Regression Models Course Project

Daniel Pont

2019-02-10

Executive Summary

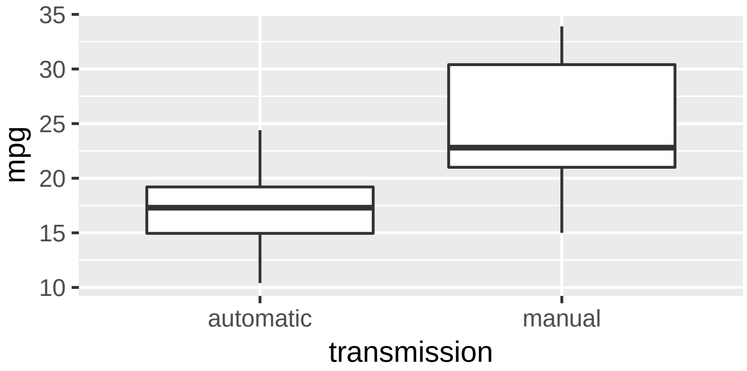
This report is the final product of the Regression Models Course Project. We try to answer the following questions about the Motor Trend Car Road Tests dataset :

- Is an automatic or manual transmission better for MPG ?
- What is the MPG difference quantification between automatic and manual transmissions ?

We'll show that a manual transmission is better than an automatic one. Quantitatively the factor by witch the MPG is multiplied when switching from an automatic to a manual transmission is in the interval [0.05,4.12] with a 85% confidence. So the type of transmission has an impact on the MPG value but the quantification is not obvious.

1) Exploratory Data Analysis

```r
mtcars$transmission <- as.factor(mtcars$am)
levels(mtcars$transmission) <- c("automatic","manual")
ggplot(mtcars, aes(x=transmission, y=mpg)) +
    geom_boxplot()
```



Manual transmisssion seems better than automatic for MPG. Now the question is how much ?

2) Simple Linear Regresion

```r
simple_fit<-lm(mpg~am,mtcars)
summary(simple_fit)$r.squared
```

```
## [1] 0.3597989
```

```r
summary(simple_fit)$coefficients
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## am           7.244939   1.764422  4.106127 2.850207e-04
```

The R-squared value is about 0.36 so the model explains only 36% of the variance which is quite low. We also notice that the slope value is 7.24. If we look up, the difference between auto vs manual is not a 7x factor. This indicates that our simple linear model doesn't fit the data very well. How can we improve it ?

3) Multiple Linear Regresion

Next we'll add variables that have an impact on MPG. Intutively weight (wt) and horsepower (hp) makes lot of sense. Let's check it :

```
multi_fit<-lm(mpg~wt+am+hp,mtcars)
summary(multi_fit)$r.squared
```

```
## [1] 0.8398903
```

```
summary(multi_fit)$coefficients
```

```
##                Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) 34.00287512 2.642659337 12.866916 2.824030e-13
## wt          -2.87857541 0.904970538 -3.180850 3.574031e-03
## am           2.08371013 1.376420152  1.513862 1.412682e-01
## hp          -0.03747873 0.009605422 -3.901830 5.464023e-04
```

Here the R-squared value is much closer to 1 (0.84) and the slope value (2.08 for mpg ~ am) much more realistic. However the standard error is high (1.38). Let's calculate the confidence interval :

```
confint(multi_fit,level=0.85)
```

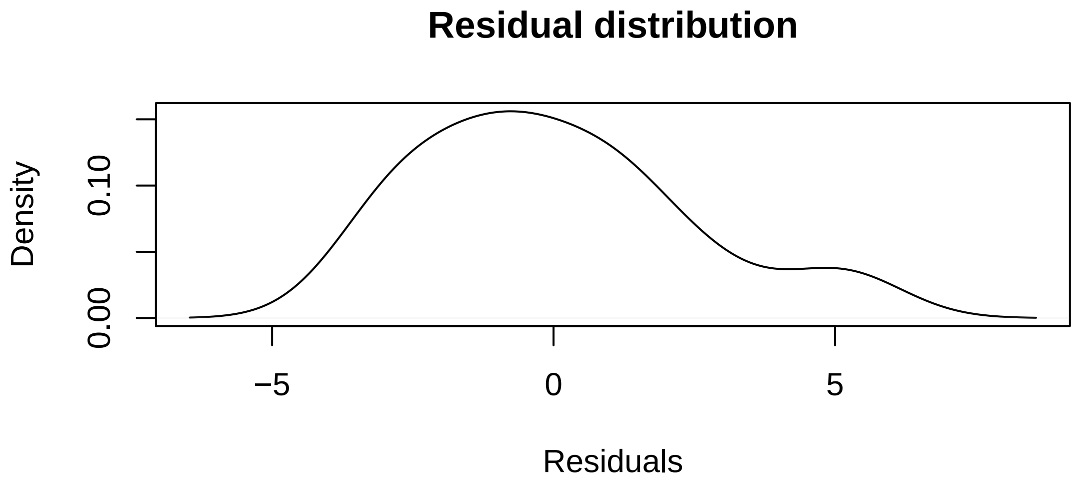
```
##                   7.5 %      92.5 %
## (Intercept) 30.09136037 37.91438987
## wt          -4.21806157 -1.53908925
## am           0.04641094  4.12100932
## hp          -0.05169613 -0.02326132
```

The 85% confidence interval is wide : [0.04641094, 4.12100932] and the lowest value is close to 0. So even if there seems to be an increase in MPG when using a manual transmission instead of an automatic one, it's difficult to quantify it. NB :

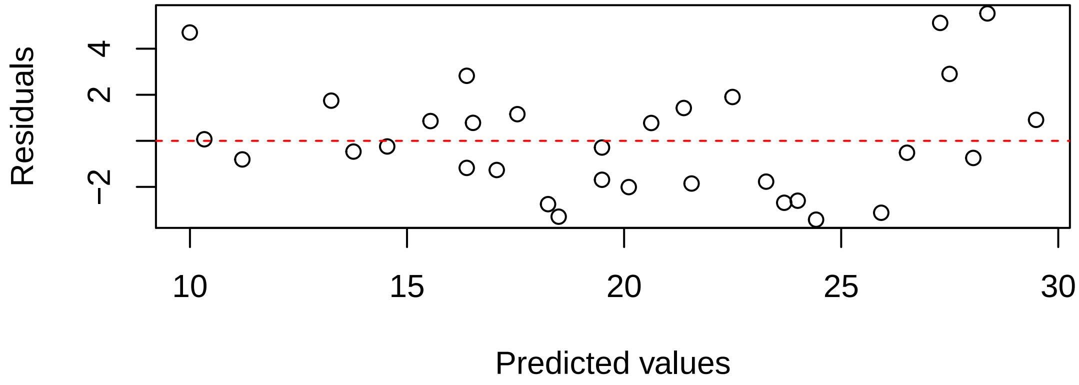
- with the use of a standard 95% confidence interval we would probably reject this effect as non significant.
- regarding residuals, the 1st plot in the annex shows that residuals distribution is approximately normal and centered on 0. So the normality assumption is valid. The 2nd plot above shows that heteroskedasticity doesn't seem to be an issue.

Annex : residual plots

```
resid <- residuals(multi_fit)
fitted <- fitted.values(multi_fit)
plot(density(resid), xlab = "Residuals", ylab = "Density", main = "Residual distribution")
```



```
plot(fitted, resid, xlab = "Predicted values", ylab = "Residuals")
abline(h = 0, col = "red", lty = "dashed")
```


Residuals
Predicted values