Regression Models Course Project

José Manuel Peñas

28 de julio de 2018

Contents

1 Analysis

MTCARS is the data set. . A new factor column *transmission* is added with *Automatic* and *manual* values instea}d of 0 and 1 of *am* for an easier manipulation later.

Looking at the boxplot at the appendix, it seems clear that Manual transmission cars do consumes less fuel than automatic ones. The quantity will be meassured by linear regression models.

1.1 Simple Model

First, the analysis creates a simple model that studies the relationship between mpg, the outcome, and transmission the predictor.

```
##
## Call:
## lm(formula = mpg ~ transmission, data = newcars)
##
## 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 ***
## transmissionManual   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
```

The estimate intercept represents the coefficient of Automatic cars (**17.15**) and the *transmissionManual* the difference between automatic and manual: **7.24**. The P-value is lower enough to discard the null hypothesis on both cases, so transmission has a relationship with gasoline consumption. As you can see on the plot generated on Appendix, the residuals plots and QQ plot are good and don't show abnormals patterns. The R-Squared indicates that the prediction of mpg only is influenced by 33-36 % by the transmission of the car. It is mandatory to check with with other variables in a multivariate model.

1.2 Multivariate Model

First, the variables to be included in the model have to be found.

```
##        mpg        cyl       disp         hp       drat         wt
##  1.0000000 -0.8521620 -0.8475514 -0.7761684  0.6811719 -0.8676594
##       qsec         vs         am       gear       carb
##  0.4186840  0.6640389  0.5998324  0.4802848 -0.5509251
```

