RegressionClassProject

Bassam Abdelnabi

9/9/2019

Synopsis

Motor Trend is a magazine that focuses on the automobile industry. Looking at a data set of a collection of cars, they are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). They are particularly interested in the following two questions:

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

This study shows that the autmoatic trasnmissions is better with an average saving of 1.4-2. It may be as high as 2.9 mpg. The models showing this conclusion have an adjusted RMS of 0.45-0.83. All the models agree that there is a benefit and saving. However, they disagree in the specific value. The author of this report recommends taking more data for validation. Please check the rmd code in the github repo to check the code.

Loading necessary libraries and data

Exploring and cleaning the data

Data structure is showing a lot of numerical while they should be listed as factors so we will correct that. We will check the data for NA, zero covariates or near zero. We will check collinearity. Let us check how this is done

```
## '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 ...
```

```
## '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 : Factor w/ 3 levels "4","6","8": 2 2 1 2 3 2 3 1 1 2 ...
##  $ 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  : Factor w/ 2 levels "0","1": 1 1 2 2 1 2 1 2 2 2 ...
##  $ am  : Factor w/ 2 levels "0","1": 2 2 2 1 1 1 1 1 1 1 ...
##  $ gear: Factor w/ 3 levels "3","4","5": 2 2 2 1 1 1 1 2 2 2 ...
##  $ carb: Factor w/ 6 levels "1","2","3","4",..: 4 4 1 1 2 1 4 2 2 4 ...
```

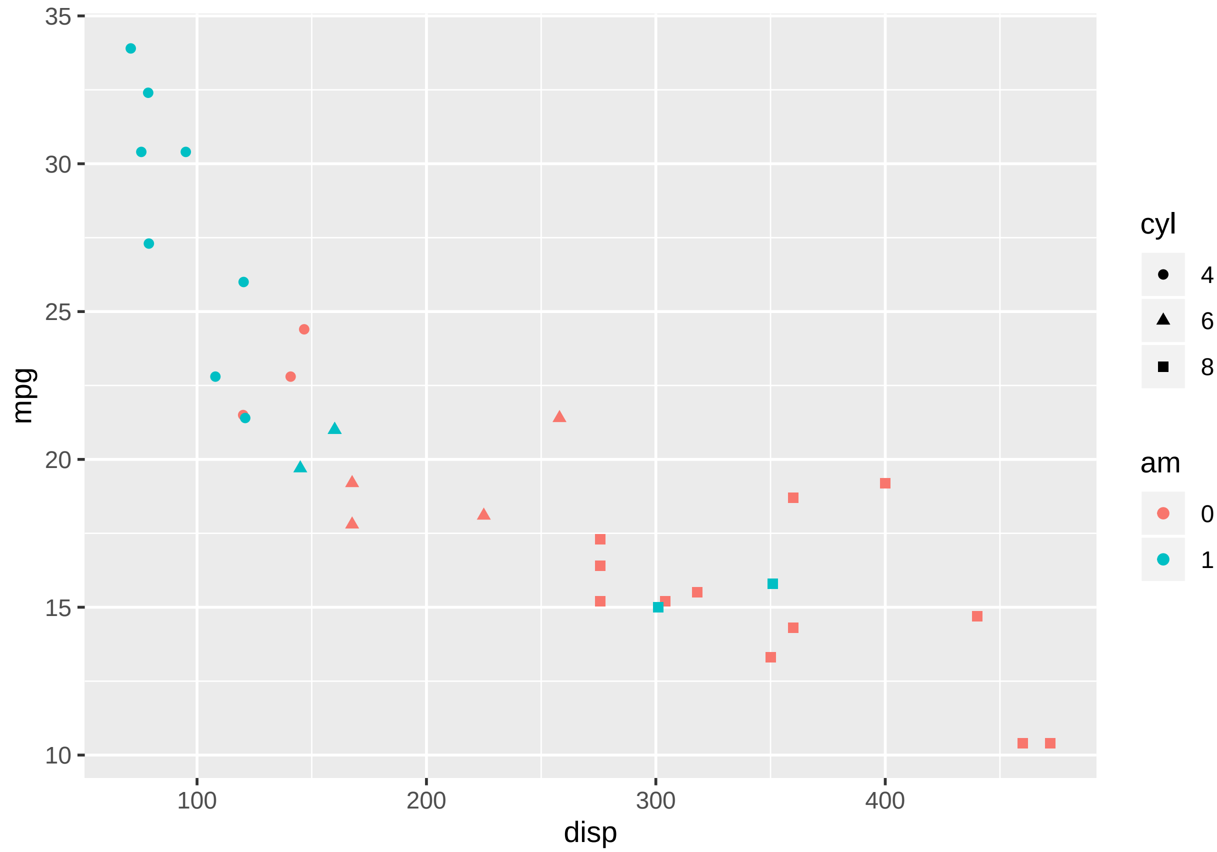
```
## [1] 0
```

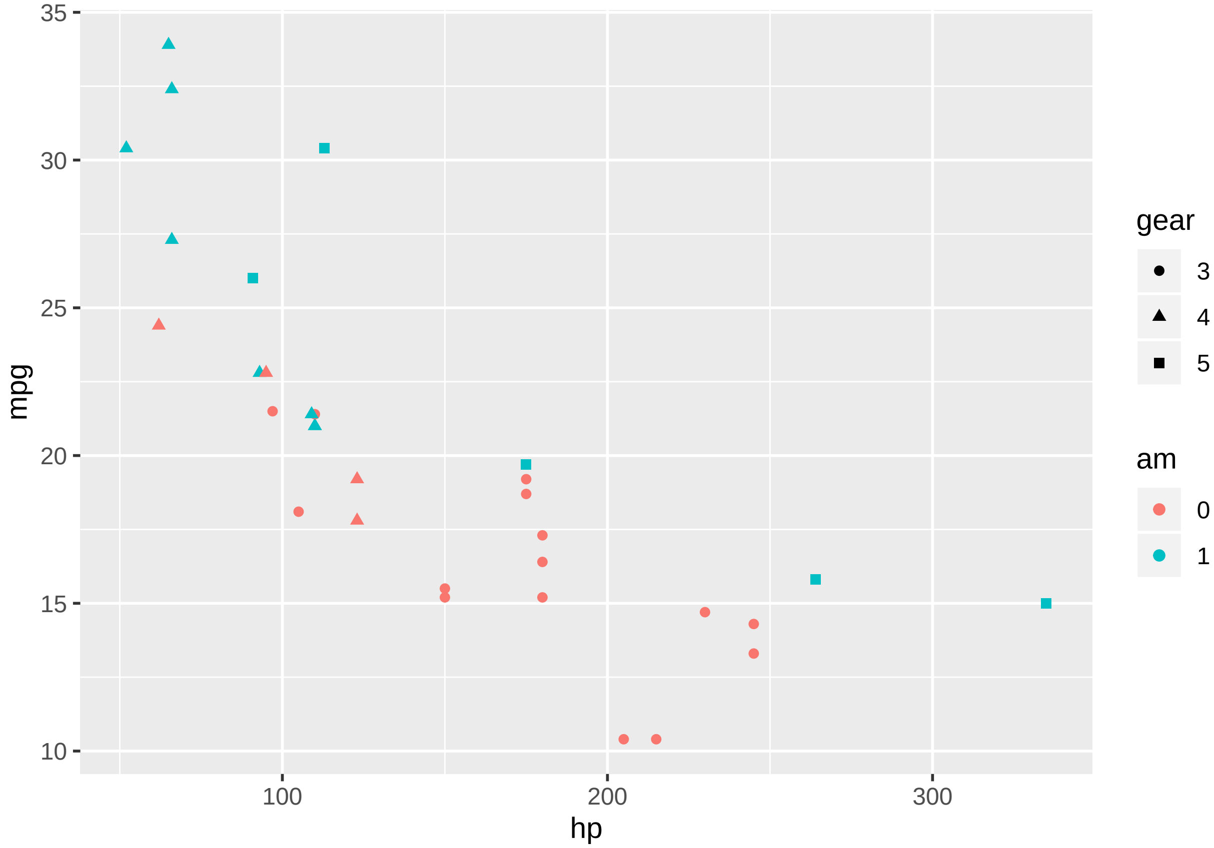
```
## [1] freqRatio     percentUnique zeroVar       nzv
## <0 rows> (or 0-length row.names)
```

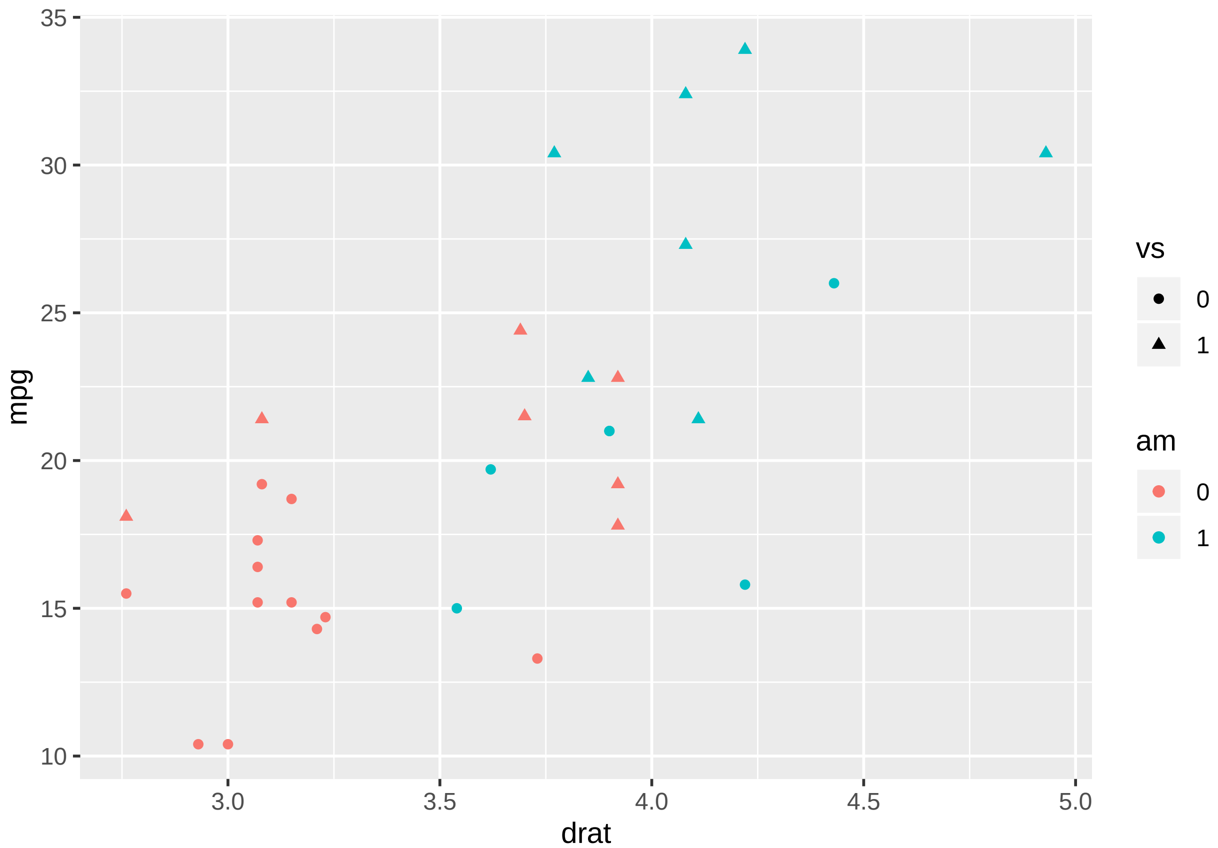
```
## [1] freqRatio     percentUnique zeroVar       nzv
## <0 rows> (or 0-length row.names)
```

Prelimanry estimates

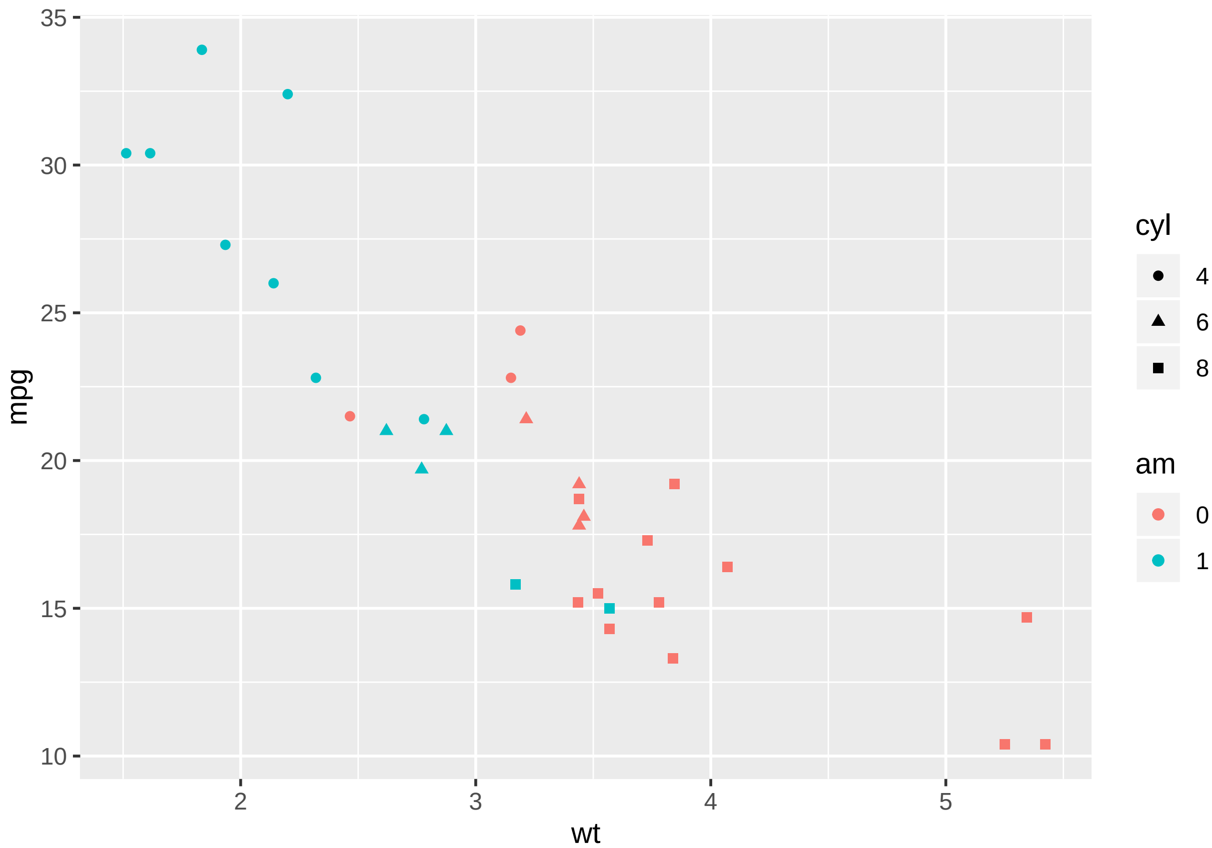
We have prepared the data set for analysis. We will start by a simple t test since the number of data is limited. We will do a plot of both groups as a visulaization of the problem. We will also do a plot of all variables to have a feel of the probelm under study. Let us see how to do that.

