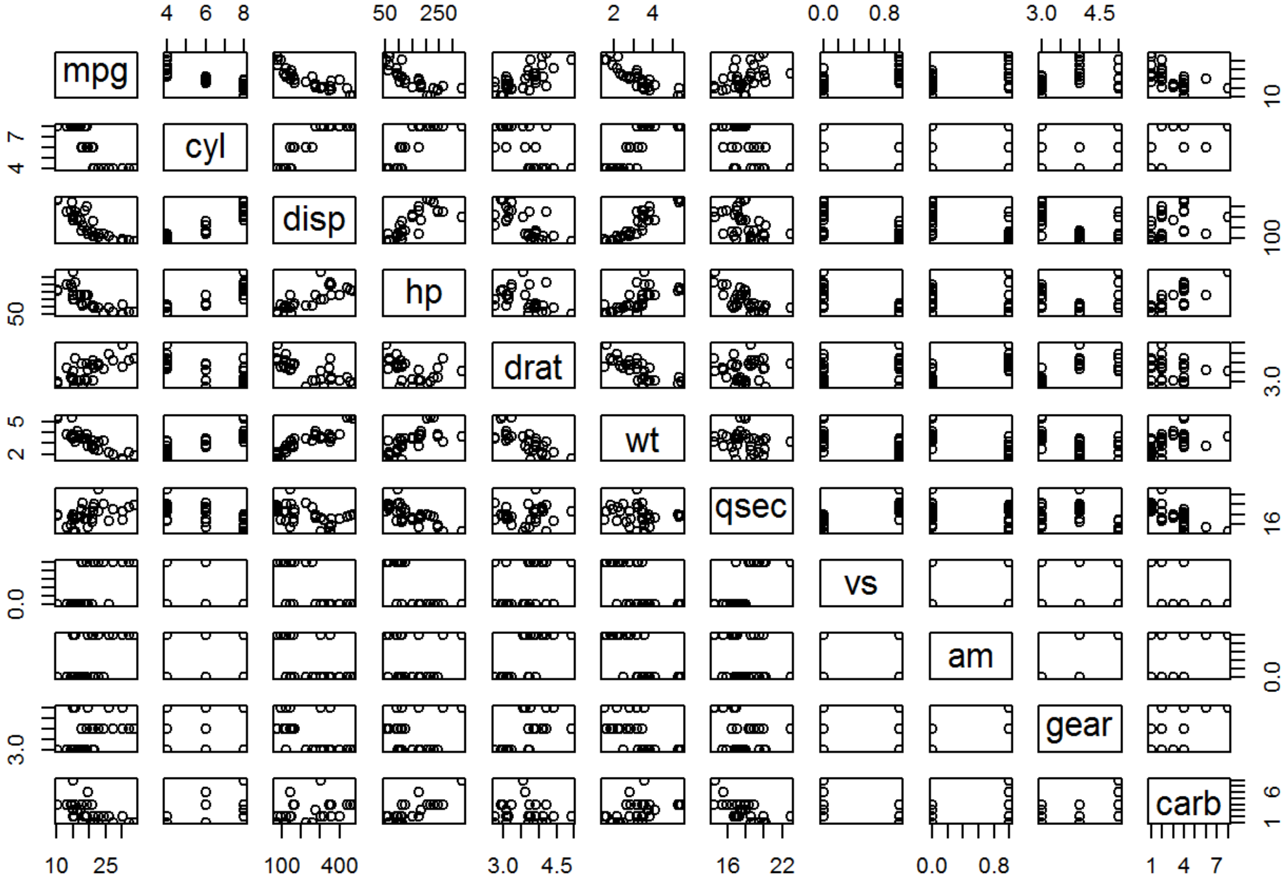
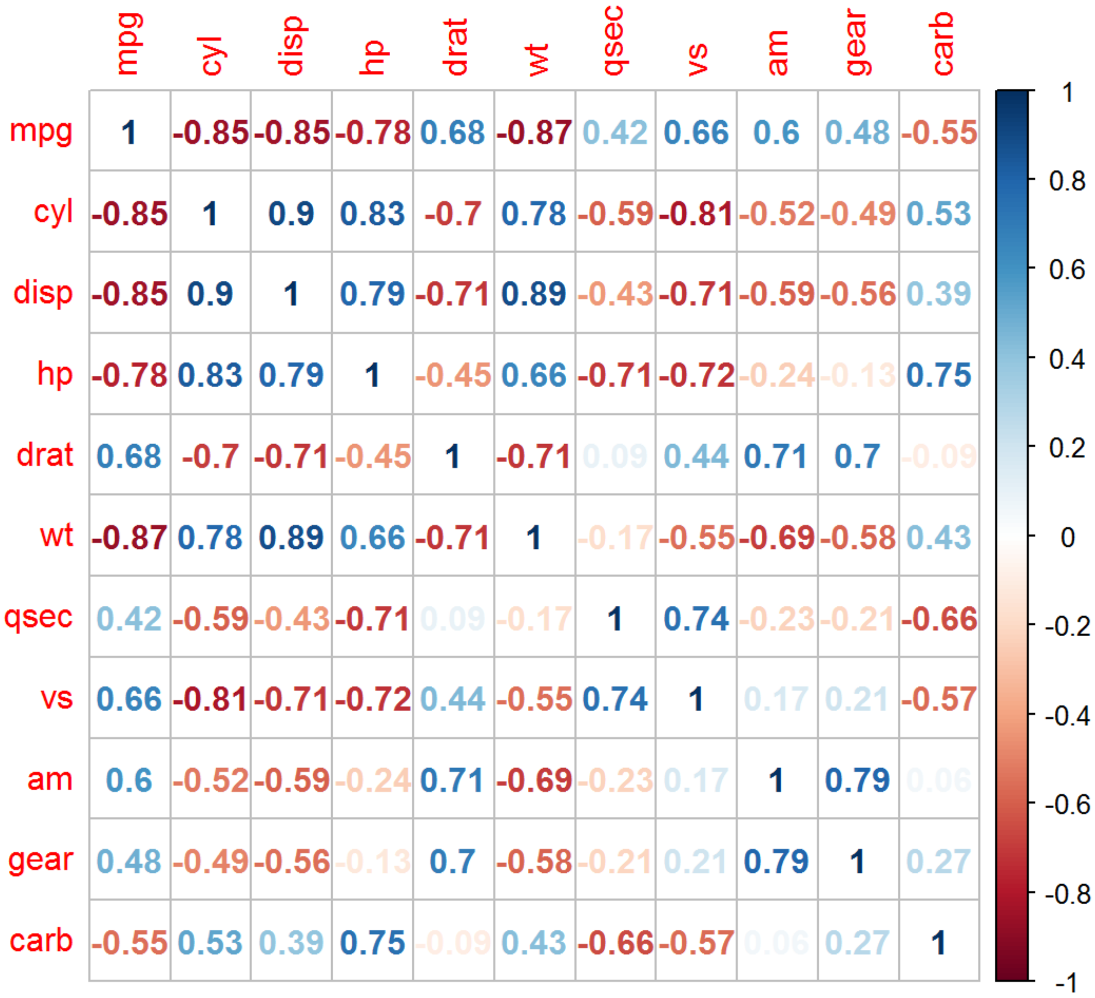

