Regression Model

Luiz Bergo

24/11/2019

Summary

This is a project assignment as per request to complete the Regression Model course from Johns Hopkins University powered by Coursera. The main idea is to analyse the ***mtcars*** dataset provided in R Packages and answer the question that raised:

"Is an automatic or manual transmission better for MPG?"

Also is asked to "Quantify the MPG difference between automatic and manual transmissions".

At the following ssesions the data is discussed to properly answer the question.

Exploratory Data Analysis

The **mtcars** dataset is composed for 32 samples and 11 variables.

```
data("mtcars")
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
```

```
dim(mtcars)
```

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

The idea is to predict the influence of type of transmission (automatic or manual) in the fuel consumption **mpg** (Miles/(US) gallon). The remaining variable are descripted as follow:

cyl - Number of cylinders
disp - Displacement (cu.in.)
hp - Gross horsepower
drat - Rear axle ratio
wt - Weight (1000 lbs)
qsec - 1/4 mile time
vs - Engine (0 = V-shaped, 1 = straight)
am - Transmission (0 = automatic, 1 = manual)
gear - Number of forward gears
carb - Number of carburetors

Correlation of outcome with possible predictors:

```r
cor(mtcars)[1,]
```

```
##        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
```

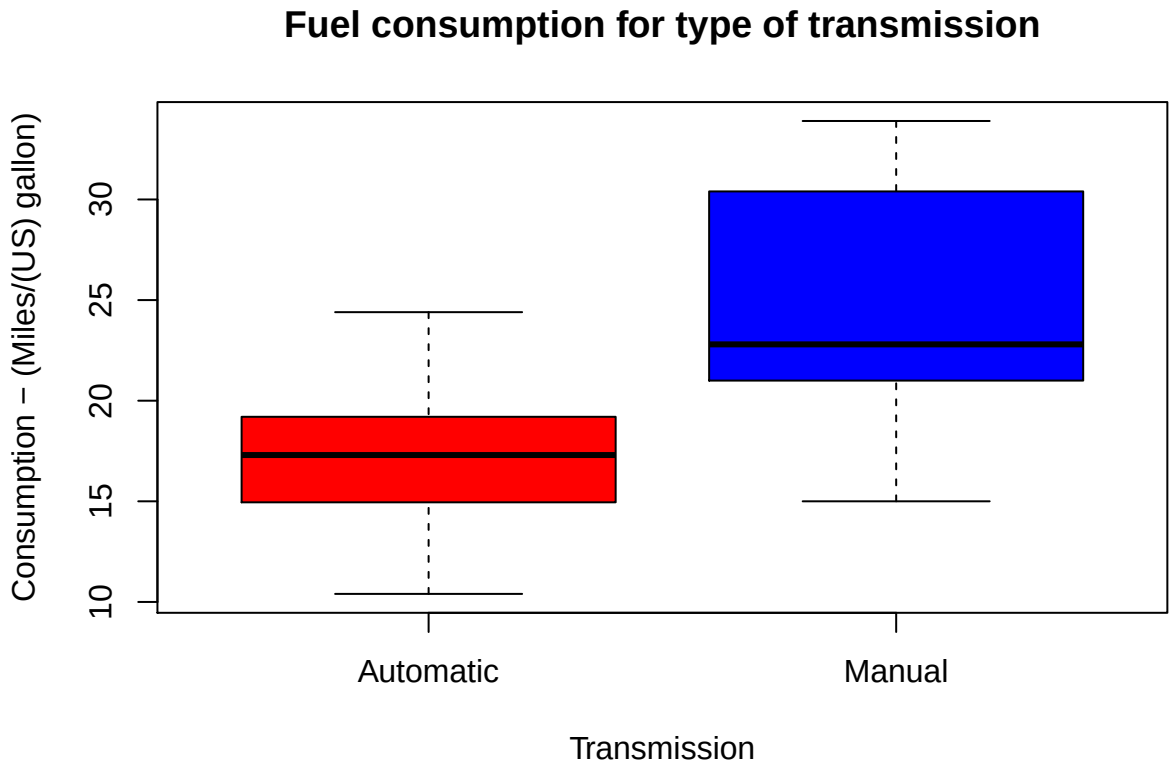
A comparison between the variables are depicted in the **Appendix**.

