How transmission type affects fuel consumption?

Regression Models Project by Florin Toth

Executive Summary

In this report we are looking at a data set of a collection of cars (mtcars). We are interested in exploring the relationship between a set of variables and miles per gallon (MPG). We are particularly interested in answering the following two questions: “Is an automatic or manual transmission better for MPG?” and “Quantify the MPG difference between automatic and manual transmissions?” We use exploratory data analyses and regression models to answer these questions. The transmission is **automatic** when variable **am**=0 and **manual** when **am**=1.

At first we find that cars with manual transmission have on average 7.245 miles per gallon more than those with automatic transmission. Then, we fit several other linear regression models that include more independent variables and select the one with the highest adjusted R-squared value. We find that weight (“wt”) and quarter mile time (“qsec”) along with transmission (“am”) are together better predictors of fuel efficiency (“mpg”) though the influence of the manual transmission on the fuel efficiency is reduced to 2.936 miles per gallon on average.

Exploratory Data Analysis

Before delving into regression analysis we need to have an overview of the data set. First we load and check the data set.

```
data(mtcars)
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 ...
```

Then we plot the data. We use a boxplot to check the relationship between the **mpg** and **am**, our variables of interest, i.e. between miles per gallon and transmission type. We notice manual transmission corresponds to higher values of miles per gallon.

Using a pairs graph we then check the relationships among all variables in the data set. Here we find strong correlations between (independent) variables “wt”, “disp” and “cyl”. Please see **Figure 1** and **Figure 2** in the **Appendix** below.

Regression Analysis

First we check a simple model to see if transmission type (“am”, the independent variable) has a significant impact on the miles per gallon (“mpg”, the dependent variable)

```
library(broom) # package for better display of summary statistics
modelSimple <- lm(mpg ~ am, data=mtcars)
glance(modelSimple)[c(1:3,11,5)]
```

```
##   r.squared adj.r.squared    sigma df.residual      p.value
## 1 0.3597989     0.3384589 4.902029          30 0.0002850207
```

```
tidy(modelSimple)
```

```
##          term  estimate std.error statistic      p.value
## 1 (Intercept) 17.147368  1.124603 15.247492 1.133983e-15
## 2          am  7.244939  1.764422  4.106127 2.850207e-04
```

This model shows that transmission type has significant influence on miles per gallon given the low p-values. It also shows that if the transmission is manual, the “mpg” has an average increase of 7.245 (also as expected from the boxplot discussed above).

This model has an adjusted R-squared value of 0.3385 meaning that the model can explain only about 34% of the variance of the “mpg” variable. The sigma (the residual standard error) of the model is 4.902 on 30 degrees of freedom.

Now we will check a model to see how all variables impact the miles per gallon (“mpg”).

```
modelFull <- lm(mpg ~ ., data=mtcars)
glance(modelFull)[c(1:3,11,5)]
```

```
##   r.squared adj.r.squared    sigma df.residual      p.value
## 1 0.8690158     0.8066423 2.650197          21 3.793152e-07
```

```
tidy(modelFull)
```

```
##           term    estimate   std.error  statistic    p.value
## 1  (Intercept) 12.30337416 18.71788443  0.6573058 0.51812440
## 2          cyl -0.11144048  1.04502336 -0.1066392 0.91608738
## 3         disp  0.01333524  0.01785750  0.7467585 0.46348865
## 4           hp -0.02148212  0.02176858 -0.9868407 0.33495531
## 5         drat  0.78711097  1.63537307  0.4813036 0.63527790
## 6           wt -3.71530393  1.89441430 -1.9611887 0.06325215
## 7         qsec  0.82104075  0.73084480  1.1234133 0.27394127
## 8           vs  0.31776281  2.10450861  0.1509915 0.88142347
## 9           am  2.52022689  2.05665055  1.2254035 0.23398971
## 10        gear  0.65541302  1.49325996  0.4389142 0.66520643
## 11        carb -0.19941925  0.82875250 -0.2406258 0.81217871
```

The adjusted R-squared value is 0.807, explaining around 80% of the variance of the “mpg” variable. However, none of the coefficients are significant at 0.05 significant level. The residual standard error (sigma) of the model is 2.65 on 21 degrees of freedom.