mpg
cyl
disp
hp
drat
wt
qsec
vs
am
gear
carb
4 6 8
50 250
2 4
0.0 0.8
3.0 4.5
10
4 7
100
50
3.0
2 5
16
0.0
3.0
1 6
10 25
100 400
3.0 4.5
16 22
0.0 0.8
1 4 7

Figure 1



	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
mpg	1	-0.85	-0.85	-0.78	0.68	-0.87	0.42	0.66	0.6	0.48	-0.55
cyl	-0.85	1	0.9	0.83	-0.7	0.78	-0.59	-0.81	-0.52	-0.49	0.53
disp	-0.85	0.9	1	0.79	-0.71	0.89	-0.43	-0.71	-0.59	-0.56	0.39
hp	-0.78	0.83	0.79	1	-0.45	0.66	-0.71	-0.72	-0.24	-0.13	0.75
drat	0.68	-0.7	-0.71	-0.45	1	-0.71	0.09	0.44	0.71	0.7	-0.09
wt	-0.87	0.78	0.89	0.66	-0.71	1	-0.17	-0.55	-0.69	-0.58	0.43
qsec	0.42	-0.59	-0.43	-0.71	0.09	-0.17	1	0.74	-0.23	-0.21	-0.66
vs	0.66	-0.81	-0.71	-0.72	0.44	-0.55	0.74	1	0.17	0.21	-0.57
am	0.6	-0.52	-0.59	-0.24	0.71	-0.69	-0.23	0.17	1	0.79	0.06
gear	0.48	-0.49	-0.56	-0.13	0.7	-0.58	-0.21	0.21	0.79	1	0.27
carb	-0.55	0.53	0.39	0.75	-0.09	0.43	-0.66	-0.57	0.06	0.27	1

