Regression Models Course Project

Kike Sedes

MrCars Data Analysis.

Instructions

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:

1. Is an automatic or manual transmission better for MPG
2. Quantify the MPG difference between automatic and manual transmissions

Exploratory Analysis

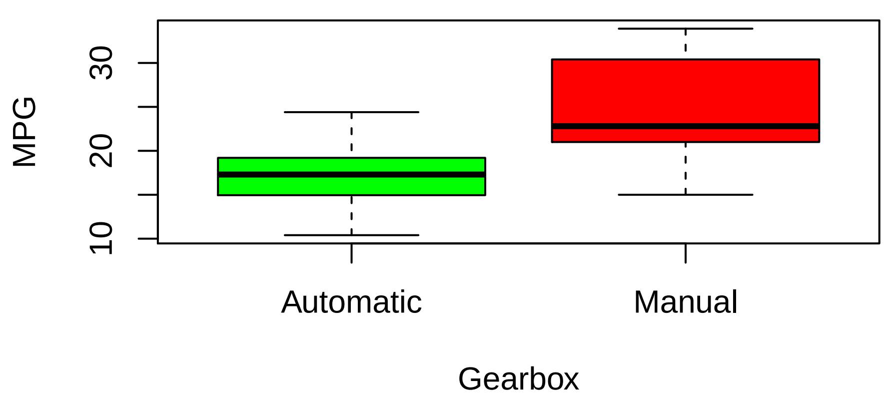
Let's start by loading the mtcars dataset and perform some basic exploratory analysis

```
head(mtcars,3)
```

```
##                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
```

As the am column is not a factor, let's create a new variable for type of gearbox

```
mtcars$vs <- factor(mtcars$vs)
mtcars$gearbox <- factor(mtcars$am, labels=c("Automatic","Manual"))
boxplot(mpg ~ gearbox, data = mtcars, col = (c("green","red")), ylab = "MPG", xlab = "Gearbox")
```



```r
means<-mtcars %>%
  group_by(gearbox) %>%
  summarise(avg_mpg = mean(mpg))
means
```

```
## # A tibble: 2 x 2
##   gearbox   avg_mpg
##   <fct>       <dbl>
## 1 Automatic    17.1
## 2 Manual       24.4
```

According to our exploratory analysis, data leads us to think that vehicles with a manual transmission provide higher MPG

Regression Analysis

To start with, let's check our basic exploratory analysis. Is zero the difference between Manual and Automatic Transmission (H0)?

```r
model_basic <- lm(mpg ~ gearbox, data=mtcars)
summary(model_basic)
```

```
##
## Call:
## lm(formula = mpg ~ gearbox, 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 ***
## gearboxManual    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
```

Looking at the summary, T-Test rejects the null hypothesis H0 (p-valor is very small), so type of transmission plays a role in the regression. Nevertheless, R-squared is only 0,35, that means that there are other variables that we are not taking into account (only a 35% of variance in MPG is associated with gearbox). This is why we need a more complex linear model. Let's explore it:

```r
data(mtcars)
model_allVars <- aov(mpg ~ ., data = mtcars)
summary(model_allVars)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## cyl          1  817.7   817.7 116.425 5.03e-10 ***
```

```
## disp         1   37.6    37.6   5.353  0.03091 *
## hp           1    9.4     9.4   1.334  0.26103
## drat         1   16.5    16.5   2.345  0.14064
## wt           1   77.5    77.5  11.031  0.00324 **
## qsec         1    3.9     3.9   0.562  0.46166
## vs           1    0.1     0.1   0.018  0.89317
## am           1   14.5    14.5   2.061  0.16586
## gear         1    1.0     1.0   0.138  0.71365
## carb         1    0.4     0.4   0.058  0.81218
## Residuals   21  147.5     7.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
cor(mtcars)[1,]
```

```
##        mpg        cyl       disp         hp       drat         wt
##  1.0000000 -0.8521620 -0.8475514 -0.7761684  0.6811719 -0.8676594
##       qsec         vs         am       gear       carb
##  0.4186840  0.6640389  0.5998324  0.4802848 -0.5509251
```

Obtained p-values suggest that we should also consider cyl, disp and wt in our model. Moreover, cor functions introduces horsepower ("hp") as another relevant variable. Let's create a new model wiht these variables and check if R-squared increases compared to our first model (mpg ~ gearbox)

