Regression Models Course Project

Davy Meesemaecker

21/1/2018

Executive summary

You work for Motor Trend, a magazine about the automobile industry. Looking at a data set of a collection of cars, they are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). They are particularly interested in the following two questions:

° “Is an automatic or manual transmission better for MPG”

° “Quantify the MPG difference between automatic and manual transmissions”

In the report, we fit three different models. The first model is a simple model with only one variable (auto/manual), which suggests that manual transmission’s MPG is 7.245 miles higher than auto transmission, but the model only explained 33.85% of the variance in MPG. The second model is using all variables does not have any coefficients which appear to be significant. The last model, a model with 4 variables (cylinder, horsepower, weight and auto/manual) which explains 86% of the variance in MPG, suggests that manual transmission’s MPG is still higher than auto transmission’s MPG, even though the miles per gallon saved is only higher by 1.81 miles compared to auto transmission after including the three other variables - cylinder, horsepower and weight.

Data analysis

The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models).

```r
library(ggplot2)
library (dplyr)
data("mtcars")
str(mtcars)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

```r
sum(is.na(mtcars))
```

```
## [1] 0
```

By taking a quick look at the data, we learn our data doesn’t contain any NA and that all variables are numeric for now but some of them need be converted to factors

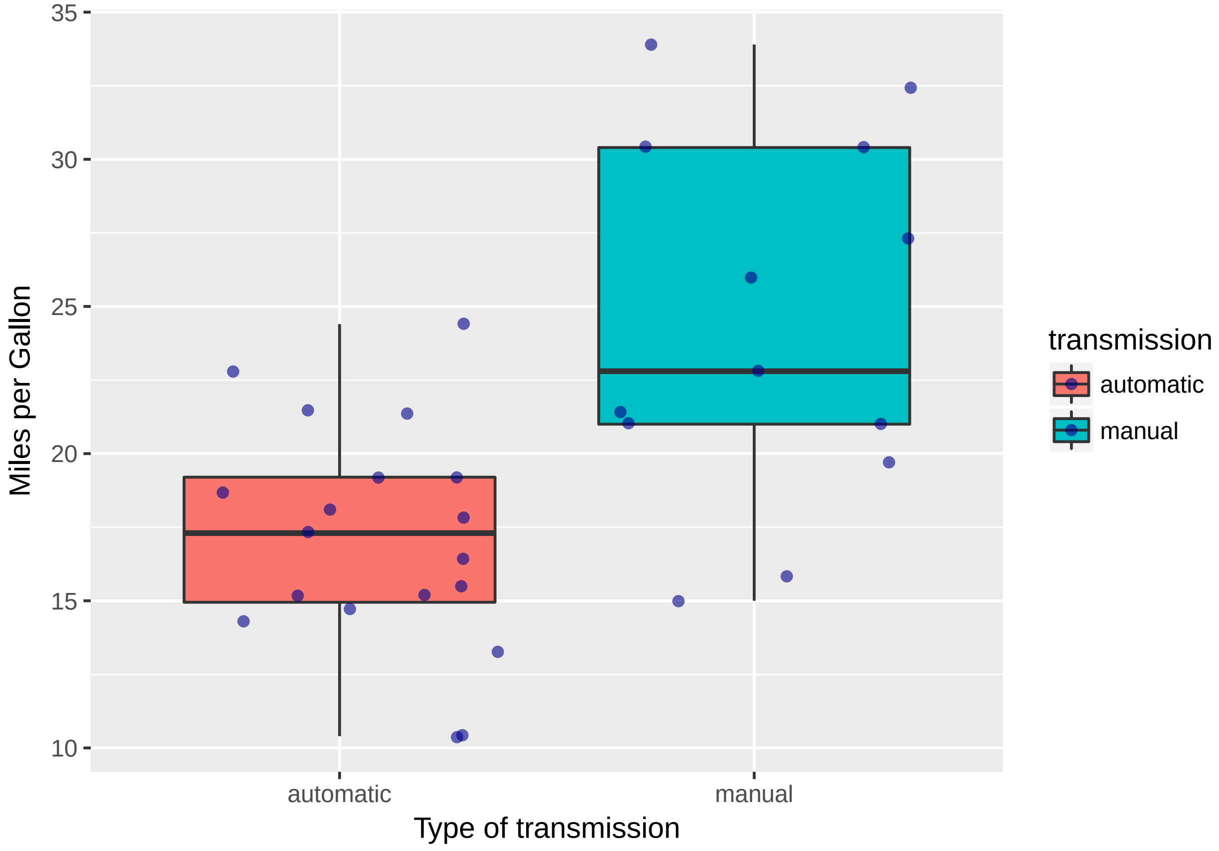
```
mtcars$cyl <- as.factor(mtcars$cyl)
mtcars$vs <- as.factor(mtcars$vs)
mtcars$am <- as.factor(mtcars$am)
mtcars$gear <- as.factor(mtcars$gear)
mtcars$carb <- as.factor(mtcars$carb)
```

Now it's done, we can get a first look on how transmission type impacts the fuel consumption

