Analysing the Effect of Transmission Type on Fuel Consumption

Regression Models Course Project

Marco Pasin - 24 June 2017

Executive Summary

In this analysis we investigated the mtcars dataset, a collection of cars data extracted from the 1974 Motor Trend US magazine. In particular we were interested in answering the following two questions:

- Is an automatic or manual transmission better for fuel consumption (MPG variable)?
- Can we quantify the MPG difference between automatic and manual transmissions?

We noticed a statistically significant difference between automatic and manual transmission: cars endowed with manual transmissions tend to have higher levels of MPG (miles per gallon). When bulding a simple linear model with the variable `am` as the only one predictor, manual transmission cars present about 7.2 miles per gallon more than automated transmission cars. However when we included other variables into our model, the effect of transmission seem to be much lower, and manual transmission get just around 2 miles per gallon more. We finally concluded that there are more important variables in the dataset that can help predicting fuel consumption, such as weight and gross horsepower instead focusing only on car transmission type.

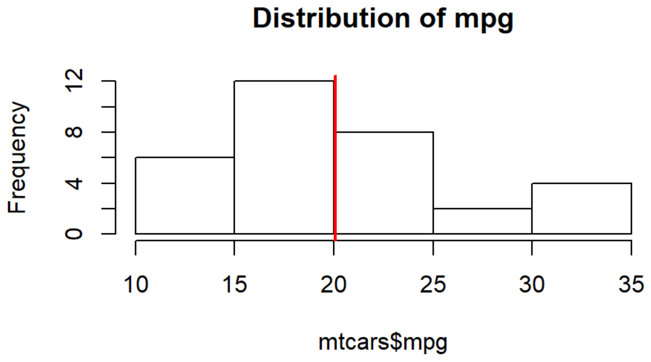
Explore the dataset

```
head(mtcars)
```

```
##                    mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

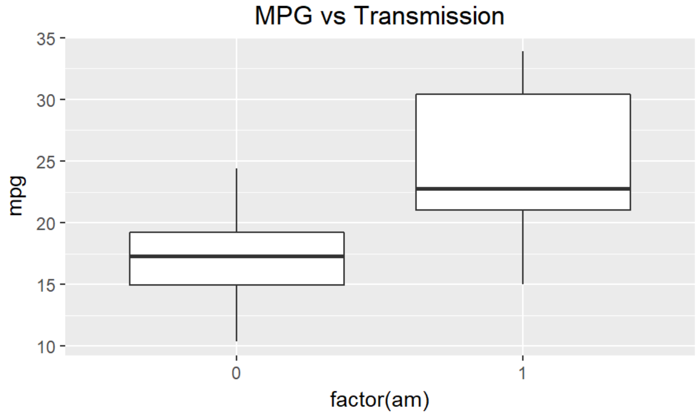
In the histogram below we can see how our outcome variable `mpg` it's distributed. Most of cars in the dataset have a `mpg` consumption between 15-20. Mean (red vertical line) is around 20 mpg.

```
hist(mtcars$mpg, main = "Distribution of mpg")
abline(v=mean(mtcars$mpg), col="red",lwd=2)
```



We can also visualize with a boxplot the relationship between transmission type and miles per gallon. Data about transmission are stored in the variable `am` where 0 indicates automatic transmission while 1 indicates that it is manual (in R type `?mtcars` for more details).

```
ggplot(mtcars, aes(x=factor(am),y=mpg)) + geom_boxplot() +
    ggtitle("MPG vs Transmission") +
    theme(plot.title = element_text(hjust = 0.5))
```



Cars with automated transmission seem to have lower levels of mpg, that is a lower fuel efficiency than manual transmission cars do. However this pattern might happen by random chance, hence we have to perform a statistical test. To check if the two groups have different means we can use the *Student t-test*.

We set the null hypothesis H0 as there is no difference between mps of manual and automatic transmission (no relationship between `mpg ~ am`). On the other hand, the alternative hypothesis H1 states that there is a relationship between transmission type and mpg.

