Motor Trends: Regression Models Course Project

Javier Chang

13/9/2020

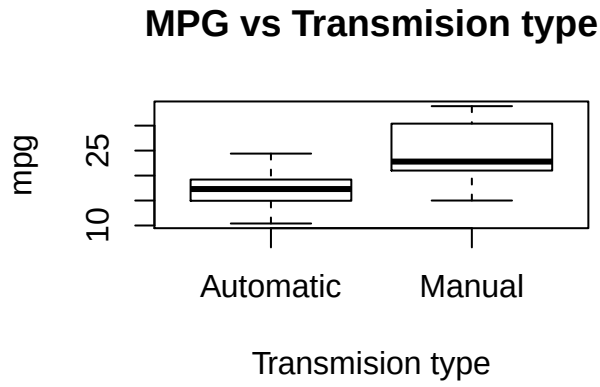
Executive summary

This is the Regresion Models Course Project from Coursera. It is a work for Motor Trend, a magazine about the automobile industry. Looking at a data set of a collection of cars, they were interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). The conclusions of this analysis were as follows:

- Manual transmission is statistically significant better for fuel efficiency (MPG) than automatic transmission, with a significance level of 0.05, comparing engines with the same number of cylinders and horsepower.
- There is a 95% probability that the MPG difference between automatic and manual transmissions is between **1.58** and **6.73** miles per gallon.

Exploratory data analysis

The *mtcars* dataset was extracted from the 1974 Motor Trend US magazine, it comprises fuel efficiency and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models). In the following boxplot graph it can be seen that there is a difference in fuel efficiency (*mpg*) between different types of transmission (*am*). The mean mpg for automatic transmision is **17.15** and the mean mpg for manual transmision is **24.39**.



However, as it can be observed on *Appendix 1* there are several other variables that can influence fuel efficiency, such as the number of cylinders (*cyl*), the gross horse power (*hp*) or the vehicle weight (*wt*), among others.

Model selection

Using a linear regression model between *mpg* and all the dataset variables (Appendix 2), we can observe that although the model is statistically significant (the p-value is less than 0.05), its coefficients are not statistically significant (Pr (> | t | are much larger than 0.05).

So, since Motor Trend is interested only in how a particular variable (am) influences fuel efficiency (mpg), we are going to use a *stepwise selection process with forward selection strategy* to select the relevant variables for the model. It is going to start with the variable *am*, then iteratively adds the most contributive predictors based on p values, and stops when the improvement is no longer statistically significant. According to the analysis of variance (anova) we select model 3, which is the last to have a Pr (> F) less than 0.01:

mpg ~ am + cyl + hp

Linear regression analysis for the selected model

The linear model for estimating fuel efficiency (mpg) based on the transmission type (factor am), number of cylinders (factor cyl) and gross horsepower (hp) is as follow:

mpg = *b0* + (*b1* * am1) + (*b2* * cyl6) + (*b3* * cyl8) + (*b4* * hp)

The linear regression results are shown in *Appendix 4*, where we observe the following:

- The model is statistically significant. The p-value of the F-statistic is **0.0026598**, which is highly significant (it is much less than 0.05), so we reject the null hyphotesis. It means that at least one predictor variable is significantly related to the outcome variable (mpg).
- All of the coefficients are significant, except *cyl8*. The Pr(>|t|) values are much less than 0.05 except for *cyl8*.
- The model is useful. The variance explained by the model (R-squared) is **82%** which is good enough for our purpose.
- The data satisfies all the assumptions for linear regression. As it can be seen in Appendix 4.2 there are no significant outliers (Residuals vs Fitted), the residuals are approximately normally distributed (Q-Q plot), the data shows homoscedasticity (Scale-Location, sqrt of standarized residuals are less than 1.5) and no auto correlation is observed (variance inflation factors are less than 2 in Appendix 4.3).

