Motor Trend Analysis

Tomasz Jaskula

5 août 2016

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

The data used for the analysis is the `mtcars` data set. The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models).

Loading the data

```r
data(mtcars)
mtcars_original <- mtcars # saving for later use
```

Exploring the data

Let's explore data size

```r
dim(mtcars)
```

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

Structure of the data:

```r
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 ...
```

As we can see all the variables are numeric. Let's describe them:

- **mpg:** Miles/(US) gallon
- **cyl:** Number of cylinders
- **disp:** Displacement (cu.in.)
- **hp:** Gross horsepower
- **drat:** Rear axle ratio
- **wt:** Weight (1000 lbs)
- **qsec:** 1/4 mile time
- **vs:** V/S
- **am:** Transmission (0 = automatic, 1 = manual)
- **gear:** Number of forward gears
- **carb:** Number of carburetors

So we are particularly interested in `mpg` (Miles/US gallon) variable as a outcome and a relationship with other variables and how they influence the outcome.

Data transformation

For better interpretability let's transform some of the variables

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

Also for our predictor variable `am` we convert it to factor levels `Automatic` and `Manual`.

```
mtcars$am <- factor(mtcars$am, labels = c("Automatic", "Manual"))
```

Figure 1 shows how miles per US gallon `mpg` relates to transmission type. This can be easily confirmed comparing averages of Miles per gallon by transmission type

```
aggregate(mpg ~ am, data = mtcars, mean)
```

```
##          am      mpg
## 1 Automatic 17.14737
## 2    Manual 24.39231
```

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.

Let's start by looking at the variables correlation to the outcome `mpg`.

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. So the final choice for the model would be to keep as predictors the following variables `am`, `wt`, `cyl` and `hp`. Except `am` variable, all the other has the negative impact on the `mpg` which is quite logical because the more important the car weight is or the horse power, the fewer miles per gallon it can make. This can be confirmed later in the automatic model selection.

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)
```

```
##
## Call:
## lm(formula = mpg ~ am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.147      1.125  15.247 1.13e-15 ***
## amManual       7.245      1.764   4.106 0.000285 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.902 on 30 degrees of freedom
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3385
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

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

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

```
##
## 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 **
## amManual     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
```

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. It is also quite expected that as more cylinders a car has the more gas it will use. The same goes for horse power. 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)
```

```
## 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
```

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**. 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.

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

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

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

These cars are present in our diagnostic plots **Figure 2**

How about leverage outlying points

```
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 and as you can see in **Figure 3** 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.

```
library(ggplot2)
```

```
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)
```

