Motor Trend - The relationship between a set of variables and miles per gallon

By Ashwin Venkatesh Prabhu

Executive Summary

We work for Motor Trend, a magazine about the automobile industry. Looking at a data set of a collection of cars, we are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome).

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

We will start with the following steps:

- Process the data
- Conduct exploratory data analysis, focusing on the two paramaters we are interested in (Transmission and MPG)
- Model selection, where we try different models to help us answer our questions
- Model examination, to see wether our best model holds up to our standards
- A Conclusion where we answer the questions based on the data

Processing

Change ‘am’ to a factor (1 = manual, 0 = automatic). Make cylinders a factor.

```
library(ggplot2)
library(GGally)
library(dplyr)
library(ggfortify)

data(mtcars)

df <- mtcars
df$am <- as.factor(df$am)
levels(df$am) <- c("automatic", "manual")

df$cyl <- as.factor(df$cyl)
df$gear <- as.factor(df$gear)
df$vs <- as.factor(df$vs)
levels(df$vs) <- c("V", "S")
```

Exploratory data analysis

Look at the dimensions & head of the dataset to get an idea

```
# Result 1
dim(df)
```

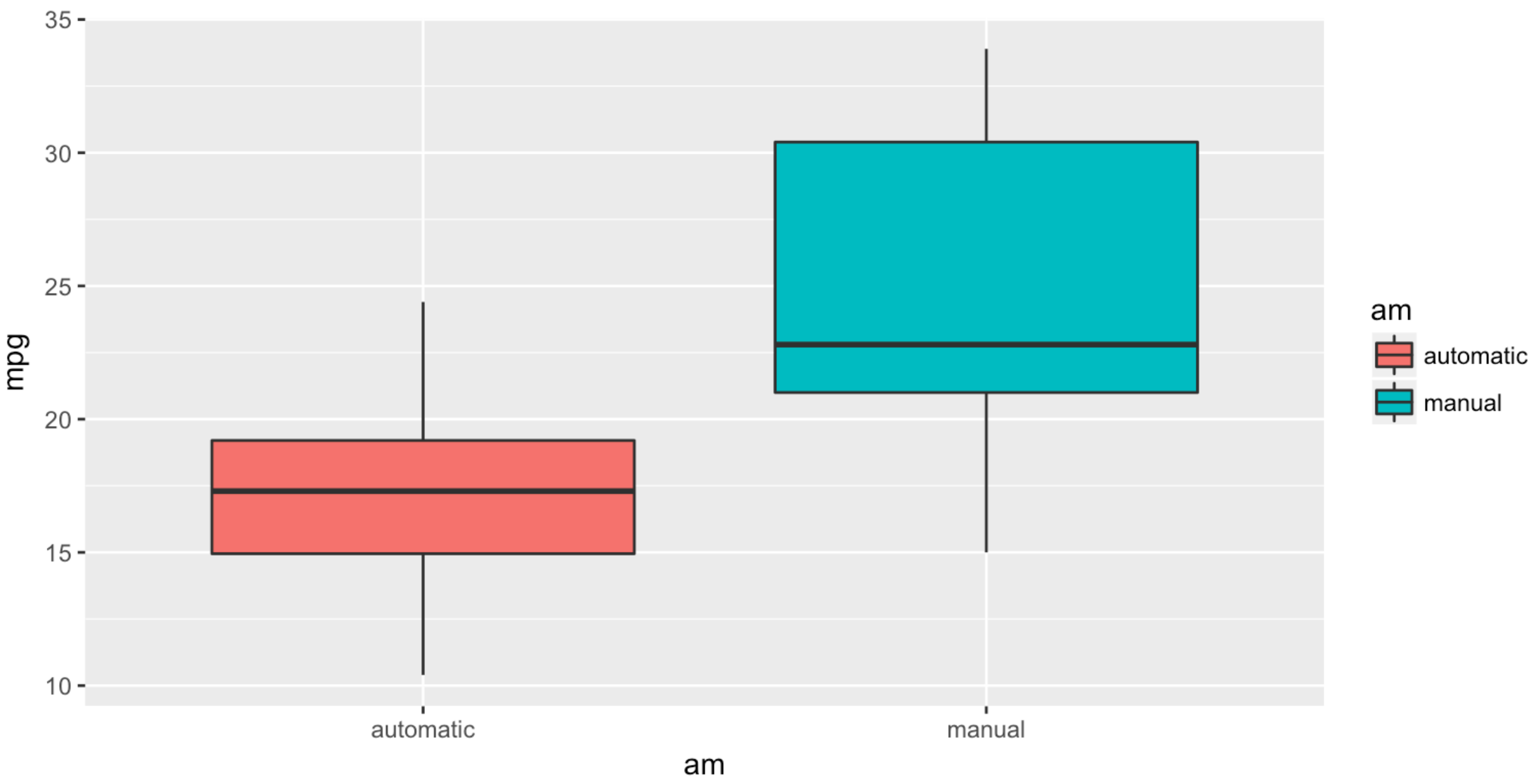
```
## [1] 32 11
```

```
# Result 2
head(df)
```

```
##                    mpg cyl disp  hp drat    wt  qsec vs        am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  V    manual    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  V    manual    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  S    manual    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  S automatic    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  V automatic    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  S automatic    3    1
```