In conclusion, the linear model between *mpg* and *am + cyl + hp* fits well to answer the question if an automatic or manual transmission is better for MPG.

Model interpretation

- Is an automatic or manual transmission better for MPG?

As *b1* p-value is **4.1578565** we reject the null hypothesis (*b1* = 0). So we can conclude that manual transmission is statistical significant better than automatic transmission for fuel efficiency (mpg).

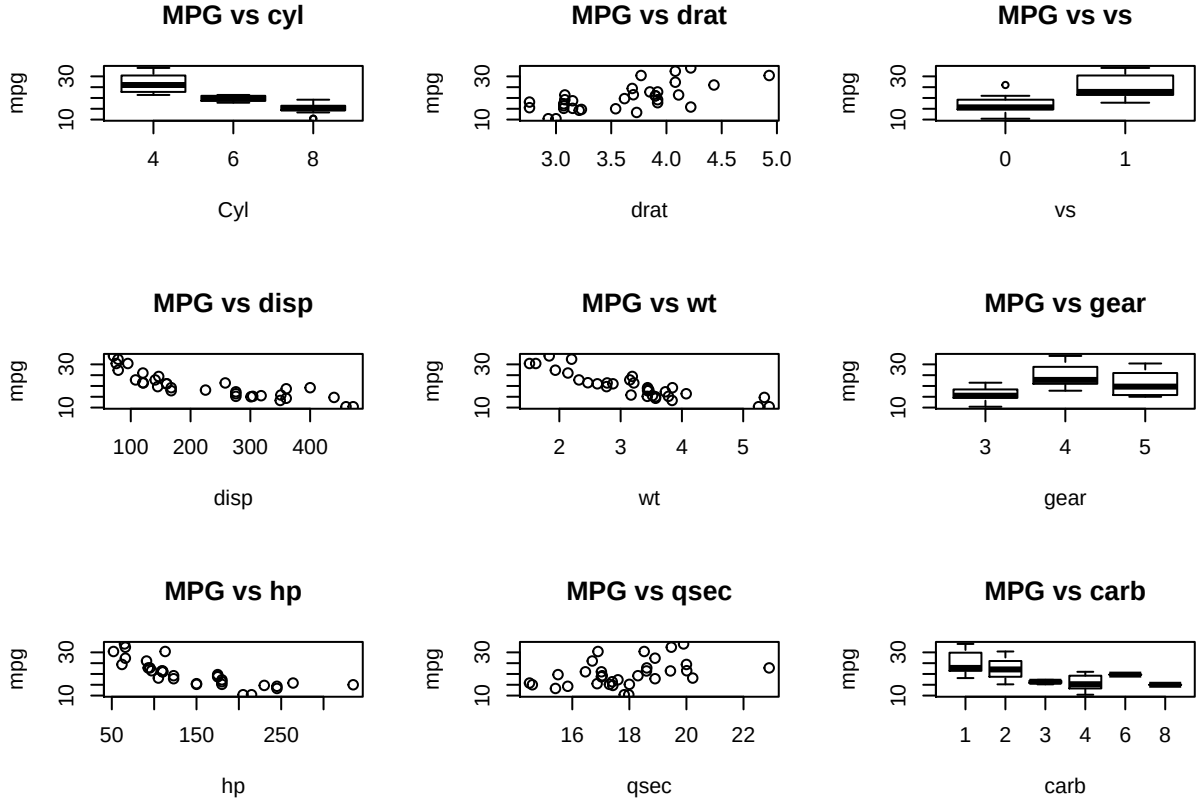
- Quantify the MPG difference between automatic and manual transmissions

The difference in *mpg* between automatic and manual transmission is between the 95% confidence interval of *b1* [**1.5796288**, **6.7360842**].