Model Adjustment

The first idea is to evaluate fuel consumption acording to the transmisison.

```r
cars <- mtcars
cars$am <- as.factor(cars$am)
levels(cars$am) <- c("Automatic", "Manual")

boxplot(mpg~am, data = cars, main = "Fuel consumption for type of transmission",
        col=c("red","blue"), xlab = "Transmission",ylab = "Consumption - (Miles/(US) gallon)")
```




```r
fit0 <- lm(mpg ~ factor(am), data=cars)
tcritical <- qt(0.975, fit0$df)
print(paste("Critical t value at 95% of confidence to reject null hypothesis :", round(tcritical, 3)))
```

```
## [1] "Critical t value at 95% of confidence to reject null hypothesis : 2.042"
```

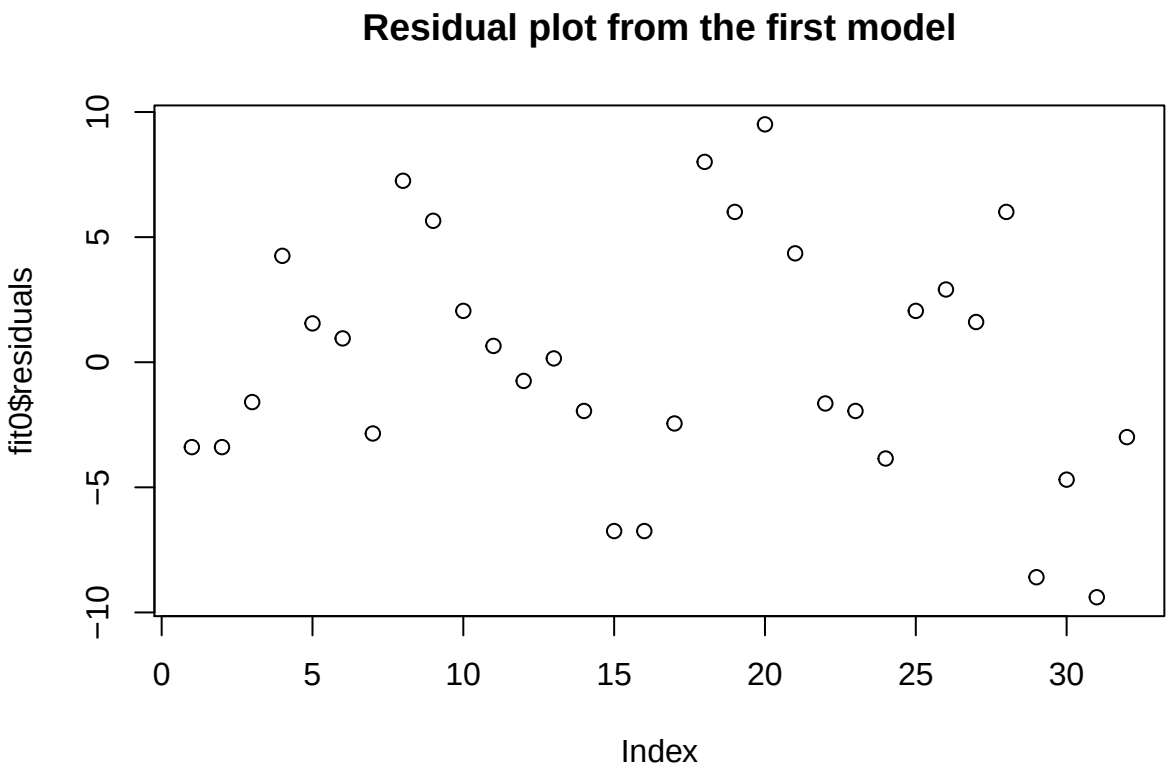
```
summary(fit0)
```

```
##
## Call:
## lm(formula = mpg ~ factor(am), data = cars)
##
## 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)Manual    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 t-values for coeficients are greater than t_c so it is validy assume some correlation between the variables and transmission explain about **34%** of variance at fuel comsuption. This lead us to interpret that **manual transmission cars** tend to consume less than **automatic**.

Considering that only **34%** of variation is explained for this model some other variables may be included for a better understandig. Such variation can either be observed in the plot of residuals.