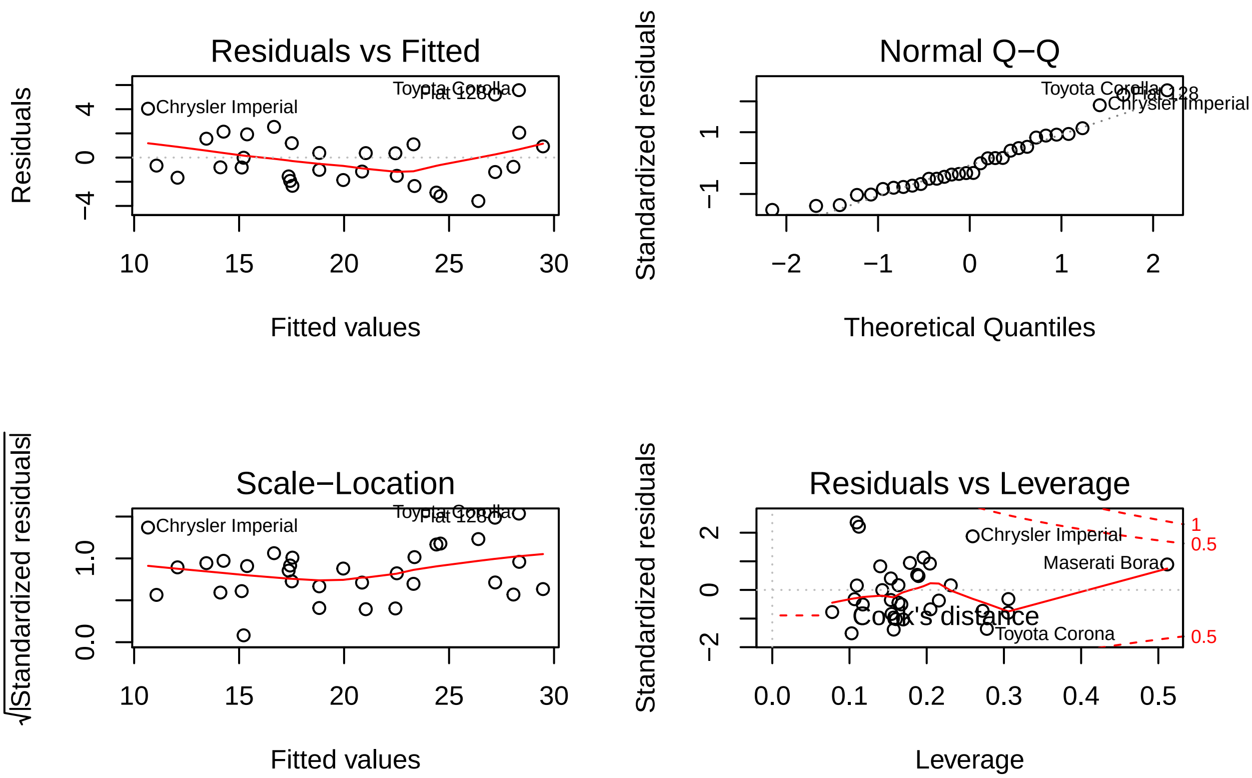
```r
model_selected <- lm(mpg ~ cyl + disp + wt + hp + am, data = mtcars)
summary(model_selected)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + disp + wt + hp + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5952 -1.5864 -0.7157  1.2821  5.5725
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 38.20280    3.66910  10.412 9.08e-11 ***
## cyl         -1.10638    0.67636  -1.636  0.11393
## disp         0.01226    0.01171   1.047  0.30472
## wt          -3.30262    1.13364  -2.913  0.00726 **
## hp          -0.02796    0.01392  -2.008  0.05510 .
## am           1.55649    1.44054   1.080  0.28984
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.505 on 26 degrees of freedom
## Multiple R-squared:  0.8551, Adjusted R-squared:  0.8273
## F-statistic:  30.7 on 5 and 26 DF,  p-value: 4.029e-10
```

With our new model, R-squared is 0,85, so the model has clearly been improved. On the other hand, the negative values in the coefficients of cyl, wt and hp show that the more power (or the more cyl, more weight) the less mpg.

Finally, let's check the ressiduals of our model:

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



In these figures, we can see:

1. Plots 1 & 3 confirms the constant variance assumption. Homocedastic
2. Normal Q-Q plot shows that the distribution of residuals is normal (close to the line)
3. There are no significant outliers.

Conclusion

According to the plots and tests we have performed, we can conclude that there is a difference in MPG based on transmission type. A manual transmission will provide a higher MPG range. Nevertheless, it seems that number of cylinders, horsepower, and weight play a more important role when determining MPG.