Regression Model - Course Project

Juan Agustín Morello

30/8/2020

Summary

In this course project I'll be looking at `Motor Trend Car Road Tests`, also known as `mtcars`, a data set of a collection of cars extracted from 1974 *Motor Trend US* magazine. Comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles. I'm interested in exploring the relationship between a set of variables and miles per gallon. I'll be trying to answer the following:

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

After doing an analysis, I conclude that while seems that manual transmission will yield better miles per gallon when compared with Automatic (on average, a manual car seems to achieve 24 mpg, versus 17 mpg for automatics), after fitting a Multiple Linear Regression, the analysis showed that both type of transmissions contributed negligibly to MPG (nor automatic nor manual transmission are better for MPG), having other variables (weight, displacement, and number of cylinders) a more significant correlation.

Exploratory Data Analysis

```
# Loading the dataset
data(mtcars)
# Showing the first 5 rows in dataset
head(mtcars)
```

```
##                   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
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

Data dictionary

1. `mpg` - Miles per gallon
2. `cyl` - Number of cylinders
3. `disp` - Displacement (cu.in.)
4. `hp` - Gross horsepower

5. `drat` - Rear axle ratio
6. `wt` - Weight (1000 lbs)
7. `qsec` - 1/4 mile time
8. `vs` - Engine (0 = V-shaped, 1 = straight)
9. `am` - Transmission (0 = automatic, 1 = manual)
10. `gear` - Number of forward gears
11. `carb` - Number of carburetors

Some data transformation

```
# Transform some variables from numeric type to factor
# Engine ('vs') and Transmision ('am')  (are binomial)
# Cylinders ('cyl'), Gears ('gear') and Carburetors ('carb')

mtcars$cyl <- factor(mtcars$cyl)
mtcars$carb <- factor(mtcars$carb)
mtcars$vs <- factor(mtcars$vs)
mtcars$gear <- factor(mtcars$gear)
# For more understanding, I'll change 0 and 1 to "Automatic" and "Manual"
mtcars$am <- factor(mtcars$am, labels=c("Automatic","Manual"))
```

Data summary

```
summary(mtcars)
```

```
##       mpg        cyl         disp             hp             drat
##  Min.   :10.40   4:11   Min.   : 71.1   Min.   : 52.0   Min.   :2.760
##  1st Qu.:15.43   6: 7   1st Qu.:120.8   1st Qu.: 96.5   1st Qu.:3.080
##  Median :19.20   8:14   Median :196.3   Median :123.0   Median :3.695
##  Mean   :20.09          Mean   :230.7   Mean   :146.7   Mean   :3.597
##  3rd Qu.:22.80          3rd Qu.:326.0   3rd Qu.:180.0   3rd Qu.:3.920
##  Max.   :33.90          Max.   :472.0   Max.   :335.0   Max.   :4.930
##        wt             qsec       vs             am     gear   carb
##  Min.   :1.513   Min.   :14.50   0:18   Automatic:19   3:15   1: 7
##  1st Qu.:2.581   1st Qu.:16.89   1:14   Manual   :13   4:12   2:10
##  Median :3.325   Median :17.71                         5: 5   3: 3
##  Mean   :3.217   Mean   :17.85                                4:10
##  3rd Qu.:3.610   3rd Qu.:18.90                                6: 1
##  Max.   :5.424   Max.   :22.90                                8: 1
```

If we check the relationship between Miles Per Gallon and Transmission (see Appendix, "Plot 1"), we'll see that Automatic transmissions have a lower MPG than Manual transmissions.

```
automaticMean <- mean(mtcars$mpg[mtcars$am=="Automatic"])
manualMean <- mean(mtcars$mpg[mtcars$am=="Manual"])
data.frame(automatic = automaticMean,
           manual = manualMean,
           difference = manualMean - automaticMean)
```

```
##   automatic   manual difference
## 1  17.14737 24.39231   7.244939
```

On average, manual transmission have 7.24 mpg more than automatic transmission. We'll check this fact doing a Regression analysis.

Regression Analysis

Single Linear Regression

Our immediate objective is to see the relationship between two variables: `mpg` (outcome) and `am` (predictor). So, the ideal regression model to use in this case is a Single Linear Regression. We'll make a model and check it:

```
lm1 <- lm(mpg ~ am, data = mtcars)
summary(lm1)
```

```
##
## Call:
## lm(formula = mpg ~ am, 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 ***
## amManual       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
```

Here we can see that between Automatic and Manual transmissions there is a difference (on average) of 7.25 `mpg` in favor to Manual. But as the R-squared value is 36% and that tells us that the model only explains that that specific percentage of the variance in `mpg` can be attributed to the transmission variable alone. So, I'll have to put aside this Single Linear Regression and look for other variables in the data that could help us develop a Multiple Linear Regression more robust.

Multiple Linear Regression

To have a more accurate model I'll have to look for other variables that are correlated to `mpg`.