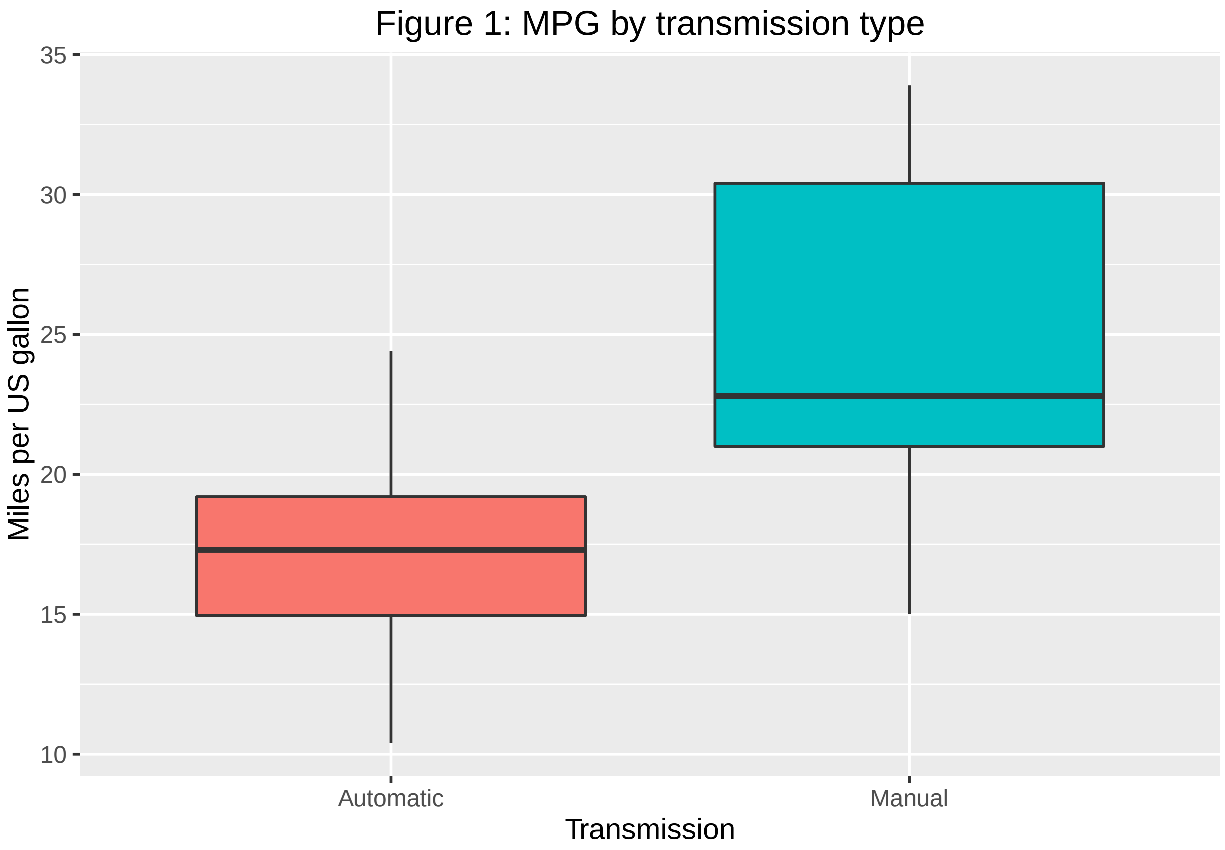


Figure 2: Diagnostic plots

```
library(ggfortify)
```

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

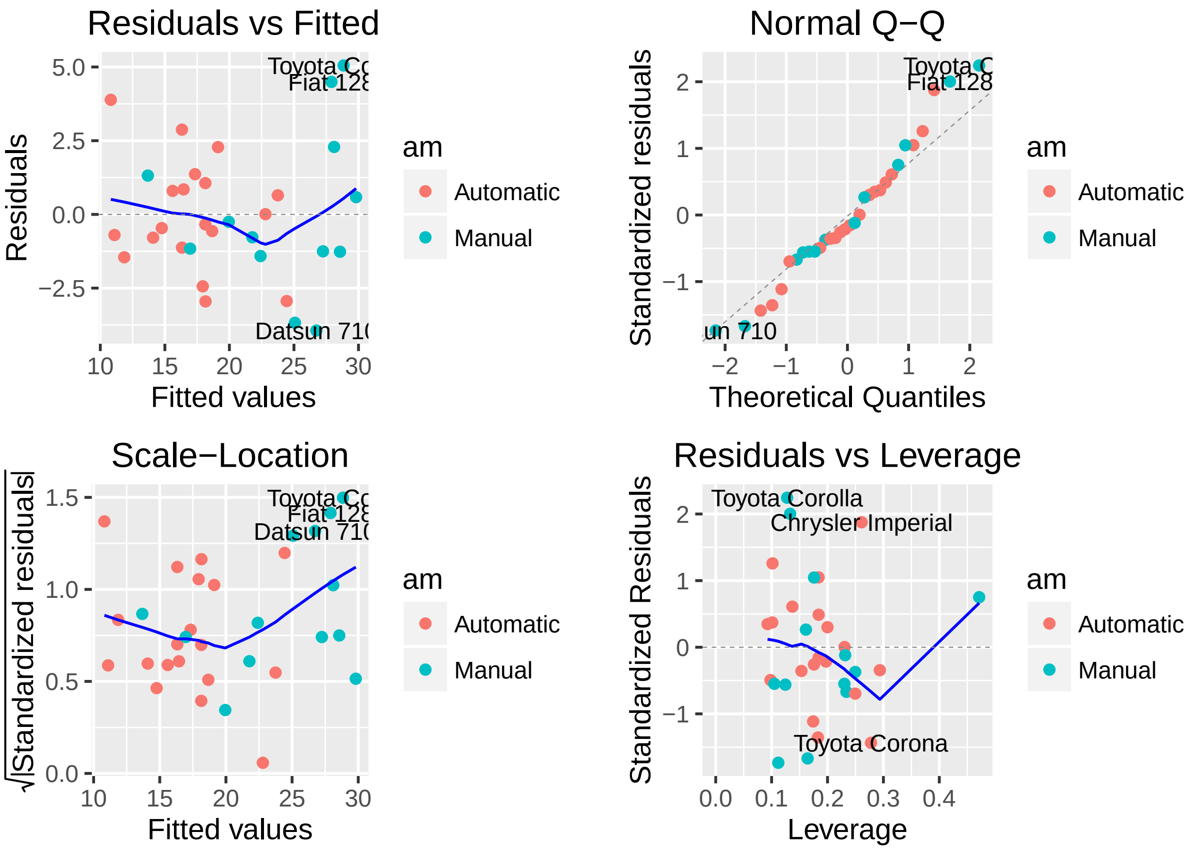
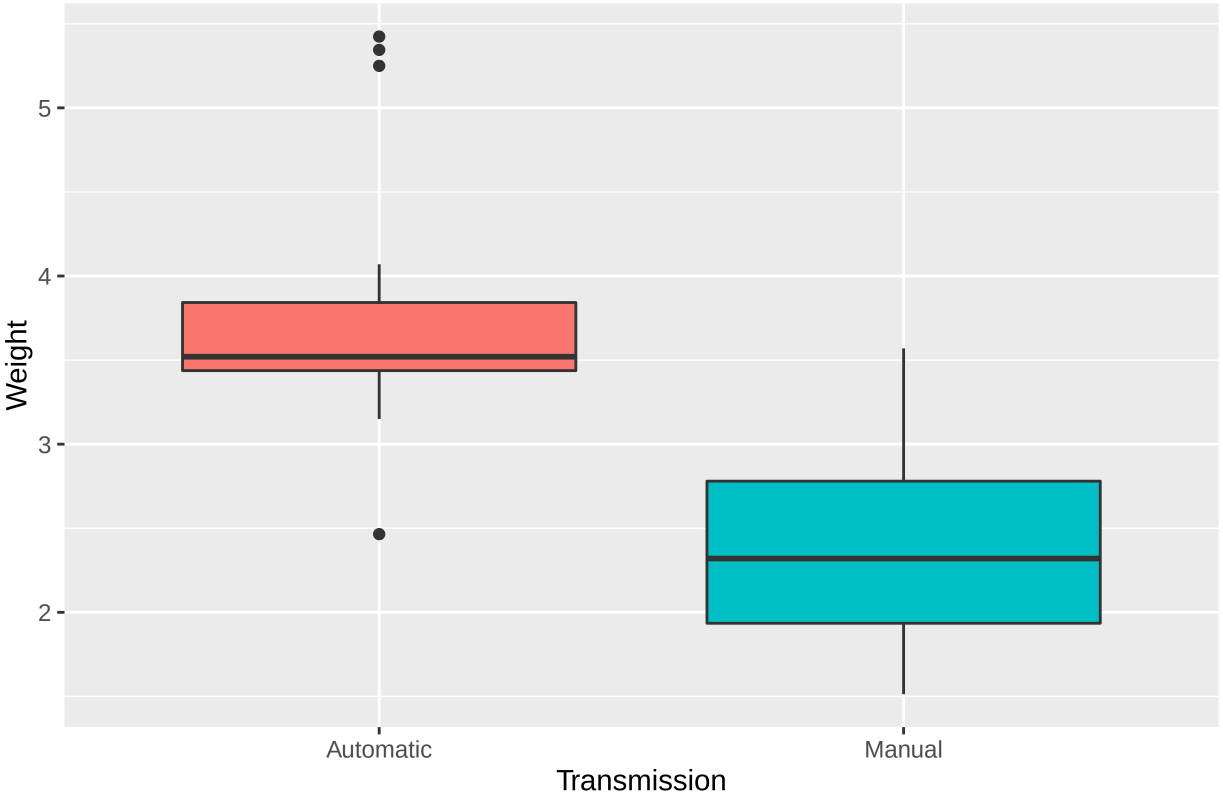


Figure 3: Weight per transmission type

```r
ggplot(mtcars, aes(x=am, y=wt, fill=am)) +
  geom_boxplot() +
  ylab("Weight") +
  xlab("Transmission") +
  ggtitle("Figure 3: Weight by transmission type") +
  guides(fill=FALSE)
```

Figure 3: Weight by transmission type


Weight
Automatic
Manual
Transmission