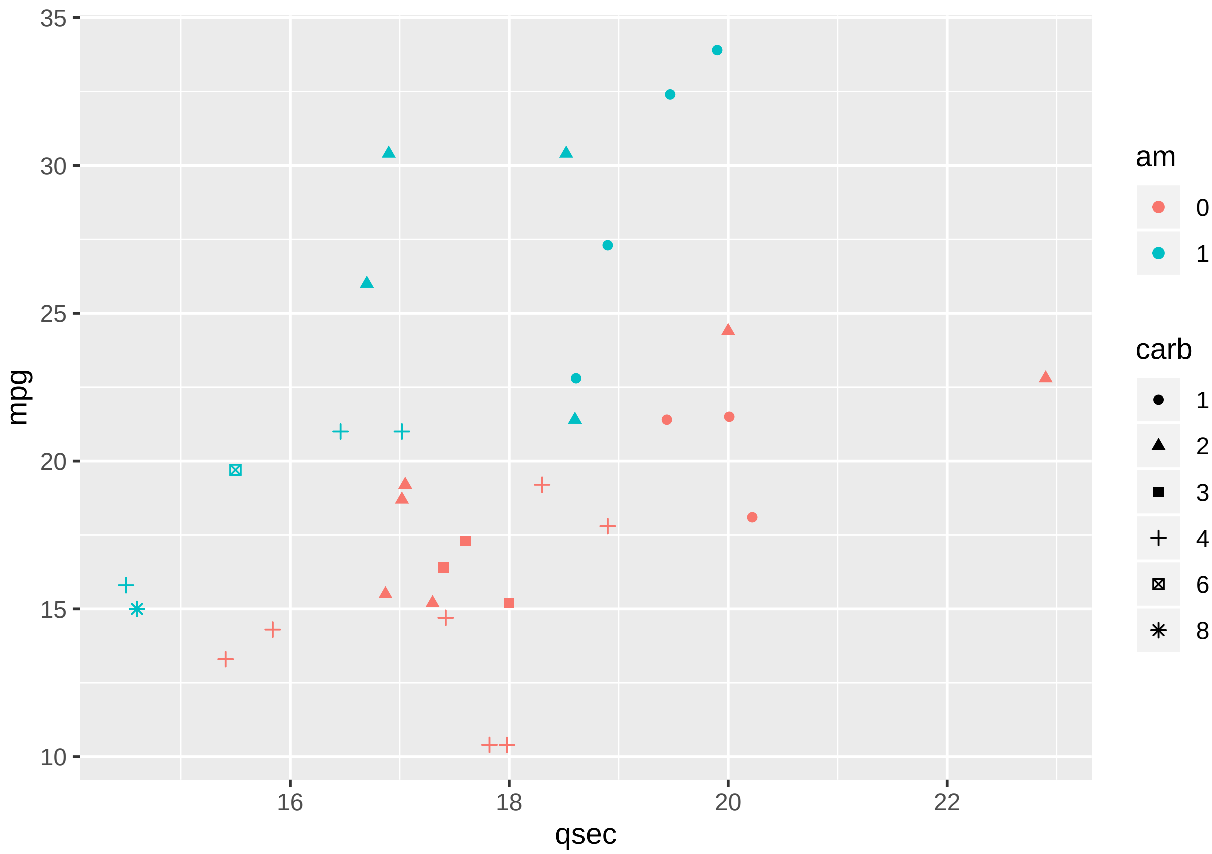



35
30
25
20
15
10
mpg
100
200
300
hp
gear
3
4
5
am
0
1



35
30
25
20
15
10
mpg
3.0
3.5
4.0
4.5
5.0
drat
vs
0
1
am
0
1


35
30
25
20
15
10
mpg
2
3
4
5
wt
cyl
4
6
8
am
0
1




```
## '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 : Factor w/ 3 levels "4","6","8": 2 2 1 2 3 2 3 1 1 2 ...
##  $ 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  : Factor w/ 2 levels "0","1": 1 1 2 2 1 2 1 2 2 2 ...
##  $ am  : Factor w/ 2 levels "0","1": 2 2 2 1 1 1 1 1 1 1 ...
##  $ gear: Factor w/ 3 levels "3","4","5": 2 2 2 1 1 1 1 2 2 2 ...
##  $ carb: Factor w/ 6 levels "1","2","3","4",..: 4 4 1 1 2 1 4 2 2 4 ...
```

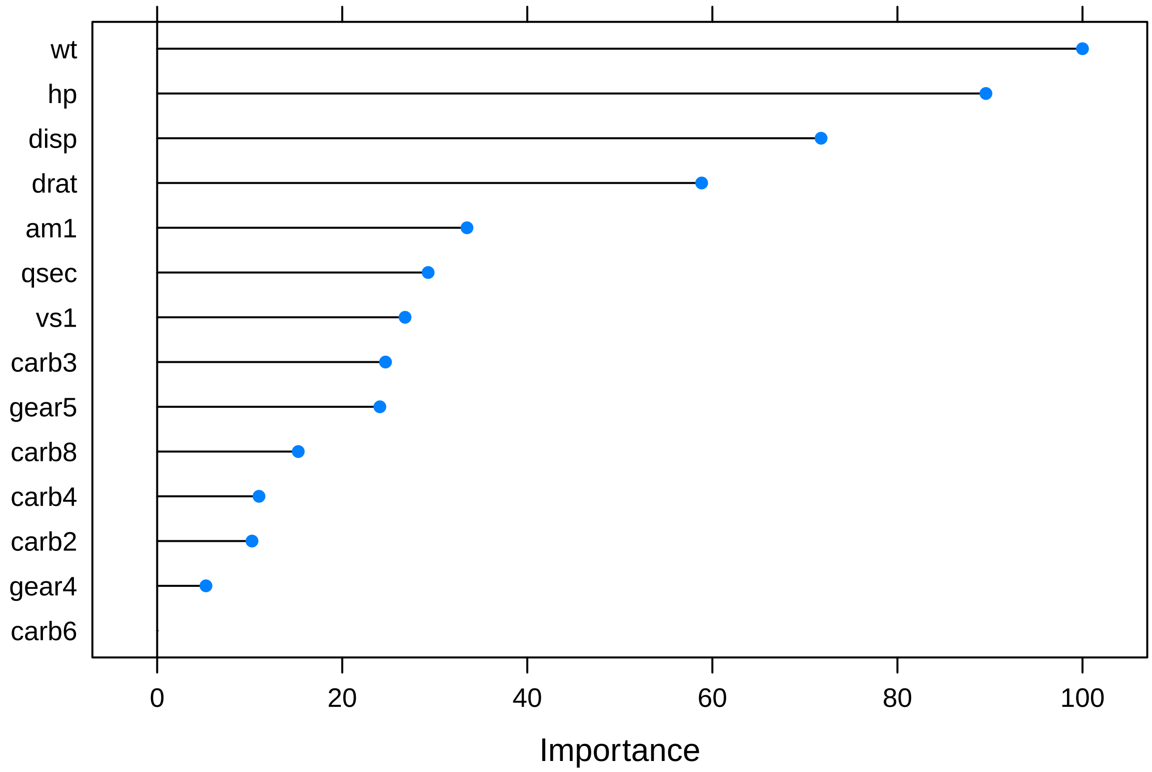
Building models and exploring them

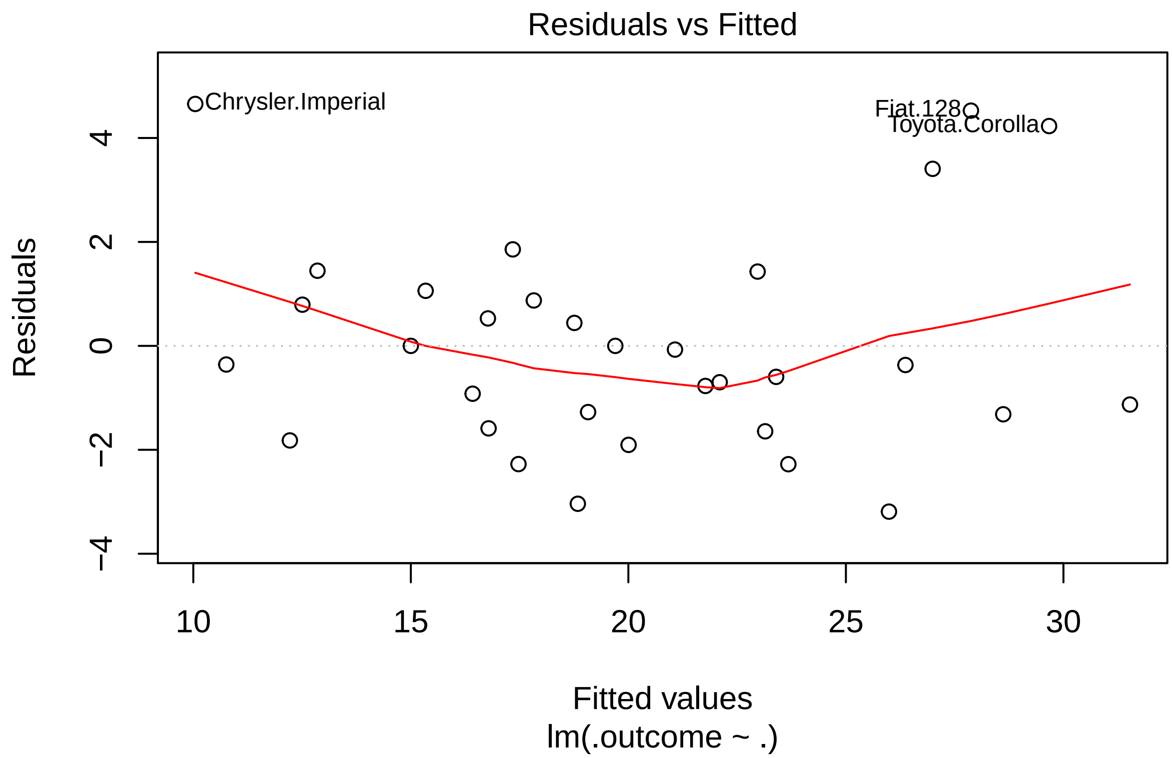
In the previous section we have seen that there is a difference in the mpg for both groups that is statistically significant. The plots show several variables correlate with the mpg, however, the displacment is sufficent to be express the cylinders number so i will remove it from the analysis.

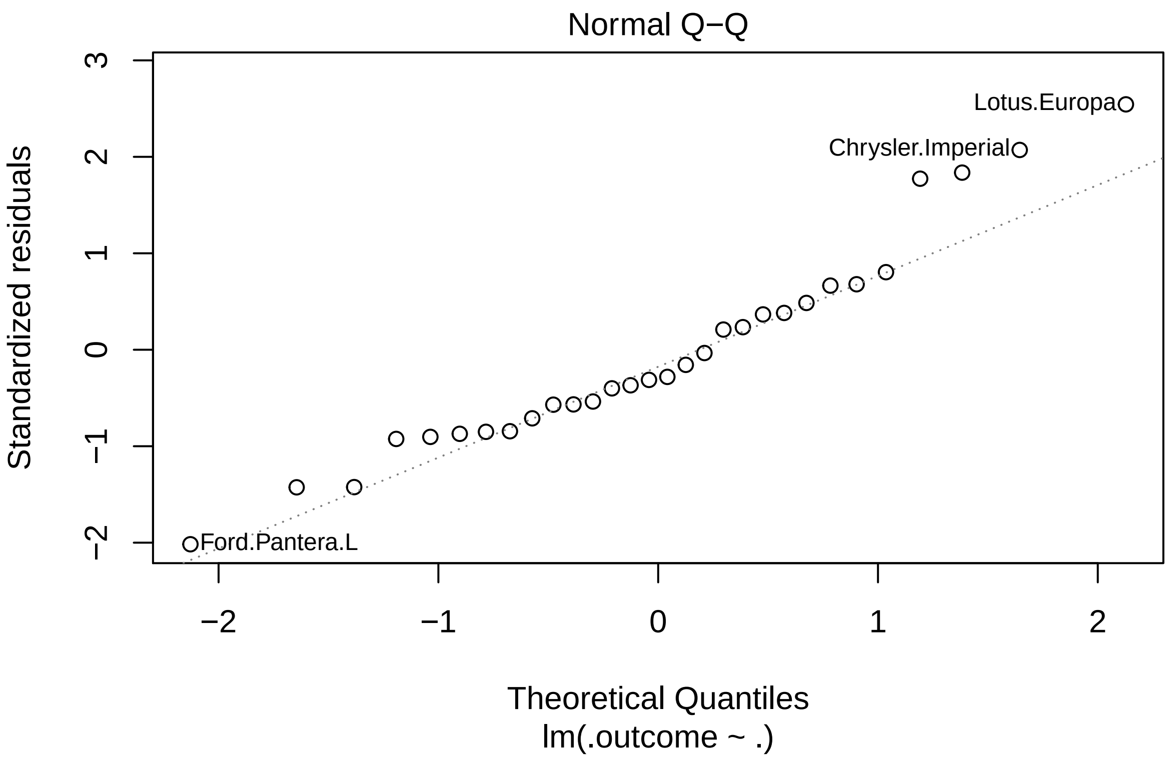
Now, we will build several models and compare them. We will do a linear model, a ridge regression, lasso and elastic net. We will also do a simple physics based model based on my experience as a combustion engineer.

