Project

Thursday, December 18, 2014

Executive Summary

Let's take a look and investigate the dataset mtcars. According to the R description, 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).

This is a data frame with 32 observations on 11 variables. See variables description via ?mtcars in R.

We have to analyze this data in the following context: "You work for Motor Trend, a magazine about the automobile industry. Looking at a data set of a collection of cars, they are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome)."

The following questions must be answered from this data:

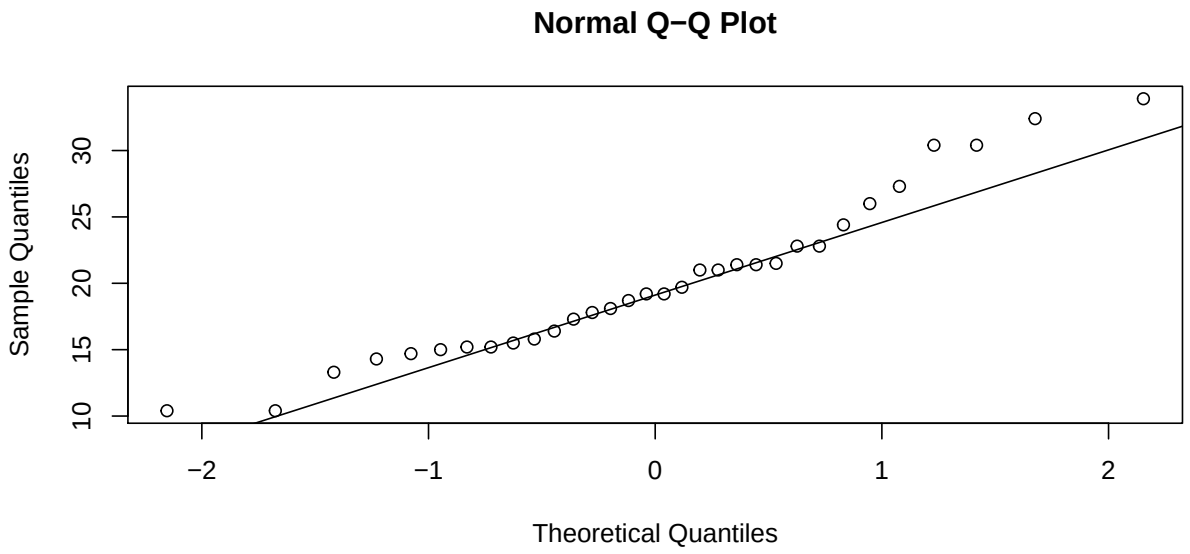
-Is an automatic or manual transmission better for MPG?

-Quantify the MPG difference between automatic and manual transmissions?

Analisys

Let's investigate the distribution and see Quantile-Quantile plot to verify normality:

```
qqnorm(mtcars$mpg)
qqline(mtcars$mpg) ## plots a line that goes from 0.25 to 0.75 quantile (2nd and 4th quantile)
```



In the graph above we can see that there are many points away from the 1st and 3rd quantile, suggesting that this distribution should not be normal. We need to use other methods to investigate this data.

Let's see the means:

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.1.2
```

```
##
## Attaching package: 'dplyr'
##
## The following object is masked from 'package:stats':
##
##     filter
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
summarise(group_by(mtcars, am), mean(mpg), sd(mpg), qty=n())
```

