Motor Trend Analysis

Tomasz Jaskula

5 août 2016

Executive Summary

The goal of the study is to explore the data set of collection of cars and answering the following questions:

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

Analysis

Exploring the data

Let's explore data size

```
dim(mtcars)
```

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

Variables of the data:

```
names(mtcars)
```

```
##  [1] "mpg"  "cyl"  "disp" "hp"   "drat" "wt"   "qsec" "vs"   "am"   "gear"
## [11] "carb"
```

Figure 1 shows how miles per US gallon `mpg` relates to transmission type `am`. We can clearly see a difference between the two. At a glance we know that Manual transmissions seem to get better gas mileage but we have to dig deeper to find out if this impact is really a transmission type or some other car characteristics.

Model selection

The model selection strategy would be to compare a simple linear model based only on `mpg` and `am` variables. Then use an automatic model selection based on the R `step` function.

Correlation

To determine which predictor variables should be included in our regression model we can build a correlation matrix and check how each of the variable is related to the `mpg` variable.

```
# we use the original mtcars with non transformed variables
sort(cor(mtcars_original)[1,])
```

```
##         wt        cyl       disp         hp       carb       qsec
## -0.8676594 -0.8521620 -0.8475514 -0.7761684 -0.5509251  0.4186840
##       gear         am         vs       drat        mpg
##  0.4802848  0.5998324  0.6640389  0.6811719  1.0000000
```

The result shows that the most correlated variables to `mpg` (except `am` that we have to include in our model) are `wt`, `cyl`, `disp` and `hp`. However it seems that `cyl` and `disp` are collinear and we shouldn't have them both included in the model.

Linear regression models

We start our model testing with a simple model and single predictor variable `am`.

```
fit1 <- lm(mpg ~ am, data = mtcars)
summary(fit1)$coef
```

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

```
summary(fit1)$r.squared
```

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

Interpreting the result we can see that cars with manual transmission have **7.245** Miles per gallon more the automatic. However our R-squared value is of 0.3598, which means that only **35.98%** of the variance is explained by the model.

We need to understand what is the impact of the other variables.

Let's try with automatic model selection

```
fit2 <- step(lm(mpg ~ ., data = mtcars), trace=0, steps=1000, direction="both")
summary(fit2)$coef
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 33.70832390 2.60488618 12.940421 7.733392e-13
## cyl6        -3.03134449 1.40728351 -2.154040 4.068272e-02
## cyl8        -2.16367532 2.28425172 -0.947214 3.522509e-01
## hp          -0.03210943 0.01369257 -2.345025 2.693461e-02
## wt          -2.49682942 0.88558779 -2.819404 9.081408e-03
## amManual     1.80921138 1.39630450  1.295714 2.064597e-01
```

```
summary(fit2)$r.squared
```

```
## [1] 0.8658799
```

We can see that the automatic model selection is based on the same variables we have chosen based on the correlation check i.e `am`, `wt`, `cyl` and `hp`. This shows that the most negative influence on the Miles per gallon has cylinders and weight. For example, each increase in weight by 1000lb (`wt`) decreases the `mpg` by **2.49683** miles. As for R-squared value we obtain 0.8659 which means that **86.59%** of the variation is explained by the model which indicates it's a robust and highly predictive model.

Comparing the model `fit1` to `fit2` using an Analysis of Variance (ANOVA) shows our second model `fit2` based on multi-variable regression is superior to the first model.

```r
anova(fit1, fit2)
```

The p-value of **1.688e-08** confirm this.

Diagnostics

Now that we have made our model selection which is `fit2`the next thing to do would be to run some diagnostics and to look at the **Residuals** plot in appendix **Figure 2**.

Let's run some more diagnostics. Are there any influential and leverage outlying points:

```r
infl <- dfbetas(fit2)
tail(sort(infl[, "amManual"]), 3)
```

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

```r
levrg <- hatvalues(fit2)
tail(sort(levrg), 3)
```

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

Again, except Maserati Bora we can see these cars present in our diagnostic plots **Figure 2** which indicates our analysis is correct.

