Regression models - Peer review

João Pedro Schmitt

31 de outubro de 2016

Introduction

A study for Motor Trend, a magazine about the automobile industry. Looking at a data set of a collection of cars, was explored the relationship between a set of variables and miles per gallon (MPG) (outcome). The particularly interested were following two questions:

- *Is an automatic or manual transmission better for MPG*
- *Quantify the MPG difference between automatic and manual transmissions*

Prepare the enviroment

For this work, was used the dataset mtcars, available in the package *datasets*. The code below demonstre the setup:

```
require(datasets)
require(graphics)
require(ggplot2)
require(MASS)
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
```

Finding the relationship

The code below creates a linear model using *mpg* as outcome and only *am* as predictor:

```
fit1 <- lm(mpg ~ am, data = mtcars)
print(summary(fit1)$coef)
```

```
##              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
```

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

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

In the summary above we can observe that for an automatic car we have the slope beeing zero with the intercept of miles per galon around 17.417 and when we have a manual car we have the increase of 7.245 in miles per galon. The P-value is significative because is less than 5% (0.05), so we can ignore the null hypothesis that the coeficients are zero. The current model explain 35.98% of the variance.

The best model is identified by *Final Model* (mpg ~ wt + qsec + am), using stepAIC that compares the p-values and r-squared of the predictors for each interaction:

```
fit2 <- stepAIC(lm(mpg ~ ., data = mtcars), direction="both")
```

```
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
```

The new model has the R-squared explained 84.97% compared with the old model that explain 35.98% doing the new model much more significative.

```
print(summary(fit2)$coef)
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  9.617781  6.9595930  1.381946 1.779152e-01
## wt          -3.916504  0.7112016 -5.506882 6.952711e-06
## qsec         1.225886  0.2886696  4.246676 2.161737e-04
## am           2.935837  1.4109045  2.080819 4.671551e-02
```

