Motor Trend Magazine - Data Analysis Report

Executive Summary

Looking at the dataset of a collection of cars, the following report explores the relationships between a set of variables and miles per gallon (**MPG**), providing information on three main aspects:

(a) Is an automatic or manual transmission better for MPG

(b) Quantification of the MPG difference between automatic and manual transmissions

(c) what model best describes MPG data

Approach outline:

1. Dataset exploration and variable transformations
2. Statistical inference related to the difference between automatic and manual transmissions of mpg data
3. Simple linear regression model with MPG as the outcome and AM as the predictor
4. Multivariable Linear Models
5. Model Comparisons
6. Residuals and diagnostics

Results

Our analysis showed that cars with manual transmission are better in terms of miles per gallon compared to cars with automatic transmission. Based on the results of a linear regression model with transmission type as the only explanatory variable, a change from manual to automatic transmission will increase Miles/(US)gallon by 7.245. However, as indicated, transmission type explains only 34% of MPG variance.

In addition to transmission variable ("am" for automatic/manual), the best subset of cofounder variables explaining our MPG data include:

(a) Number of cylinders ("cyl", can be 4,6, or 8) **(b)** Gross horsepower ("hp") **(c)** Weight ("wt",lb/1000).

The proportion of variance explained by this model is 87% (see also figure 3). Based on this model, a change in transmission from automatic to manual is related to 1.8 increase in Miles/(US)gallon. However, transmission type does not appear to be a good explanatory variable in our model, with its contribution not reaching significance, and explaining only about 12% of our mpg data. In contrast, number of cylinders, horsepower and weight appear to be significantly influential for the MPG data, and variables that should be consider over transmission for explaining MPG.

1. Exploring the Dataset and Variable Transformations

```
data(mtcars)
dim(mtcars)
```

```
## [1] 32 11
```

```
head(mtcars,5)
```

```
##                    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
```

Overall, there are 11 variables related to automobiles, and data related to the performance of 32 different automobiles. We are interested on the relationships of **mpg** variable (miles per gallon) and **am** (automatic/manual transmission) variable.

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

2. Statistical Inference

Assuming that transmission data have a normal distribution, a t-test was performed on the two subsets of data related to automatic and manual transmission. The null hypothesis that both subsets come from the same distribution of mpg data was examined.

```
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
```

```
by(mtcars$mpg, mtcars$am, sd)
```

