Motor Trend Investigates the Relationship Between Transmission Type and Miles per Gallon

Aslak

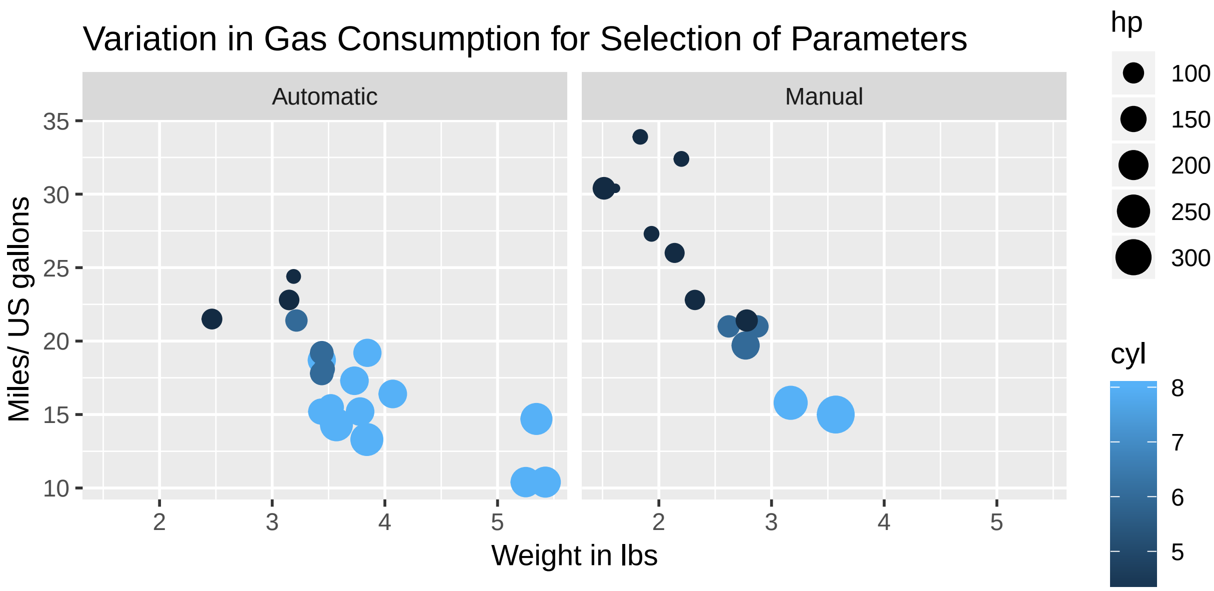
```
# This report has been written in knitr.
```

Executive Summary

The report shows that features of a car model, such as transmission type, number of cylinders and horsepower, can predict the consumption (MPG). This multivariable regression model explains approximately 80 % of the variance.

The provided data in the mtcars data set shows with 95 % confidence that manual transmission is better for MPG. However, the data has its limitation and the linear model emphasizes that MPG should not be predicted solely by transmission type.

Exploratory Data Analysis: manual better for MPG - but other features should be considered



The plot above suggests that the cars with manual transmission have more miles per gallon. However, the graph shows that there are other parameters - such as number of cylinders, weight and horsepower - that may affect miles per gallon.

Is an Automatic or Manual Transmission Better for MPG?

As a starting point, the MPG difference for transmission type is shown. One way is to calculate the different in mean MPG - for the data set provided:

Automatic	Manual
17.1	24.4

The mean MPG for automatic transmission is shown to be worse compared to manual - for the car models in the mtcars data set. Additionally, we calculate whether this can be concluded with 95 % confidence. Hence, our hypothesis H0 is that the mean for automatic transmission is equal to manual transmission.

```
## [1] -11.280194  -3.209684
## attr(,"conf.level")
## [1] 0.95
```

The t.test shows that - **given the provided data** - automatic transmission has a worse miles per gallon than manual transmission (with 95 % confidence). However, the data has its limitation. As the explanatory data analysis has indicated, a multivariable linear regression may improve the prediction of MPG. If the data was larger, it is not unlikely that the t test's conclusion would be different.

Model Selection: Transmission Type, Number of Cylinders and Horsepower

Our strategy for selecting a model is as follows: start with a simple linear regression model and add different features of the car models as predictors. It is a pragmatic approach to start with a simple linear regression, which may be provide good "value" for litte effort. However, our intuition supports the idea of a multivariable regression - there are probably other features that should be accounted for - other than transmission type.

To conclude on a model, a test of the significance/importance of each feature to predict miles per gallon is then performed (**anova test**).

