Comparison of MPG on cars with automatic and manual transmission

Stefano Masneri

13/05/2016

Executive Summary

In this document we will investigate the *mtcars* dataset and will try to answer these two questions:

- Is an automatic or manual transmission better for MPG?
- Can we quantify the MPG difference between automatic and manual transmissions?

We will create single and multiple variables linear regression models to answer the questions. We noticed when considering a single variable model that *MANUAL* transmission is better than *AUTOMATIC* in terms of Miles per Gallon. The improvement when using *MANUAL* compared to *AUTOMATIC* is about 7 miles per gallon. The results change quite a bit when considering other variables in the model, namely the number of cylinders, the weight and the displacement. In this case the increase in *mpg* due to the usage of a *MANUAL* transmission system is negligible (0.14 mpg) and statistically not significant.

Getting and cleaning Data

We will use for this project the *mtcars* dataset.

```r
data(mtcars)
str(cars)
```

```
## 'data.frame':    50 obs. of  2 variables:
##  $ speed: num  4 4 7 7 8 9 10 10 10 11 ...
##  $ dist : num  2 10 4 22 16 10 18 26 34 17 ...
```

All the variables are treated as *num*, so we have to convert some of them to factor variables. Since we are mostly concerned with the transmission, we will label the values for the Transmission (*am*) variable, so that "0" becomes *Automatic* and "1" becomes *Manual*.

```r
mtcars$am <- factor(mtcars$am,labels=c("Automatic","Manual"))
mtcars$gear <- factor(mtcars$gear)
mtcars$vs   <- factor(mtcars$vs)
mtcars$cyl  <- factor(mtcars$cyl)
mtcars$carb <- factor(mtcars$carb)
```

To get an idea on how the *mpg* can vary depending on the transmission used, we just compute some statistics differentiating the two cases, and draw a boxplot of the miles per gallon in the case of automatic and manual transmission (see Appendix, figure 1)

```
with(mtcars, tapply(mpg, am, mean))
```

```
## Automatic    Manual
##  17.14737  24.39231
```

```
with(mtcars, tapply(mpg, am, sd))
```

```
## Automatic    Manual
##  3.833966  6.166504
```

```
with(mtcars, tapply(mpg, am, summary))
```

```
## $Automatic
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.40   14.95   17.30   17.15   19.20   24.40
##
## $Manual
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15.00   21.00   22.80   24.39   30.40   33.90
```

It seems as the cars with manual transmission make more miles per gallon than the ones with automatic transmission. We will verify if that is true in the next section.

Analysis

The obvious thing to do to start the analysis is to create a linear model that fits *mpg* according to the *transmission*. We also create a model that fits *mpg* according to all other variables and that will be used later when doing anova.

```
fit1 <- lm(mpg ~ ., mtcars)

fit2 <- lm(mpg ~ am, mtcars)
summary(fit2)$coef
```

```
##              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
```

From the second model it seems that, with huge statistical significance, the transmission influences the miles per gallon. The model tells us that using a car with *manual* transmission, the miles per gallon increase by more than 7. We can compute a 95% confidence interval for the increase in *mpg*

```
est <- summary(fit2)$coef["amManual", "Estimate"]
err <- summary(fit2)$coef["amManual", "Std. Error"]
numSamples <- length(mtcars$mpg)
tstat <- qt(0.975, numSamples - 2)
est + c(-1, 1) * (err * tstat)
```

```
## [1]  3.64151 10.84837
```

The fact that the interval does not include zero (and the low p-value) means that we can reject the null hypothesis (that is, *mpg* is not influenced by the transmission type) and conclude that the cars with manual transmission are more efficient than the ones with automatic transmission. Figure 2 of the Appendix shows plots for this model.

Next, we want to select the variables which influence the *mpg* value the most and to do so we ran anova on the first model.

```
anova(fit1)
```

```
## Analysis of Variance Table
##
## Response: mpg
##           Df Sum Sq Mean Sq F value    Pr(>F)
## cyl        2 824.78  412.39 51.3766 1.943e-07 ***
## disp       1  57.64   57.64  7.1813   0.01714 *
## hp         1  18.50   18.50  2.3050   0.14975
## drat       1  11.91   11.91  1.4843   0.24191
## wt         1  55.79   55.79  6.9500   0.01870 *
## qsec       1   1.52    1.52  0.1899   0.66918
## vs         1   0.30    0.30  0.0376   0.84878
## am         1  16.57   16.57  2.0639   0.17135
## gear       2   5.02    2.51  0.3128   0.73606
## carb       5  13.60    2.72  0.3388   0.88144
## Residuals 15 120.40    8.03
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We will then create a model including *cyl*, *disp* and *wt*. We can already notice, via the analysis of variation, that when including all the variables, the transmission value is not statistically significant. We extract the coefficients and test once again whether the results obtained for the transmission coefficients are significative.