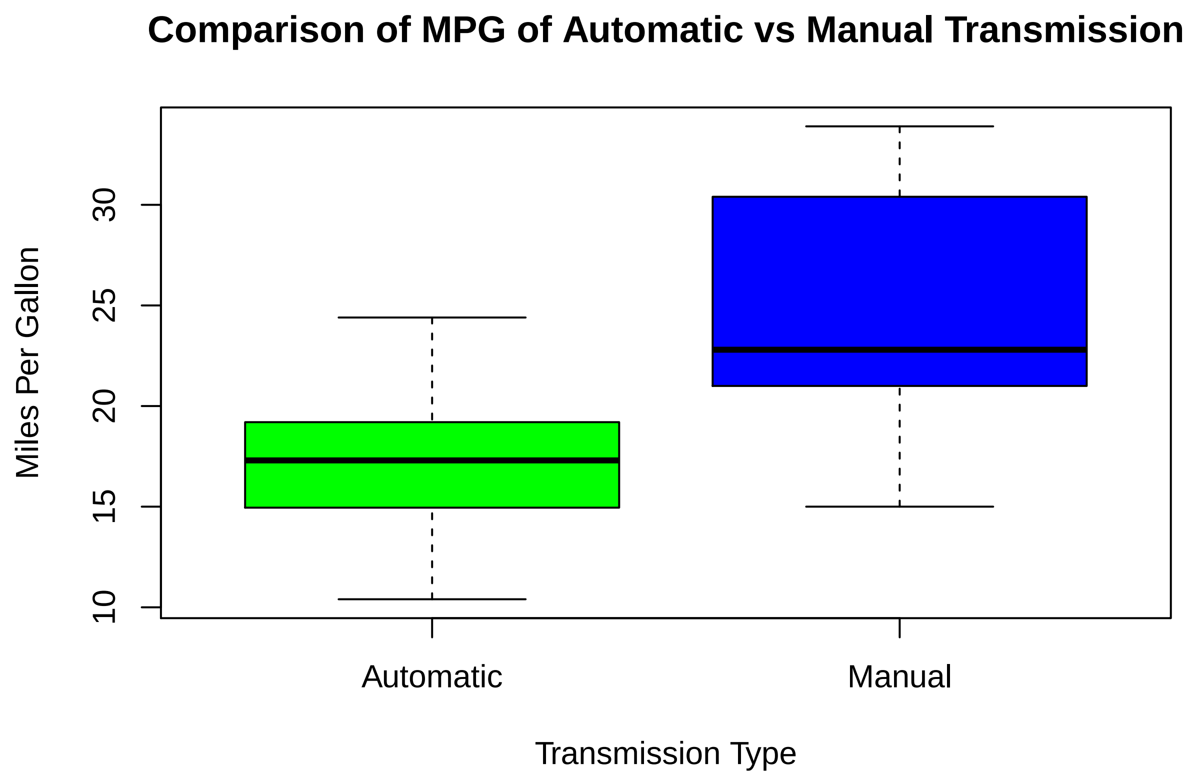
```
## mtcars$am: 0
## [1] 3.833966
## --------------------------------------------------------
## mtcars$am: 1
## [1] 6.166504
```

Results

Both data subsets come from different distributions of the mpg data. Based on the results, we see a significant difference between automatic and manual transmission (p = 0.001374), with manual transmission related to significantly higher MPG (M=24.39, SD = 6.16) compared to the automatic transmission (M = 17.15, SD = 3.83). Hence, we reject the null hypothesis supporting that both subsets come from the same mpg data distribution.

The figure below represents the distribution of mpg for each level of transmission (Automatic or Manual). Manual transmission appears to be related to higher MPG values.

```
boxplot(mpg ~ am, data = mtcars, col = (c("green","blue")), ylab = "Miles Per Gallon", xlab = "Transmis
```



These results are further analyzed and discussed using simple and multivariable linear regression models.

3. Simple Linear Regression Model

In this part we start exploring the relationships between the variables. We start by fitting a simple linear regression model, with MPG as the outcome, and AM as the predictor. This model will also serve as our base for comparisons with multivariable regression models.

```
fit <- lm(mpg ~ am, data = mtcars)
summary(fit)$coef
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## am1          7.244939   1.764422  4.106127 2.850207e-04
```

```
summary(fit)$r.squared
```

```
## [1] 0.3597989
```

Results reveal that a change from automatic to manual transmission will result in 7.245 increase of Miles/(US)gallon. The transmission type significantly contributes to the mpg data, however the model explains only 36% of variance in mpg data.

4. Multivariable Linear Regression Model

We fit a multivariable model including all variables in the dataset and mpg as the outcome.

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

```
##
## 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
```

As shown, the model that includes all the variables explains 78% of the variance in mpg data. Even though the overall model is significant (p = 0.000124), none of the variables appears to reach a level of significance for predicting mpg data.

In order to define the best subset of variables that best describe our MPG data, we performed stepwise regression using the R "step" function. This particular function runs multiple linear regression models and presents the best subset of variables, using both forward selection and backward elimination.