Appendix

Appendix 1 Exploratory data analysis

In the following graph we can see the relationship between mpg and the rest of the variables of the mtcars dataset



Appendix 2 Linear regression model: Fuel efficiency (mpg) vs all variables

```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars2)
##
## 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
## am1          1.21212    3.21355   0.377   0.7113
## 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
## ---
## 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
```

Appendix 3 Analysis of variance (Anova) of the models selected by stepwise process

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + cyl
## Model 3: mpg ~ am + cyl + hp
## Model 4: mpg ~ am + cyl + hp + wt
## Model 5: mpg ~ am + cyl + hp + wt + disp
## Model 6: mpg ~ am + cyl + hp + wt + disp + vs
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     28 264.50  2    456.40 38.2752 3.419e-08 ***
## 3     27 197.20  1     67.30 11.2874  0.002603 **
## 4     26 151.03  1     46.17  7.7445  0.010331 *
## 5     25 150.41  1      0.62  0.1035  0.750511
## 6     24 143.09  1      7.32  1.2275  0.278870
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

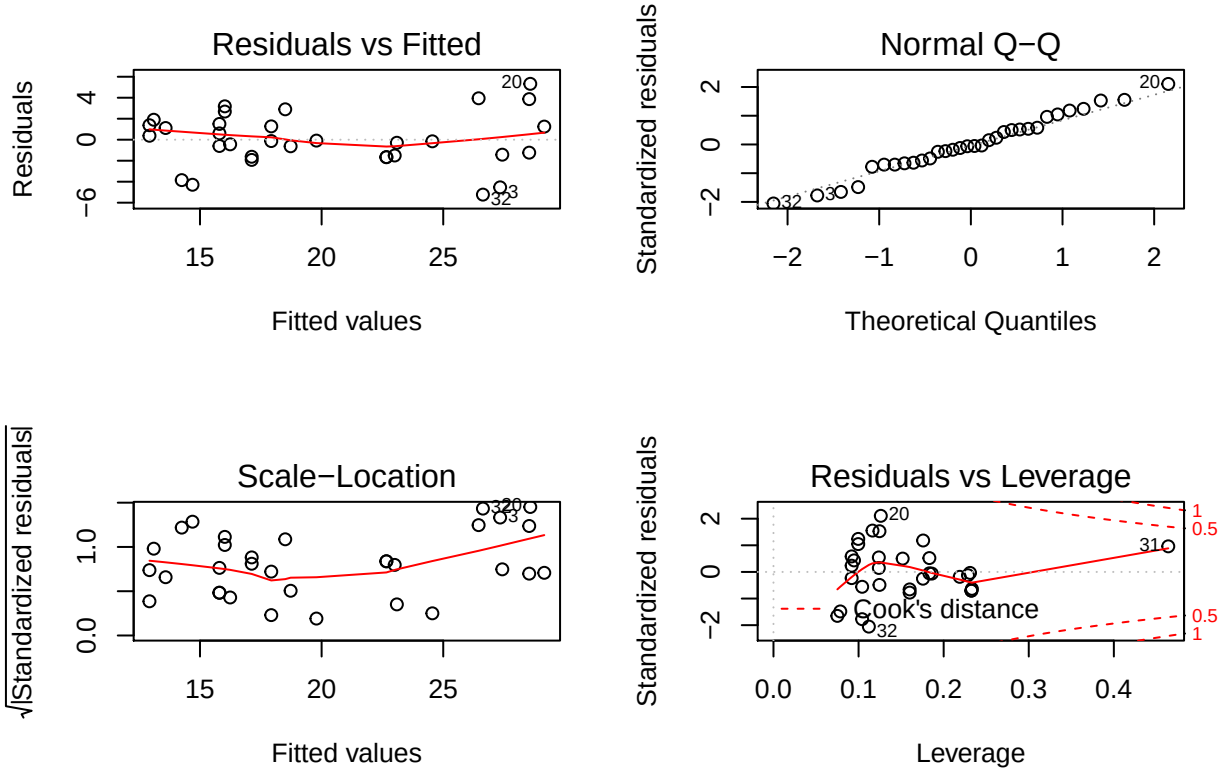
Appendix 4 Linear regression of the selected model (mpg ~ am + cyl + hp)