```
## Source: local data frame [2 x 4]
##
##   am mean(mpg)  sd(mpg) qty
## 1  0  17.14737 3.833966  19
## 2  1  24.39231 6.166504  13
```

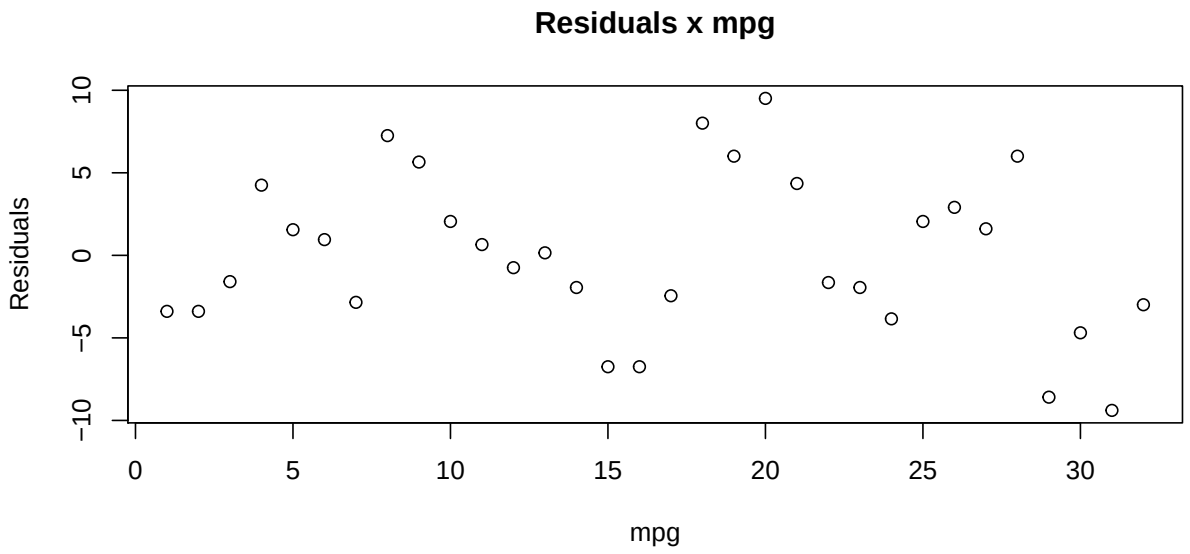
At first sight, the mpg mean for automatics are lower than those that are manual. So automatics seem to be more inefficient but we cannot conclude that yet.

```
summary(lm(data=mtcars, formula=mpg ~ 0 + factor(am)))
```

```
##
## Call:
## lm(formula = mpg ~ 0 + 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|)
## factor(am)0   17.147      1.125   15.25 1.13e-15 ***
## factor(am)1   24.392      1.360   17.94  < 2e-16 ***
## ---
## 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:  0.9487, Adjusted R-squared:  0.9452
## F-statistic: 277.2 on 2 and 30 DF,  p-value: < 2.2e-16
```

If we try to create a linear regression model like above, we can see above that the coefficients are the means of both automatic and manual measured mpg. We can see that this regression can explain about 94.87% of the variance of the data. This is quite look since we evaluated a single vaiable (am) of the 10 possible (not considering mpg, since it is the outcome).

```
plot(resid(lm(data=mtcars, formula=mpg ~ 0 + factor(am))),
     ylab="Residuals", xlab="mpg", main="Residuals x mpg")
```



The residuals are in the range [-10, 10], as we can see in the y-axis. There is a residual standard error of 4.90.

Now let's switch to the full model, evaluate all the 11 variables through a linear regression to see if there is improvement. Let's see what happens when the try a regression similar to the former, but using all of the variables:

```
lm_full<-lm(mpg~ 0 + . -am +factor(am), data=mtcars)
summary(lm_full)
```

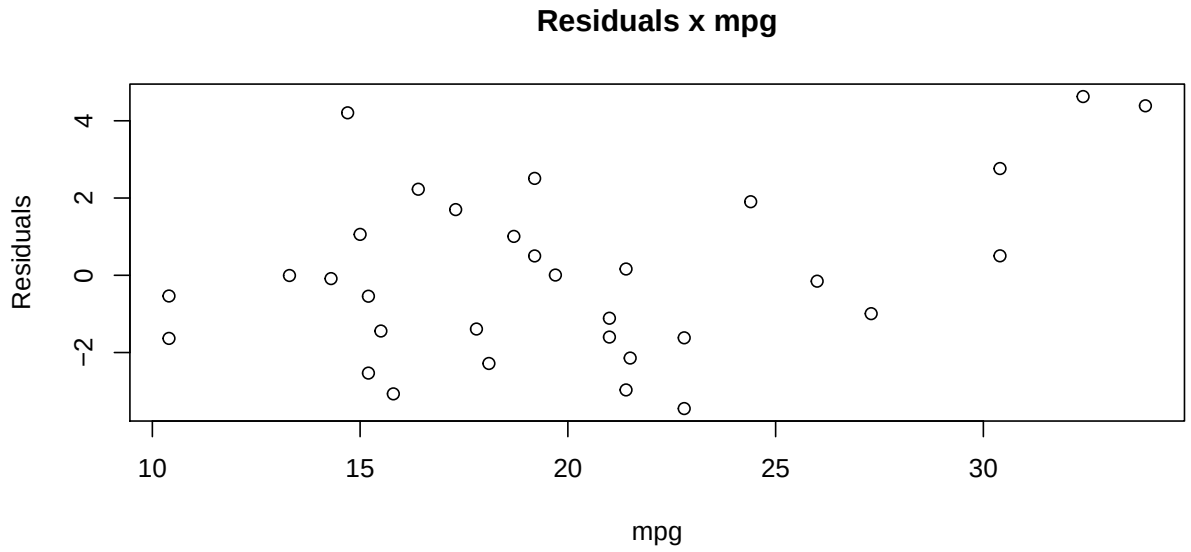
```
##
## Call:
## lm(formula = mpg ~ 0 + . - am + factor(am), data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4506 -1.6044 -0.1196  1.2193  4.6271
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## cyl         -0.11144    1.04502  -0.107   0.9161
## disp         0.01334    0.01786   0.747   0.4635
## hp          -0.02148    0.02177  -0.987   0.3350
## drat         0.78711    1.63537   0.481   0.6353
## wt          -3.71530    1.89441  -1.961   0.0633 .
## qsec         0.82104    0.73084   1.123   0.2739
## vs           0.31776    2.10451   0.151   0.8814
## gear         0.65541    1.49326   0.439   0.6652
## carb        -0.19942    0.82875  -0.241   0.8122
## factor(am)0 12.30337   18.71788   0.657   0.5181
## factor(am)1 14.82360   18.35265   0.808   0.4283
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.65 on 21 degrees of freedom
```

