Regression models - Course Project

Fabrice Tereszkiewicz

16 Oct 2015

Introduction

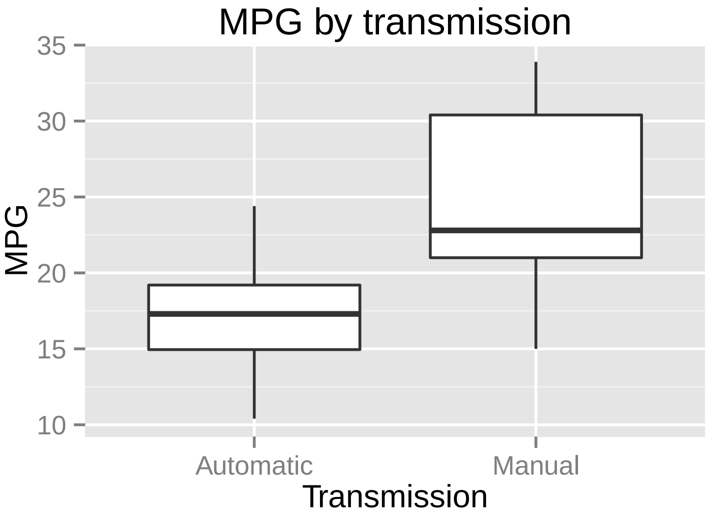
We work for Motor Trend, a magazine about the automobile industry. Looking at a data set of a collection of cars, we are asked to explore the relationship between a set of variables and miles per gallon (MPG). More especially, we are interested in the impact of the transmission type on the MPG.

Executive summary

We were asked to look into the effect of the transmission type on the MPG for a dataset of vehicles. We have first conducted an exploration of the data and an inference analysis and we can confirm that the transmission has an impact. Then using linear regression and different models, we established that the MPG of a vehicle is mostly determined by the weight of the car, the 1/4 mile time and the transmission type. **We can conclude with 95% confidence that a manual transmission results in an increase of the MPG between 0.05 and 5.83.**

Exploratory Analysis

The dataset contains 32 observations and 11 variables. 19 vehicles with automatic transmission and 13 with manual transmission.



This plot shows that the MPG values seem to be higher for the manual transmissions. To first determin if the transmission has an impact on the MPG, let H_0 be the null hypothesis that it has no impact : $H_0 : \mu_{auto} = \mu_{manual}$

The variance difference of the 2 groups is not negligible (14.7 for automatic cars versus 38.03 for manual cars), we will assume the 2 groups to have unequal variance for the T test.

As the 95% interval `[-11.28 -3.21]` doesn't contain 0 and the P-value (0.0013736) is < 0.05, the hypothesis that the transmission is not important can be rejected.

Regression

We can try a marginal linear regression first to see the effect of the transmission type on MPG, holding all other variables constant.

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

We can see that on average manual transmission vehicules can do 7.245 miles per gallon more than automatic transmission vehicules. The R-squared error shows that only 35.98% of the variance in MPG is explained by the type of transmission. We must therefore look at other parameters in order to explain the MPG change.

We can do a step search to find the most optimal model to fit the MPG value.

```
mdl.opt = step(lm(data = mtcars, mpg ~ .),direction = "both")
```

```
summary(mdl.opt)$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
## amManual     2.935837  1.4109045  2.080819 4.671551e-02
```

The MPG can be better explained by including the weight (`wt`) and the 1/4 mile time (`qsec`) to the model. With this model, a manual transmission increases the MPG by 2.936 on average holding all other variables constant.

```
anova(mdl.mar, mdl.opt)
```

```
## 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
```

The P-value is significant (almost 0), which confirms that the model `mdl.opt` is more accurate than the marginal model `mdl.mar`.

Residuals and Diagnostics

The plots are in appendix :

- there are no signs of correlation between the residuals and fitted values, which is good for a homoscedastic linear model with normally distributed errors.
- the QQ plot support normality of the residuals, as all points are close to the line.
- the points look randomly distributed on the Scale-Location plot, so we can assume constant variance.
- all the residuals are well away from the 0.5 Cook's distance, which doesn't indicate excessive leverage.