```
plot(fit0$residuals, main = "Residual plot from the first model")
```



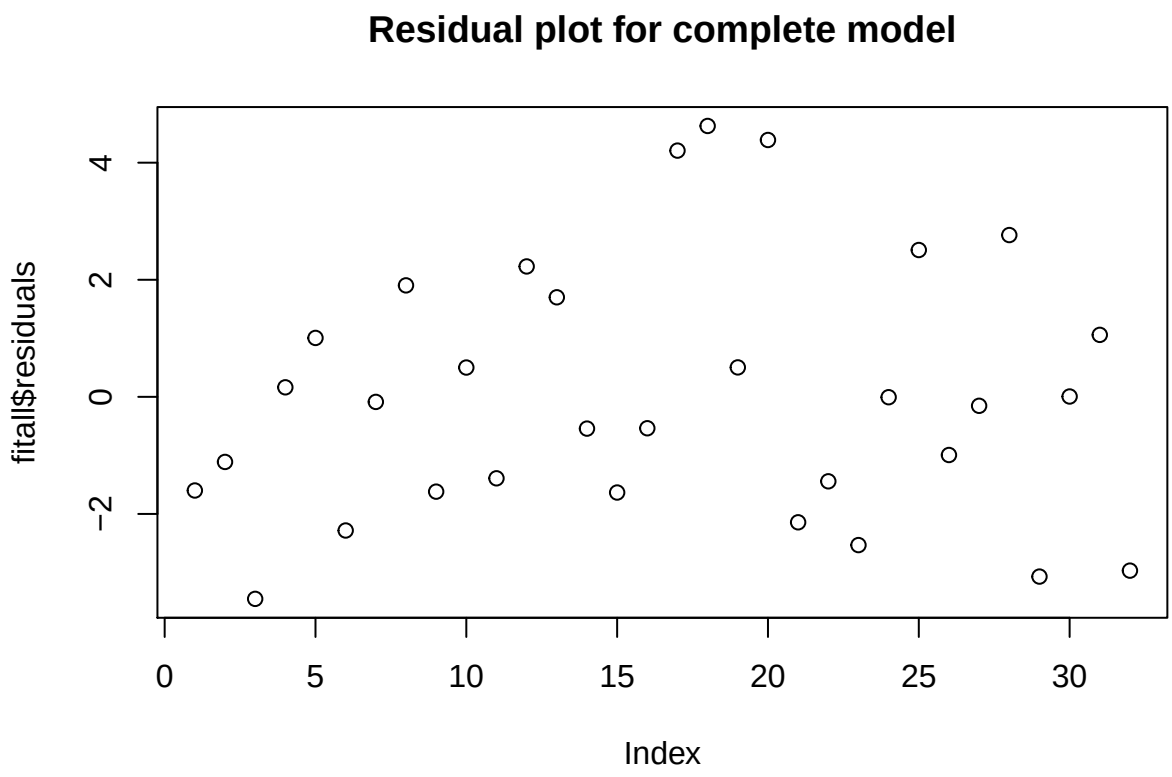
The first approach is to evaluate a linear fit including all the variables available as predictors.

```
fitall <- lm(mpg ~ ., data=cars)
summary(fitall)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = cars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4506 -1.6044 -0.1196  1.2193  4.6271
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.30337   18.71788   0.657   0.5181
## 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
## amManual     2.52023    2.05665   1.225   0.2340
## gear         0.65541    1.49326   0.439   0.6652
## carb        -0.19942    0.82875  -0.241   0.8122
```

```
## ---
## 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:  0.869,  Adjusted R-squared:  0.8066
## F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```

```
plot(fitall$residuals, main = "Residual plot for complete model")
```



The model including all variables change has a better result than the first model with just **am** (transmition) as a predictor with about **81%** of variance explained.

Although it is possible to observe a high variation inflation, suggesting the model can be simplified.

```
vif(fitall)
```

```
##       cyl      disp        hp      drat        wt      qsec        vs
## 15.373833 21.620241  9.832037  3.374620 15.164887  7.527958  4.965873
##        am      gear      carb
##  4.648487  5.357452  7.908747
```

Observing the plot and correlation between variables depicted at the *Appendix* it is possible to propose new fits observing that onde more correlated to mpg.

```r
fit1 <- lm(mpg ~ factor(am) + wt, data=cars)
vif(fit1)
```

```
## factor(am)         wt
##   1.921413   1.921413
```

```r
summary(fit1)$adj.r.squared
```

```
## [1] 0.7357889
```

With this model **74%** of variance is explained.