Let's take a look at the realtionship between the two parameters which we are intereseted in.

```
plot1 <- ggplot(df, aes(am, mpg))
plot1 + geom_boxplot(aes(fill = am))
```



The above plot shows clearly that the manual transmissions have higher mpg's there could be a bias in the dataset that we are overlooking. Before creating a model we should look at which paramters to include besides 'am'. So we look at all correlations of parameters and take only those higher then the 'am' correlation.

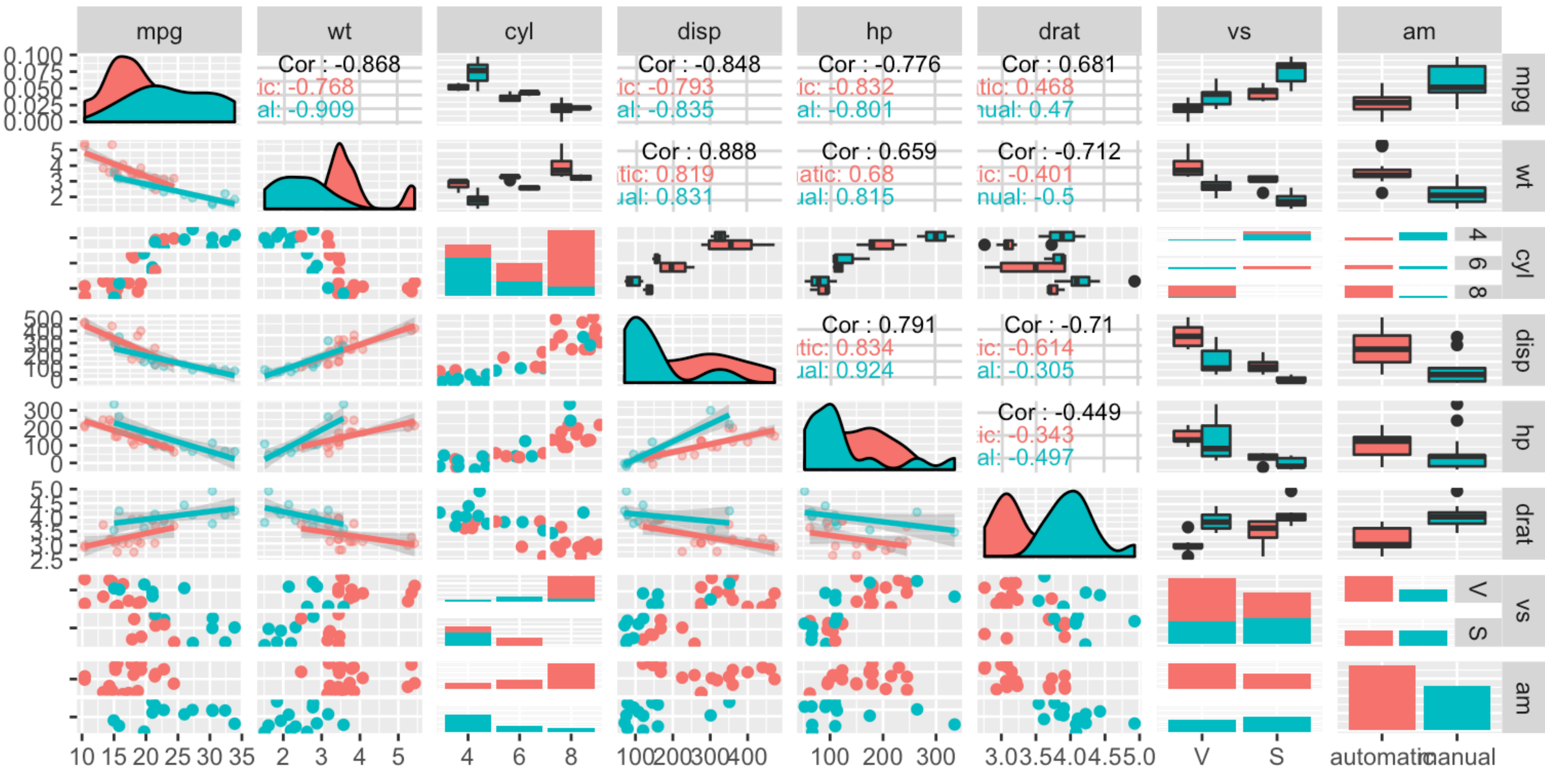
```
# Result 3
cors <- cor(mtcars$mpg, mtcars)
orderedCors <- cors[,order(-abs(cors[1,]))]
orderedCors
```

```
##         mpg         wt        cyl       disp         hp       drat         vs
am       carb
##  1.0000000 -0.8676594 -0.8521620 -0.8475514 -0.7761684  0.6811719  0.6640389  0.59
98324 -0.5509251
##       gear       qsec
##  0.4802848  0.4186840
```

```
# Result 4
amPos <- which(names(orderedCors)=="am")
subsetColumns <- names(orderedCors)[1:amPos]
subsetColumns
```

```
## [1] "mpg"  "wt"   "cyl"  "disp" "hp"   "drat" "vs"   "am"
```

```
df[,subsetColumns] %>%
   ggpairs(
     mapping = ggplot2::aes(color = am),
     upper = list(continuous = wrap("cor", size = 3)),
     lower = list(continuous = wrap("smooth", alpha=0.4, size=1), combo = wrap("dot")
)
   )
```



