Regression models project

RDSN

21 October 2015

Executive summary

In this report, we investigate the relationship between MPG (Miles per Gallon) and several variables within the dataset **mtcars**. 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). We will first show that using a manual transmission rather than an automatic one seems to lead to a larger MPG. Though, when we try to find a clear relationship between MPG and the transmission, we cannot just stop there and we must investigate further, considering a lot of other features that are at stake. The conclusion is, wether in the case of the relationship between MPG and the am variable alone, or with 2 other variables, the cars with manual transmission in this study have on average significantly higher MPG than the cars with automatic transmission.

Loading the **mtcars** data

```
data(mtcars)
```

1. Is an automatic or manual transmission better for MPG

First let's have a look at the data

```
head(mtcars)
```

```
str(mtcars)
```

By looking at the structure of the data, we see that all variables are numerical.

Now let's have a look at the relation between MPG and transmission.

- See Appendix - Figure 1 for the plot and code -

As we see on this plot, we can suppose that manual transmission (am =1) is associated with a larger MPG.

Let's fit a simple linear regression to check this assumption.

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

- See Appendix - Figure 2 for code results -

The coefficients of this model mean : - if the transmission is automatic, am = 0, the prediction is 17.147, which is the mean of MPG for am == 0; - if the transmission is manual, am = 1, then the prediction is 17.147 + 7.245 = 24.392 which is the mean of MPG for am == 1; This is a quite simple model.

Let's calculate a 95% confidence interval for Beta1.

```
m <- coef(summary(fit1))[2,1]
se <- coef(summary(fit1))[2,2]
m + c(-1,1)*qt(.975,30)*se # (n = 32 so n-2 = 30)
```

```
## [1]  3.64151 10.84837
```

The confidence interval does not include 0. p-Value for Beta1 is small (2e10-4 < 0.05). The confidence interval is positive. So we can reject the null hypothesis and so assume that the means of the 2 groups are significantly different at alpha = 0.05, and furthermore that the mean of the sample with am = 1 is likely to be larger than the mean of the sample with am = 0.

But, as we are aware of, MPG is not only a function of the type of transmission, but it depends on a lot of different features. We are going to investigate those features in the next question.

2. Quantify the MPG difference between automatic and manual transmissions

So, we have seen before that a manual transmission seems to lead to a larger MPG. So to model the relation between MPG and all of those features, we are going to create several models and compare them to each other.

Let's see the correlation between the features to choose features to include into a model.

```
cor(mtcars) # results not shown here
```

The variable the more correlated with MPG is `wt`. Then we are going to add the variable the less corelated with `wt` which is `qsec`. And then we add the variable of interest here which is `am`.

Let's build those models.

```
fit1 <- lm(mpg ~ wt, data=mtcars)
fit2 <- lm(mpg ~ wt + qsec, data=mtcars)
fit3 <- lm(mpg ~ wt + qsec + factor(am), data=mtcars)
anova(fit1,fit2, fit3)
```

- See Appendix - Figure 3 for code results -

As we can see with this comparison, the 2nd and the third models seem to be significant, with a p_value very small, which leads us to reject the null hypothesis, and so to suppose that these models lead to an improvement in comparison of the model 1.

Let's go further with the 3rd model.

```
summary(fit3)
```

- See Appendix - Figure 4 for code results -

All the variables are significant in this model. This summary shows that if `wt` and `qsec` are maintened constant, then a car with a manual transmission add 2.94 more MPG on average than cars with automatic transmission.

Let's calculate a 95% confidence interval for Beta3.

```
m <- coef(summary(fit3))[4,1]
se <- coef(summary(fit3))[4,2]
m + c(-1,1)*qt(.975,30)*se # (n = 32 so n-2 = 30)
```

```
## [1] 0.05438576 5.81728862
```

The confidence interval does not include 0. p-Value for Beta1 is small (4e10-2 < 0.05). The confidence interval is positive. So we can reject the null hypothesis and so assume that the means of the 2 groups are significantly different at alpha $= 0.05$, and furthermore that the mean of the sample with am $= 1$ is likely to be larger than the mean of the sample with am $= 0$.

Let's plot the fit3 to see the residuals

```
par(mfrow=c(2,2))
plot(fit3)
```

- See Appendix - Figure 5 for the plot and code -

As we can see from those plots :

- The Residual vs Fitted and the Scale-Location plots show that there is a slight curve, indicating a slight pattern (to be investigated). And several points seem to be outliers, exercing an influence over the curve, for example Toyota Corolla (row 20)
- the normal Q-Q plot indicates that the residuals tend to follow a normal distribution, so that the points lie on the line, except for the outliers at the top-right.
- Finally, the Residuals vs Leverage points out those outliers, but indicates that those outliers are within the confidence interval (so not really outliers).