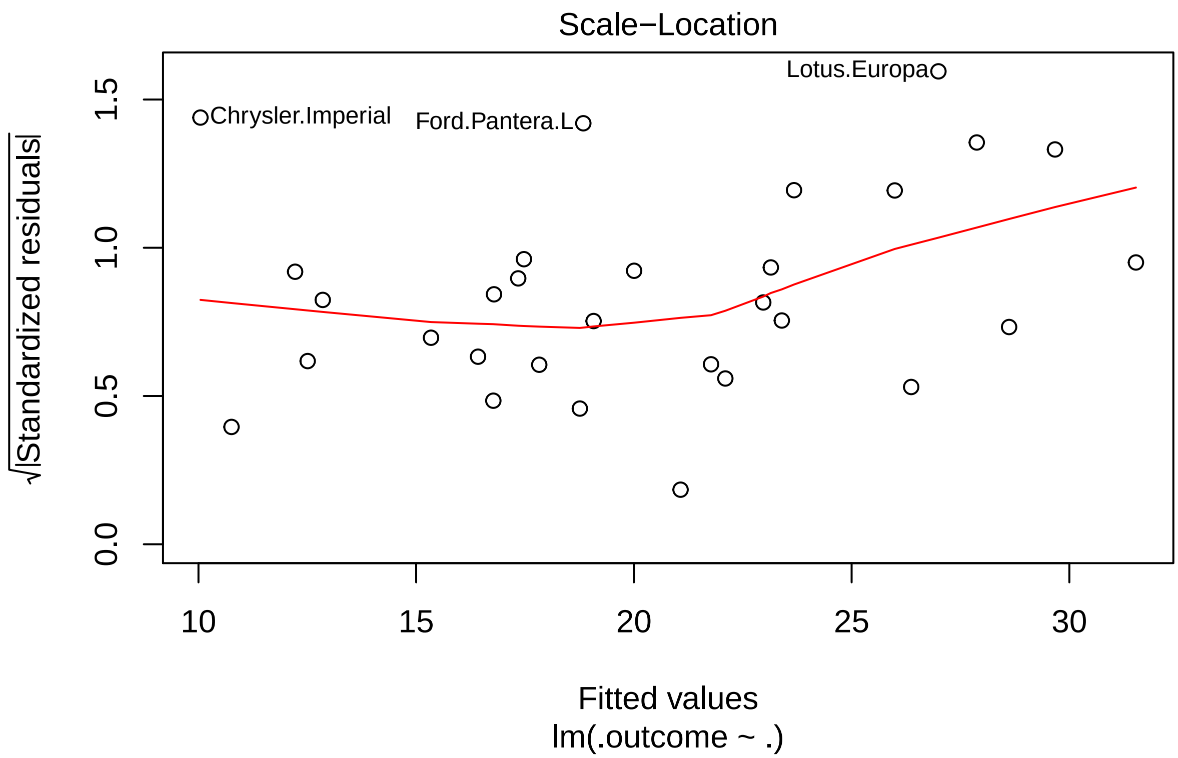
```
##
## Call:
```

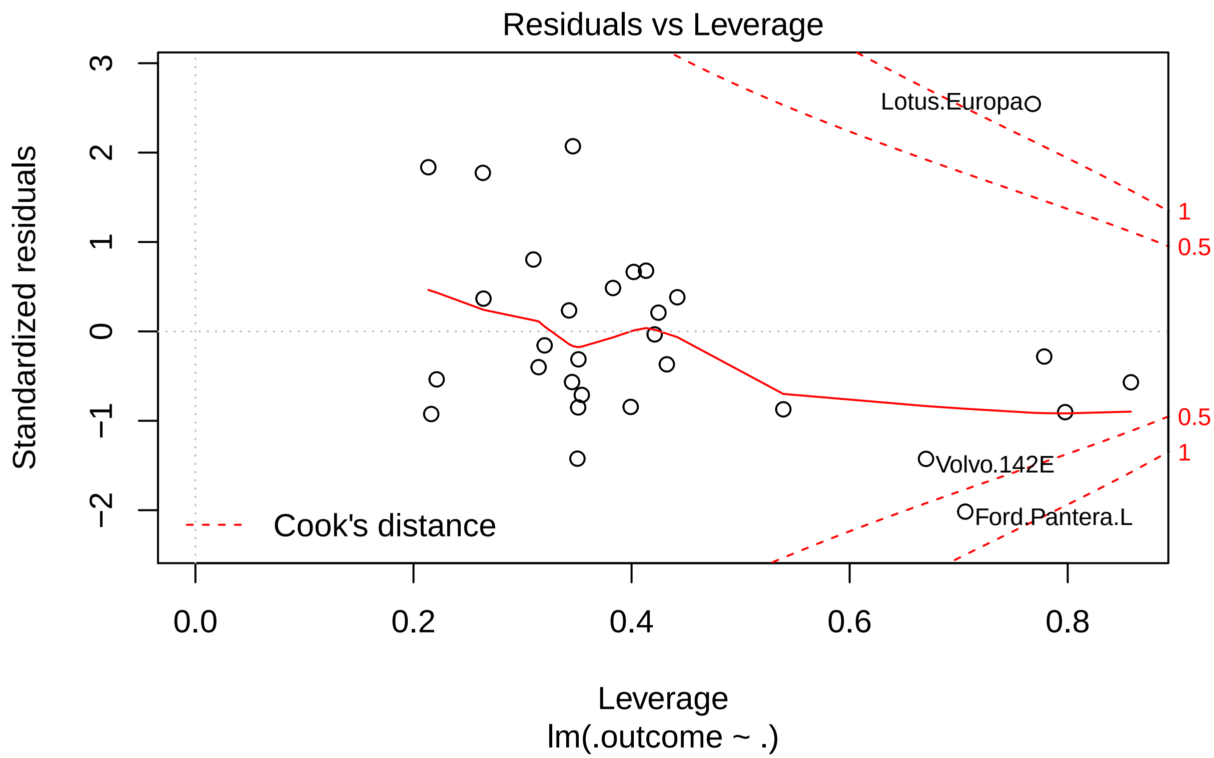
```
## lm(formula = .outcome ~ ., data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1897 -1.3843 -0.3634  0.9201  4.6548
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.87276   15.43382   0.834    0.416
## disp         0.03120    0.02420   1.289    0.215
## hp          -0.04916    0.03139  -1.566    0.136
## drat         2.39306    2.19773   1.089    0.291
## wt          -3.89980    2.25665  -1.728    0.102
## qsec         0.52619    0.83550   0.630    0.537
## vs1          1.52262    2.57613   0.591    0.562
## am1          1.95166    2.80814   0.695    0.496
## gear4        0.88681    3.45338   0.257    0.800
## gear5        1.97540    3.60004   0.549    0.590
## carb2       -0.75477    2.25913  -0.334    0.742
## carb3        2.08122    3.72838   0.558    0.584
## carb4       -1.30096    3.76183  -0.346    0.734
## carb6        0.96288    5.50575   0.175    0.863
## carb8        3.04608    7.39745   0.412    0.686
##
## Residual standard error: 2.779 on 17 degrees of freedom
## Multiple R-squared:  0.8834, Adjusted R-squared:  0.7873
## F-statistic: 9.197 on 14 and 17 DF,  p-value: 2.359e-05
```


wt
hp
disp
drat
am1
qsec
vs1
carb3
gear5
carb8
carb4
carb2
gear4
carb6
0
20
40
60
80
100
Importance


Residuals vs Fitted
Chrysler.Imperial
Fiat.128
Toyota.Corolla
Residuals
4
2
0
−2
−4
10
15
20
25
30
Fitted values
lm(.outcome ~ .)

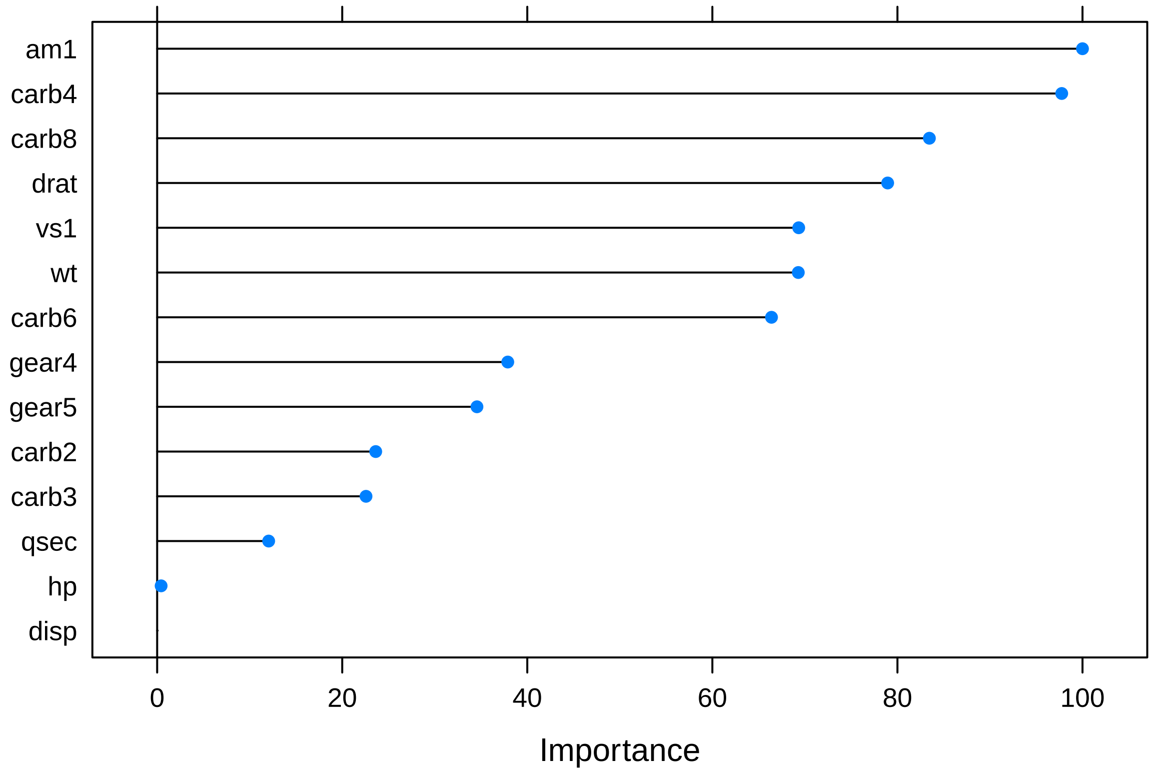

Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(.outcome ~ .)
Lotus.Europa
Chrysler.Imperial
Ford.Pantera.L
3
2
1
0
−1
−2
−2
−1
0
1
2

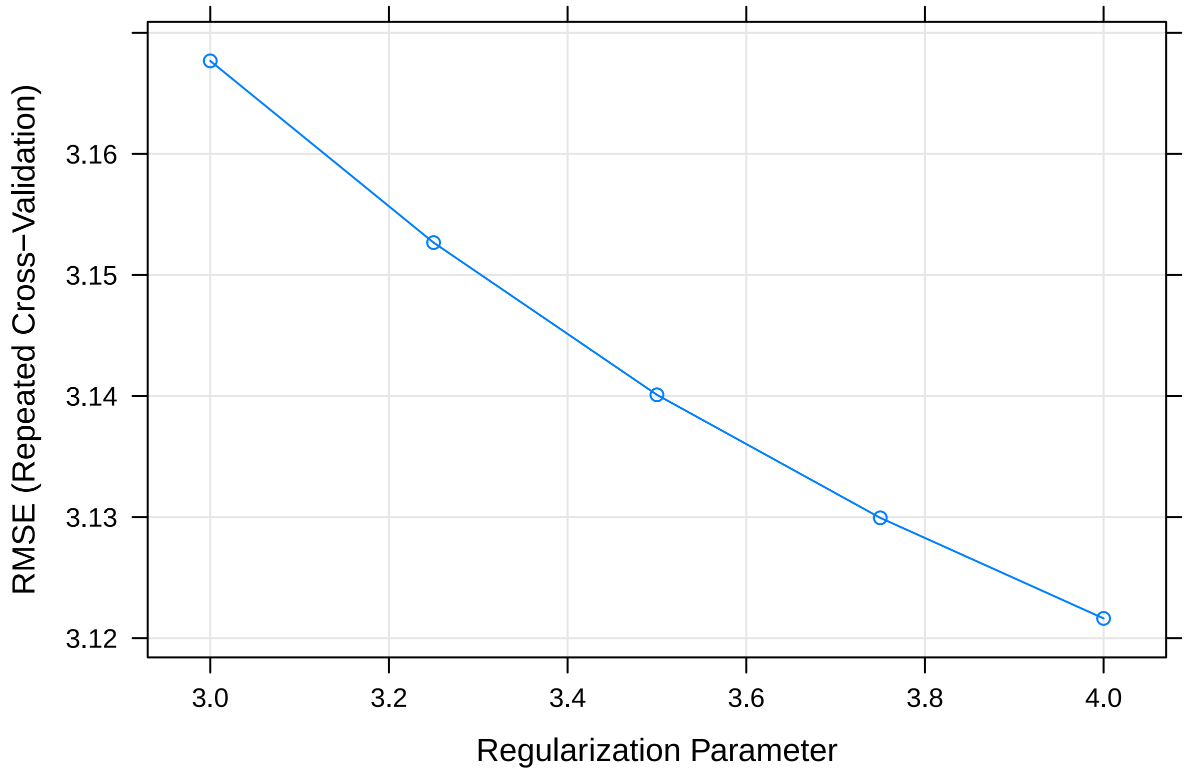

Scale−Location
Lotus.Europa
Chrysler.Imperial
Ford.Pantera.L
√|Standardized residuals|
0.0
0.5
1.0
1.5
10
15
20
25
30
Fitted values
lm(.outcome ~ .)