Model selection

We have seen that mpg has many other (stronger) correlations than just ‘am’ we can guess that a model predicting the mpg solely on this parameter will not be the most accurate model. Let’s check this out.

First we start with the basic model

```
# Result 5
fit1 <- lm(mpg ~ am, df)
summary(fit1)
```

```
##
## Call:
## lm(formula = mpg ~ am, data = df)
##
## 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
```

The p-values are actually quite low, the R-squared is problematic however. Now go to the other side of the spectrum by fitting all parameters of mtcars.

```
# Result 6
fit2 <- lm(mpg ~ ., df)
summary(fit2)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2015 -1.2319  0.1033  1.1953  4.3085
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 15.09262   17.13627   0.881   0.3895
## cyl6        -1.19940    2.38736  -0.502   0.6212
## cyl8         3.05492    4.82987   0.633   0.5346
## disp         0.01257    0.01774   0.708   0.4873
## hp          -0.05712    0.03175  -1.799   0.0879 .
## drat         0.73577    1.98461   0.371   0.7149
## wt          -3.54512    1.90895  -1.857   0.0789 .
## qsec         0.76801    0.75222   1.021   0.3201
## vsS          2.48849    2.54015   0.980   0.3396
## ammanual     3.34736    2.28948   1.462   0.1601
## gear4       -0.99922    2.94658  -0.339   0.7382
## gear5        1.06455    3.02730   0.352   0.7290
## carb         0.78703    1.03599   0.760   0.4568
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.616 on 19 degrees of freedom
## Multiple R-squared:  0.8845, Adjusted R-squared:  0.8116
## F-statistic: 12.13 on 12 and 19 DF,  p-value: 1.764e-06
```

The R-squared has improved, but the p-values hardly show any significance anymore. Perhaps this is due to overfitting. We now have to meet somewhere in the middle. Let’s iterate using the step method.

```
# Result 7
fit <- step(fit2, direction="both",trace=FALSE)
summary(fit)
```