```
fit3 <- lm(mpg ~ am + wt + cyl + disp, data = mtcars)
summary(fit3)$coef
```

```
##                 Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) 33.816067258 2.91427152 11.6036090 8.792094e-12
## amManual     0.141212000 1.32675115  0.1064344 9.160547e-01
## wt          -3.249175911 1.24909839 -2.6012170 1.512685e-02
## cyl6        -4.304782473 1.49235501 -2.8845566 7.774612e-03
## cyl8        -6.318405700 2.64765755 -2.3864135 2.458094e-02
## disp         0.001632161 0.01375694  0.1186427 9.064703e-01
```

```
est3 <- summary(fit3)$coef["amManual", "Estimate"]
err3 <- summary(fit3)$coef["amManual", "Std. Error"]
tstat <- qt(0.975, numSamples - 5) # 5 because we have 4 variables in this case
est3 + c(-1, 1) * (err3 * tstat)
```

```
## [1] -2.581057  2.863481
```

In this case the interval contains 0, so we fail to reject the null hypothesis that the transmission directly influences the miles per gallon value. Figure 3 contains the plot for this model.

Appendix

Figure 1:

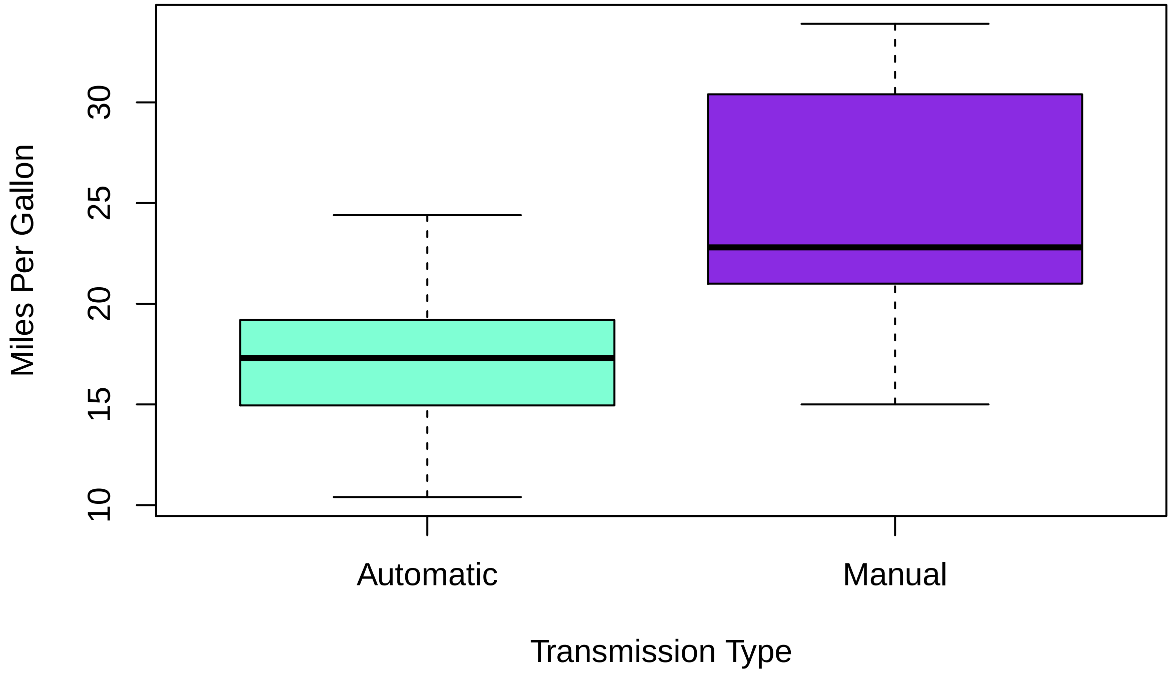


Figure 2: Plot of linear model of mpg ~ am

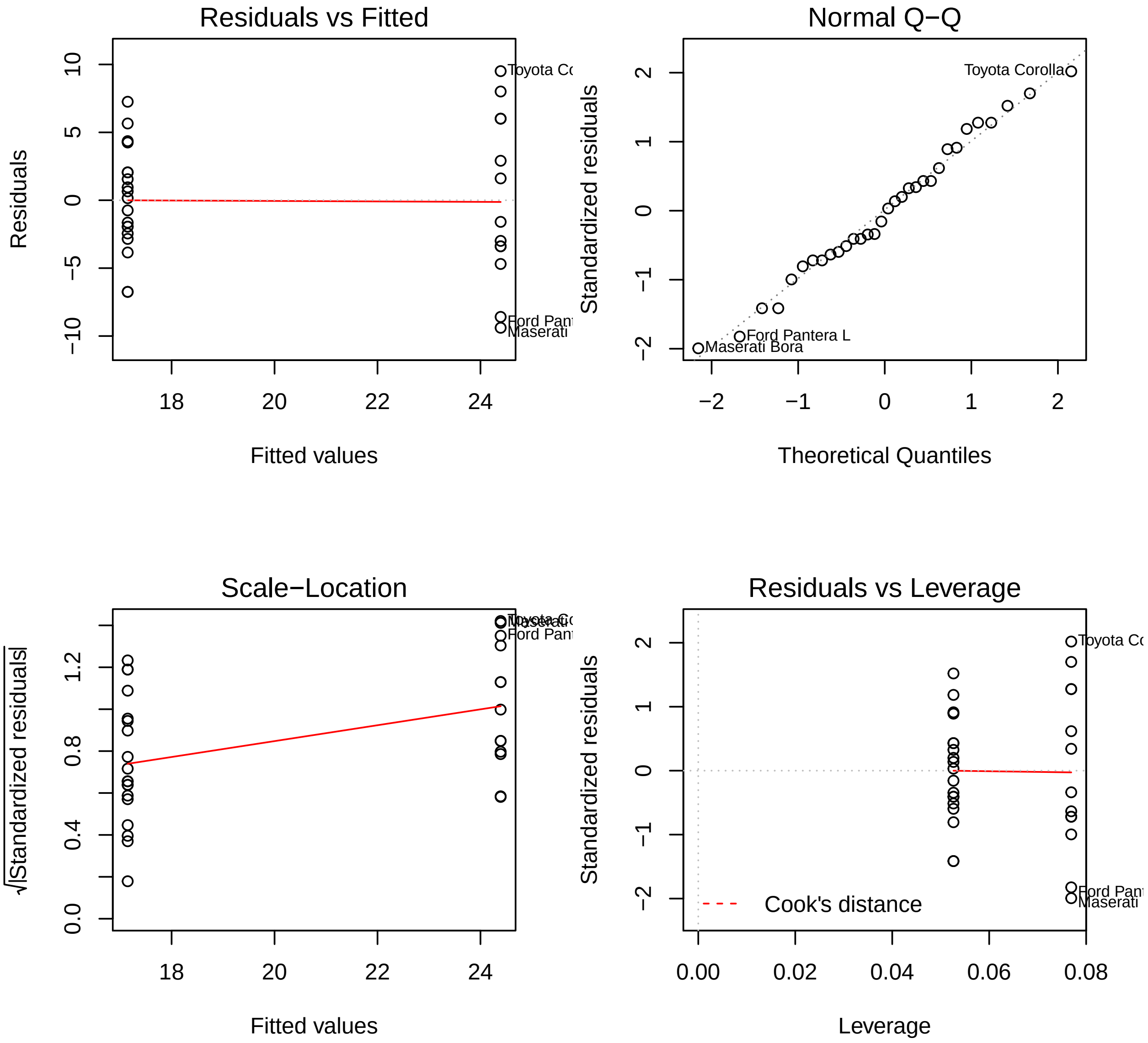
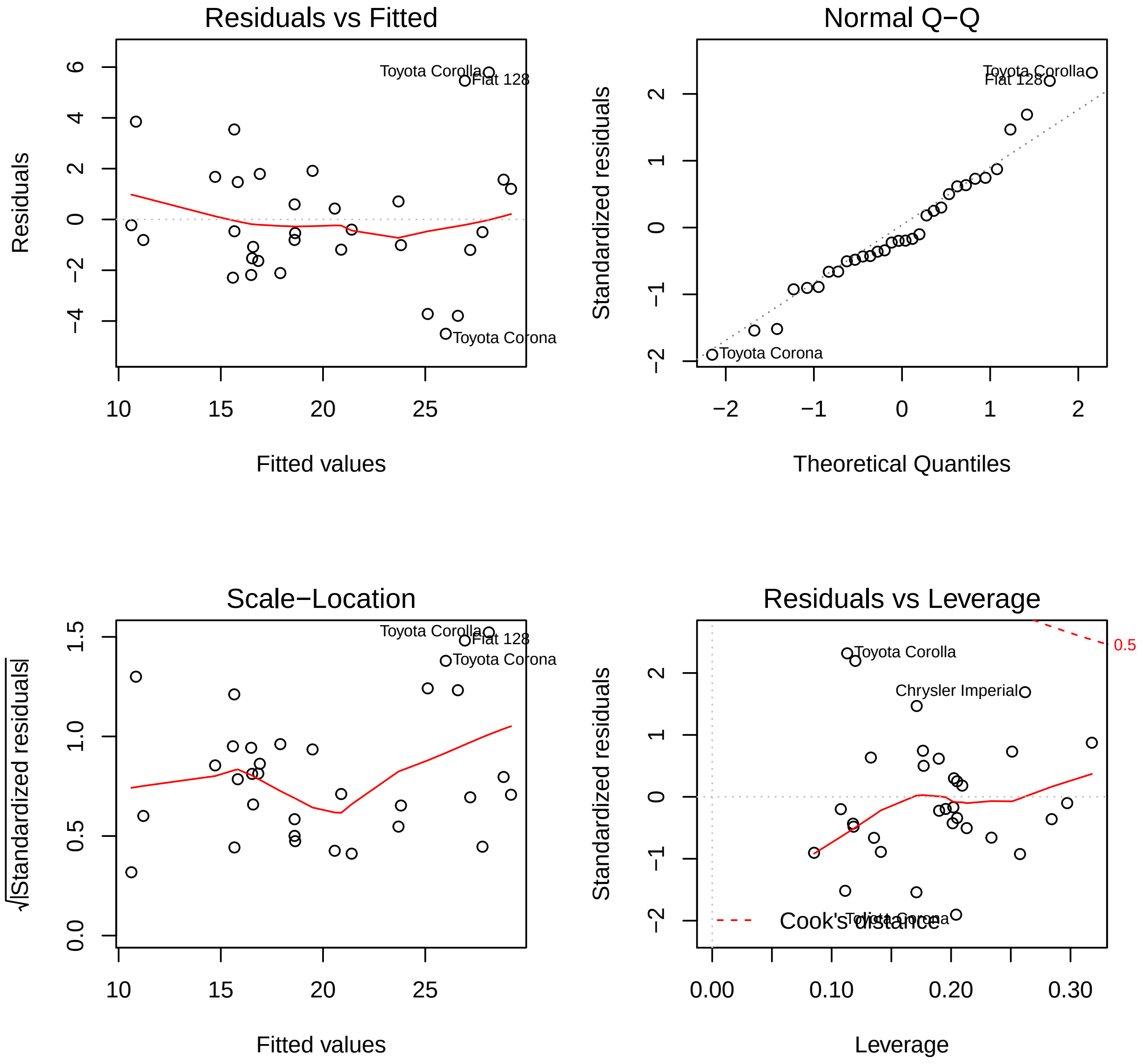


Figure 3: Plot of linear model of mpg ~ am + wt + cyl + disp


Residuals vs Fitted
Residuals
Fitted values
Toyota Corolla
Fiat 128
Toyota Corona
6
4
2
0
−2
−4
10
15
20
25
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Toyota Corolla
Fiat 128
Toyota Corona
2
1
0
−1
−2
−2
−1
0
1
2
Scale−Location
√|Standardized residuals|
Fitted values
Toyota Corolla
Fiat 128
Toyota Corona
1.5
1.0
0.5
0.0
10
15
20
25
Residuals vs Leverage
Standardized residuals
Leverage
Toyota Corolla
Chrysler Imperial
Toyota Corona
Cook's distance
0.5
2
1
0
−1
−2
0.00
0.10
0.20
0.30