To improve the model the **hp** predictor was included and the variance explained is **82%**, better than the model with all the predictors.

```r
fit2 <- lm(mpg ~ factor(am) + wt + hp, data=cars)
vif(fit2)
```

```
## factor(am)         wt         hp
##   2.271082   3.774838   2.088124
```

```r
summary(fit2)$adj.r.squared
```

```
## [1] 0.8227357
```

Other predictors were tested aiming increase in the adjusted R square factor, and the final model includes **am**, **wt**, **hp** and **cyl** as predictors and the variance explained is **83%**.

A caveat is the variance inflation observed in the **hp** predictor since it is very correlated to **cyl** predictor.

```r
fit3 <- lm(mpg ~ factor(am) + wt + hp + cyl, data=cars)
vif(fit3)
```

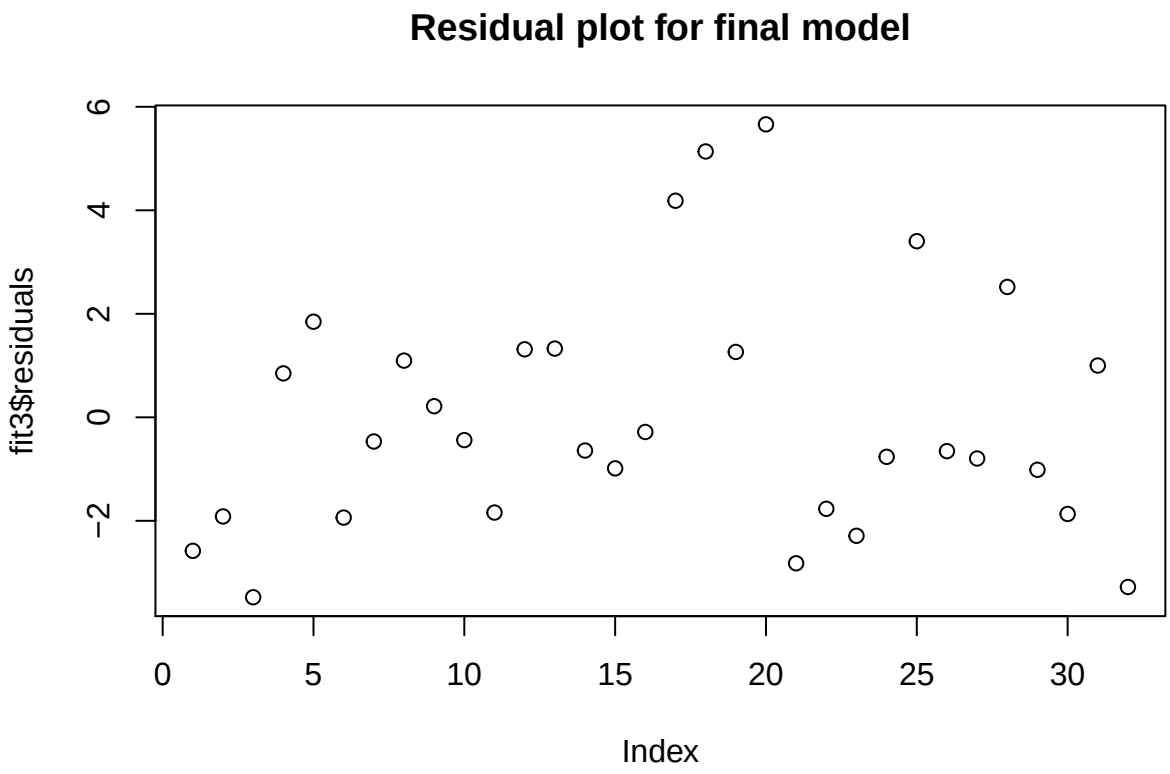
```
## factor(am)         wt         hp        cyl
##   2.546159   3.988305   4.310029   5.333685
```

```r
summary(fit3)$adj.r.squared
```