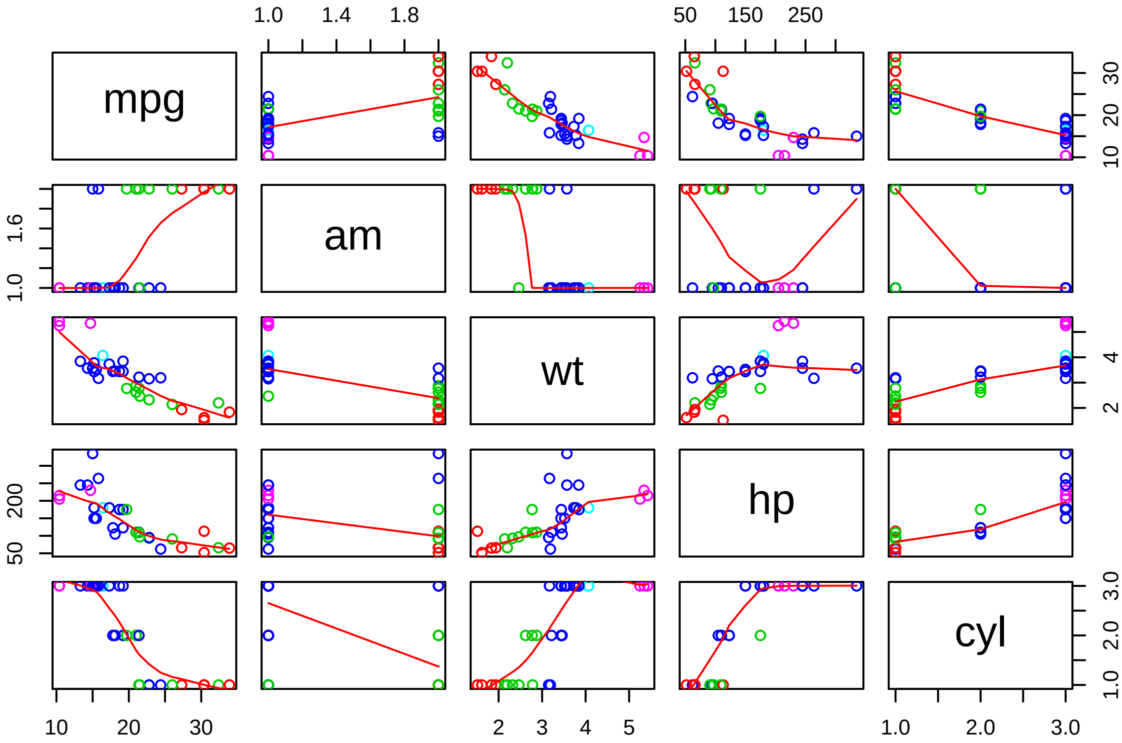
```
#Stepwise Regression
library(MASS)
#stepwise model selection by exact AIC
bestfit <- stepAIC(fullfit, direction = "both") #results are not presented
```

```
summary(bestfit)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + hp + wt + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9387 -1.2560 -0.4013  1.1253  5.0513
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.70832    2.60489  12.940 7.73e-13 ***
## cyl6        -3.03134    1.40728  -2.154  0.04068 *
## cyl8        -2.16368    2.28425  -0.947  0.35225
## hp          -0.03211    0.01369  -2.345  0.02693 *
## wt          -2.49683    0.88559  -2.819  0.00908 **
## am1          1.80921    1.39630   1.296  0.20646
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.41 on 26 degrees of freedom
## Multiple R-squared:  0.8659, Adjusted R-squared:  0.8401
## F-statistic: 33.57 on 5 and 26 DF,  p-value: 1.506e-10
```

The plots below represent the correlations for each variable in our model.

```
pairs(mpg ~ am + wt + hp + cyl, data = mtcars, panel = panel.smooth, col = 9 + mtcars$wt)
```