Inference

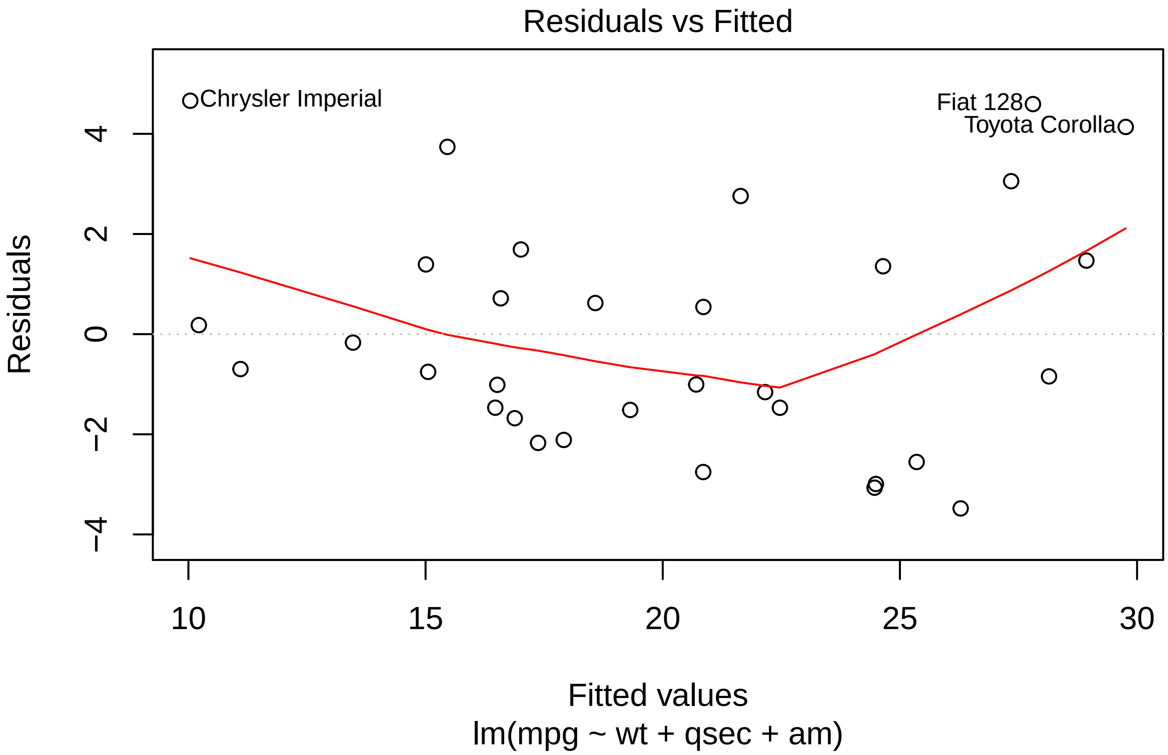
A T-test on the transmission coefficient shows that with 95% confidence, the transmission is responsible for an increase in MPG between 0.05 and 5.83.