4.1 Linear regression model

```
##
## Call:
## lm(formula = mpg ~ am + cyl + hp, data = mtcars2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.231 -1.535 -0.141  1.408  5.322
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 27.29590    1.42394  19.169  < 2e-16 ***
## am1          4.15786    1.25655   3.309  0.00266 **
## cyl6        -3.92458    1.53751  -2.553  0.01666 *
## cyl8        -3.53341    2.50279  -1.412  0.16943
## hp          -0.04424    0.01458  -3.035  0.00527 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.703 on 27 degrees of freedom
## Multiple R-squared:  0.8249, Adjusted R-squared:  0.7989
## F-statistic: 31.79 on 4 and 27 DF,  p-value: 7.401e-10
```

4.2 Residual analysis



4.3 Variance inflation factors

```
##         GVIF Df GVIF^(1/(2*Df))
## am  1.668652  1        1.291763
## cyl 5.486920  2        1.530496
## hp  4.238903  1        2.058860
```

Appendix 5 Code for this report

```
knitr::opts_chunk$set(echo = FALSE); library(dplyr); library(olsrr); library(car)
## Load dataset and transform factor variables
data(mtcars);
mtcars2 <- mtcars %>%
        mutate(am=factor(am), cyl=factor(cyl),vs=factor(vs), gear=factor(gear), carb=factor(carb)) %>%
        select(mpg, am, cyl, disp, hp, drat, wt, qsec, vs, gear, carb)

## Boxplot mgp vs am
plot(mtcars2$am, mtcars2$mpg, main="MPG vs Transmision type", xlab="Transmision type", ylab="mpg",
     names=c("Automatic", "Manual"))
## Models generation
fitall <- lm(mpg ~ ., data=mtcars2)
fit1 <- lm(mpg ~ am, data=mtcars2)
fit2 <- lm(mpg ~ am + cyl, data=mtcars2)
fit3 <- lm(mpg ~ am + cyl + hp, data=mtcars2)
fit4 <- lm(mpg ~ am + cyl + hp + wt, data=mtcars2)
fit5 <- lm(mpg ~ am + cyl + hp + wt + disp, data=mtcars2)
fit6 <- lm(mpg ~ am + cyl + hp + wt + disp + vs, data=mtcars2)
## Plot mpg vs the rest of variables
par(mfcol = c(3, 3))
plot(factor(mtcars$cyl), mtcars$mpg, main="MPG vs cyl", xlab="Cyl", ylab="mpg")
plot(mtcars$disp, mtcars$mpg, main="MPG vs disp", xlab="disp", ylab="mpg")
plot(mtcars$hp, mtcars$mpg, main="MPG vs hp", xlab="hp", ylab="mpg")
plot(mtcars$drat, mtcars$mpg, main="MPG vs drat", xlab="drat", ylab="mpg")
plot(mtcars$wt, mtcars$mpg, main="MPG vs wt", xlab="wt", ylab="mpg")
plot(mtcars$qsec, mtcars$mpg, main="MPG vs qsec", xlab="qsec", ylab="mpg")
plot(factor(mtcars$vs), mtcars$mpg, main="MPG vs vs", xlab="vs", ylab="mpg")
plot(factor(mtcars$gear), mtcars$mpg, main="MPG vs gear", xlab="gear", ylab="mpg")
plot(factor(mtcars$carb), mtcars$mpg, main="MPG vs carb",xlab="carb", ylab="mpg")
## Shows summary of mpg vs all variables
summary(fitall)
## Analysis of variance
anova(fit1, fit2, fit3, fit4, fit5, fit6)
## Shows selected model
summary(fit3)
## Residual analysis
par(mfrow = c(2,2))
plot(fit3)
vif(fit3)
```