```
# I am interested more in analysis of variance than making a linear model.
# The 'aov' wrapper function helps
analysis <- aov(mpg ~ ., data = mtcars)
summary(analysis)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## cyl          2  824.8   412.4  51.377 1.94e-07 ***
## disp         1   57.6    57.6   7.181   0.0171 *
## hp           1   18.5    18.5   2.305   0.1497
## drat         1   11.9    11.9   1.484   0.2419
## wt           1   55.8    55.8   6.950   0.0187 *
## qsec         1    1.5     1.5   0.190   0.6692
## vs           1    0.3     0.3   0.038   0.8488
## am           1   16.6    16.6   2.064   0.1714
## gear         2    5.0     2.5   0.313   0.7361
## carb         5   13.6     2.7   0.339   0.8814
## Residuals   15  120.4     8.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In this summary, We can see that the more significant variables, looking at those that have a P-value < 0.05, are `cyl`, `disp`, `wt`. Those variables have the strongest correlation with `mpg`.For the sake of it, I'll add the `hp` variable because, compared with the leftover variables, has the lowest P-value. I'll build a new model using these variables plus `am` (I relate `cyl` and `am` because both are factors) .

```
lm2 <- lm(mpg ~ cyl*am + disp + wt + hp, data = mtcars)
summary(lm2)
```

```
##
## Call:
## lm(formula = mpg ~ cyl * am + disp + wt + hp, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2422 -1.3508 -0.2468  1.3024  4.8152
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   32.82228    3.06747  10.700  2.1e-10 ***
## cyl6          -1.81302    2.06118  -0.880   0.3882
## cyl8          -1.94504    3.10378  -0.627   0.5370
## amManual       2.87997    1.89552   1.519   0.1423
## disp           0.00254    0.01317   0.193   0.8487
## wt            -2.59279    1.23073  -2.107   0.0463 *
## hp            -0.03139    0.01817  -1.728   0.0974 .
## cyl6:amManual -2.44037    2.60453  -0.937   0.3585
## cyl8:amManual -1.04684    3.19692  -0.327   0.7463
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.51 on 23 degrees of freedom
## Multiple R-squared:  0.8714, Adjusted R-squared:  0.8266
## F-statistic: 19.47 on 8 and 23 DF,  p-value: 1.6e-08
```

The summary tell us that adding `cyl`, `disp`, `wt` and `hp`, being their correlation with `mpg` more significant, affected the correlation between `mpg` and `am`. The model explains 86% of the variance of `mpg`.

The coefficients conclude that :

- Having 4 cylinders, manual transmission has 2.8 `mpg` than automatic transmission.
- Increasing the number of cylinders from 4 to 6 decreases `mpg` by 1.8 having automatic transmission and decreases by 2.4 having manual transmission.
- Increasing the cylinders to 8 decreases `mpg` by 1.94 and decreases by 1 having manual transmission.
- Seems that increasing the displacement doesn't change `mpg`
- Seems that increasing the horsepower decreases `mpg` 0.3 per every 100 horsepower.
- Seems that Weight decreases `mpg` by 2.6 for each 1000 lb increase.

In conclusion, we cannot hold that the difference between automatic and manual transmission is of 7.24 mpg. The Multiple Linear Regression show us that factors as weight, displacement, horsepower and number of cylinders make the true difference in MPG consumption in automobiles.

Manual transmission has 2.8 higher mpg than automatic transmission when there are 4 cylinders. When there are 6 cylinders, there is a 0.7 mpg difference in favor of automatic transmission; and when there are 8 cylinders, there is a 1 mpg difference in favor of manual transmission. Those differences are almost negligible: transmission do not influences in the Miles Per Gallon of automobiles.