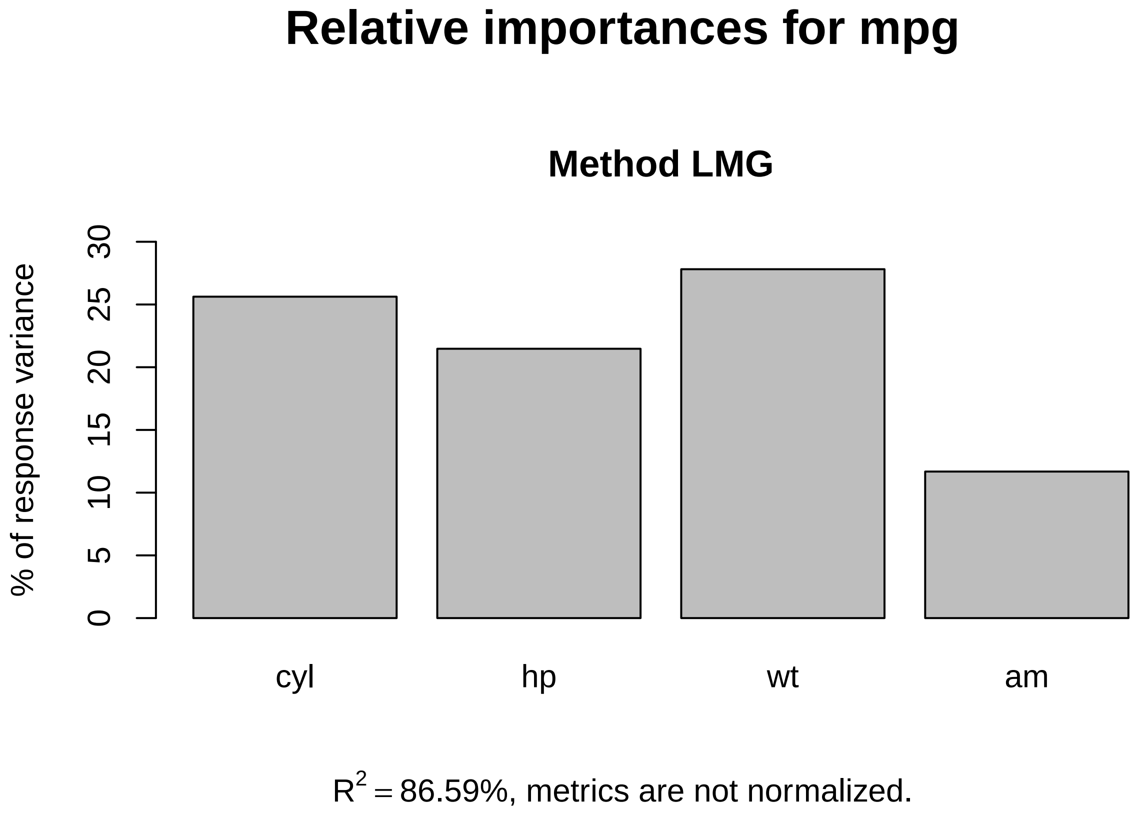


The p-value associated with the transmission type ("am"") variable is way above our alpha level (p = 0.206), suggesting a non-significant contribution of the transmission type variable in our model.In contrast, number of cylinders, horsepower and weight appear to be significantly influential for the MPG data. Details related to the calculation of relative importance for each predictor in our model are depicted below.The plot shows the accountability of each predictor in our model.

```
library(relaimpo)
calc.relimp(bestfit)
```

```
## Response variable: mpg
## Total response variance: 36.3241
## Analysis based on 32 observations
##
## 5 Regressors:
## Some regressors combined in groups:
##         Group  cyl : cyl6 cyl8
##
##  Relative importance of 4 (groups of) regressors assessed:
##  cyl hp wt am
##
## Proportion of variance explained by model: 86.59%
## Metrics are not normalized (rela=FALSE).
##
## Relative importance metrics:
##
##           lmg
## cyl 0.2562195
## hp  0.2147171
## wt  0.2781240
## am  0.1168192
##
## Average coefficients for different model sizes:
##
##              1group     2groups     3groups     4groups
## cyl6  -6.92077922 -5.45978507 -3.84697398 -3.03134449
## cyl8 -11.56363636 -8.21975665 -4.26613908 -2.16367532
## hp    -0.06822828 -0.03823319 -0.03494749 -0.03210943
## wt    -5.34447157 -4.14541848 -3.06985908 -2.49682942
## am     7.24493927  2.60447460  2.13055657  1.80921138
```

```r
plot(calc.relimp(bestfit))
```