Conclusion

Our analysis allowed to answer the question if the manual or automatic transmissions has a better MPG (Miles per gallon). The cars with manual transmissions tend to have a better gas millage on average. Our best model `fit2` explained **86%** of the variance but there is still some amount of uncertainty. The most important influence seems to have the weight of the car. It could be just that the cars with automatic transmission tend to be heavier. In our analysis we also quantified the MPG difference between automatic and manual transmissions.

Appendix

Figure 1: MPG by transmission type

The first idea would be to visualize the difference of how `mpg` usage relates to the transmission.

```r
library(ggplot2)
```

```r
ggplot(mtcars, aes(x=am, y=mpg, fill=am)) +
  geom_boxplot() +
  ylab("Miles per US gallon") +
  xlab("Transmission") +
  ggtitle("Figure 1: MPG by transmission type") +
  guides(fill=FALSE)
```

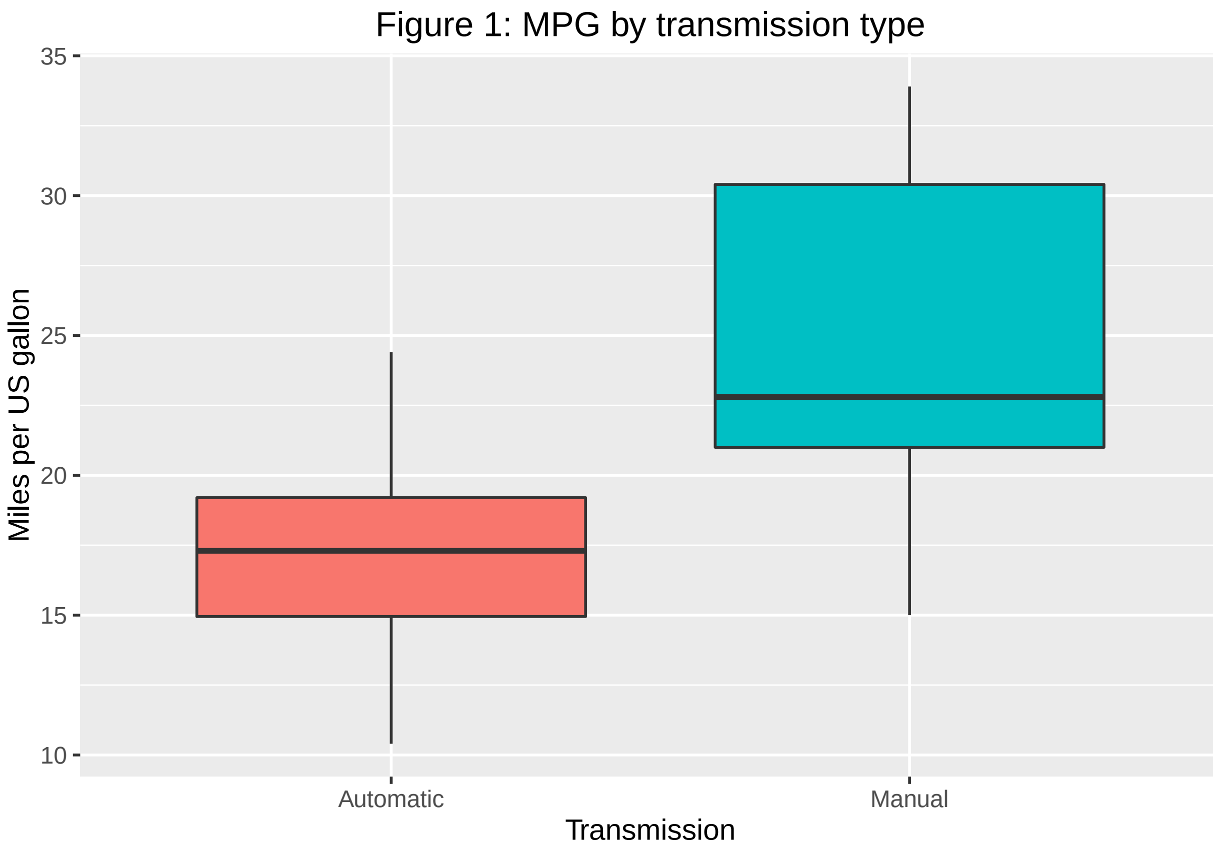
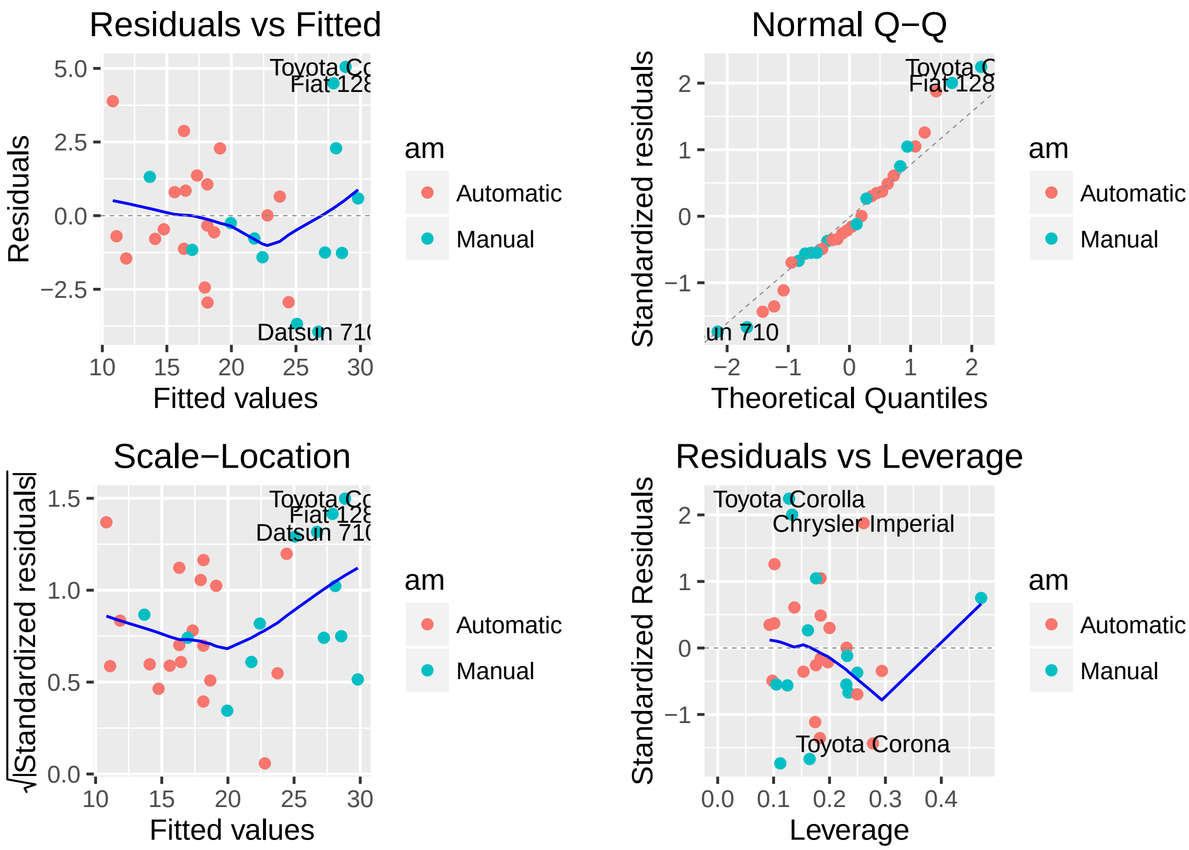


Figure 2: Diagnostic plots

The normal Q-Q plot shows residual points located mostly near the line implying the residuals are normally distributed. The Residuals vs. Fitted plot show randomly scattered points above and below the 0 line. We cannot see any pattern which means it show normality and no evidence of heteroskedasticity.

```
library(ggfortify)
```

```
autoplot(fit2, data = mtcars,
         colour = 'am', label.size = 3)
```


Residuals vs Fitted
Residuals
Fitted values
Toyota Co
Fiat 128
Datsun 710
am
Automatic
Manual
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Toyota C
Fiat 128
un 710
Scale−Location
√|Standardized residuals|
Residuals vs Leverage
Standardized Residuals
Leverage
Toyota Corolla
Chrysler Imperial
Toyota Corona