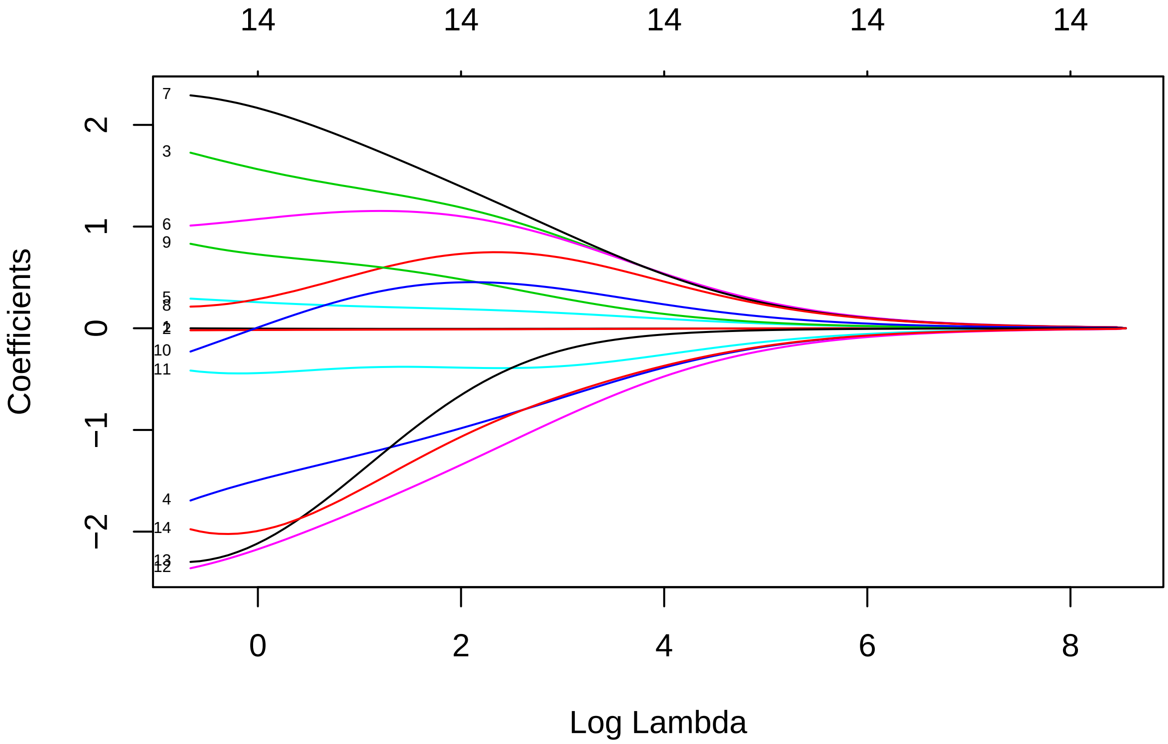


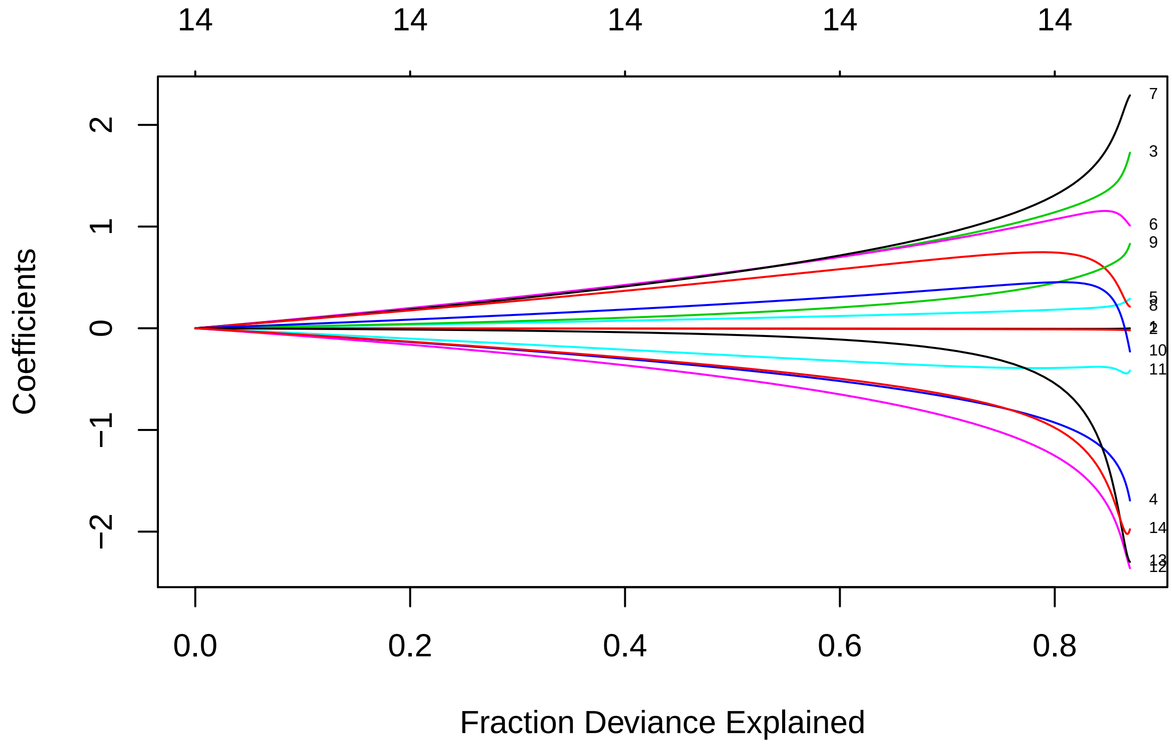

```
##   intercept    RMSE  Rsquared      MAE    RMSESD RsquaredSD     MAESD
## 1      TRUE 5.16461 0.5215343 3.983717 0.9120602  0.1119934 0.5317417
```

```
##             Length Class      Mode
## a0           100   -none-     numeric
## beta        1400   dgCMatrix  S4
## df           100   -none-     numeric
## dim            2   -none-     numeric
## lambda       100   -none-     numeric
## dev.ratio    100   -none-     numeric
## nulldev        1   -none-     numeric
## npasses        1   -none-     numeric
## jerr           1   -none-     numeric
## offset         1   -none-     logical
## call           5   -none-     call
## nobs           1   -none-     numeric
## lambdaOpt      1   -none-     numeric
## xNames        14   -none-     character
## problemType    1   -none-     character
## tuneValue      2   data.frame list
## obsLevels      1   -none-     logical
## param          0   -none-     list
```

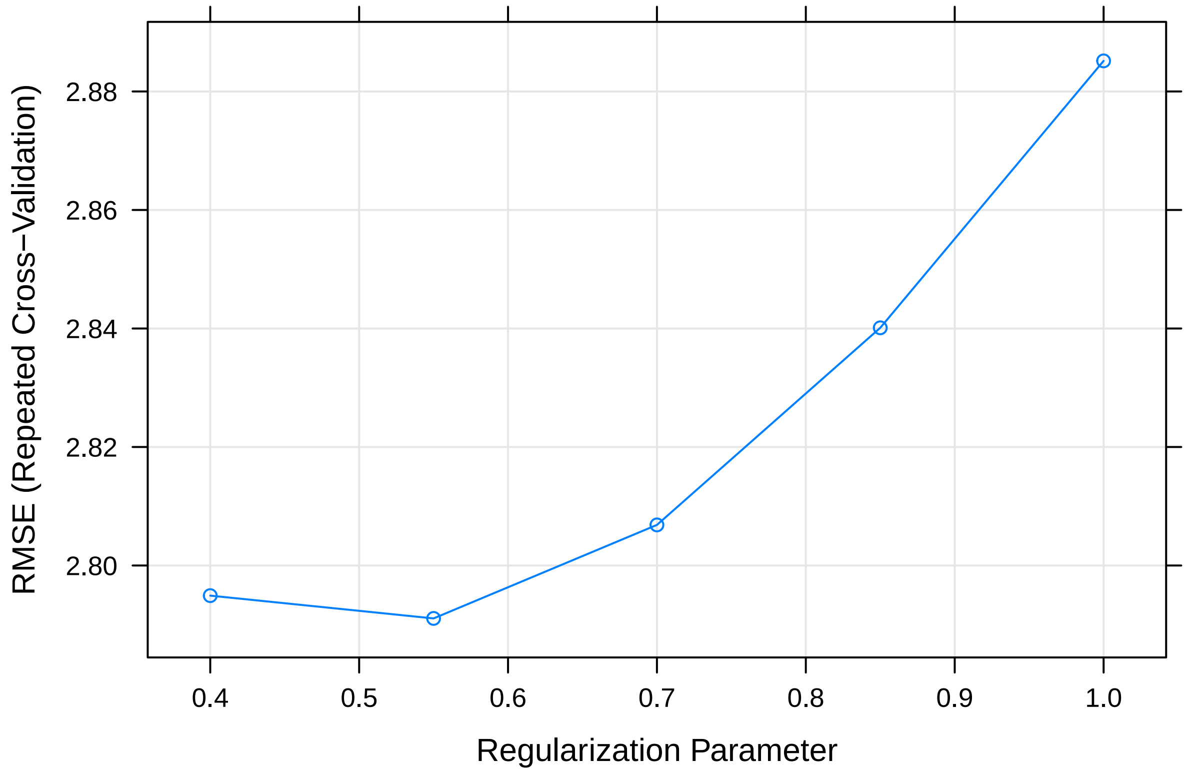

am1
carb4
carb8
drat
vs1
wt
carb6
gear4
gear5
carb2
carb3
qsec
hp
disp
0
20
40
60
80
100
Importance


RMSE (Repeated Cross−Validation)
3.16
3.15
3.14
3.13
3.12
3.0
3.2
3.4
3.6
3.8
4.0
Regularization Parameter

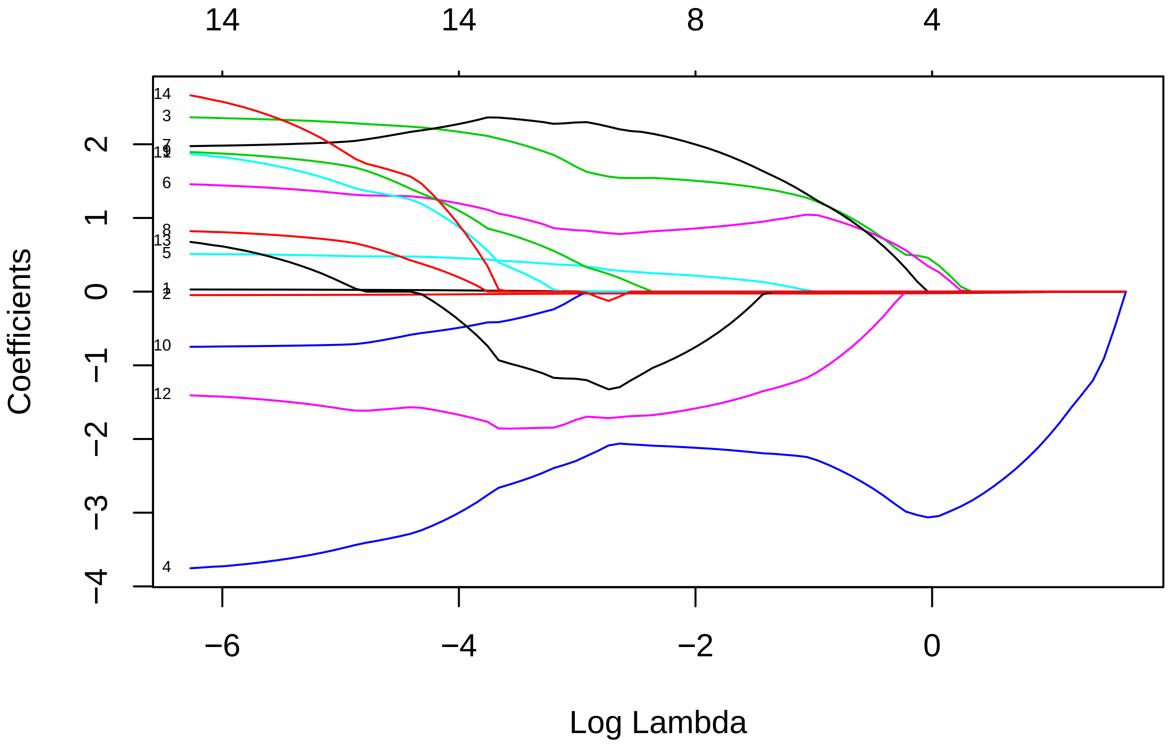

14
14
14
14
14
Coefficients
2
1
0
−1
−2
7
3
6
9
5
8
1
2
10
11
4
14
13
12
0
2
4
6
8
Log Lambda