Considering the non-significant contribution of our transmission type, it would be wise to check also for difference between our best-fit model and the model that does not contain the transmission type variable

```r
newfit <- lm(mpg ~ cyl + hp + wt, data = mtcars )
summary(newfit)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + hp + wt, data = mtcars)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2612 -1.0320 -0.3210  0.9281  5.3947
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 35.84600    2.04102  17.563 2.67e-16 ***
## cyl6        -3.35902    1.40167  -2.396 0.023747 *
## cyl8        -3.18588    2.17048  -1.468 0.153705
## hp          -0.02312    0.01195  -1.934 0.063613 .
## wt          -3.18140    0.71960  -4.421 0.000144 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.44 on 27 degrees of freedom
## Multiple R-squared:  0.8572, Adjusted R-squared:  0.8361
## F-statistic: 40.53 on 4 and 27 DF,  p-value: 4.869e-11
```

Simplifying the above mentioned model, we also consider to include only weight (“wt”) and number of cylinders (“cyl”) as explanatory variables for MPG data

```r
simplefit <- lm(mpg ~ wt + cyl, data = mtcars)
summary(simplefit)
```

```
##
## Call:
## lm(formula = mpg ~ wt + cyl, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5890 -1.2357 -0.5159  1.3845  5.7915
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  33.9908     1.8878  18.006  < 2e-16 ***
## wt           -3.2056     0.7539  -4.252 0.000213 ***
## cyl6         -4.2556     1.3861  -3.070 0.004718 **
## cyl8         -6.0709     1.6523  -3.674 0.000999 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.557 on 28 degrees of freedom
## Multiple R-squared:  0.8374, Adjusted R-squared:   0.82
## F-statistic: 48.08 on 3 and 28 DF,  p-value: 3.594e-11
```

5. Model Comparisons

Comparisons between all mentioned models, in order to define the best model explaining mpg data.

Models:

1. Base model - transmission as explanatory variable (simple linear regression)

2. multivariable model - all variables included

3. best-fit model - model defined using stepwise regression

4. best-fit model excluding transmission type (weight, cylinders, and horsepower as explanatory variables)

5. simplified model, including cylinder and weight only as the explanatory variables

First, the best-fit model is compared to the base model

```
anova(fit, bestfit)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ cyl + hp + wt + am
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     26 151.03  4    569.87 24.527 1.688e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Results reveals a significant difference with a small p-value (close to 0), suggesting that the multivariable model is more accurate compared to the simple model.

Second, we compare our best-fit model with the model that does not contain the transmission type.