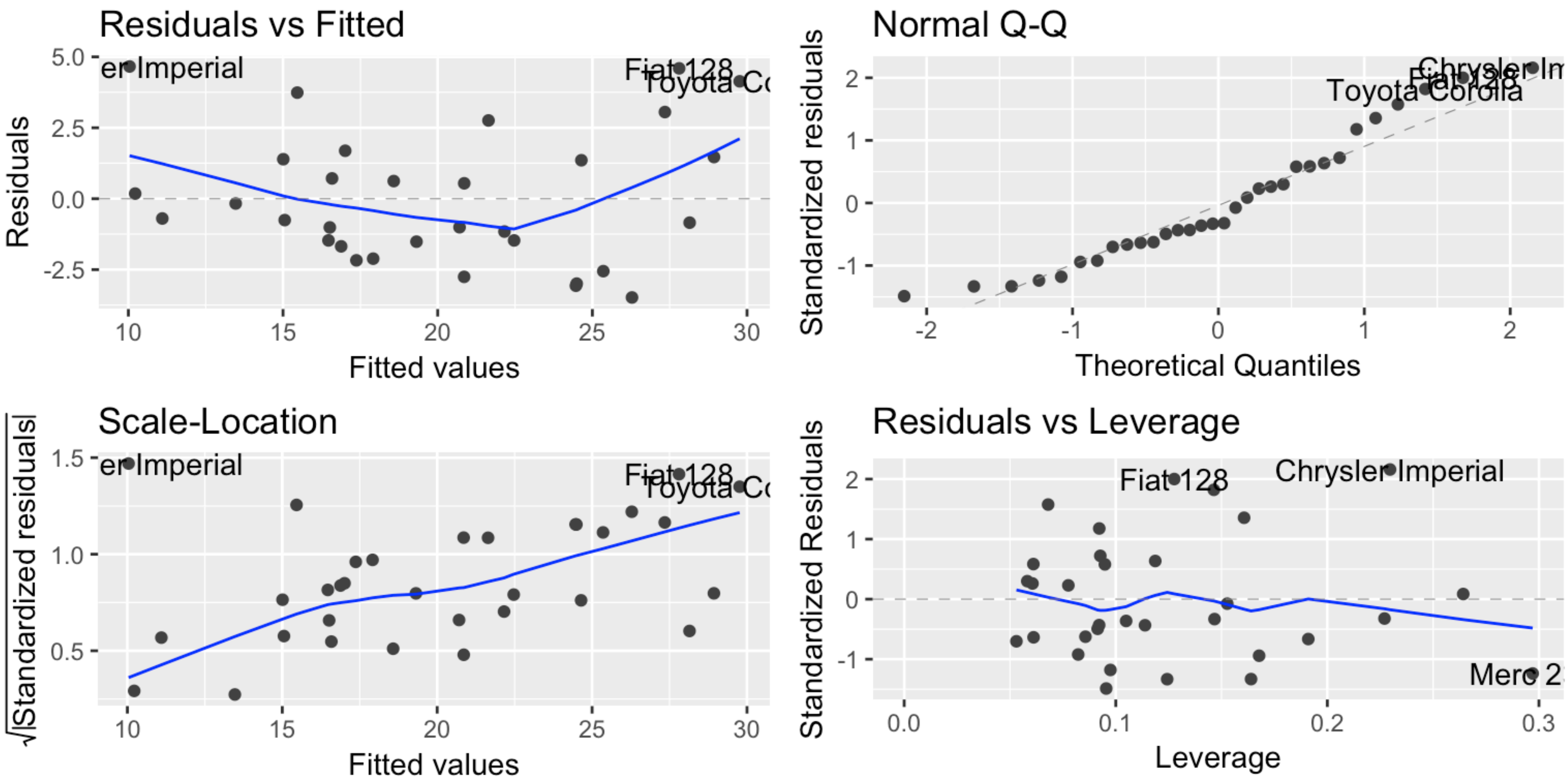
```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.6178     6.9596   1.382 0.177915
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## ammanual      2.9358     1.4109   2.081 0.046716 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.459 on 28 degrees of freedom
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8336
## F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

Model examination

The resulting best model `mpg ~ wt + qsec + am` is actually dependant on the transmission (am), but also weight (wt) and 1/4 mile time (qsec). All have significant p-values. The R-squared is pretty good to (0.85)

Now let's look (amongst others) at the Residuals vs Fitted

```
autoplot(fit)
```



The ‘Normal Q-Q’ plot looks ok, but the ‘Residuals vs Fitted’ and ‘Scale-Location’ both show worrysome trends.

Conclusion

The question “Is an automatic or manual transmission better for MPG” can be answered because all models (#Result 5, #Result 6 and #Result 7) show that, holding all other paramters constant, manual transmission will increase your MPG.

The question “Quantify the MPG difference between automatic and manual transmissions” is harder to answer.

Based on the ‘fit’ (#Result 7) model `mpg ~ wt + qsec + am` we could conclude that (with a $p < 0.05$ confidence) cars with manual transmission have 2.9358 (say 3) more miles per gallon than automatic transmissions. The model seems clean with a $p < 0.05$ and R squared of 0.85

The residuals vs fitted chart however warns us that there is something missing in our model. The real problem I think is that we only have 32 observations to train on (#Res1) and that observations hardly have overlap on the parameters ‘wt’ and ‘qsec’ (amongst others) if we look at the diagonal in the matrix chart

Although the conclusion of ca. 3 mpg better performance on manual transmissions seems feasible, I cannot with confidence conclude that this model will fit all future observations.