We need to find a better model, with less independent variables, less potential multicollinearity issues and significant coefficients. We can either try removing variables selectively on trial and error basis or we can use the “step” function that chooses a model in a stepwise algorithm.

```
modelStep <- step(modelFull)
summary(modelStep)
```

```
glance(modelStep)[c(1:3,11,5)]
```

```
##   r.squared adj.r.squared    sigma df.residual      p.value
## 1 0.8496636     0.8335561 2.458846          28 1.210446e-11
```

```
tidy(modelStep)
```

```
##          term  estimate std.error statistic      p.value
## 1 (Intercept)  9.617781 6.9595930  1.381946 1.779152e-01
## 2          wt -3.916504 0.7112016 -5.506882 6.952711e-06
## 3        qsec  1.225886 0.2886696  4.246676 2.161737e-04
## 4          am  2.935837 1.4109045  2.080819 4.671551e-02
```

The model selected by the “step” function above is “mpg ~ wt + qsec + am”. The adjusted R-squared value is 0.8336, meaning that the model can explain about 83% of the variance of the “mpg” variable, more than the previous models and all of the coefficients are significant at 0.05 significant level. The residual standard error (sigma) is 2.459 on 28 degrees of freedom.

Therefore, if we take into account the weight (“wt”) and quarter mile time (“qsec”) variables, the manual transmission translates into only 2.936 miles per gallon (“mpg”) more than the cars with automatic transmission.

The last model seems to be the best but we can also use “anova” function to choose between the modelStep and modelFull as compared to modelSimple (the first one with only “am” as independent variable). “Anova” tests the models against each other in the order specified.

```
anova(modelSimple, modelStep, modelFull)
```

```
## 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
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     28 169.29  2    551.61 39.2687 8.025e-08 ***
## 3     21 147.49  7     21.79  0.4432    0.8636
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As results show, the modelStep is selected and also has the biggest adjusted R-squared value. **Figure 3** in the **Appendix** shows a comparison between the estimates of the variables from the modelFull vs modelStep selected. It shows that the estimates in the modelStep remain with the same sign after adding the potential errors as opposed the the full model where the estimates change sign frequently.

Finaly we check a diagnostic plot of our modelStep regression in **Figure 4** from the **Appendix**. The residual plots do not show any consistent pattern, seem randomly and normally distributed and do not have significant oultiers.

Appendix

Figure 1: Boxplot of MPG vs Manual/Automatic