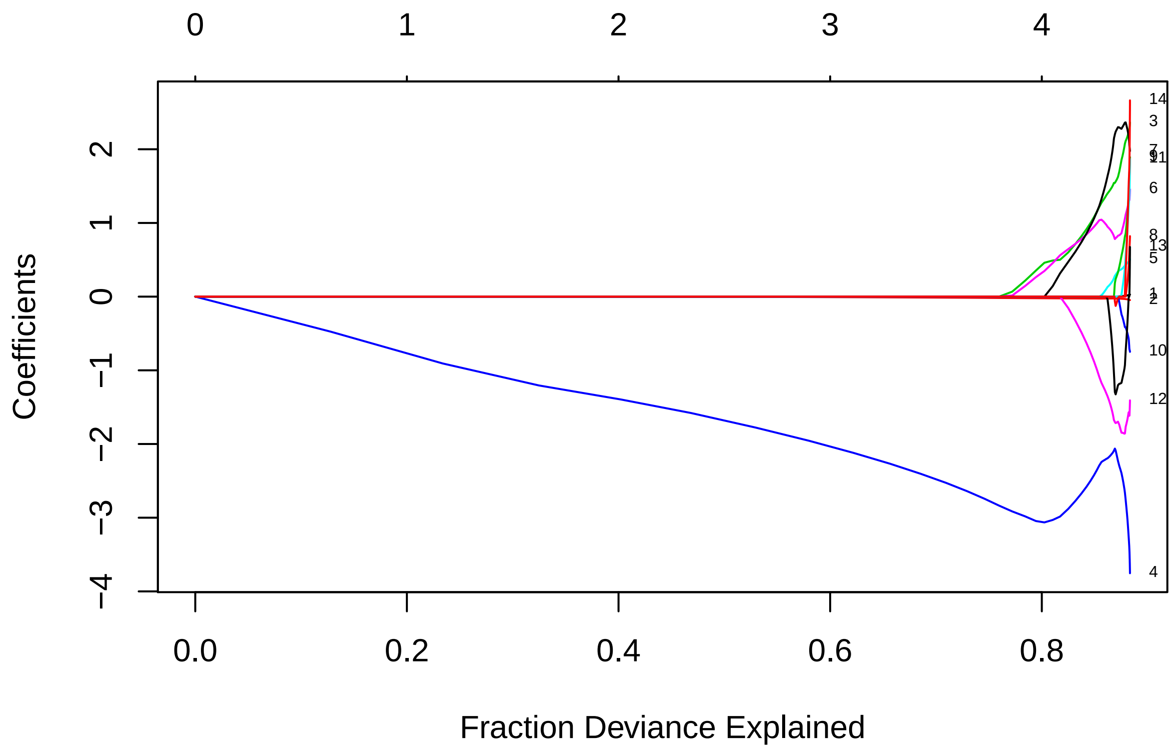


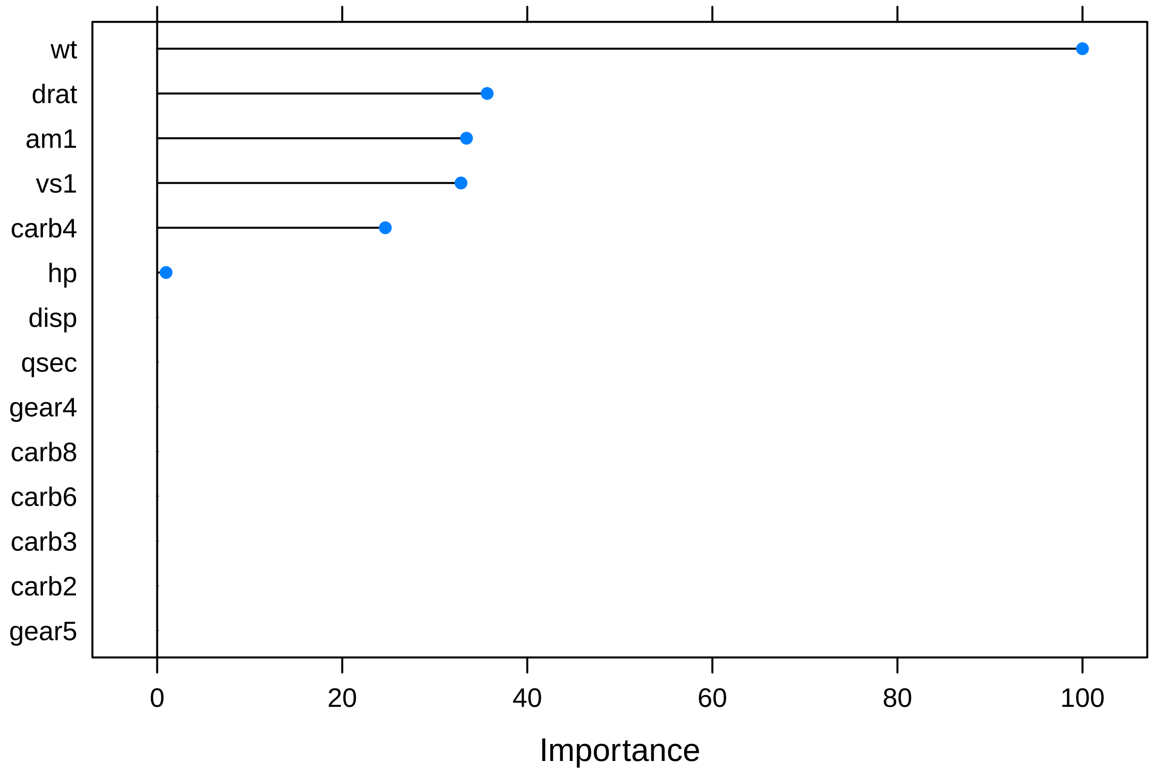
14
14
14
14
14
Coefficients
2
1
0
−1
−2
0.0
0.2
0.4
0.6
0.8
Fraction Deviance Explained
7
3
6
9
5
8
1
2
10
11
4
14
13
12

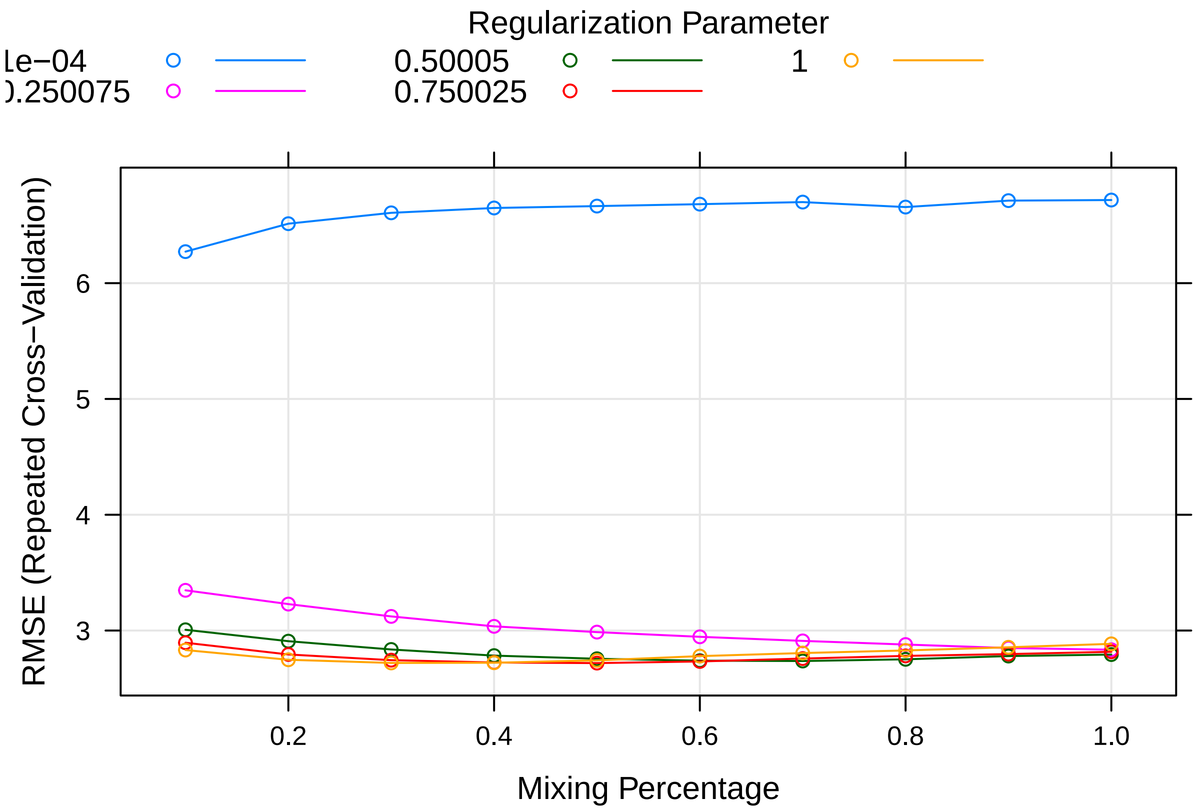


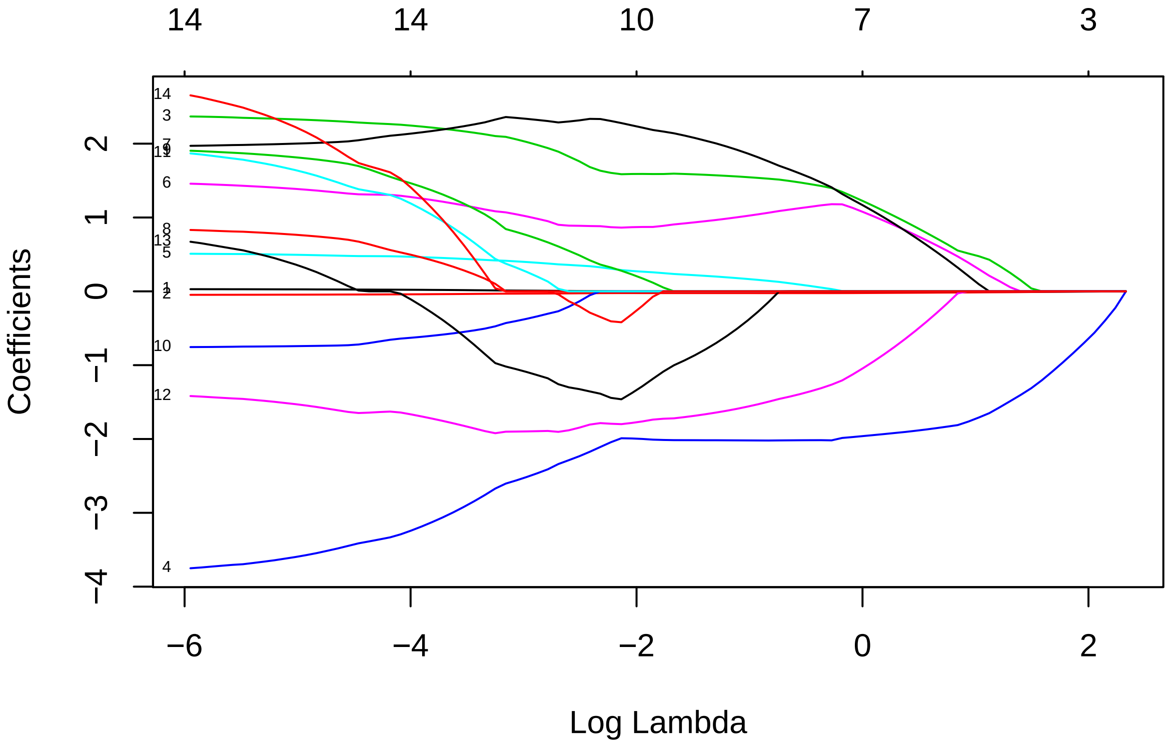
RMSE (Repeated Cross−Validation)
2.88
2.86
2.84
2.82
2.80
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Regularization Parameter

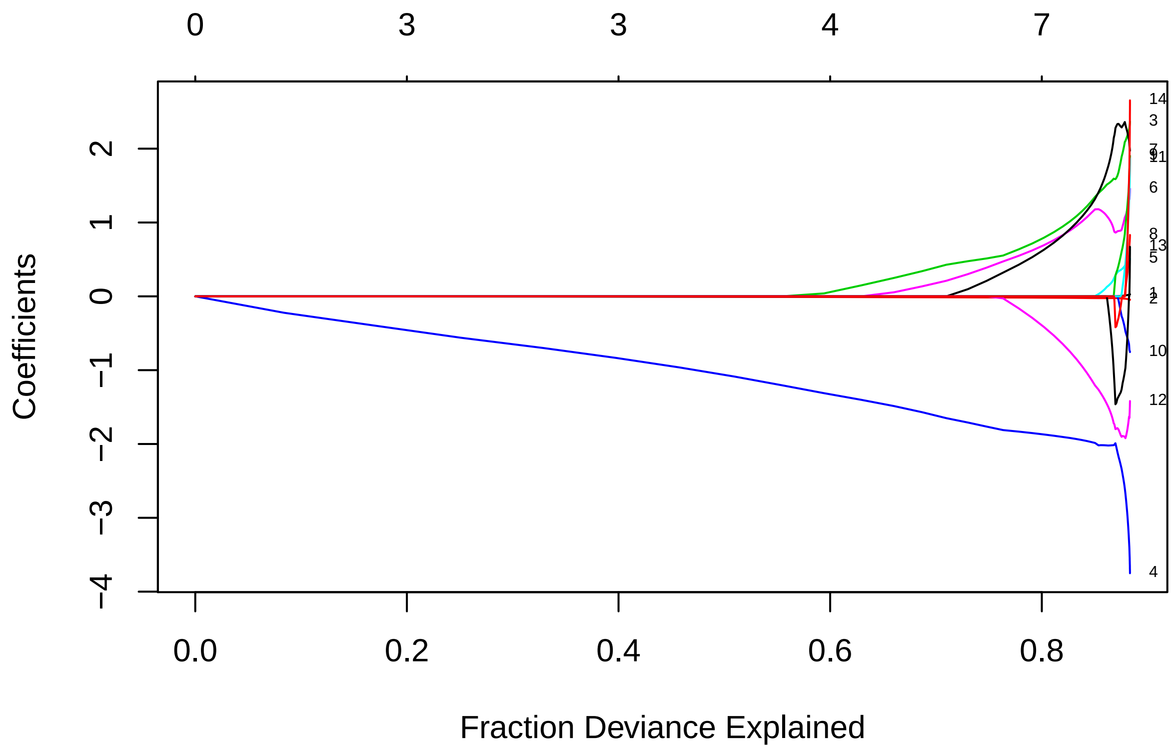

14
14
8
4
Coefficients
Log Lambda
−6
−4
−2
0
2
1
0
−1
−2
−3
−4
14
3
7
9
11
6
8
13
5
1
2
10
12
4