```
t.test(mpg~am,paired=FALSE,var.equal=TRUE,data=mtcars)
```

```
##
##  Two Sample t-test
##
## data:  mpg by am
## t = -4.1061, df = 30, p-value = 0.000285
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -10.84837  -3.64151
## sample estimates:
## mean in group 0 mean in group 1
##        17.14737        24.39231
```

The test looks statistically significant since the p-value (the probability that the resulting difference between the two groups happened by chance) is much lower than 0.05, our alpha significance level. Hence we can say that there is a statistically significant difference between automatic and manual transmission. Cars endowed with automatic transmissions tend to have lower levels of mpg (keeping other variables fixed).

According to our 95% interval, we can be quite confident that the true difference in mpg is something between 3.6 and 10.8.

Simple Linear Regression Model

On the basis of what we found earlier through the t-test, we can investigate the relationship between mpg and `am` further by fitting a simple linear regression model. This simple model will try to explain mpg (dependent variable) as a function of only `am` (independent variable or predictor).
This will allows also to quantify the MPG difference between automatic and manual transmission.

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

```
##
## Call:
## lm(formula = mpg ~ factor(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 ***
## factor(am)1    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
```

According to p-value, again the relationship seems to be statistically significant and the output shows that the average mpg for cars with manual transmission is about 7.2 mpg higher than automatic (the mean mpg for automatic transmission is 17.14). That's the difference we were looking to quantify.

However the R squared value is 0.36 which means that **this simple model only explains only about 36% of the total variance in the data**. Modeling with just one predictor might not be sufficient to explaining our response variable. In the following section we will try to come up with a more complete model by including other important variables for predicting mpg. And quantify with more precision the mpg difference due to `am` .

Multivariable Linear Regression Model

In order to identify other key variables that can explain mpg, a commong methodogy is to run a correlation test using the co.test function in R.

In the appendix we can see that the variables `wt` , `cyl` , `disp` and `hp` present the strongest correlation with mpg. Therefore it might make sense to try multivariable regression models including some of these variables.

However, some of these potential predictors might be correlated among eahch others (i.e. **cofounders** in statistics). For example, it might be that cars with higher weight `wt` might have also higher volume `disp` . Or that cars that present higher weight `wt` it's because thay have a higher number of cylinders `cyl` (in the appendix we explored these relationships). It's important avoid including such cofounders in a regression model as they can compromise the stability of its estimates.

The correlation matrix available in the appendix, shows in particular that weight is highly correlated with both the number of cylinders and displacement. Cylinders and displacement have strong correlations too.

Hence, we decided not to include `cyl` and `disp` variables in our multivariable model, but only `am` , `wt` and `hp` as follows:

```
fit2<- lm(mpg~am + wt + hp, data = mtcars)
```

Since the simple linear model is **nested** into the multivariable (two models are considered nested when they both contain the same terms and one has at least one additional term, i.e. moel one is the reduced model and model 2 is the full model), we can use **Anova test** if there is a difference between them. In other words Anova will compare them and tell whether the full model contributes significant additional information for explaining our response variable.

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

```
## Analysis of Variance Table
## 
## Model 1: mpg ~ factor(am)
## Model 2: mpg ~ am + wt + hp
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)    
## 1     30 720.90                                  
## 2     28 180.29  2    540.61 41.979 3.745e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value is very small, therefore we can state the full multivariable model is significantly different from previous simple model. Let's get a summary of it:

```
summary(fit2)
```

```
## 
## Call:
## lm(formula = mpg ~ am + wt + hp, data = mtcars)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -3.4221 -1.7924 -0.3788  1.2249  5.5317 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 34.002875   2.642659  12.867 2.82e-13 ***
## am           2.083710   1.376420   1.514 0.141268    
## wt          -2.878575   0.904971  -3.181 0.003574 ** 
## hp          -0.037479   0.009605  -3.902 0.000546 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 2.538 on 28 degrees of freedom
## Multiple R-squared:  0.8399, Adjusted R-squared:  0.8227 
## F-statistic: 48.96 on 3 and 28 DF,  p-value: 2.908e-11
```

The multivariable model explain 84% of total variance in `am` **variable has now a smaller effect on mpg**: its coefficient shows that, on average, cars with manual transmission have **only 2.08 `mpg`** more than automatic (it was around 7 in the simple model).

