Motor Trend Transmission Analysis

Emanuele Di Saverio

19 June 2015

Executive Summary

Analyzing the data by the Motor Trend magazine it's possible to design a linear model that supports the hypotesis that there is **positive effect** of ~1.9 mpg (Miles per Gallon) on the fuel consumption that is **statistically significantly** attributable to the **different transmission type** - automatic or manual.

In more technical terms, we aim to esimate the *coefficient* β_1, am which is the **amount of mpg change** that I could resonably expect when upgrading from automatic to manual gearshifting, leaving all other regressors constant.

Exploratory Analysis

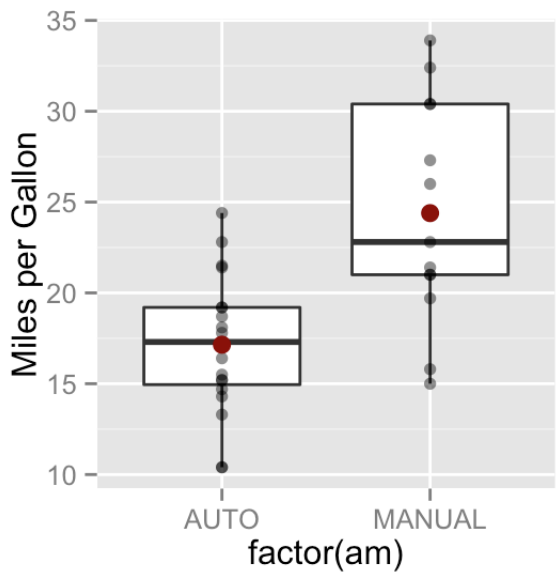
While the mtcars data is already in good form, we start by doing some preliminary preprocessing to get an idea of the data and format values in a readable way

```
library(car); library(ggplot2); data(mtcars); correlationMatrix <- cor(mtcars);
vifs <- vif(lm(mpg ~ ., data = mtcars))
options(width=120); library(dplyr, warn.conflicts = F)
names <- as.data.frame(row.names(mtcars)); colnames(names) <- c("name");
mtcars <- mutate(bind_cols(names, mtcars), am = factor(am, labels =c("AUTO","MANUAL")))
mtcars <- mutate(mtcars, vs = factor(vs, labels =c("VEE","STRAIGHT")))
```

This allows us to have proper factor variables and car name in the data frame (instead of named rows).

Simple direct plot shows that there is some difference in the mpg among the two transmission modes.

```
g <- ggplot(mtcars, aes(x=factor(am), y=mpg))  + geom_boxplot() + geom_point(alpha = 0.5)
g <- g + stat_summary(fun.y = mean, geom="point",colour="darkred", size=3)
print(g + ylab("Miles per Gallon"))
```



But this difference may as well be correlated to the other factors involved, and that's what we're going to investigate.

Model Selection

We're going to perform model selection by selecting different sets of regressors, trying to increase the amount of variance predicted by the model without increasing the variance of the estimated coefficients too much.

To boostrap the model selection beyond the am regressor (which is part of the required questions to be answered), we can look at the correlation mpg vector.

```
correlationMatrix["mpg",-1]
```

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

we can see that the regressor more correlated with the outcome are cyl, disp and wt. But these variables are also strongly correlated among themselves, so they are probably redundant to be included all in a linear model.

This is also hinted by the Variance Inflaction Factors which are higher in value

```
vifs
```

```
##       cyl      disp        hp      drat        wt      qsec        vs        am      ge
ar      carb
## 15.373833 21.620241  9.832037  3.374620 15.164887  7.527958  4.965873  4.648487  5.3574
52  7.908747
```

Let's first fit a baseline model, which ignores other regressors

```
fitam <- lm(mpg ~ am, data = mtcars)
```

Every regressor added will increase the R-squared score of the model, but the more the new regressor is linearly dependent form the ones already in the model (collinear), the less effective it will be and it will just increse the variance of the estimated factors.

We can try to include the terms more correlated to our outcome:

```
fitam_wt <- lm(mpg ~ am + wt, data = mtcars)
fitam_wt_cyl <- update(fitam_wt, mpg ~ am + wt + cyl)
fitam_wt_cyl_hp <- update(fitam_wt_cyl, mpg ~ am + wt + cyl + hp)
#P-values of fitam VS fitam_wt VS fitam_wt_cyl
anova(fitam, fitam_wt, fitam_wt_cyl, fitam_wt_cyl_hp)["Pr(>F)"]
```

```
##      Pr(>F)
## 1
## 2 < 2.2e-16 ***
## 3  0.000916 ***
## 4  0.078553 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
vifs['hp']
```

```
##       hp
## 9.832037
```

Including hp does not make the model significantly different at $\alpha = 0.05$ while still increasing variance ten-fold.

Another possible strategy would be starting from the regressors less correlated to am, carb and qsec:

```
fitam_wt_carb_qsec <- update(fitam_wt, mpg ~ am + wt + carb + qsec)
anova(fitam, fitam_wt, fitam_wt_carb_qsec)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + wt
## Model 3: mpg ~ am + wt + carb + qsec
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     29 278.32  1    442.58 74.1059 3.187e-09 ***
## 3     27 161.25  2    117.07  9.8012 0.0006308 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The anova tests indicate that is **appropriate to use a model** including:

- Transmission
- Weight
- Number of Carburetors
- Acceleration time