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

```
## [1] 0.8496636
```

Comparing the relevance of the new model, the new model is much more significative (p-value is 1.55e-09) and we can ignore the null hypothesis of use the old model.

```
anova(fit1, fit2)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ wt + qsec + am
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     30 720.90
## 2     28 169.29  2    551.61 45.618 1.55e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusion

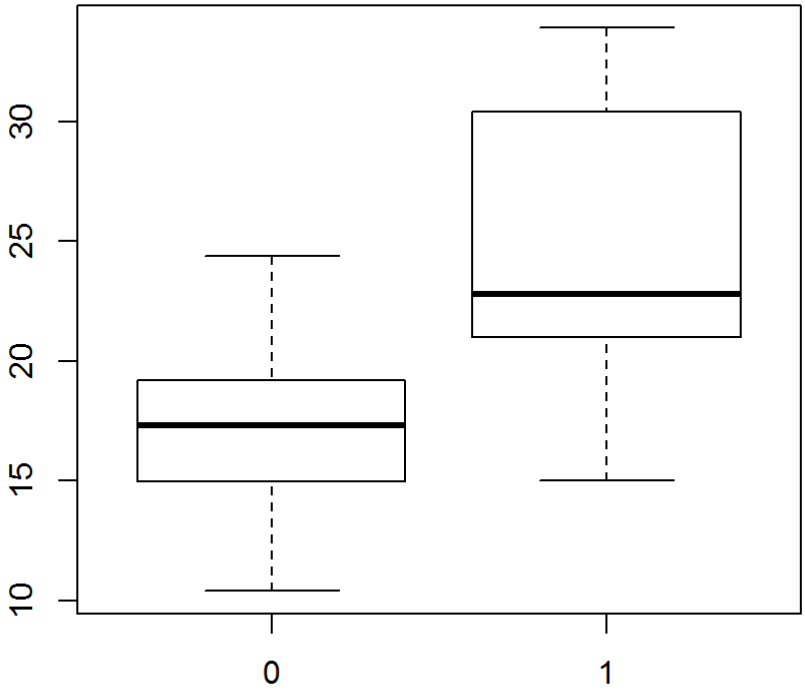
We concluded that for estimate the consumption of miles per galon for a car, not only the transmission is relevant but the 1/4 mile time and the weight of the car are very important to obtain more precise results. Using the best model we can see that cars that has automatic transmission have a increase of 2.9358 at miles per gallon.

Appendix