Figure 1

```
library(ggplot2)
g <- ggplot(mtcars, aes(x = factor(am), y = mpg)) + geom_violin()
g <- g + ggtitle("MPG vs Transmisison (am)")
g
```

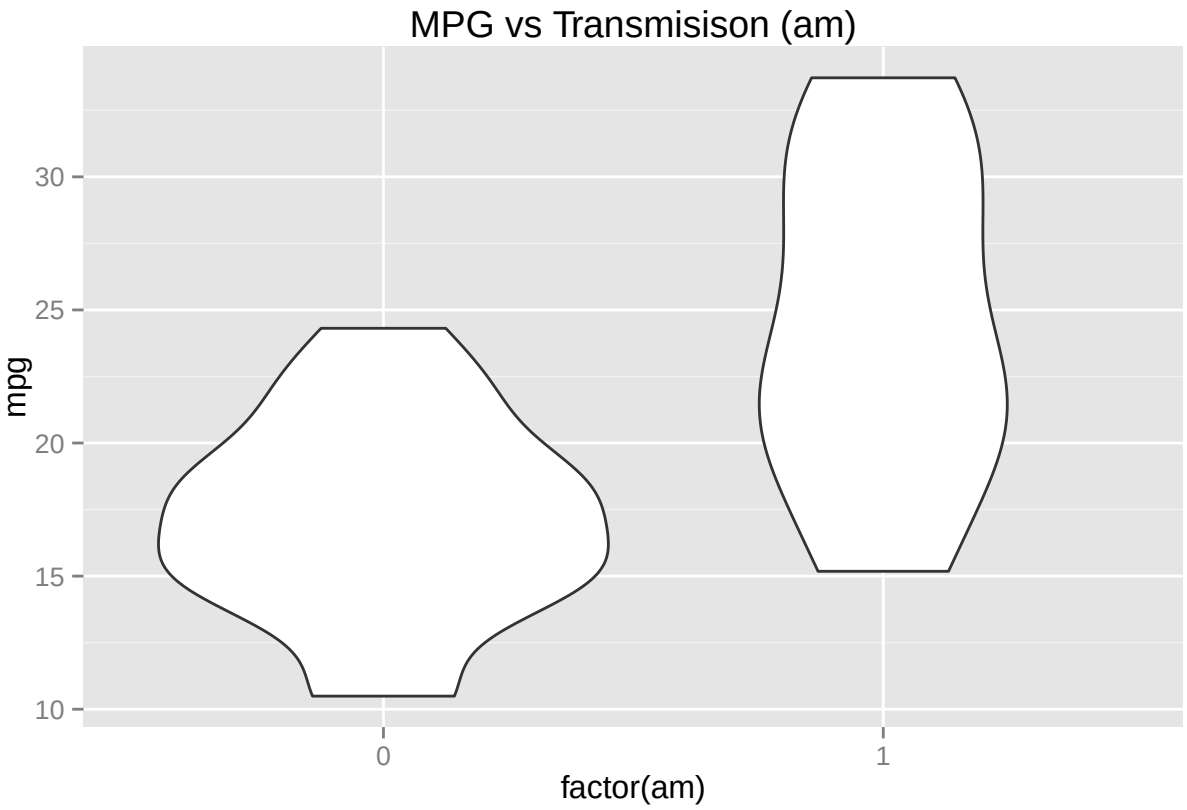

Figure 2

```
##
## Call:
## lm(formula = mpg ~ wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5432 -2.3647 -0.1252  1.4096  6.8727
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  37.2851     1.8776  19.858  < 2e-16 ***
## wt           -5.3445     0.5591  -9.559 1.29e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.046 on 30 degrees of freedom
## Multiple R-squared:  0.7528, Adjusted R-squared:  0.7446
## F-statistic: 91.38 on 1 and 30 DF,  p-value: 1.294e-10
```

Figure 3

```
## Analysis of Variance Table
##
## Model 1: mpg ~ wt
## Model 2: mpg ~ wt + qsec
## Model 3: mpg ~ wt + qsec + factor(am)
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 278.32
## 2     29 195.46  1    82.858 13.7048 0.0009286 ***
## 3     28 169.29  1    26.178  4.3298 0.0467155 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 4

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + factor(am), data = mtcars)
##
## 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 ***
## factor(am)1   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
```

Figure 5

```
par(mfrow=c(2,2))
plot(fit3)
```