```
library(ggplot2)
g <- ggplot(mtcars, aes(x=factor(am), y=mpg))
g <- g + geom_boxplot()
g <- g + xlab("Manual (am=1) / Automatic (am=0)") + ylab("Miles per Gallon")
g <- g + ggtitle("Boxplot of MPG vs Manual/Automatic")
g
```

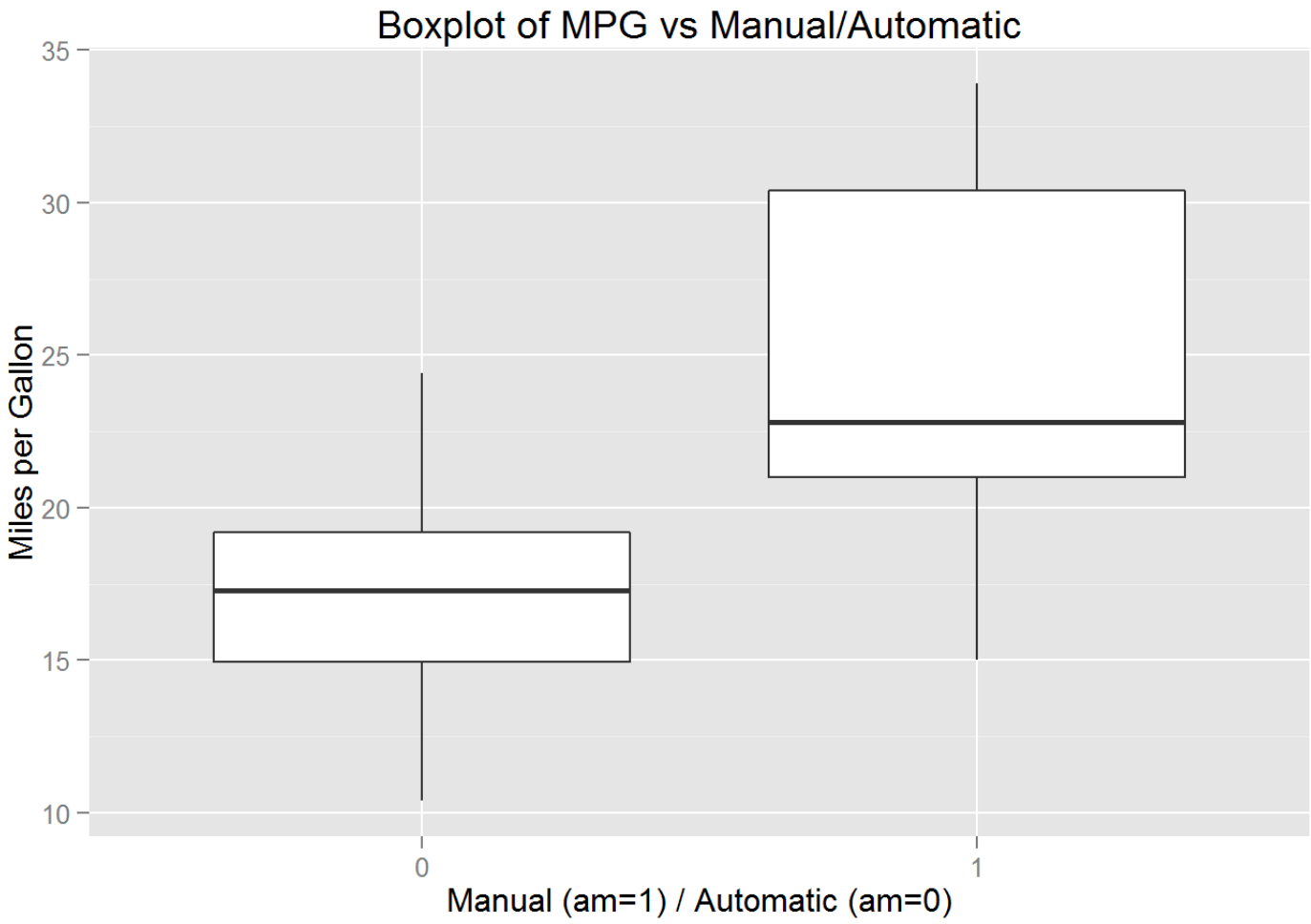


Figure 2: Pairs Graph of Variables