The anova test shows (details in appendix) that the **miles per gallon consumption** predicted by **transmission type** should include the features: **number of cylinders** and **horsepower**. The anova test shows significance for the inclusion of these features. The transmission type explains approximately 34 % of the variance (adjusted R-squared) in a simple linear regression model. In the model with the features cylinders and horsepower - approximately 80 % of the variance is explained.

Residuals and Diagnostics supports Multivariable Linear Model (details in appendix)

After selecting the multivariable linear model that we believe can predict the consumption - we assess the residuals and perform some diagnostics. This can aid us in finding limitations and weaknesses of the model.

The residuals of the linear model does not imply a systematic pattern - which supports the choice of a multivariable linear model. Additionally, some diagnostics are simulated to reveal any other weaknesses in the linear model.

The diagnostics show that the multivariable linear regression model with transmission type, number of cylinders and horsepower is a useful model for predicting miles per gallon. These graphs are included in the appendix.

Uncertainty in Conclusions

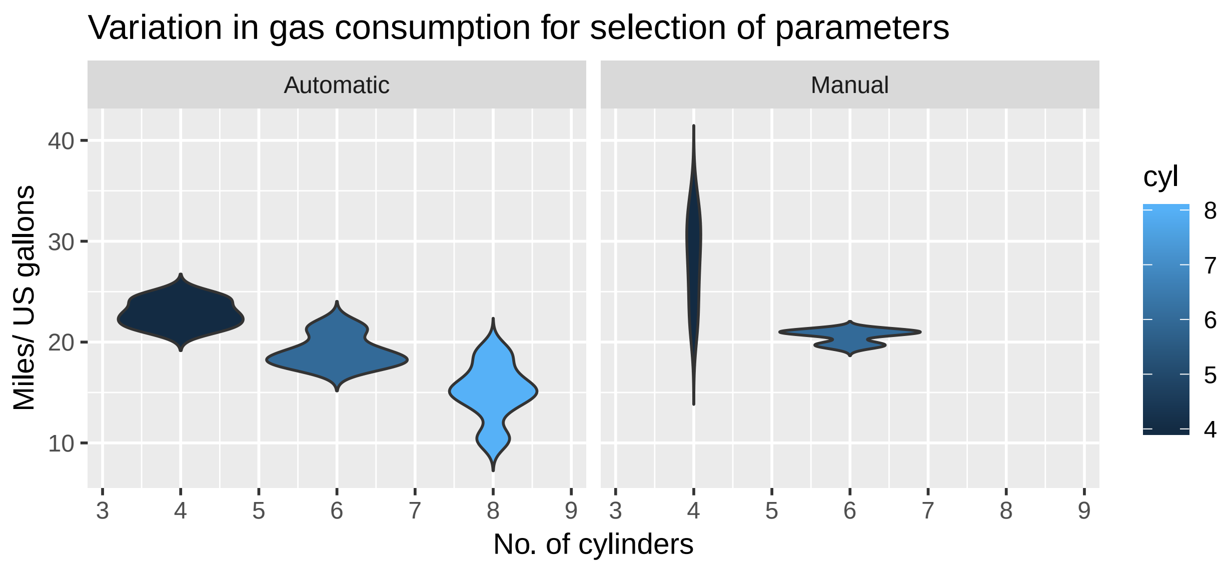
Lastly, we illustrate the uncertainty in our conclusion of the model by calculating the confidence interval (uncertainty around the mean predictions). The intercept corresponds to automatic transmission in the table below.