However we can also notice that the p-value for `am` is much larger than 0.05 so not significant. This means that transmission type is redundant with one or both the other predictors and provides no additional information for prediction. Therefore we believe it makes sense to **remove `am` from predictors**. A model containing only `wt` and `hp` as predictors should be more stable as we can notice from the below summary.

```
best_fit<- lm(mpg ~ wt + hp, data = mtcars)
summary(best_fit)
```

```
##
## Call:
## lm(formula = mpg ~ wt + hp, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.941 -1.600 -0.182  1.050  5.854
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 37.22727    1.59879  23.285  < 2e-16 ***
## wt          -3.87783    0.63273  -6.129 1.12e-06 ***
## hp          -0.03177    0.00903  -3.519  0.00145 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.593 on 29 degrees of freedom
## Multiple R-squared:  0.8268, Adjusted R-squared:  0.8148
## F-statistic: 69.21 on 2 and 29 DF,  p-value: 9.109e-12
```

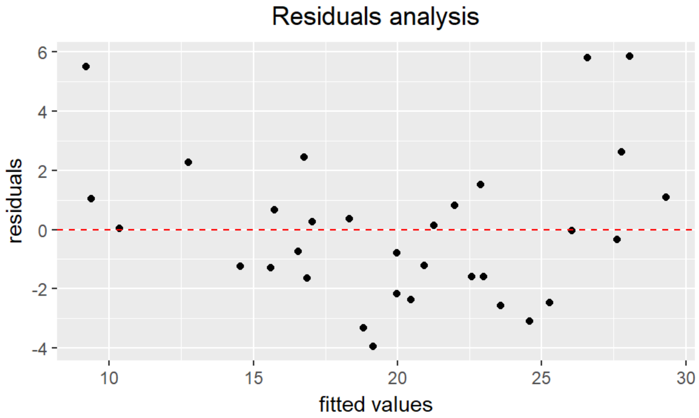
R squares is now 0.82: **this model explains around 82% of variation in the response variable mpg**. And both coefficients look statistical significant.

Residuals analysis

Residuals represent the variation left unexplained by a model. Analysing them is very useful for diagnosing potential issues in the model like for istance heteroskedasticity.

From the residual plot below we can see that points are randomly dispersed around the horizontal axis (no evidence of a pattern) which means that **our linear multivariable regression model seem to be appropriate for the data**.

```
res<-ggplot(best_fit, aes(.fitted, .resid))+geom_point()
res<-res+geom_hline(yintercept=0, col="red", linetype="dashed")
res + xlab("fitted values") + ylab("residuals") +
    ggtitle ("Residuals analysis") +
    theme(plot.title = element_text(hjust = 0.5))
```



Appendix

Exploratory Analysis

```
dim(mtcars)
```

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

```
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 ...
```

Correlation between variables

Below we will perform the test for each variable in the dataset versus mpg, using a for loop.For each test we get both a correlation estimate and a p-value which again will tell us the probability of the relationship happening by chance (the smaller the p-value and more significant the correlation is).

The resulting tables reports variables sorted by their correlation (from higher to smaller in absolute values) and with the respective p-value.