0
1
2
3
4
14
3
7
9
11
6
8
13
5
1
2
10
12
4
Coefficients
2
1
0
−1
−2
−3
−4
0.0
0.2
0.4
0.6
0.8
Fraction Deviance Explained


wt
drat
am1
vs1
carb4
hp
disp
qsec
gear4
carb8
carb6
carb3
carb2
gear5
0
20
40
60
80
100
Importance

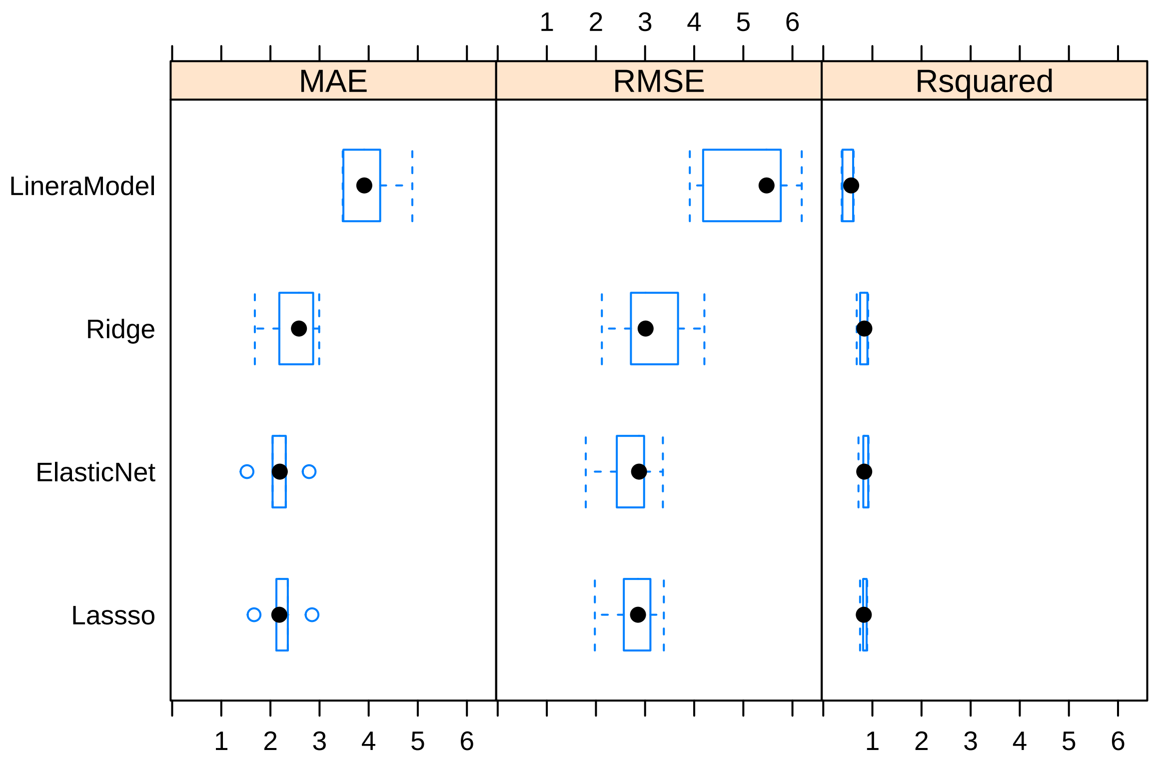

Regularization Parameter
1e−04
0.250075
0.50005
0.750025
1
RMSE (Repeated Cross−Validation)
6
5
4
3
0.2
0.4
0.6
0.8
1.0
Mixing Percentage

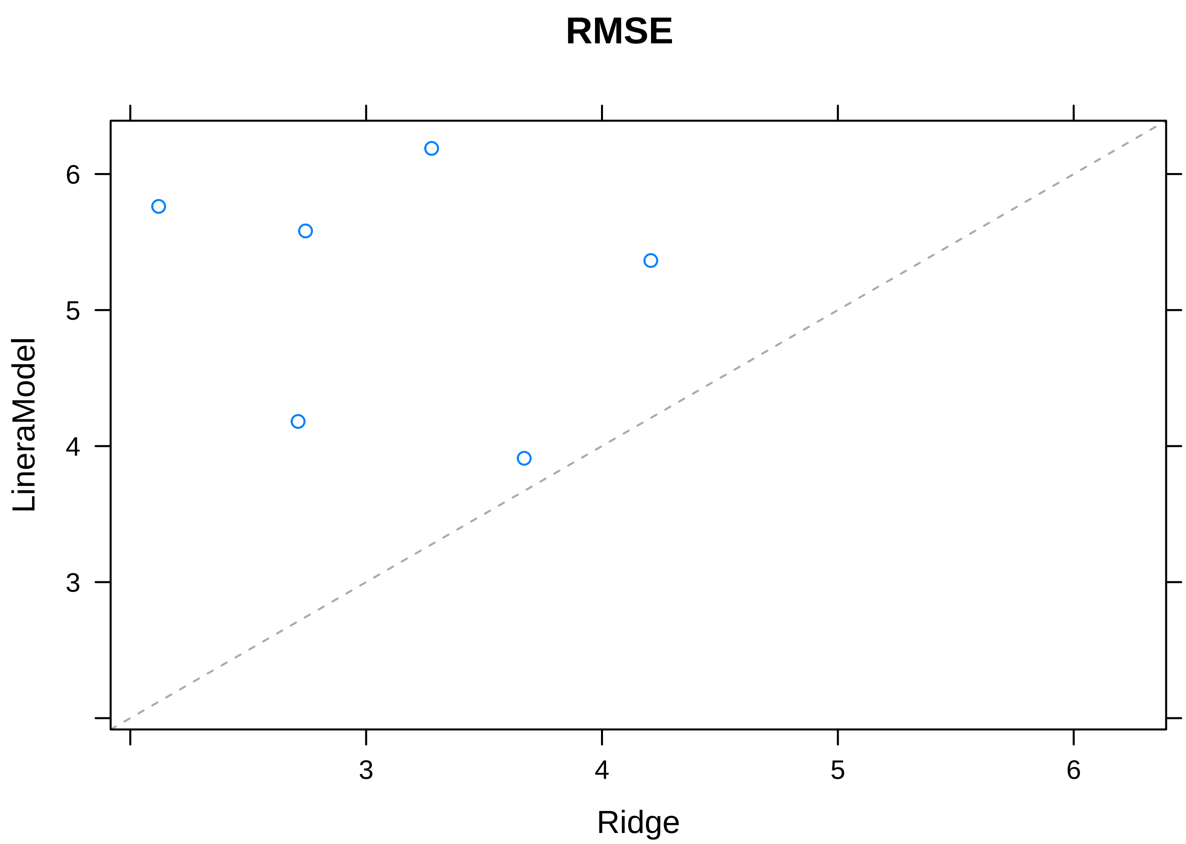

14
14
10
7
3
Coefficients
2
1
0
−1
−2
−3
−4
14
3
7
9
11
6
8
13
5
1
2
10
12
4
−6
−4
−2
0
2
Log Lambda




```
##
## Call:
## summary.resamples(object = res)
##
## Models: LineraModel, Ridge, Lassso, ElasticNet
## Number of resamples: 6
##
## MAE
##                  Min.  1st Qu.   Median     Mean  3rd Qu.     Max. NA's
## LineraModel 3.472335 3.573274 3.910761 3.983717 4.171707 4.888976    0
## Ridge       1.683005 2.213993 2.580965 2.481542 2.865176 2.992733    0
## Lassso      1.666780 2.123752 2.180514 2.225096 2.323506 2.846763    0
## ElasticNet  1.523257 2.056244 2.190967 2.175113 2.307276 2.788103    0
##
## RMSE
##                  Min.  1st Qu.   Median     Mean  3rd Qu.     Max. NA's
## LineraModel 3.910259 4.476914 5.472838 5.164610 5.716972 6.188433    0
## Ridge       2.120229 2.719470 3.010408 3.121628 3.572100 4.206900    0
## Lassso      1.977563 2.620160 2.856157 2.791072 3.064148 3.381581    0
## ElasticNet  1.792260 2.521580 2.877348 2.718877 2.970037 3.362381    0
##
## Rsquared
##                   Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## LineraModel 0.3753897 0.4254009 0.5689848 0.5215343 0.6076227 0.6178047
## Ridge       0.6808755 0.7605909 0.8363998 0.8194268 0.8947637 0.9133462
## Lassso      0.7488458 0.8130018 0.8247706 0.8309293 0.8696729 0.8934698
```

```
## ElasticNet  0.7168003 0.8192341 0.8342059 0.8393412 0.8956547 0.9204539
##             NA's
## LineraModel    0
## Ridge          0
## Lassso         0
## ElasticNet     0
```






```
##     alpha   lambda
## 24    0.5 0.750025
```

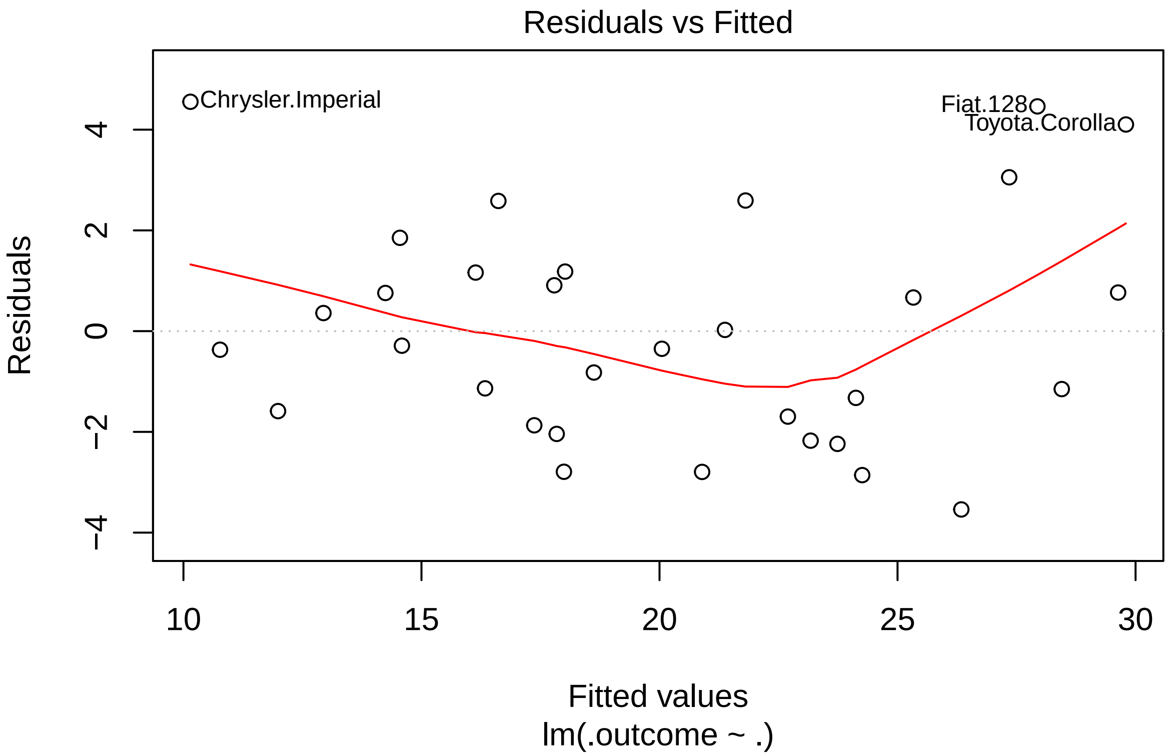
```
## 15 x 1 sparse Matrix of class "dgCMatrix"
##                        1
## (Intercept) 23.80985063
## disp         .
## hp          -0.02433772
## drat         1.40278754
## wt          -2.01802624
## qsec         0.03381991
## vs1          1.17773664
## am1          1.41930640
## gear4        .
## gear5        .
## carb2        .
## carb3        .
## carb4       -1.27168023
## carb6        .
## carb8        .
```

