Exploring mtcars dataset

Valentin Goverdovsky

26 January 2016

Executive summary

This report explores **mtcars** dataset and in particular the relationship between the type of transmission and fuel economy. To analyse the dataset, linear regression modelling methods are employed. Using these we establish that the transmission type has limited effect on fuel economy if other variables are taken into account.

Exploring the dataset

Firstly we check if it is possible to visually observe any relation between fuel consumption and the type of transmission. It's easiest to visualize using a boxplot, see *Figure 1* in the **Appendix**. There is clear difference between the two boxplots suggesting that cars with automatic transmission have lower fuel efficiency compared to those with manual transmission.

Analysis

To quantify the observed relation between **am** and **mpg** we can fit a linear regression with one variable and check the coefficients:

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

The intercept coefficient shows the mean miles per gallon for cars with automatic transmission, while **am** coefficient indicates increase in the mean mpg in cars with manual transmission compared to those with automatic transmission. It suggests that the cars with manual transmission do 7 (+/- 1.76) miles per gallon more compared to those with automatic transmission. Both coefficients are statistically significant.

The observed relationship is not exactly intuitive, thus further investigation is required to establish if there are any confounding variables. Confounding variables would exhibit strong association with the predictor and we can find these by fitting a linear model between **am** and each of the other variables and finding only those with statistically significant coefficients:

```
## [1] "cyl"  "disp" "drat" "wt"   "gear"
```

Now we can add these variables to the linear model and check the coefficients:

```
##                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) 42.849311191 7.74036062  5.5358288 9.380637e-06
## am           1.181575927 1.87872450  0.6289245 5.351034e-01
## cyl         -1.740815446 0.65474013 -2.6587884 1.348144e-02
## disp         0.005705375 0.01248138  0.4571109 6.515388e-01
## drat         0.331102911 1.61859955  0.2045614 8.395711e-01
## wt          -3.425362195 1.22604961 -2.7938202 9.851503e-03
## gear        -1.072365153 1.15316450 -0.9299325 3.612992e-01
```

This time we observe that the **am** variable has limited effect on fuel economy - its coefficient is not statistically significant. To check if there are any outliers in this model we investigate the residuals using the standard residual plots, see *Figure 2* in the **Appendix**. These don't exhibit any particular abnormalities, thus we conclude that **fit2** is a good model.

Finally we attempt to investigate if including other variables in the model would be beneficial by fitting the model with all the variables included and using analysis of variance to verify if adding all the variables to the model produced a better fit:

```
fitAll<-lm(mpg~.,data=mtcars)
anova(fit1,fit2,fitAll)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + cyl + disp + drat + wt + gear
## 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     25 182.10  5    538.80 15.3426 2.127e-06 ***
## 3     21 147.49  4     34.61  1.2318    0.3276
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value for the model with all the variables included when compared to fit2 is not significant, implying that **am, cyl, disp, wt, gear and drat** variables explain the majority of the variation in the **mpg**.

Conclusion

When taken in isolation the transmission type seems to exhibit strong correlation with the fuel economy of the cars, but when other variables, such as weight, rear axle ratio, the number of forward gears, etc. are also included in the linear model the effect of the transmission type is significantly reduced implying limited or no effect on the fuel economy.

Appendix

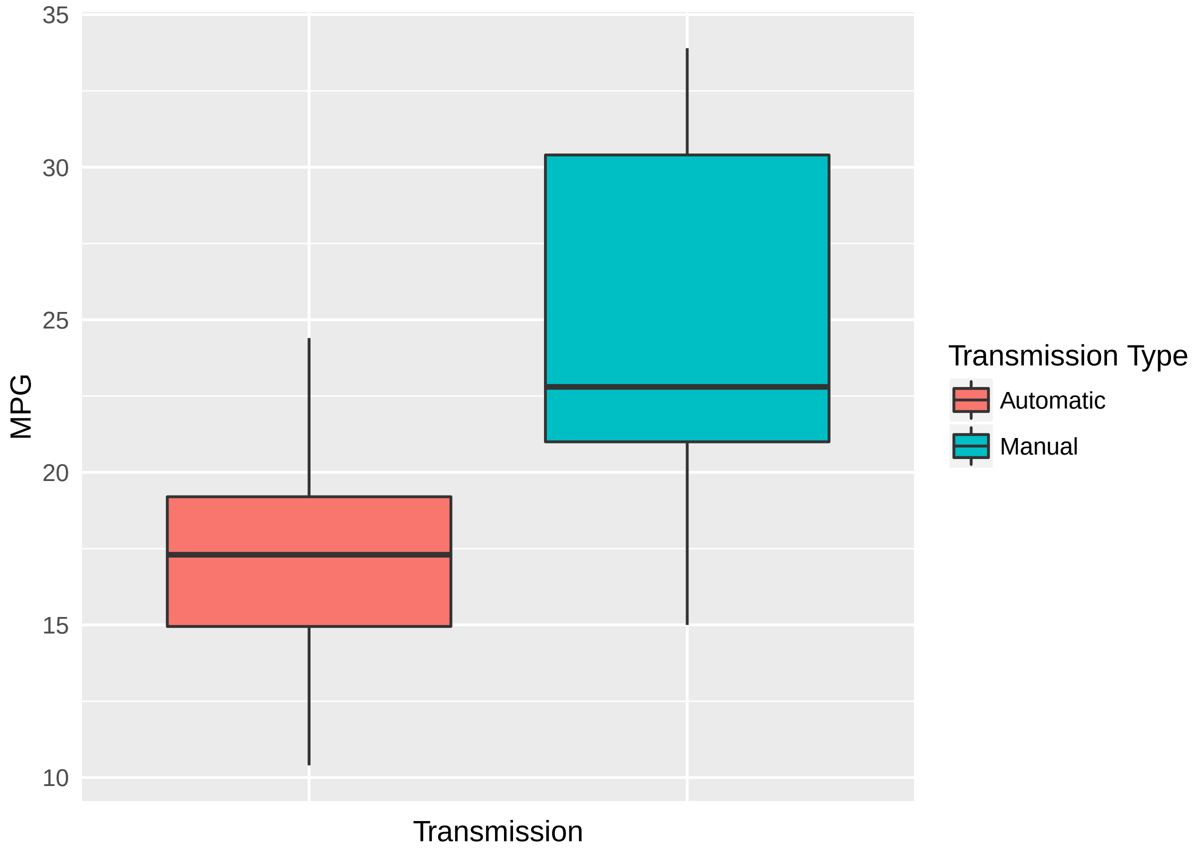

Figure 1: MPG vs Transmission

Code for finding variables which have statistically significant influence on **am** variable:

```
pval <- numeric(dim(mtcars)[2] - 2)
subcars <- subset(mtcars, select = c(-mpg,-am))
for (i in seq(dim(mtcars)[2] - 2)) {
    fit<-lm(mtcars[,names(mtcars) == 'am'] ~ subcars[,i])
    pval[i]=summary(fit)$coef[2,4]
}
names(subcars)[pval<0.05]
```

```
## [1] "cyl"  "disp" "drat" "wt"   "gear"
```

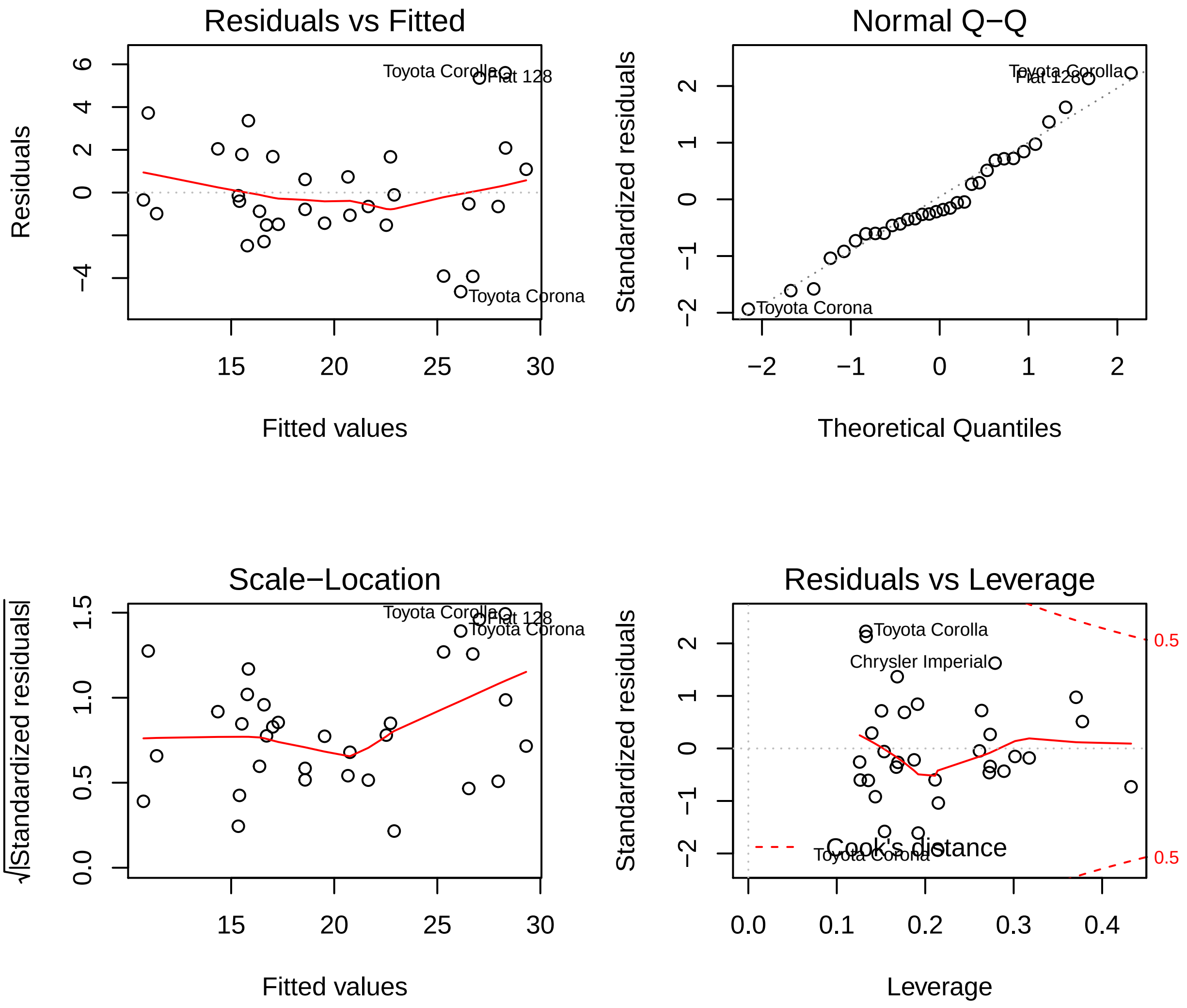

Figure 2: Residual analysis plot