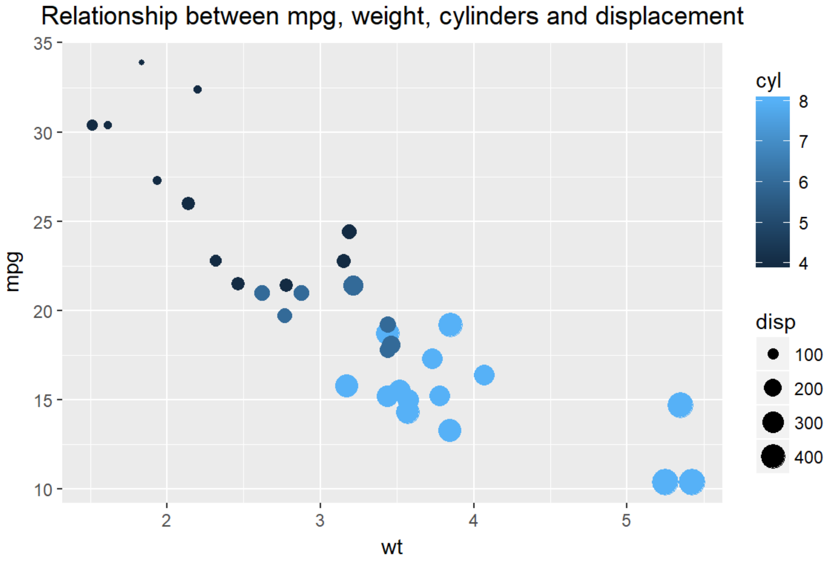
```
p<-2:11
df_cor<-data.frame()
for(i in p) {
  cor<- cor.test(mtcars$mpg,mtcars[, i])
  correlation<-cor$estimate
  variable<-names(mtcars[i])
  p_value<-cor$p.value
  df_tmp<-data.frame(variable,correlation,p_value)
  df_cor<-rbind(df_cor,df_tmp)
  df_cor<-arrange(df_cor,desc(abs(correlation)))
}

kable(df_cor)
```

variable	correlation	p_value
wt	-0.8676594	0.0000000
cyl	-0.8521620	0.0000000
disp	-0.8475514	0.0000000
hp	-0.7761684	0.0000002
drat	0.6811719	0.0000178
vs	0.6640389	0.0000342
am	0.5998324	0.0002850
carb	-0.5509251	0.0010844
gear	0.4802848	0.0054009
qsec	0.4186840	0.0170820

The code below shows how mpg is associated with some of the most correlated variables in the dataset.

```
library(ggplot2)
ggplot(mtcars, aes(x=wt,y=mpg, col=cyl, size=disp)) + geom_point() +
    ggtitle ("Relationship between mpg, weight, cylinders and displacement") +
    theme(plot.title = element_text(hjust = 0.5))
```



Correlation matrix to spot cofounders.

```
kable(cor(mtcars))
```

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	car
mpg	1.0000000	-0.8521620	-0.8475514	-0.7761684	0.6811719	-0.8676594	0.4186840	0.6640389	0.5998324	0.4802848	-0.550925
cyl	-0.8521620	1.0000000	0.9020329	0.8324475	-0.6999381	0.7824958	-0.5912421	-0.8108118	-0.5226070	-0.4926866	0.526988
disp	-0.8475514	0.9020329	1.0000000	0.7909486	-0.7102139	0.8879799	-0.4336979	-0.7104159	-0.5912270	-0.5555692	0.394976
hp	-0.7761684	0.8324475	0.7909486	1.0000000	-0.4487591	0.6587479	-0.7082234	-0.7230967	-0.2432043	-0.1257043	0.749812
drat	0.6811719	-0.6999381	-0.7102139	-0.4487591	1.0000000	-0.7124406	0.0912048	0.4402785	0.7127111	0.6996101	-0.090789
wt	-0.8676594	0.7824958	0.8879799	0.6587479	-0.7124406	1.0000000	-0.1747159	-0.5549157	-0.6924953	-0.5832870	0.427605
qsec	0.4186840	-0.5912421	-0.4336979	-0.7082234	0.0912048	-0.1747159	1.0000000	0.7445354	-0.2298609	-0.2126822	-0.656249
vs	0.6640389	-0.8108118	-0.7104159	-0.7230967	0.4402785	-0.5549157	0.7445354	1.0000000	0.1683451	0.2060233	-0.569607
am	0.5998324	-0.5226070	-0.5912270	-0.2432043	0.7127111	-0.6924953	-0.2298609	0.1683451	1.0000000	0.7940588	0.057534
gear	0.4802848	-0.4926866	-0.5555692	-0.1257043	0.6996101	-0.5832870	-0.2126822	0.2060233	0.7940588	1.0000000	0.274072
carb	-0.5509251	0.5269883	0.3949769	0.7498125	-0.0907898	0.4276059	-0.6562492	-0.5696071	0.0575344	0.2740728	1.000000