```
pairs(mtcars, panel=panel.smooth, main="Pairs Graph of Variables")
```

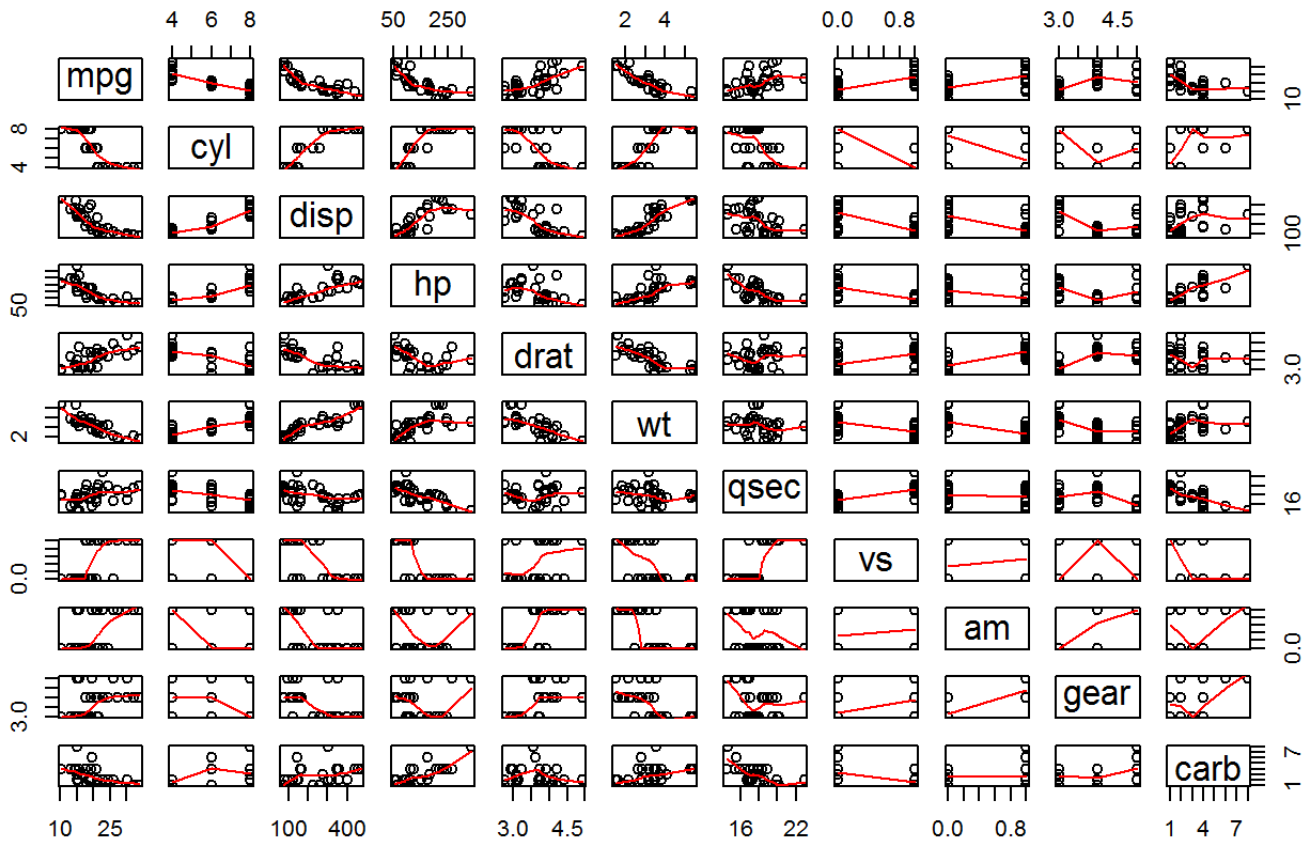


Figure 3: Comparison Chart between Full Model & Step Model Estimates

```
library(gridExtra)
```

```
## Loading required package: grid
```

```
modelPlot <- tidy(modelFull, conf.int = TRUE)
g1 <- ggplot(modelPlot, aes(estimate, term, color = term))
g1 <- g1 + geom_point()
g1 <- g1 + geom_errorbarh(aes(xmin = conf.low, xmax = conf.high))
g1 <- g1 + geom_vline()
g1 <- g1 + xlab("Estimates") + ylab("Variables")
g1 <- g1 + ggtitle("Plot of Full Model Estimates")

modelPlot <- tidy(modelStep, conf.int = TRUE)
g2 <- ggplot(modelPlot, aes(estimate, term, color = term))
g2 <- g2 + geom_point()
g2 <- g2 + geom_errorbarh(aes(xmin = conf.low, xmax = conf.high))
g2 <- g2 + geom_vline()
g2 <- g2 + xlab("Estimates") + ylab("Variables")
g2 <- g2 + ggtitle("Plot of Step Model Estimates")