Exploratory data analisys about Miles per gallon vs transmission type.

The plot shows that apparently the manual transmission has a better consumption of miles per gallons compared with a automatic transmission:

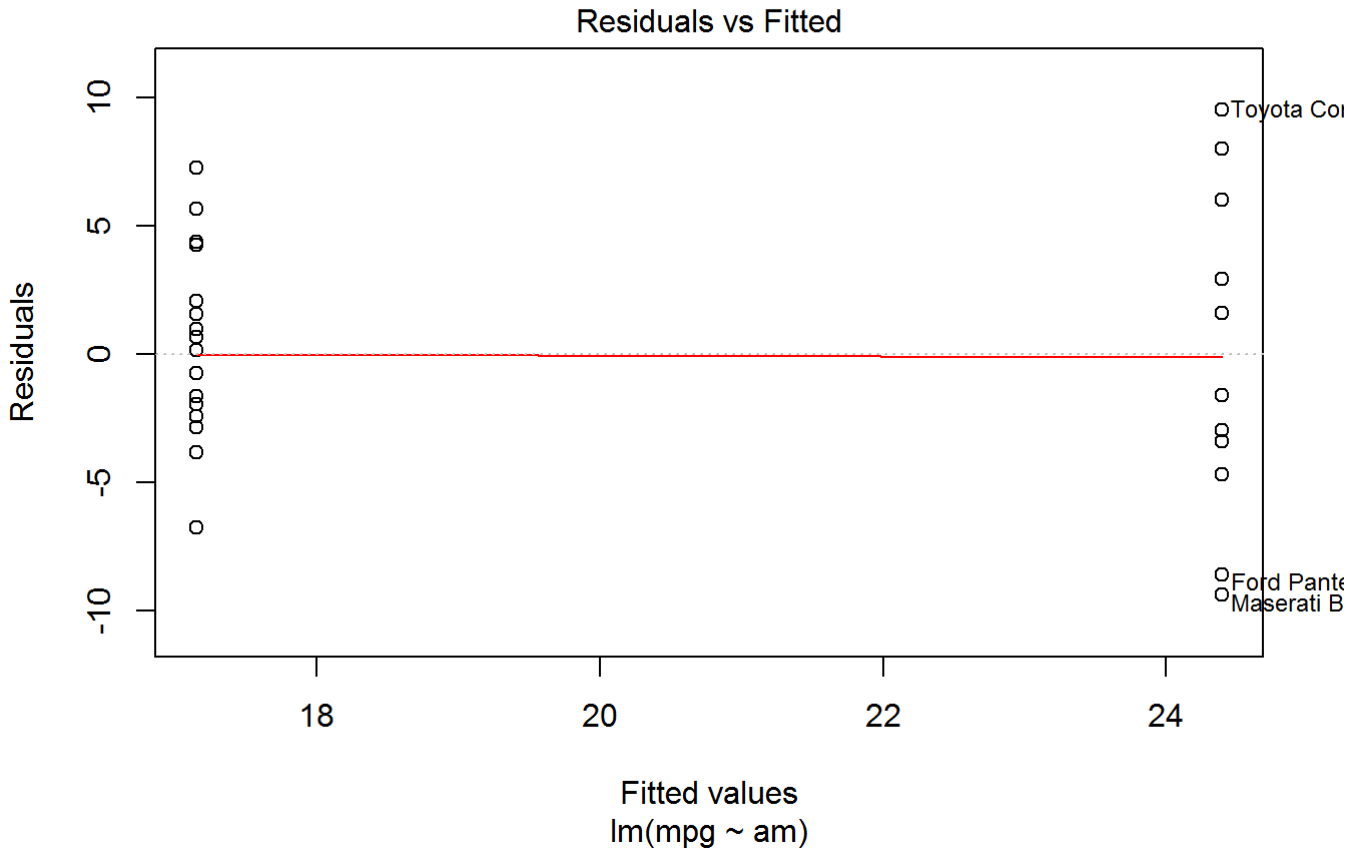
```
boxplot(mpg ~ am, mtcars)
```



Residuals of model (mpg ~ am)

The analysis of the residuals in the plot below, has some points that are over others points, and in this linear model it is difficult to estimate a good mpg for a given car

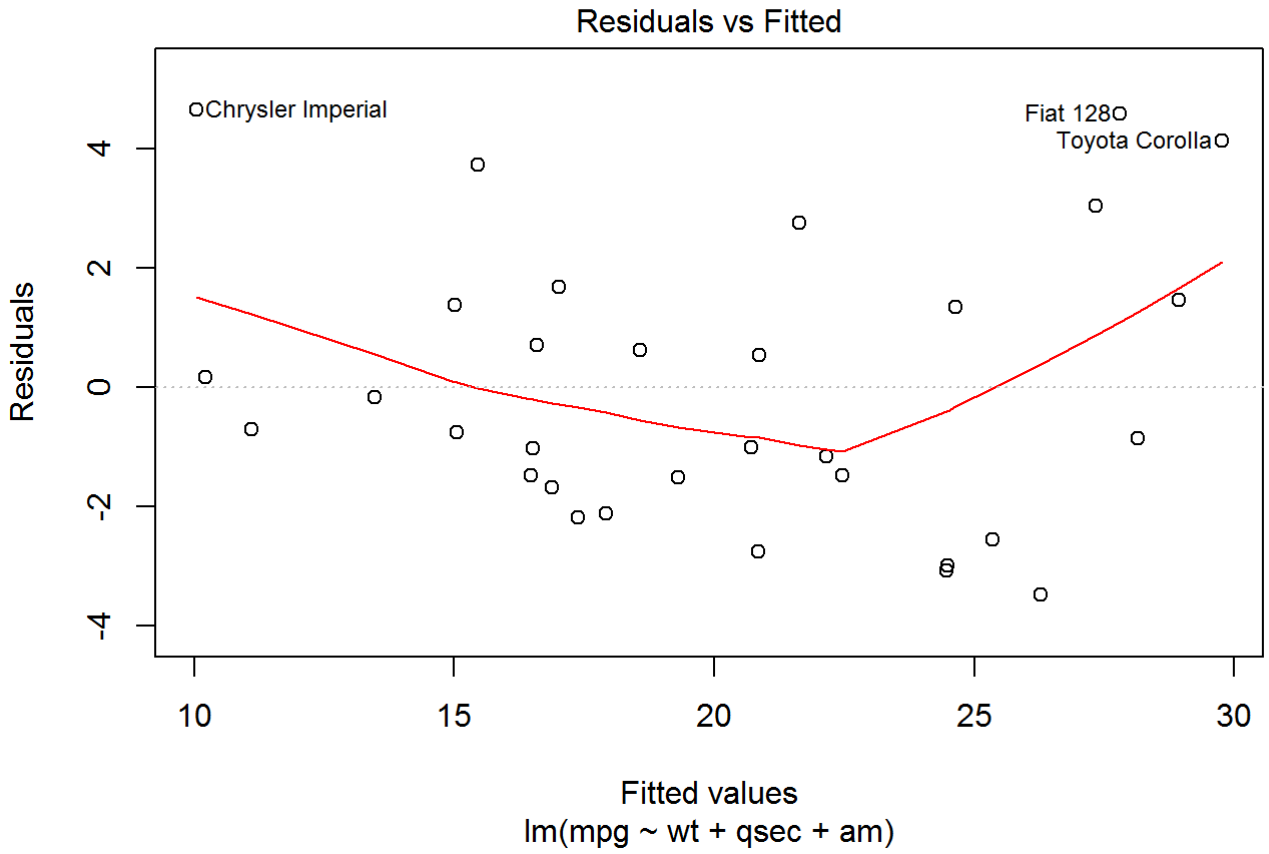
```
plot(fit1, which = 1)
```



Residuals of model (mpg ~ wt + qsec + am)

This residual plot is more reasonable with the less difference in the values compared with model (mpg ~ am), explainning better the fitted values.

```
plot(fit2, which = 1)
```


Residuals vs Fitted
Chrysler Imperial
Fiat 128
Toyota Corolla
Residuals
4
2
0
-2
-4
10
15
20
25
30
Fitted values
lm(mpg ~ wt + qsec + am)