The variables to include are the ones whose absolute value is bigger than **0.75**: *cyl*, *disp*,*hp*, *wt*. Now, the multivariate model can be generated and compared with Anova test.

```
##
## Call:
## lm(formula = mpg ~ cyl + disp + hp + wt + transmission - 1, data = newcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5952 -1.5864 -0.7157  1.2821  5.5725
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## cyl                   -1.10638    0.67636  -1.636  0.11393
## disp                   0.01226    0.01171   1.047  0.30472
## hp                    -0.02796    0.01392  -2.008  0.05510 .
## wt                    -3.30262    1.13364  -2.913  0.00726 **
## transmissionAutomatic 38.20280    3.66910  10.412 9.08e-11 ***
## transmissionManual    39.75929    2.92165  13.608 2.46e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.505 on 26 degrees of freedom
## Multiple R-squared:  0.9884, Adjusted R-squared:  0.9857
## F-statistic: 368.7 on 6 and 26 DF,  p-value: < 2.2e-16
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
30	720.8966	NA	NA	NA	NA
26	163.1199	4	557.7767	22.22629	0

It has bigger R-Squared **0.98** and if you check the plot of the model, the 4 graphics are correct with no abnomal patterns. The Anova test shows a very important decrease on the Residual Sum of Squares: **720** to **163**

2 Conclussions

2.1 Is an automatic or manual transmission better for MPG

So the boxplot as the simple and multivariates models show that the **manual transmission is better** for MPG.

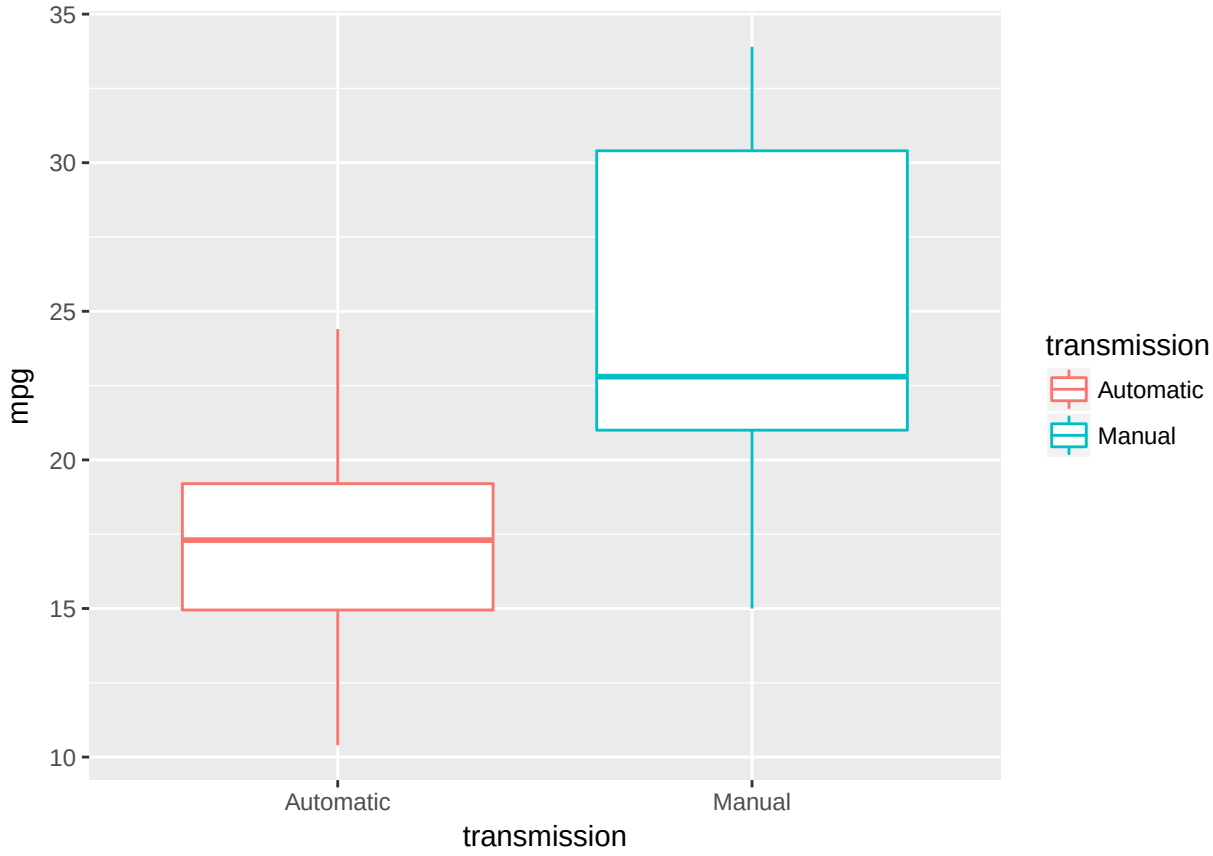
2.2 Quantify the MPG difference between automatic and manual transmissions

The multivariate model gives **1.55** the difference between automatic and manual transmissions.

3 Appendix Figures

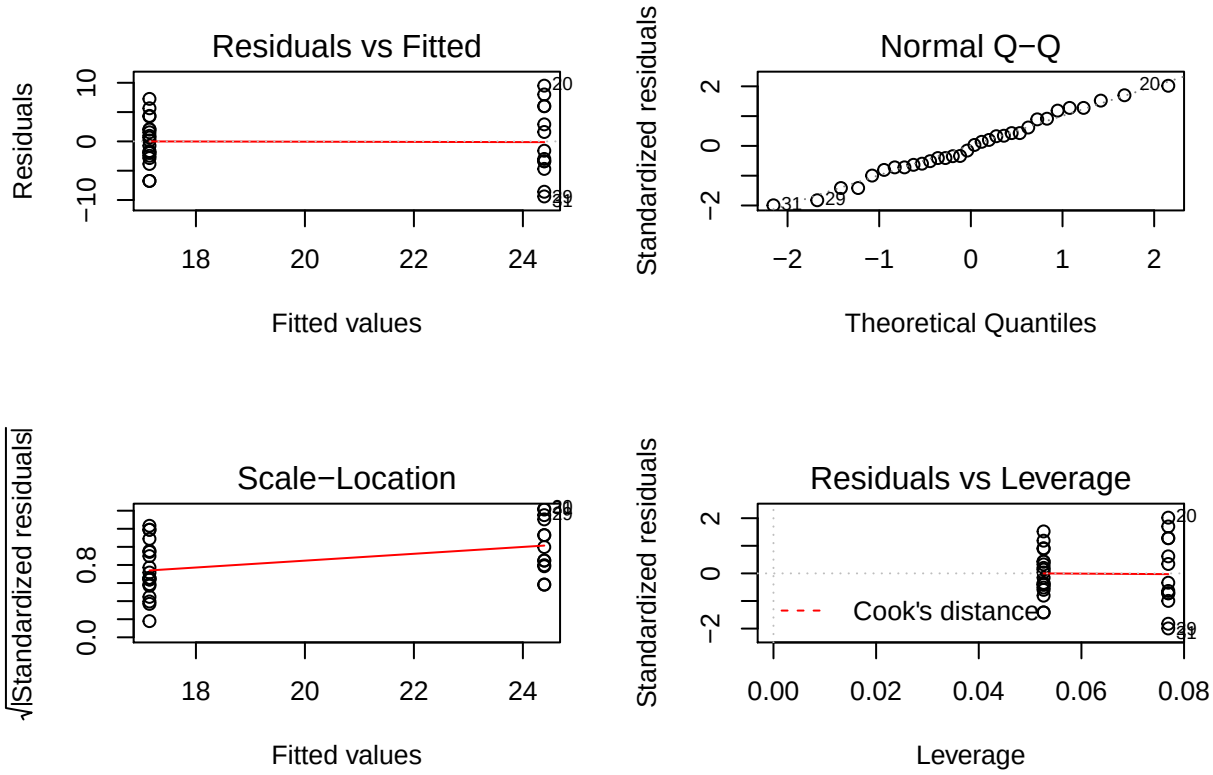
3.1 Boxplot:

```
ggplot(newcars,aes(transmission, mpg, color=transmission)) +
      geom_boxplot(outlier.colour="red", outlier.shape=16,
             outlier.size=2, notch=FALSE)
```



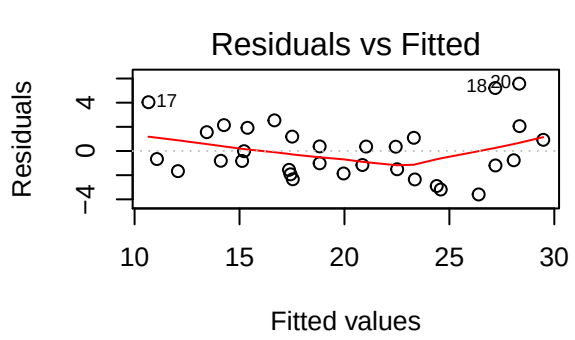
3.2 Simple Model:

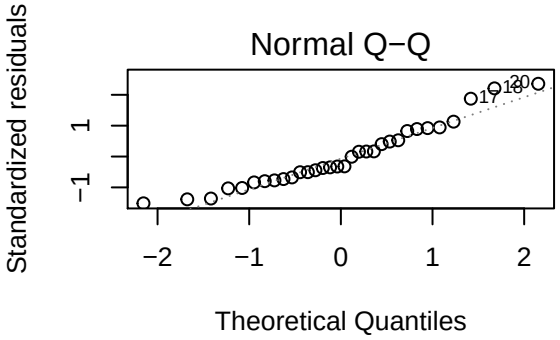
```
par(mfrow = c(2,2));plot(simpleModel)
```

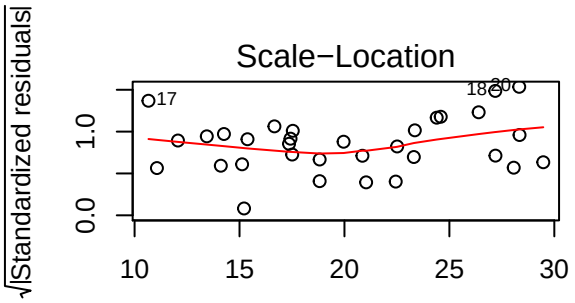


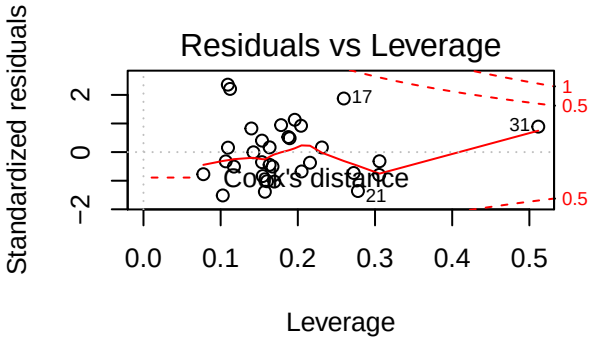
3.3 Multivariate Model:

```
par(mfrow = c(2,2));plot(multiModel)
```


Residuals vs Fitted
Residuals
Fitted values
17
18
20


Normal Q−Q
Standardized residuals
Theoretical Quantiles
17
18
20


Scale−Location
√|Standardized residuals|
Fitted values
17
18
20


Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
17
31
21
1
0.5
0.5