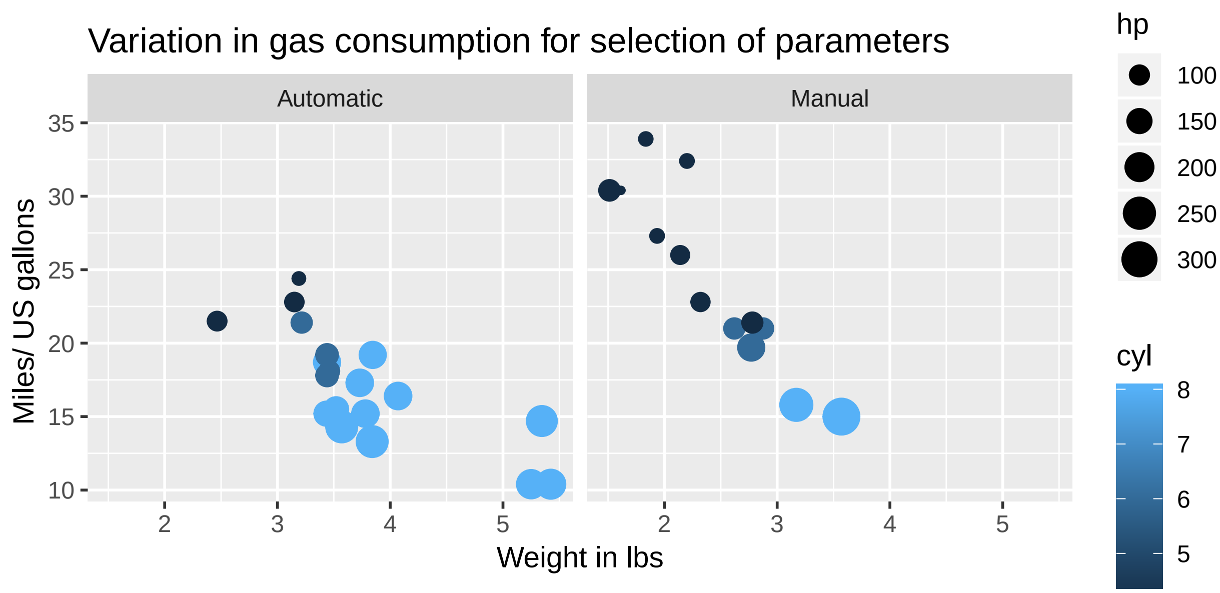
```
##                     2.5 %      97.5 %
## (Intercept)  24.37421028 30.21758829
## factor(am)1   1.57962875  6.73608418
## factor(cyl)6 -7.07929845 -0.76985854
## factor(cyl)8 -8.66871103  1.60188319
## hp           -0.07415067 -0.01433722
```

Appendix: Supporting Figures

Initiation

Exploratory Data Analysis





Is an Automatic or Manual Transmission Better for MPG?

Mean MPG for automatic (0) and manual (1) transmission.

```
## # A tibble: 2 x 2
##      am  Mean
##   <dbl> <dbl>
## 1     0  17.1
## 2     1  24.4
```

T test with hypothesis H0 that the mean for automatic and manual transmission is equal.

```
## [1] -11.280194  -3.209684
## attr(,"conf.level")
## [1] 0.95
```