grid.arrange(g1, g2, nrow = 2)
```

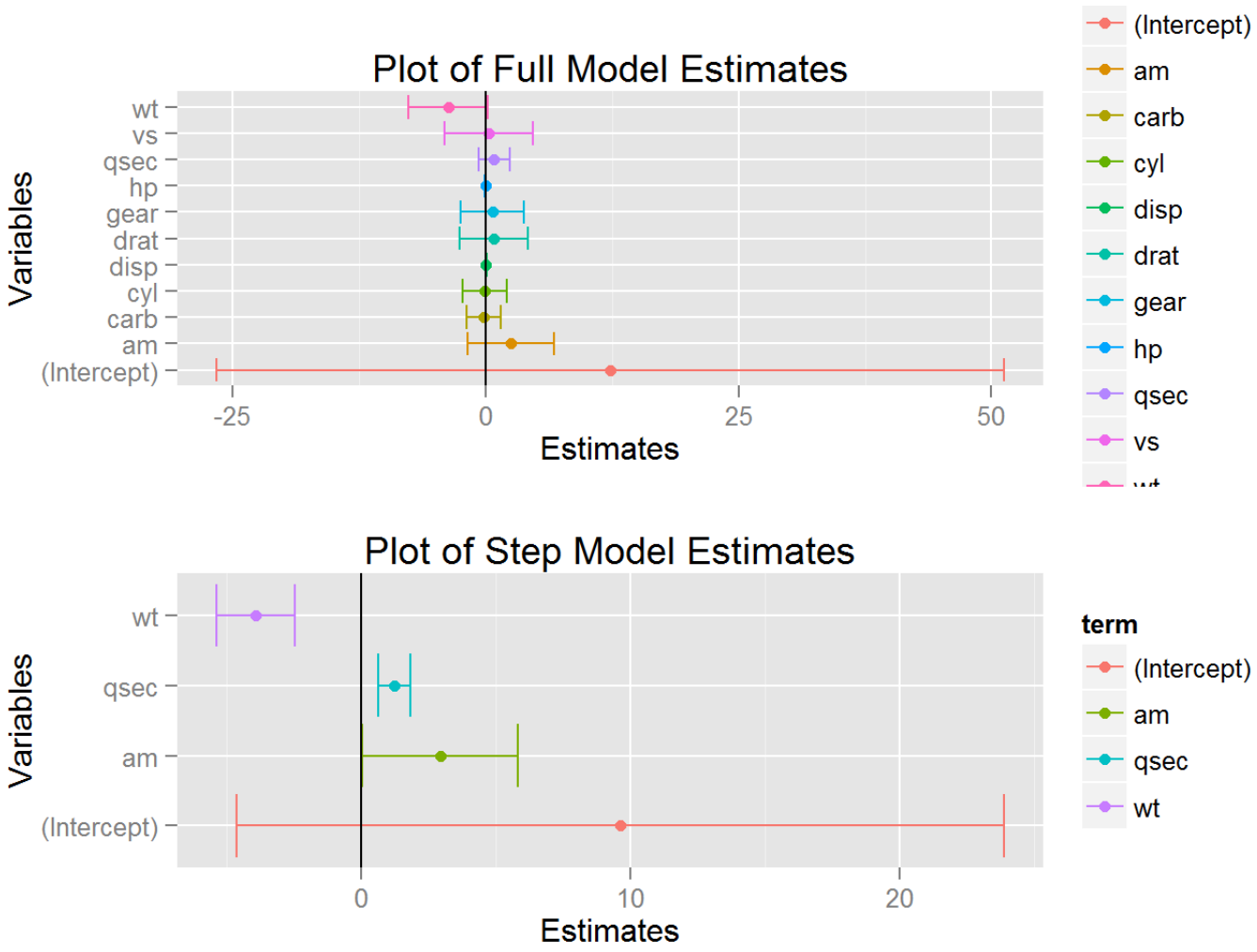
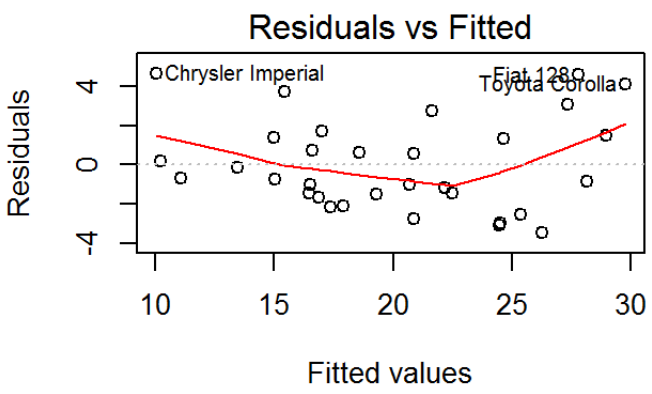
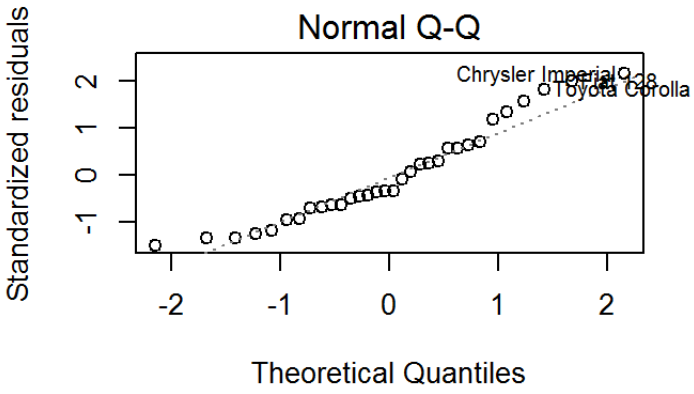
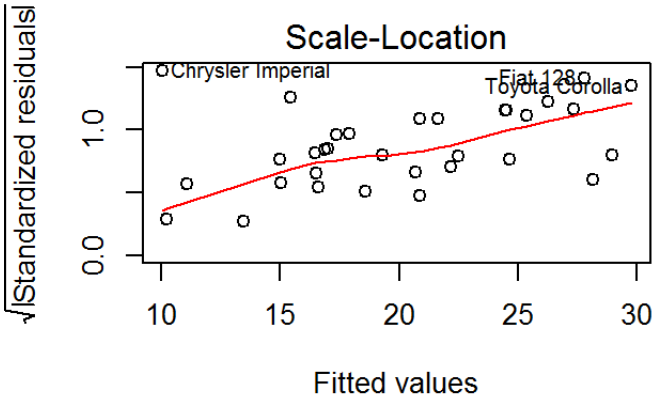


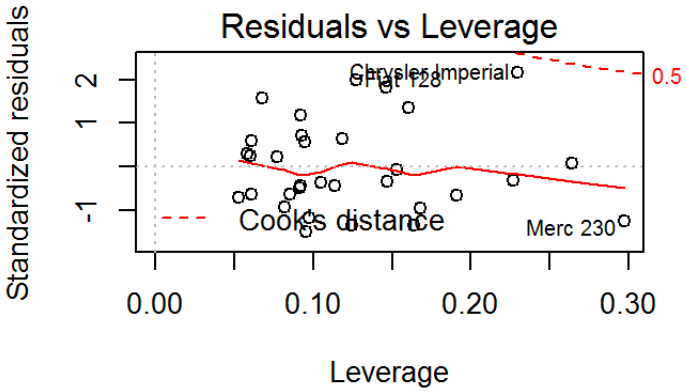
Figure 4: Diagnostic Graphics of modelStep

```
par(mfrow = c(2, 2))
plot(modelStep)
```


Residuals vs Fitted
Residuals
Fitted values
Chrysler Imperial
Fiat 128
Toyota Corolla
10
15
20
25
30
-4
0
4


Normal Q-Q
Standardized residuals
Theoretical Quantiles
Chrysler Imperial
Fiat 128
Toyota Corolla
-2
-1
0
1
2


Scale-Location
√|Standardized residuals|
Fitted values
Chrysler Imperial
Fiat 128
Toyota Corolla
10
15
20
25
30
0.0
1.0


Residuals vs Leverage
Standardized residuals
Leverage
Chrysler Imperial
Fiat 128
Cook's distance
Merc 230
0.5
0.00
0.10
0.20
0.30
-1
1
2