```
library(plyr)
```

```
## ------------------------------------------------------------------------

## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)

## ------------------------------------------------------------------------

##
## Attaching package: 'plyr'

## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

```
mtcars$transmission <- revalue(mtcars$am, c("0" = "automatic", "1" = "manual"))
mtcars <- mtcars[, -c(9)]
ggplot(mtcars, aes(x = transmission, y = mpg, fill = transmission)) + geom_boxplot() + geom_jitter(colo
```



The plot suggests an impact of the type of transmission on the fuel consumption as Automatic cars seem to be less fuel efficient than manual cars. Fitting a simple model could help us validate this trend.

```
mdl <- lm(mpg ~ transmission, data = mtcars)
summary(mdl)
```

```
##
## Call:
## lm(formula = mpg ~ transmission, data = mtcars)
##
## 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
```

So what do we learn ? First, our manuel transmission estimate is 7.245, implying a manual car can drive 7.245 miles per gallon more than an automatic car on average. It's approuved by our very low p-value, making our model is significant at a 0.001 level. However, we only cover 35% of mpg's variability and it seems obvious as we only had one predictor in our model. What if we add all variables in our model ?

```
mdl2 <- lm(mpg ~. , data = mtcars)
summary(mdl2)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5087 -1.3584 -0.0948  0.7745  4.6251
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        23.87913   20.06582   1.190   0.2525
## cyl6               -2.64870    3.04089  -0.871   0.3975
## cyl8               -0.33616    7.15954  -0.047   0.9632
## disp                0.03555    0.03190   1.114   0.2827
## hp                 -0.07051    0.03943  -1.788   0.0939 .
## drat                1.18283    2.48348   0.476   0.6407
## wt                 -4.52978    2.53875  -1.784   0.0946 .
## qsec                0.36784    0.93540   0.393   0.6997
## vs1                 1.93085    2.87126   0.672   0.5115
## gear4               1.11435    3.79952   0.293   0.7733
## gear5               2.52840    3.73636   0.677   0.5089
## carb2              -0.97935    2.31797  -0.423   0.6787
## carb3               2.99964    4.29355   0.699   0.4955
## carb4                1.09142    4.44962   0.245   0.8096
## carb6                4.47757    6.38406   0.701   0.4938
## carb8                7.25041    8.36057   0.867   0.3995
## transmissionmanual   1.21212    3.21355   0.377   0.7113
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.833 on 15 degrees of freedom
## Multiple R-squared:  0.8931, Adjusted R-squared:  0.779
## F-statistic:  7.83 on 16 and 15 DF,  p-value: 0.000124
```

Now we cover 89% of mpg's variability but unfortunately, none of the predictors is significant at 0.05. We need to find a better model by removing some noisy variables.

```r
bestformula <- step(mdl2)
```

Now we can fit our final model

```r
bestmdl <- lm(mpg ~ cyl + hp + wt + transmission, data = mtcars)
anova(mdl2, bestmdl)
```