Figure 2

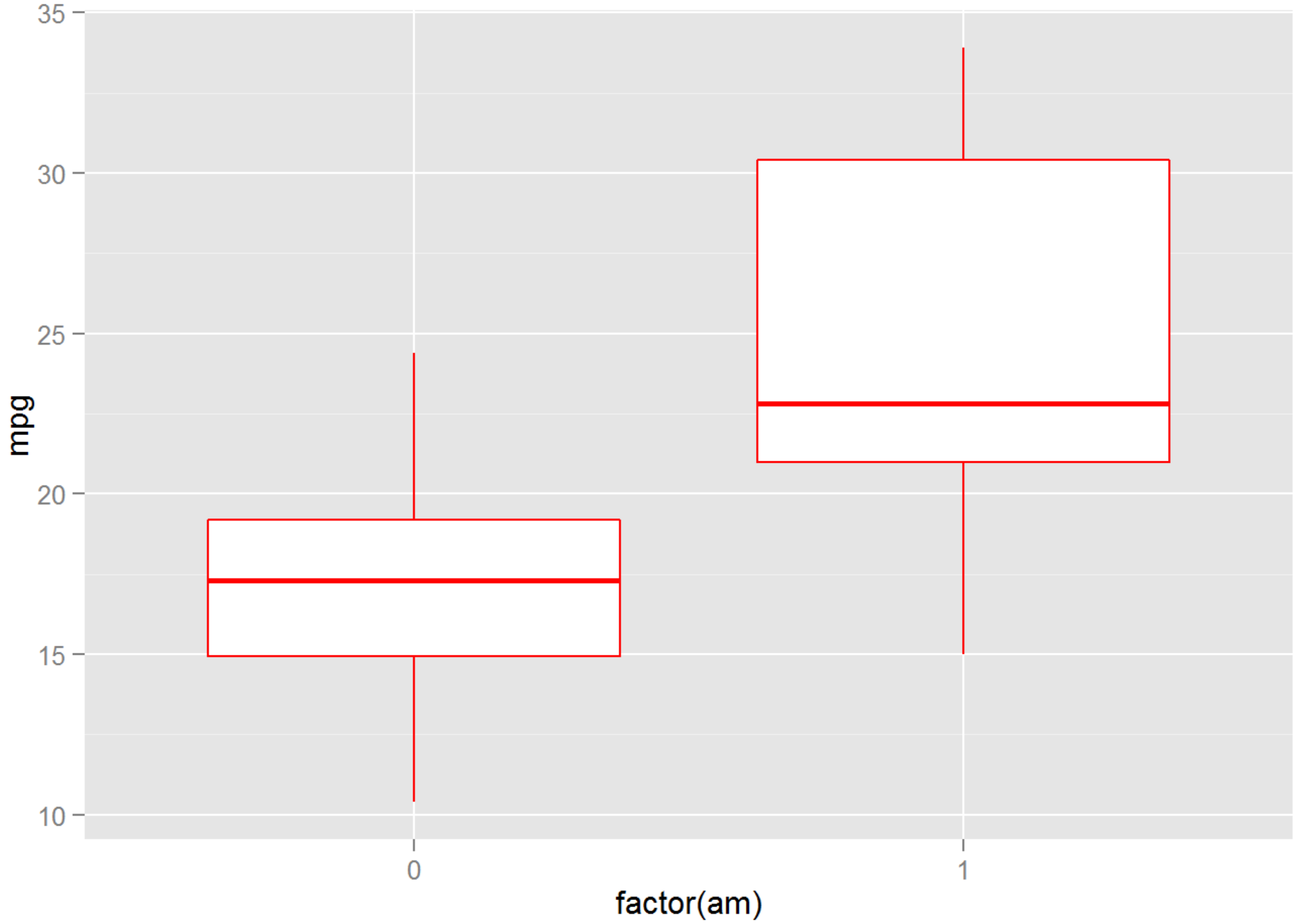


Figure 3

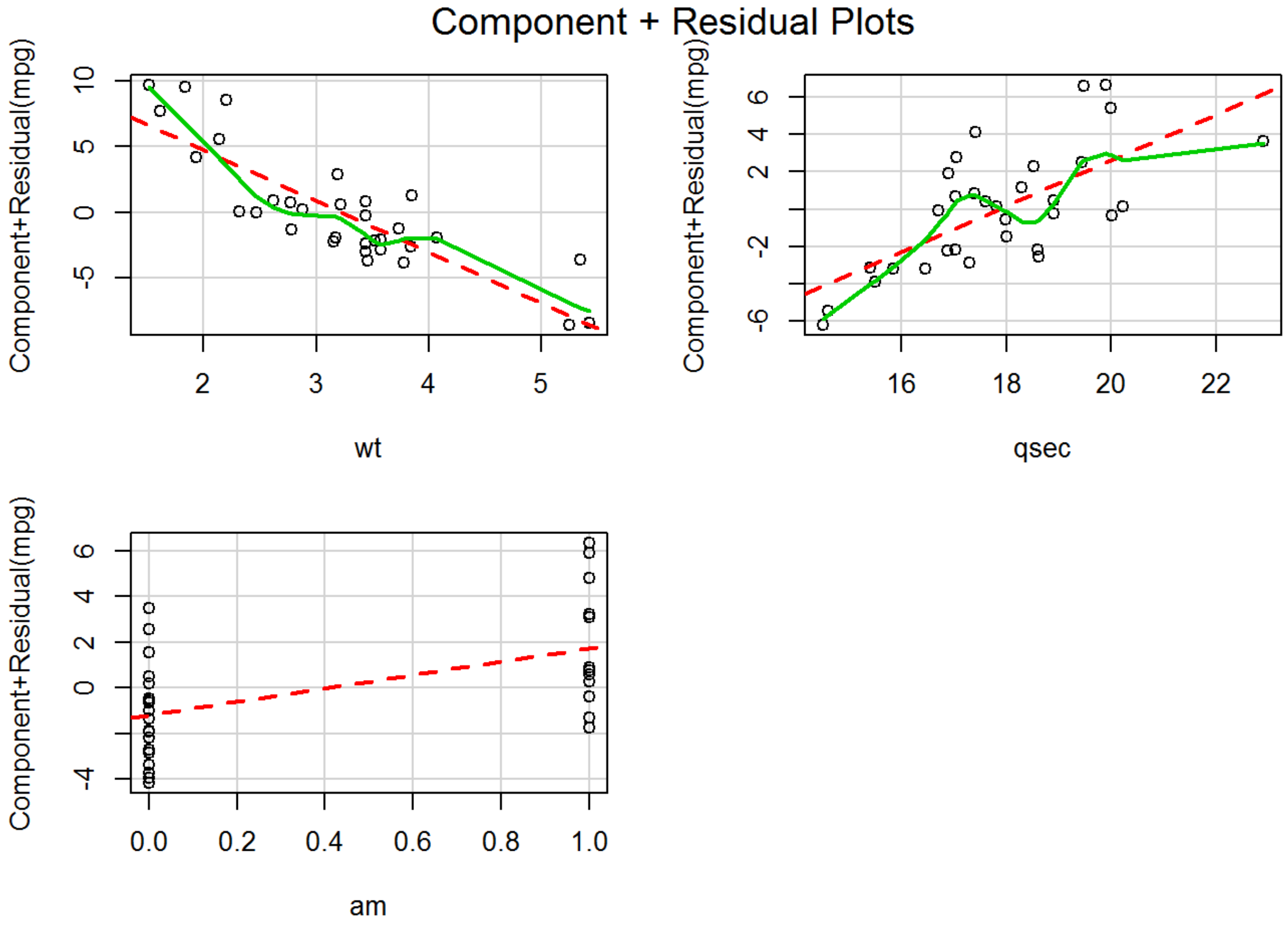

Component + Residual Plots
Component+Residual(mpg)
10
5
0
-5
2
3
4
5
wt
Component+Residual(mpg)
6
4
2
-2
-6
16
18
20
22
qsec
Component+Residual(mpg)
6
4
2
0
-4
0.0
0.2
0.4
0.6
0.8
1.0
am