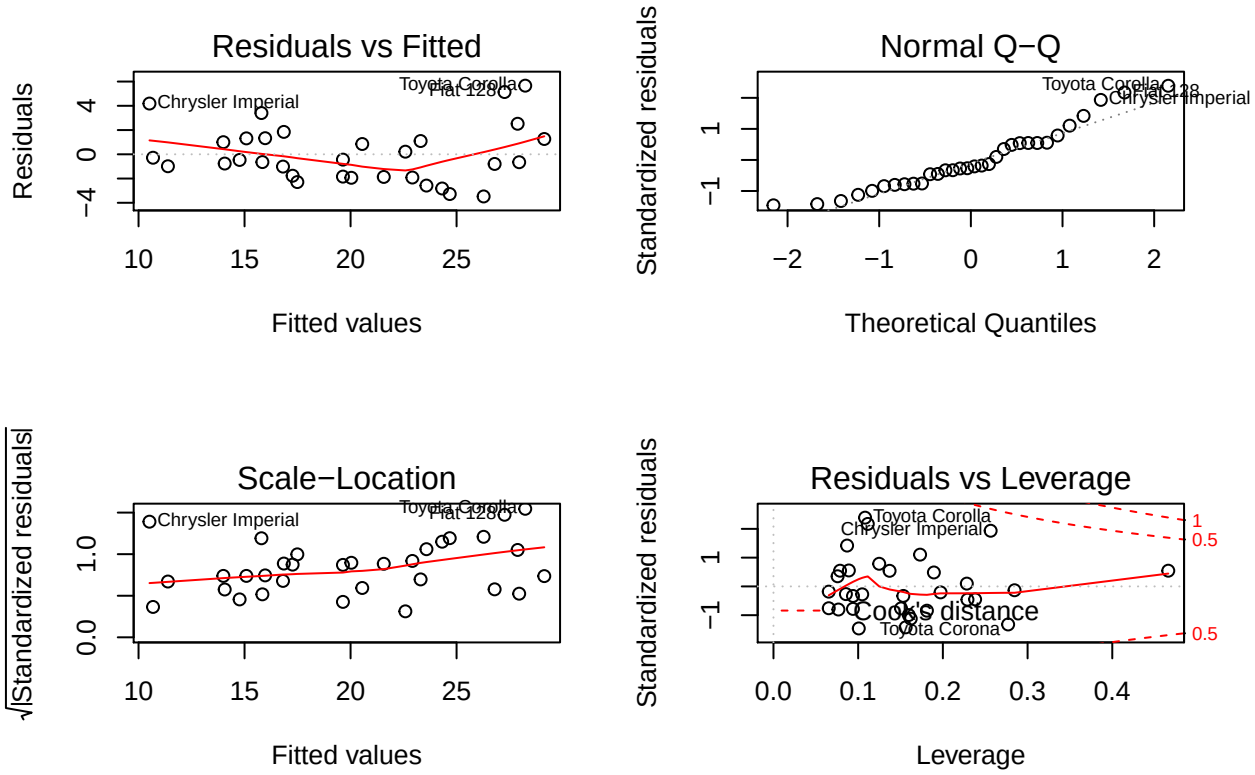
```
## [1] 0.8266657
```

It´s worthy to note the residual plot is quite near from that obtained using all the predictors.

```r
plot(fit3$residuals, main = "Residual plot for final model")
```




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



The coeficients for the final model are:

```
summary(fit3)$coef
```

```
##                    Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)      36.14653575 3.10478079 11.642218 4.944804e-12
## factor(am)Manual  1.47804771 1.44114927  1.025603 3.141799e-01
## wt               -2.60648071 0.91983749 -2.833632 8.603218e-03
## hp               -0.02495106 0.01364614 -1.828433 7.855337e-02
## cyl              -0.74515702 0.58278741 -1.278609 2.119166e-01
```

Conclusion

The confidence interval for the coeeficients are:

```
confint(fit3)
```

```
##                         2.5 %       97.5 %
## (Intercept)      29.77605177 42.517019733
## factor(am)Manual -1.47894635  4.435041763
## wt               -4.49383134 -0.719130075
## hp               -0.05295064  0.003048517
## cyl              -1.94093802  0.450623969
```

Suposing the same values for predictors **wt**, **hp** and **cyl**, manual cars are at average **4%** more economic than automatic cars.

Based on the final model it´s possible to conclude that cars with manual transmition are slightly economic than automatic car, so they are a better regarding this aspect.

Apendix

```
pairs(cars)
```