```
summary(fitam_wt_carb_qsec)$coefficients
```

```
##               Estimate Std. Error   t value    Pr(>|t|)
## (Intercept) 12.8971970  7.4724922  1.725957 0.095783676
## amMANUAL     3.5113918  1.4874727  2.360643 0.025720752
## wt          -3.4343216  0.8200199 -4.188095 0.000268619
## carb        -0.4886034  0.4212171 -1.159980 0.256211958
## qsec         1.0191298  0.3377635  3.017288 0.005507044
```

This allows us to have accurate estimation of the am **coefficient**, which is the amount of mpg change that I could resonably expect when upgrading from automatic to manual gearshifting, leaving all other regressors constant.

The adjusted R-squared score for the model is also very good:

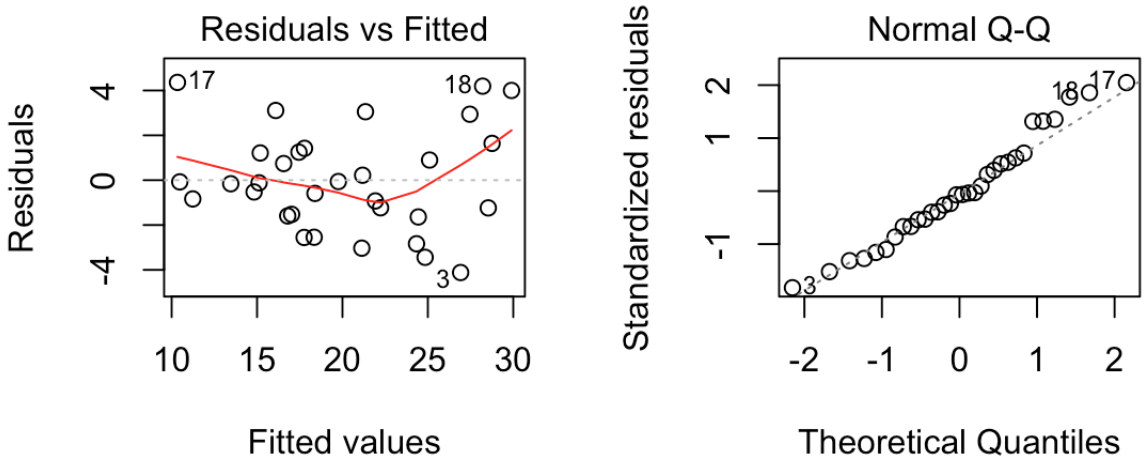
```
summary(fitam_wt_carb_qsec)$adj.r.squared
```

```
## [1] 0.8355852
```

Diagnostics and Outliers

We can inspect the model by plotting the residuals and Quantile plots, which show a good fit between the model and the data.

```
par(mfrow=c(1,2))
plot(fitam_wt_carb_qsec, which=1)
plot(fitam_wt_carb_qsec, which=2)
```



The plot highligh points 3, 17 and 18 as possible outliers. 17 and 18 are Chrysler Imperial, a very powerful car that has surprisingly low mpg, and Fiat 128, and italian car which may not fit the intended audience of Motor Trend magazine.

It may be safe to delete those two points, but we decide not to since the data set is very small and the point itself are still quite close to the rest.

Hypotheses Testing

Using our model, fitted over the clean data, we associate statistical uncertainty to our findings via hypotheses testing:

H_0 There is no difference in fuel consumption between and automatic and manual gearshifting

H_a There is statistically significant difference in fuel consumption between and automatic and manual gearshifting

```
coeff <- summary(fitam_wt_carb_qsec)$coefficients
delta <- coeff[2,1] + c(-1,1) * qt(.975,df = fitam_wt_carb_qsec$df) * coeff[2,2]
delta
```

```
## [1] 0.459350 6.563434
```

Given that the confidence interval doesn't include 0, we can reject H_0 at $\alpha = 0.05$ and say with **95% percent confidence** that the using a manual transmission car over an automatic **improves** the fuel efficiency by

```
coeff[2,1]
```

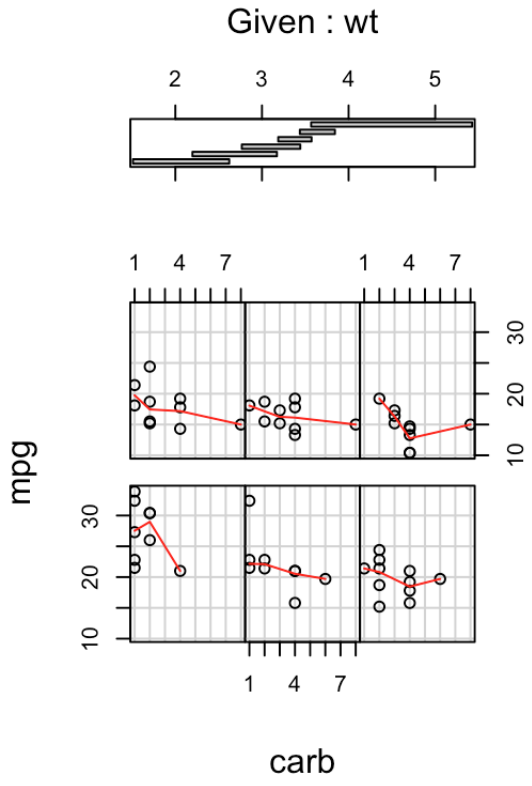
```
## [1] 3.511392
```

miles per gallon.

Notes

In this analysis we applied the strategy of isolating regressors and removing collinearity, to find a good simple model to answer the question. We didn't explore the possibility of variables interacting with each other, since the data set is very small and the VIF very large. An example of such analysis would have included some residuals conditioning plot to check for example if the carburators influence of mpg change at different weights

```
coplot(mpg ~ carb | wt,panel=panel.smooth, mtcars)
```



Since the data set is small, is difficult to extract trends from charts like this. But if such a relationship would be highlights, we would have included interaction element `carb * wt` in the formula instead the only linear `carb + wt`.