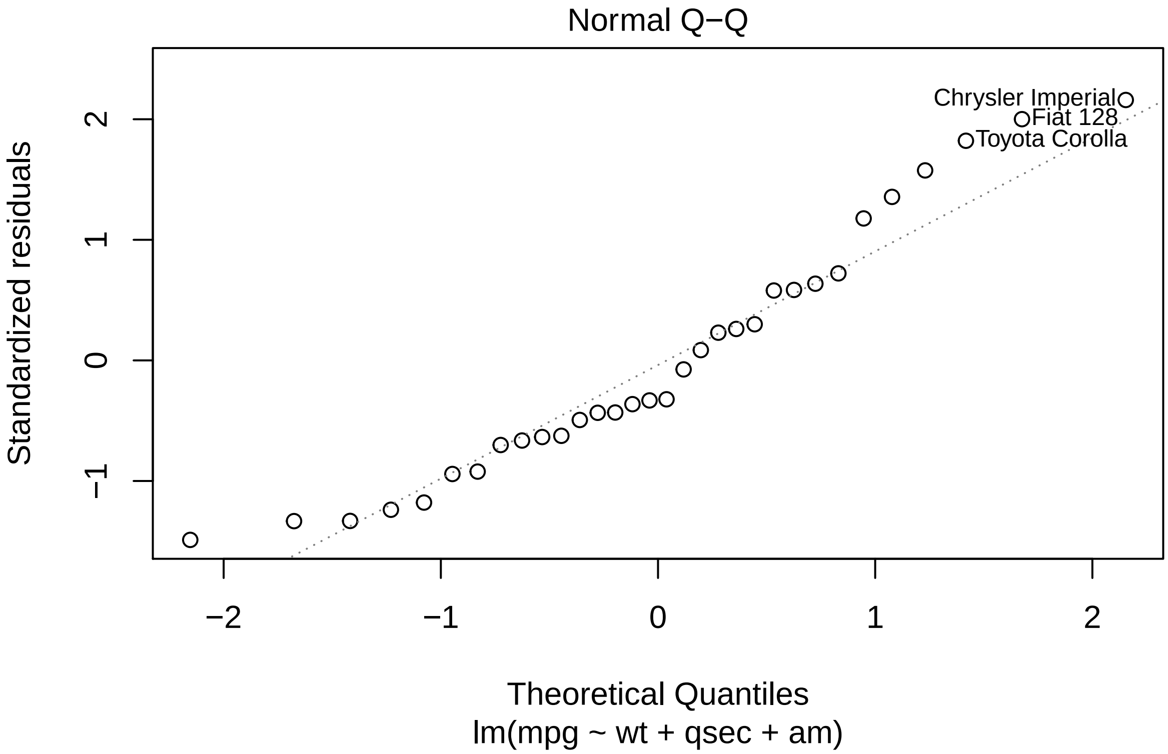
```r
sumc <- summary(mdl.opt)$coefficients
sumc[4,1] + c(-1, 1)*qt(.975, mdl.opt$df ) * sumc[4,2]
```