```
## Multiple R-squared:  0.9895, Adjusted R-squared:  0.984
## F-statistic: 179.8 on 11 and 21 DF,  p-value: < 2.2e-16
```

We have a R^2 of about 98,95% which is higher than 94.87% from the previous one. Let's stick with this one and investigate further.

The residuals now are below:

```
plot(mtcars$mpg, resid(lm(data=mtcars, formula=mpg~ 0 + . -am +factor(am))),
     ylab="Residuals", xlab="mpg", main="Residuals x mpg")
```



The new interval in the y-axis are about [-4,5] which is much lower than the previous one of [-10, 10]. This model tells us that an manual car can run in average more 2.52 miles per galon than a automatic car (this information is in am line in the summary above). There is a residual standard error of 2.65 (lower than the previous one of 4.90).

However, we have a very huge standard error in the slope am(0) and am(1) of about 18. Aside from that, if we compare F-Statistic value, we can verify that the former model is much more relevant that the last one. This tells us that the difference between the groups is more relevant when we use the simplfied model. So the variable transmission (am) alone is enough or is relevant enough to quantify the difference in mpg between the groups (we have a F-statistics of about 277 from the former against 180 from the latter). In addition to this, the last model has very high p-values for each of the variables, which means that we added too much noise to the model.

In order to elaborate a little more about the slopes of the am variable, let's check the 95% confidence interval of the slopes from the former model, since it is simplified and can explain the data like the latter model.

```
confint(lm(data=mtcars, formula=mpg ~ 0 + factor(am)))
```

```
##                 2.5 %   97.5 %
## factor(am)0 14.85062 19.44411
## factor(am)1 21.61568 27.16894
```

Conclusion

In the light of the analysis done so far, is the data about the transmission sufficient do quantify the difference between the groups and tell which one is better? If we take a look at the standard error of the am variable, we'll see a number about 18.717 (automatic) and 18.352 (manual), which are quite huge in the latter model. So the former model is much better to tell the difference between the groups than the latter one (the former has a standard error of about 1.125 for automatic and 1.360 for manuals), besides explaining about 95% of the variance of the data, which is very close to the 98% of the latter. Taking into the account the 95% confidence interval calculated above and considering the former model as best fit for the proposed analysis, we can state that manual cars are more efficient than automatic ones for 95% of the cases (our 95% confidence interval for the slopes are [14.85,19.444] for automatic and [21.615, 27.168] for manual cars). Using this information, we can quantify the difference: the interval [2.171, 12.318], which is the lowest and the highest difference of the 95% confidence interval found above, will quantify the difference between automatic and manual cars (the lowest difference will be about 2.171 mpg and the highest of about 12.318 mpg in 95% of the cases in favour of the manual cars [they are more efficient]).