Figure 4

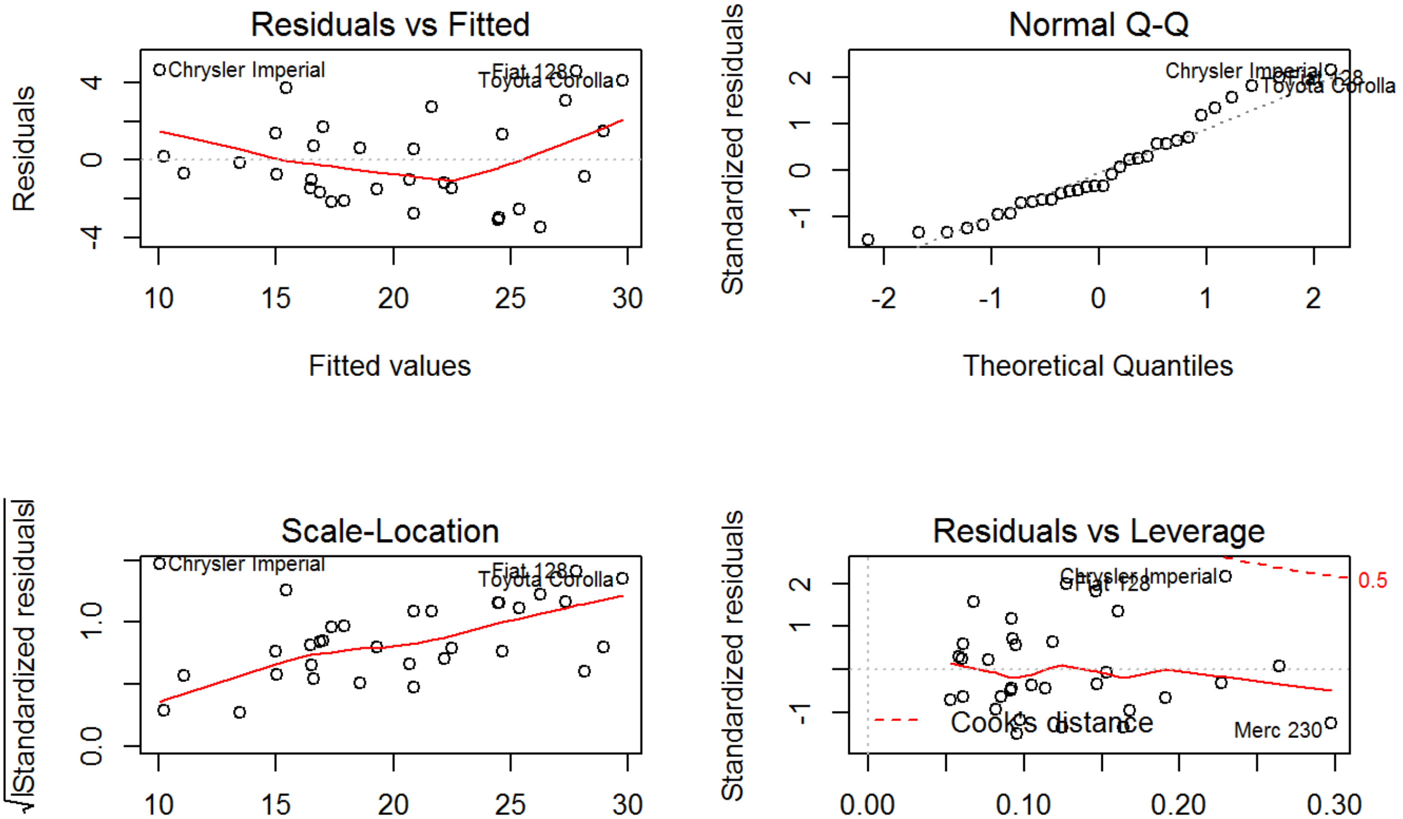

Residuals vs Fitted
Residuals
Fitted values
Chrysler Imperial
Fiat 128
Toyota Corolla
4
0
-4
10
15
20
25
30
Normal Q-Q
Standardized residuals
Theoretical Quantiles
-2
-1
0
1
2
Scale-Location
√|Standardized residuals|
0.0
1.0
Residuals vs Leverage
Leverage
Cook's distance
0.5
Merc 230
0.00
0.10
0.20
0.30

Figure 5