```
anova(bestfit, newfit)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ cyl + hp + wt + am
## Model 2: mpg ~ cyl + hp + wt
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     26 151.03
## 2     27 160.78 -1    -9.752 1.6789 0.2065
```

The difference between the models did not reach significance (p = 0.21), with the new model explaining 86% of variance in mpg data (versus 87% of variance explained by our best-fit model)

Finally, we compare our best-fit model with the final simplified model.

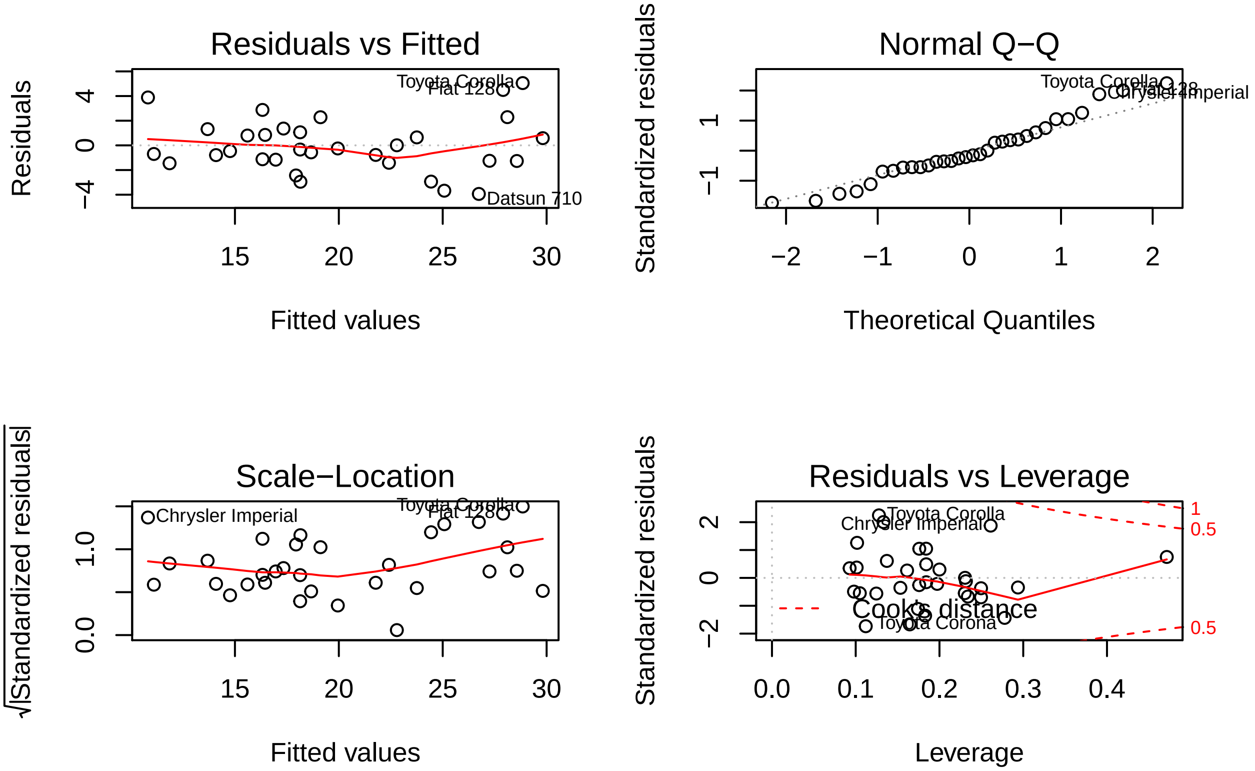
```
anova(bestfit, simplefit)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ cyl + hp + wt + am
## Model 2: mpg ~ wt + cyl
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     26 151.03
## 2     28 183.06 -2   -32.033 2.7573 0.08203 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

No difference occurs between the simplified model and our best-fit model (p = 0.08), indicating that our model is one of the most suitable for explaining MPG data.

6. Residuals and Diagnostics

```
# diagnostic plots
par(mfrow = c(2, 2))
plot(bestfit)
```



As shown in the **Residuals vs Fitted plot** data are randomly scattered on the indicating independence of the variable. The curve indicates a slight diversion from normality in the data.

The **Normal Q-Q plot** presents the points falling on the line indicating the normal distributions of the residuals.

The **Scale-Location plot** represents data points scattered in a constant band pattern, indicating constant variance.

The **Residuals vs Leverage plot** represents no leverage points, as all values fall well within the 0.5 bands.

In the following section, there are some regression diagnostics of our model to find out the top three outliers, meaning the points that are more distant from the cloud of data:

```
leverage <- hatvalues(bestfit)
tail(sort(leverage),3)
```

```
##       Toyota Corona Lincoln Continental       Maserati Bora
##           0.2777872           0.2936819           0.4713671
```

The top three influential points are also computed:

```
influential <- dfbetas(bestfit)
tail(sort(influential[,6]),3)
```

```
## Chrysler Imperial          Fiat 128     Toyota Corona
##         0.3507458         0.4292043         0.7305402
```