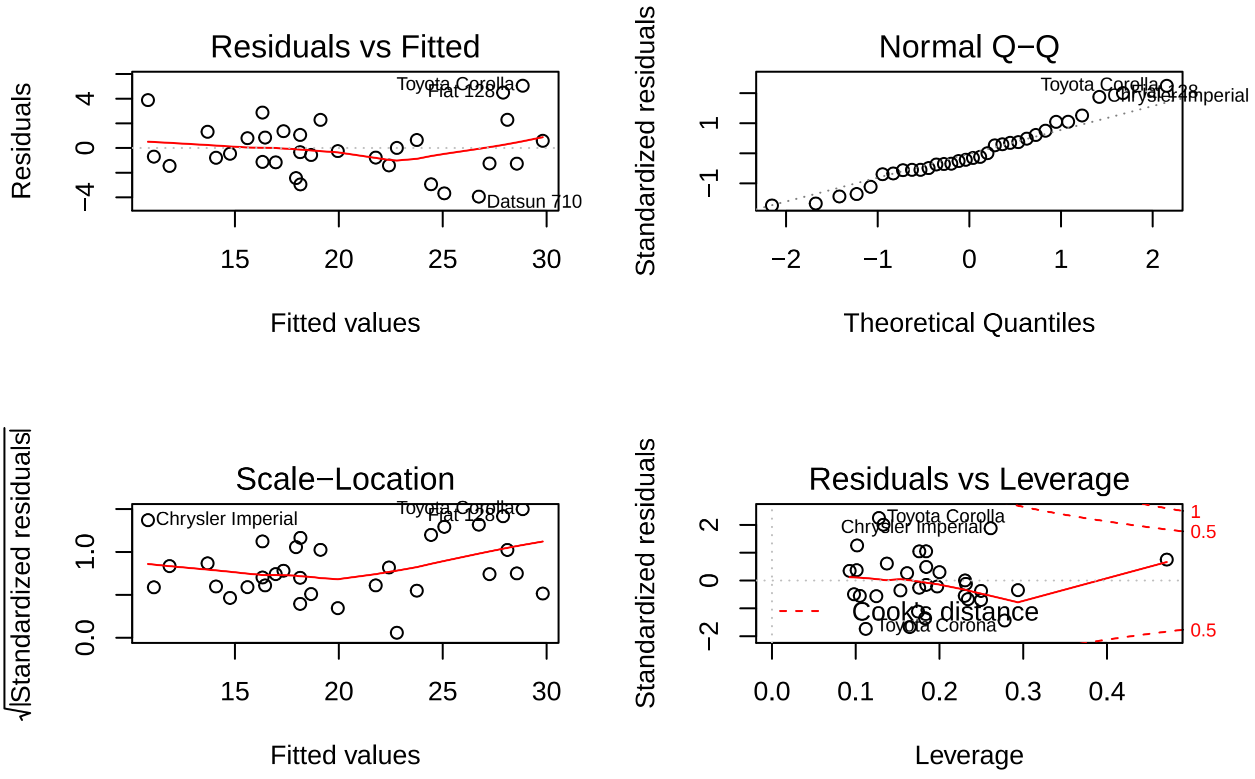
```
## Analysis of Variance Table
##
## Model 1: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + gear + carb +
##     transmission
## Model 2: mpg ~ cyl + hp + wt + transmission
##   Res.Df    RSS  Df Sum of Sq      F Pr(>F)
## 1     15 120.40
## 2     26 151.03 -11   -30.623 0.3468 0.9588
```

```r
summary(bestmdl)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + hp + wt + transmission, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9387 -1.2560 -0.4013  1.1253  5.0513
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        33.70832    2.60489  12.940 7.73e-13 ***
## cyl6               -3.03134    1.40728  -2.154  0.04068 *
## cyl8               -2.16368    2.28425  -0.947  0.35225
## hp                 -0.03211    0.01369  -2.345  0.02693 *
## wt                 -2.49683    0.88559  -2.819  0.00908 **
## transmissionmanual  1.80921    1.39630   1.296  0.20646
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.41 on 26 degrees of freedom
## Multiple R-squared:  0.8659, Adjusted R-squared:  0.8401
## F-statistic: 33.57 on 5 and 26 DF,  p-value: 1.506e-10
```

So this model covers 86% of the variance of mpg with four of all predictors, cyl(number of cylinders), hp(horsepower), weight(of the car) and transmission type. it obviously covers less than our model with all the predictors but it covers more than enough variability. Most of all, Our model p-value is very low, 1.506e-10, meaning our final model is reliable at more than 99%.

```r
par(mfrow = c(2,2))
plot(bestmdl)
```



Conclusion :

- Is an automatic or manual transmission better for MPG?

In our final model, It appears that manual transmission cars are better for MPG compared to automatic cars but less than in our very first model comparing mpg and transmission. Other variables seem more reliable to explain mpg's variability like the number of cylinders, the gross horsepower and the weight of the car.

- Quantify the MPG difference between automatic and manual transmissions :

At first sight, when simply comparing miles per gallon(mpg) to the transmission type, it seems like manual transmission was more efficient for mpg with 7.245 miles more than automatic cars on average. When comparing mpg with the number of cylinders, the gross horspower and the weight, it drops to 1.81.