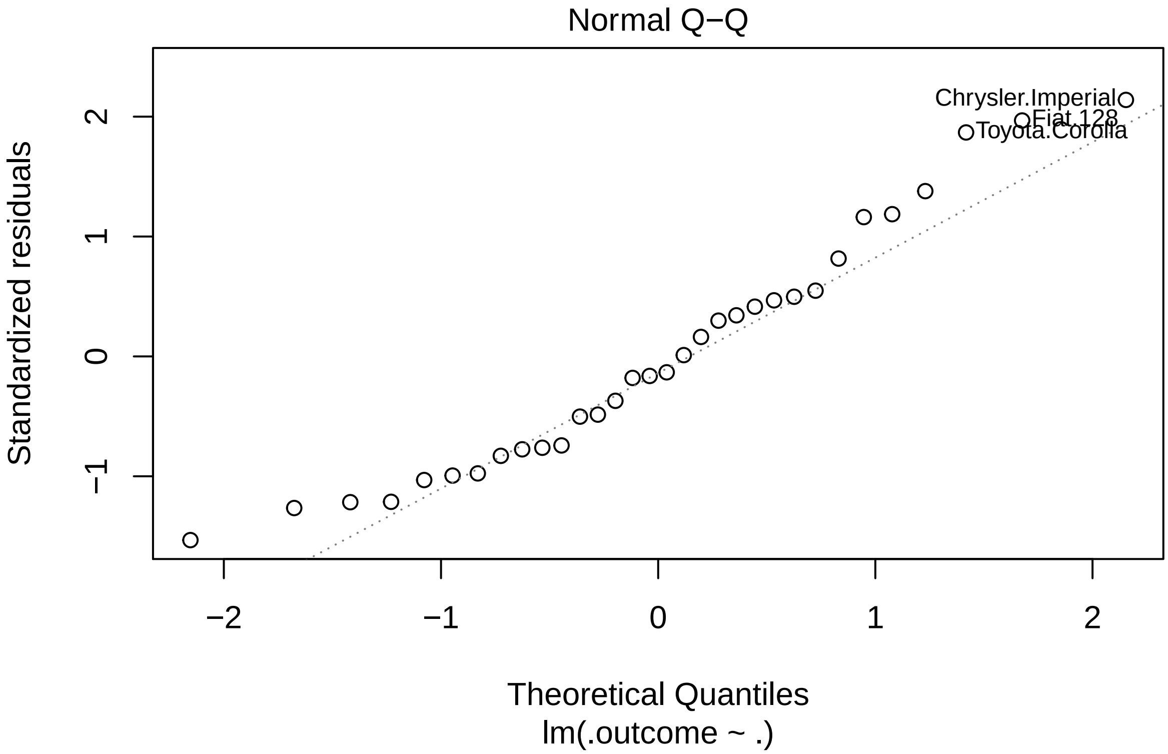
```
## [1] 1.951657
```

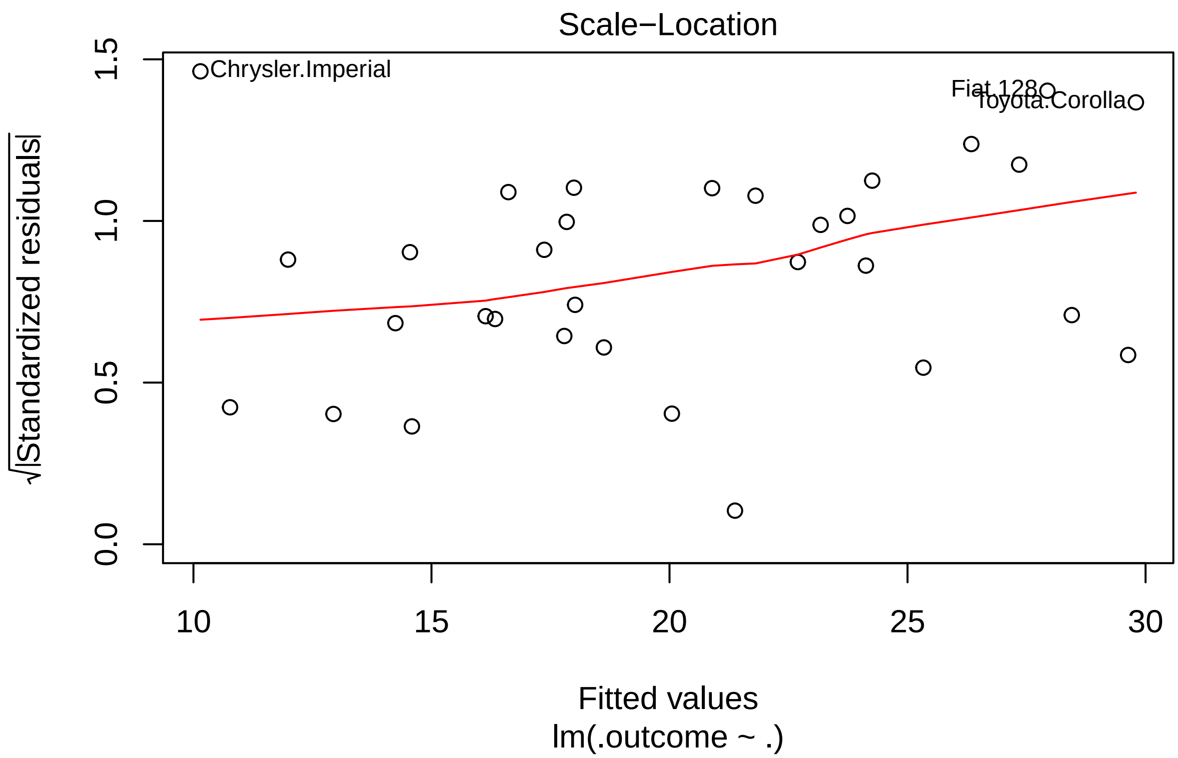
```
## [1] 1.419306
```

Simplifying the model based on physics

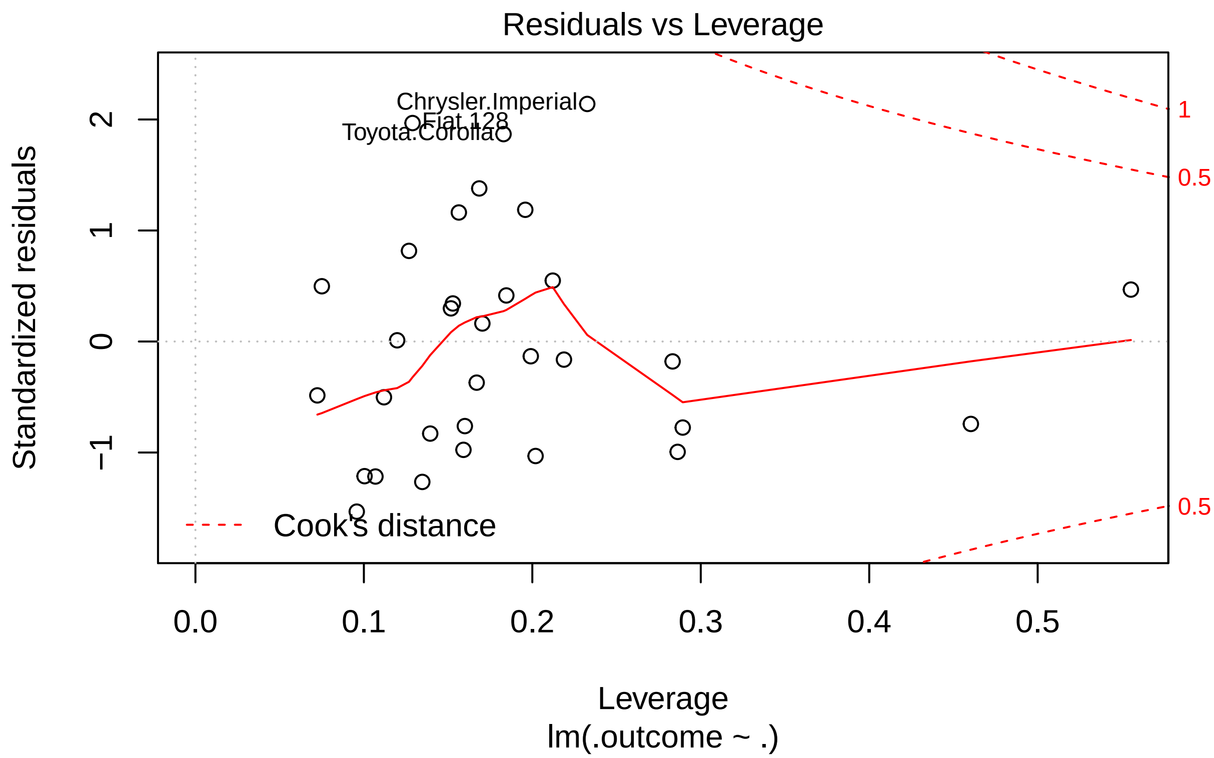
The previous analysis shows that based on several models fitted, the average saving in mpg is between 1.4 and 2 mpg as we go from manual to automatic transmission. The linear model and elastic net both suggest that mpg increase with transmission type. Let us now try a simple model based on physical sense only. In my experience, using the disp, weight, qsec, hp and transmission are enough to build a physically meaningful model.



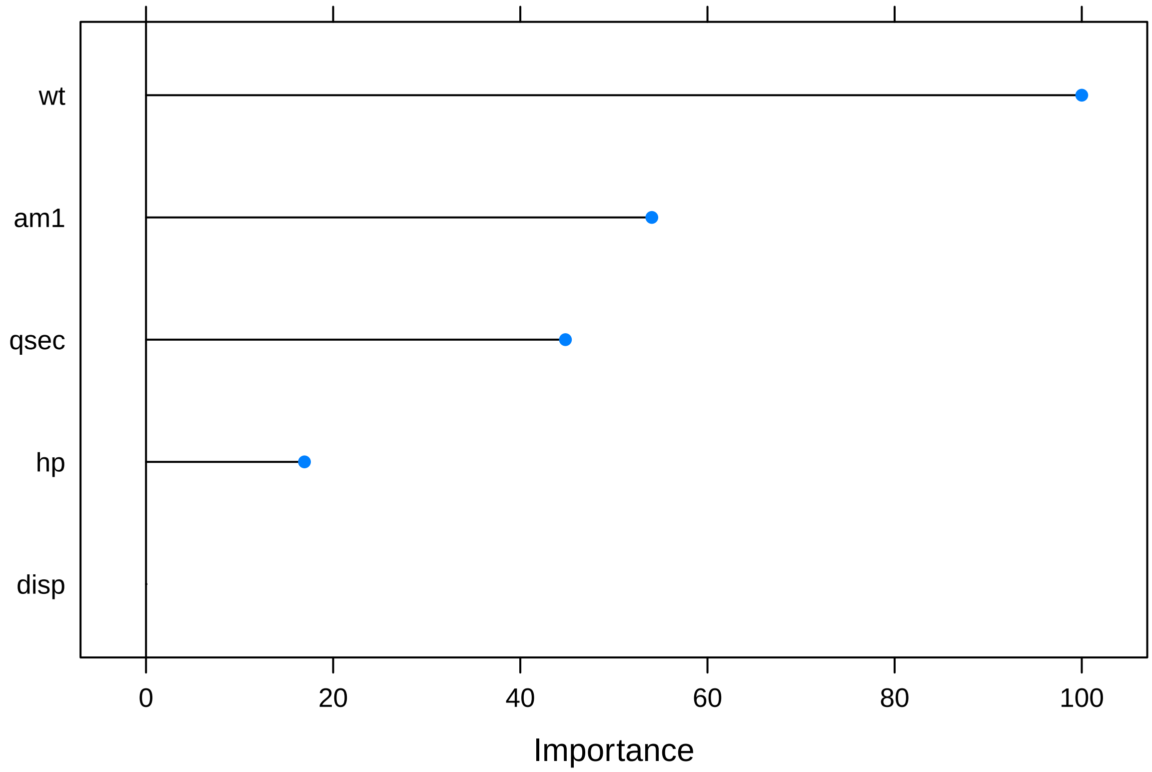

Normal Q−Q
Chrysler.Imperial
Fiat.128
Toyota.Corolla
Standardized residuals
−1
0
1
2
−2
−1
0
1
2
Theoretical Quantiles
lm(.outcome ~ .)



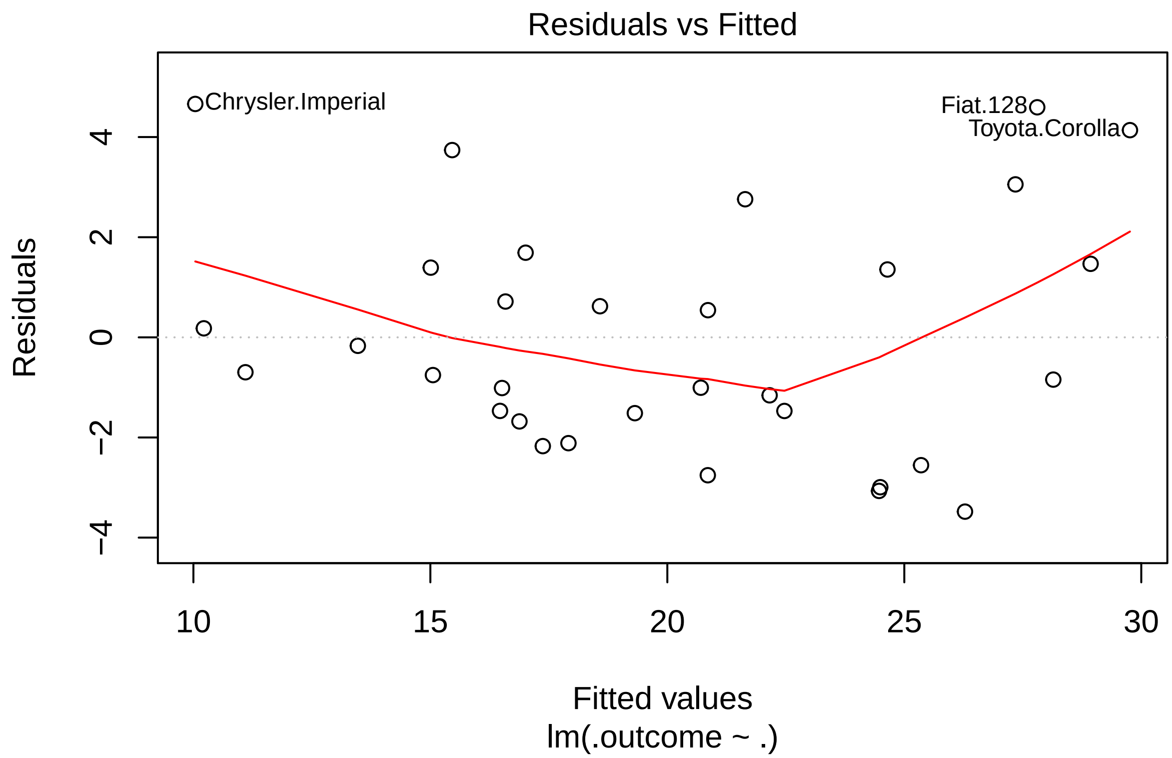
Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
10
15
20
25
30
Chrysler.Imperial
Fiat.128
Toyota.Corolla
Fitted values
lm(.outcome ~ .)