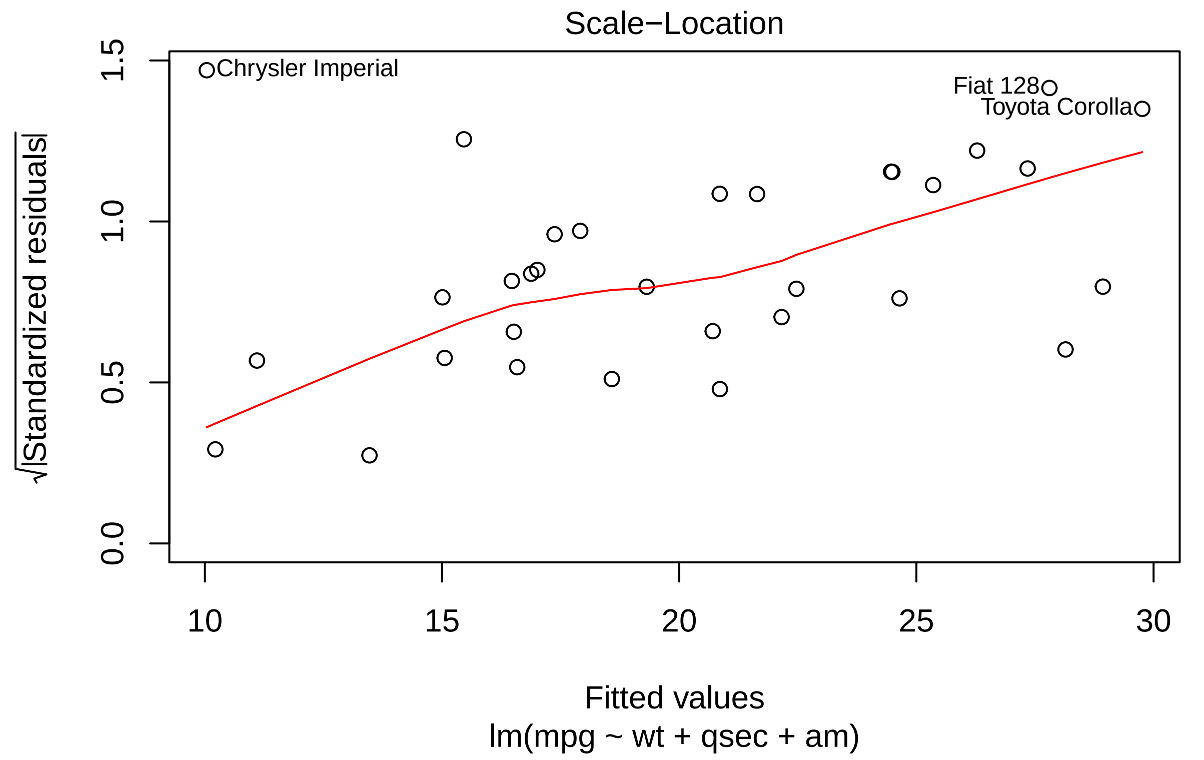
```
## [1] 0.04573031 5.82594408
```

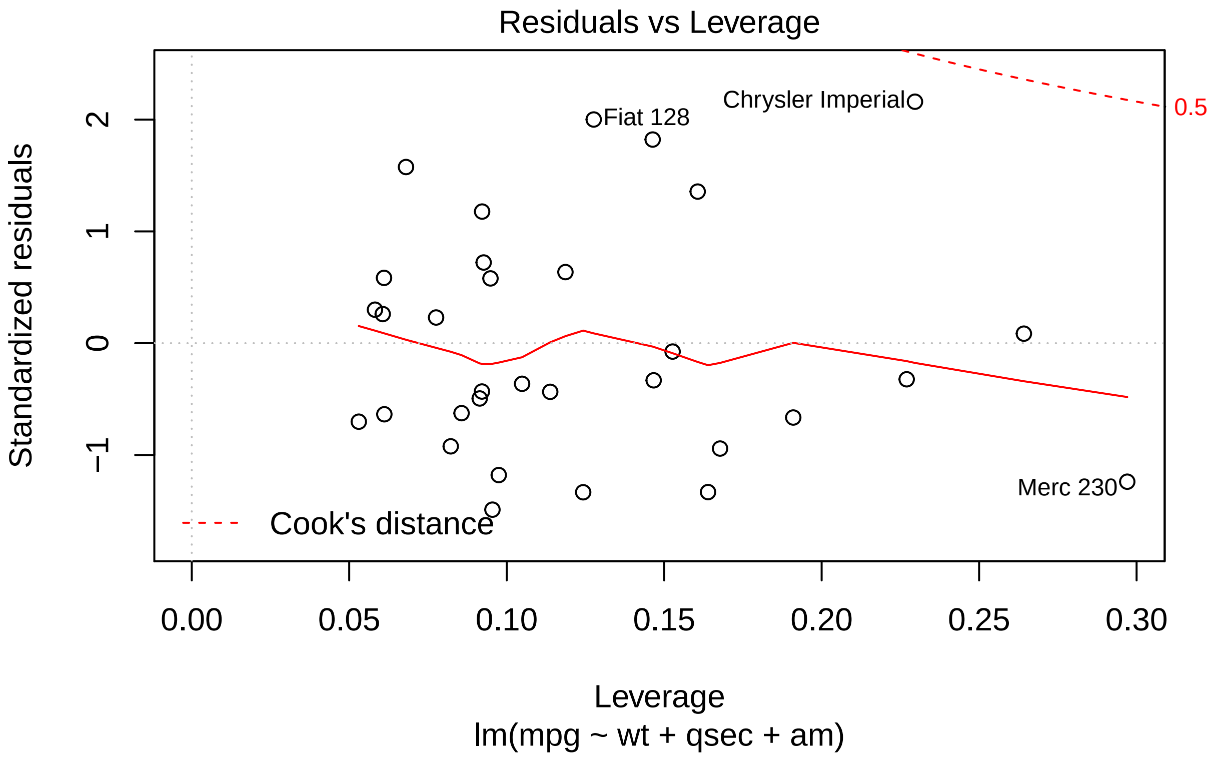
Appendix

Model residuals and diagnostic






Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
10
15
20
25
30
Fitted values
lm(mpg ~ wt + qsec + am)
Chrysler Imperial
Fiat 128
Toyota Corolla


Residuals vs Leverage
Standardized residuals
−1
0
1
2
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Leverage
lm(mpg ~ wt + qsec + am)
Fiat 128
Chrysler Imperial
Merc 230
0.5
Cook's distance