Model Selection

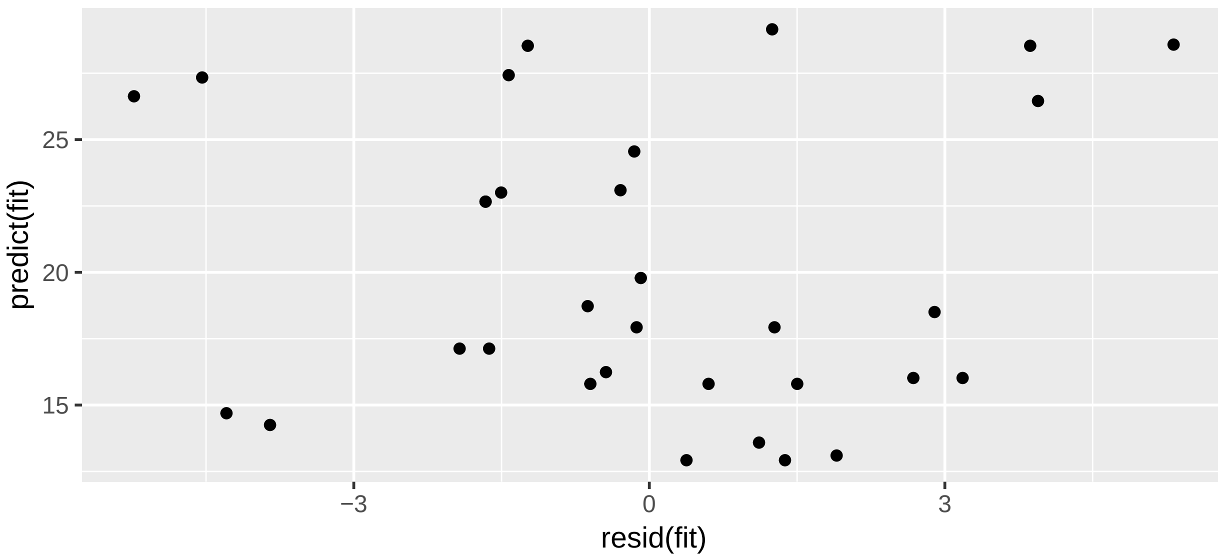
```
## Analysis of Variance Table
##
## Model  1: mpg ~ factor(am)
## Model  2: mpg ~ factor(am) + factor(cyl)
## Model  3: mpg ~ factor(am) + factor(cyl) + disp
## Model  4: mpg ~ factor(am) + factor(cyl) + disp + hp
## Model  5: mpg ~ factor(am) + factor(cyl) + disp + hp + drat
## Model  6: mpg ~ factor(am) + factor(cyl) + disp + hp + drat + wt
## Model  7: mpg ~ factor(am) + factor(cyl) + disp + hp + drat + wt + qsec
## Model  8: mpg ~ factor(am) + factor(cyl) + disp + hp + drat + wt + qsec +
##     factor(vs)
## Model  9: mpg ~ factor(am) + factor(cyl) + disp + hp + drat + wt + qsec +
##     factor(vs) + factor(gear)
## Model 10: mpg ~ factor(am) + factor(cyl) + disp + hp + drat + wt + qsec +
##     factor(vs) + factor(gear) + factor(carb)
##    Res.Df    RSS Df Sum of Sq       F   Pr(>F)
## 1      30 720.90
## 2      28 264.50  2    456.40 28.4297 7.89e-06 ***
## 3      27 230.46  1     34.04  4.2402  0.05728 .
## 4      26 183.04  1     47.42  5.9078  0.02809 *
## 5      25 182.38  1      0.66  0.0820  0.77855
## 6      24 150.10  1     32.28  4.0216  0.06331 .
## 7      23 141.21  1      8.89  1.1081  0.30916
## 8      22 139.02  1      2.18  0.2719  0.60964
## 9      20 134.00  2      5.02  0.3128  0.73606
## 10     15 120.40  5     13.60  0.3388  0.88144
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

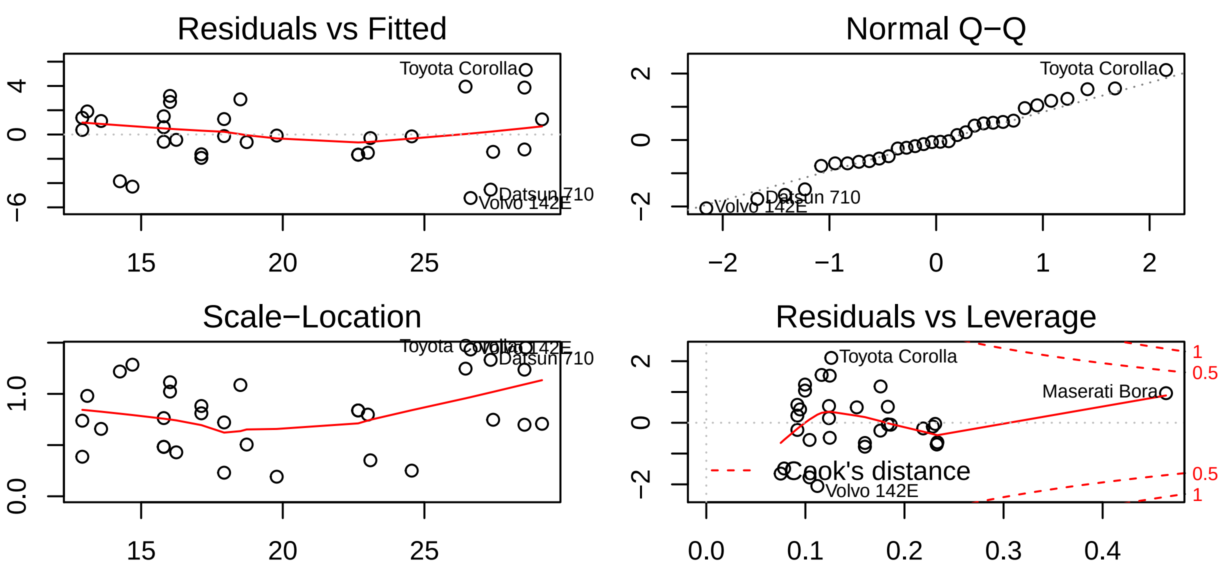
R-squared (adjusted) for the linear model and the suggested multivariable model.

```
## [1] 0.3384589
```

```
## [1] 0.7989306
```

Residuals and Diagnostics





Uncertainty in Conclusions

```
##                     2.5 %      97.5 %
## (Intercept)  24.37421028 30.21758829
## factor(am)1   1.57962875  6.73608418
## factor(cyl)6 -7.07929845 -0.76985854
## factor(cyl)8 -8.66871103  1.60188319
## hp           -0.07415067 -0.01433722
```