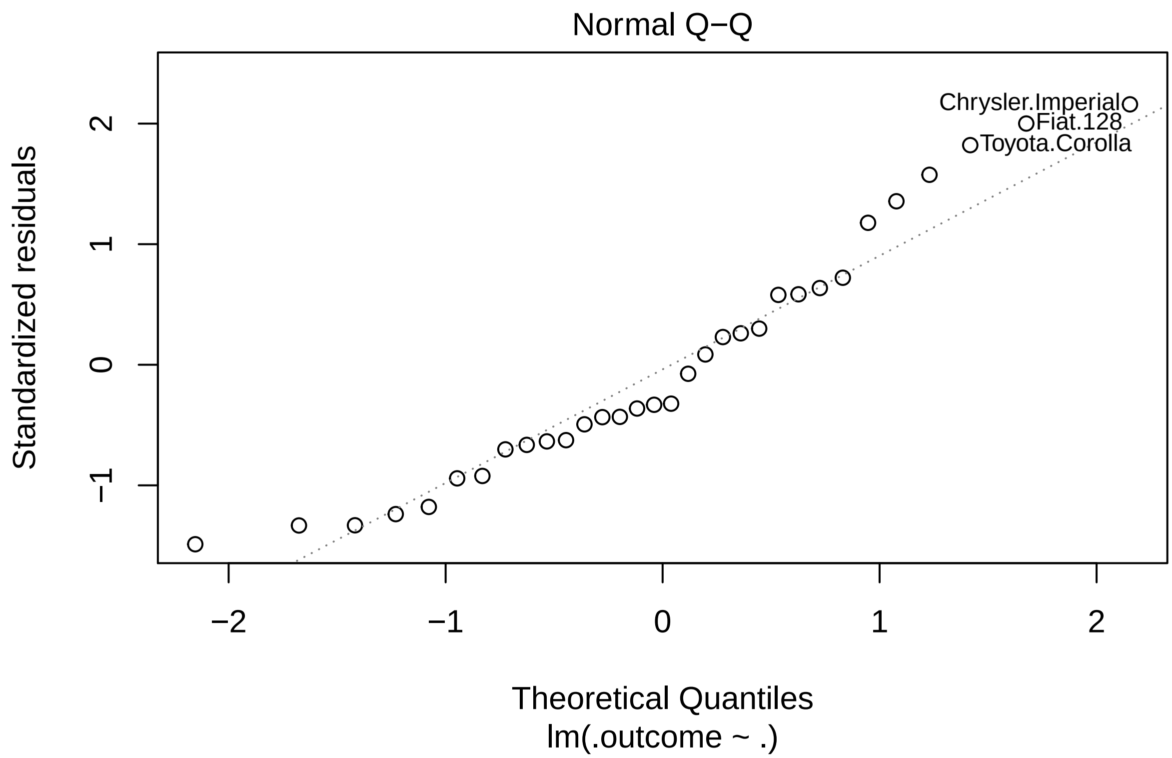


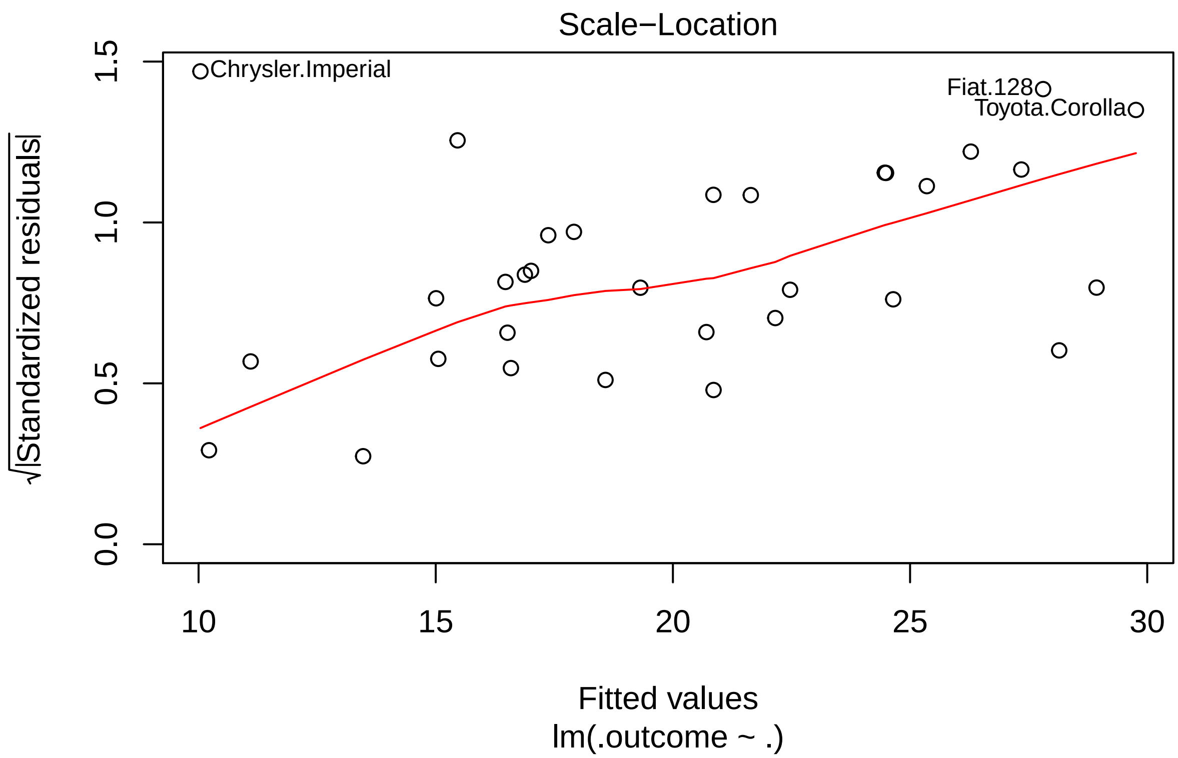

```
##   intercept     RMSE  Rsquared      MAE    RMSESD RsquaredSD     MAESD
## 1      TRUE 2.592495 0.8362622 2.153416 0.3495403 0.02274025 0.3156191
```

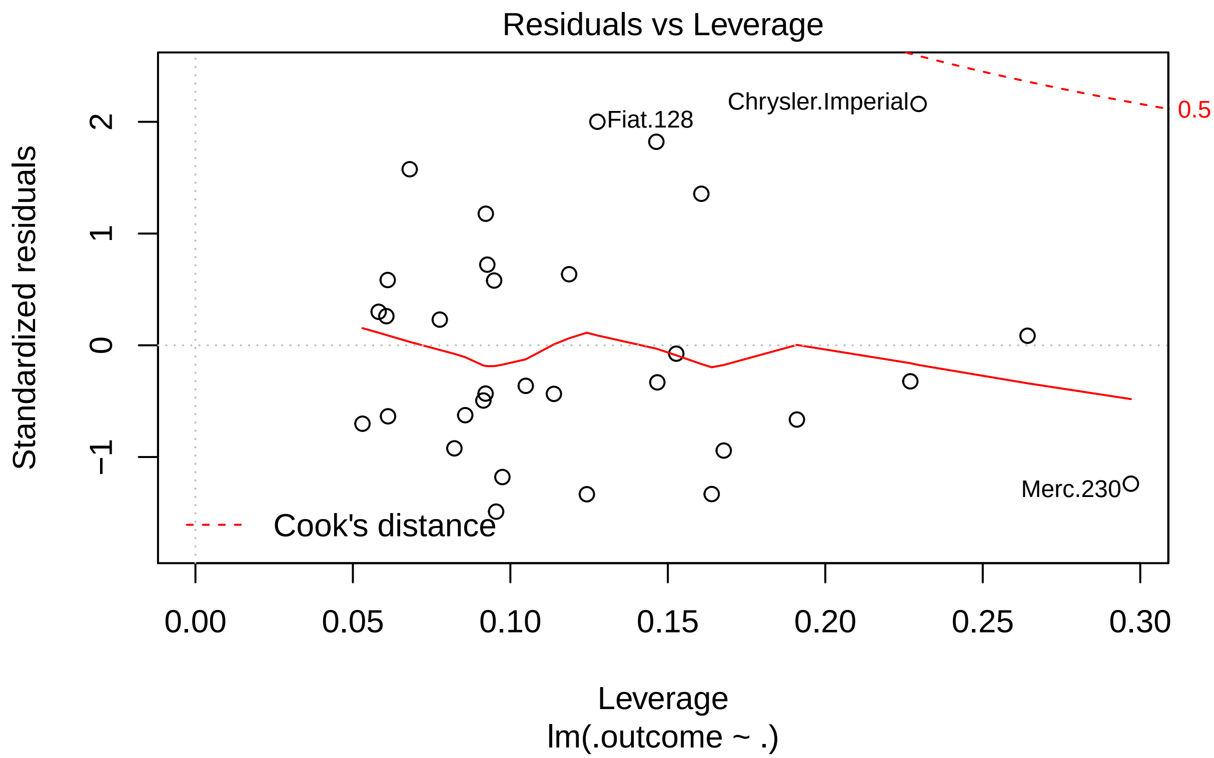



```
##
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5399 -1.7398 -0.3196  1.1676  4.5534
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.36190    9.74079   1.474  0.15238
## disp         0.01124    0.01060   1.060  0.29897
## hp          -0.02117    0.01450  -1.460  0.15639
## wt          -4.08433    1.19410  -3.420  0.00208 **
## qsec         1.00690    0.47543   2.118  0.04391 *
## am1          3.47045    1.48578   2.336  0.02749 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.429 on 26 degrees of freedom
## Multiple R-squared:  0.8637, Adjusted R-squared:  0.8375
## F-statistic: 32.96 on 5 and 26 DF,  p-value: 1.844e-10
```

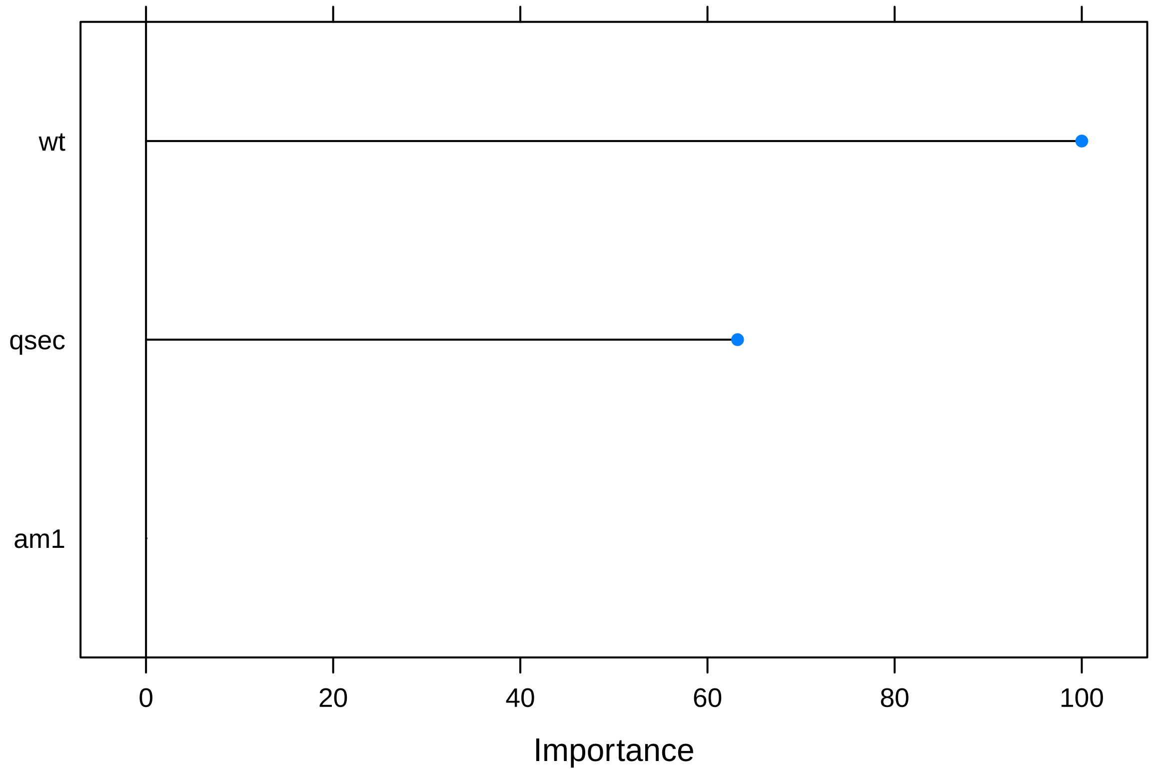

Residuals vs Fitted
Chrysler.Imperial
Fiat.128
Toyota.Corolla
Residuals
4
2
0
−2
−4
10
15
20
25
30
Fitted values
lm(.outcome ~ .)


Normal Q−Q
Chrysler.Imperial
Fiat.128
Toyota.Corolla
Standardized residuals
Theoretical Quantiles
lm(.outcome ~ .)
−2
−1
0
1
2


Scale−Location
Chrysler.Imperial
Fiat.128
Toyota.Corolla
√|Standardized residuals|
0.0
0.5
1.0
1.5
10
15
20
25
30
Fitted values
lm(.outcome ~ .)




```
##   intercept     RMSE  Rsquared      MAE   RMSESD RsquaredSD    MAESD
## 1      TRUE 2.722948 0.8503933 2.270228 0.519983 0.06128789 0.476762
```




```
##
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## 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
```

Conlsusion

Buying a car with automatic transmission will be more cost saving for the user with about 1.4 to 2 mpg.