Diagnostics

Comparing Regression Models

I'll compare the Single and Multiple Linear Regressions with the `anova` function looking at the P-value and see what model is significantly better. I'll create a new Multiple Linear Regression with the same variables of `lm2` but here I'll not consider the relationship between `cyl` and `am` to show this influences in the analysis.

```
lm3 <- lm(mpg ~ cyl + disp + wt + hp + am, data = mtcars)

anova(lm1, lm2, lm3)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ cyl * am + disp + wt + hp
## Model 3: mpg ~ cyl + disp + wt + hp + am
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     23 144.85  7    576.04 13.0666 1.112e-06 ***
## 3     25 150.41 -2     -5.56  0.4412    0.6486
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

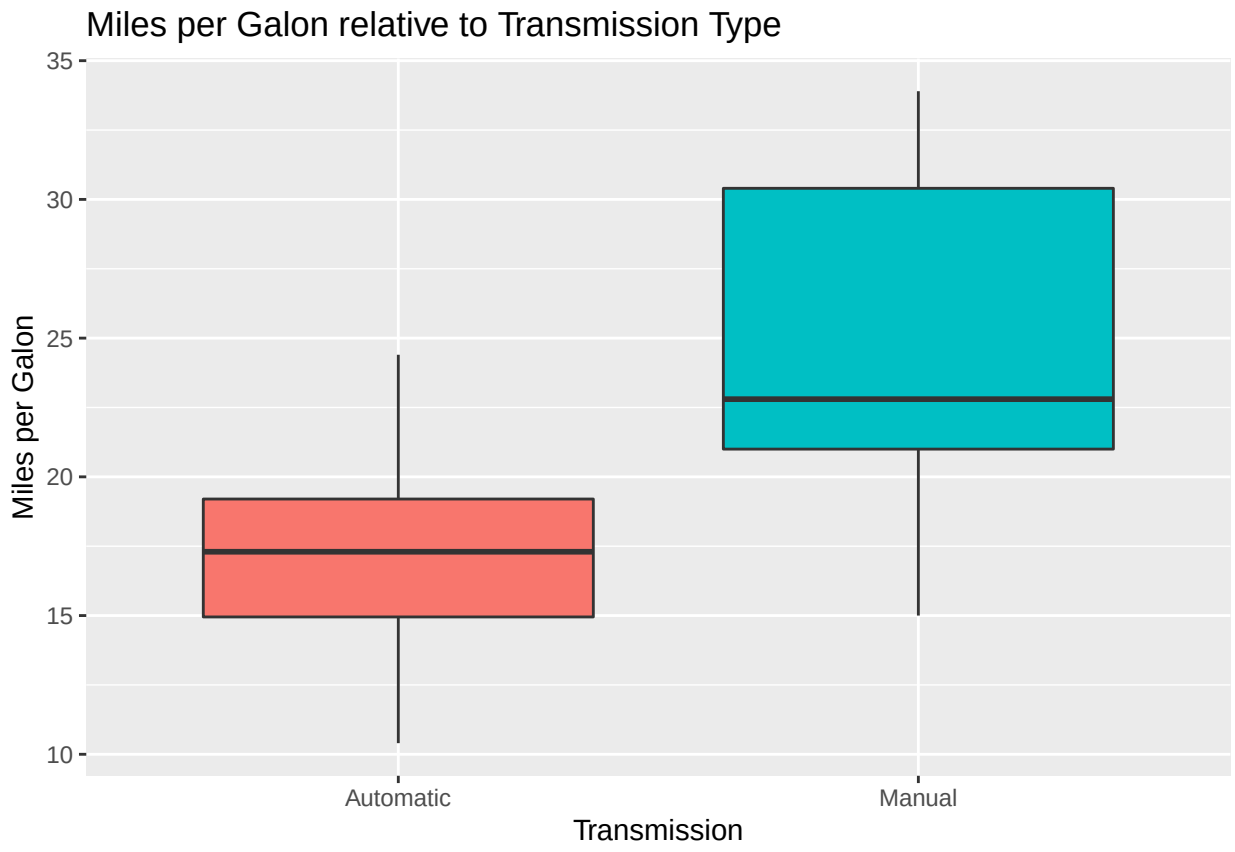
Residuals

Also, I'll make some diagnostic test on `lm2` model: I'll check the residuals of the model for non-normality (Appendix - "Plot 2").

Appendix

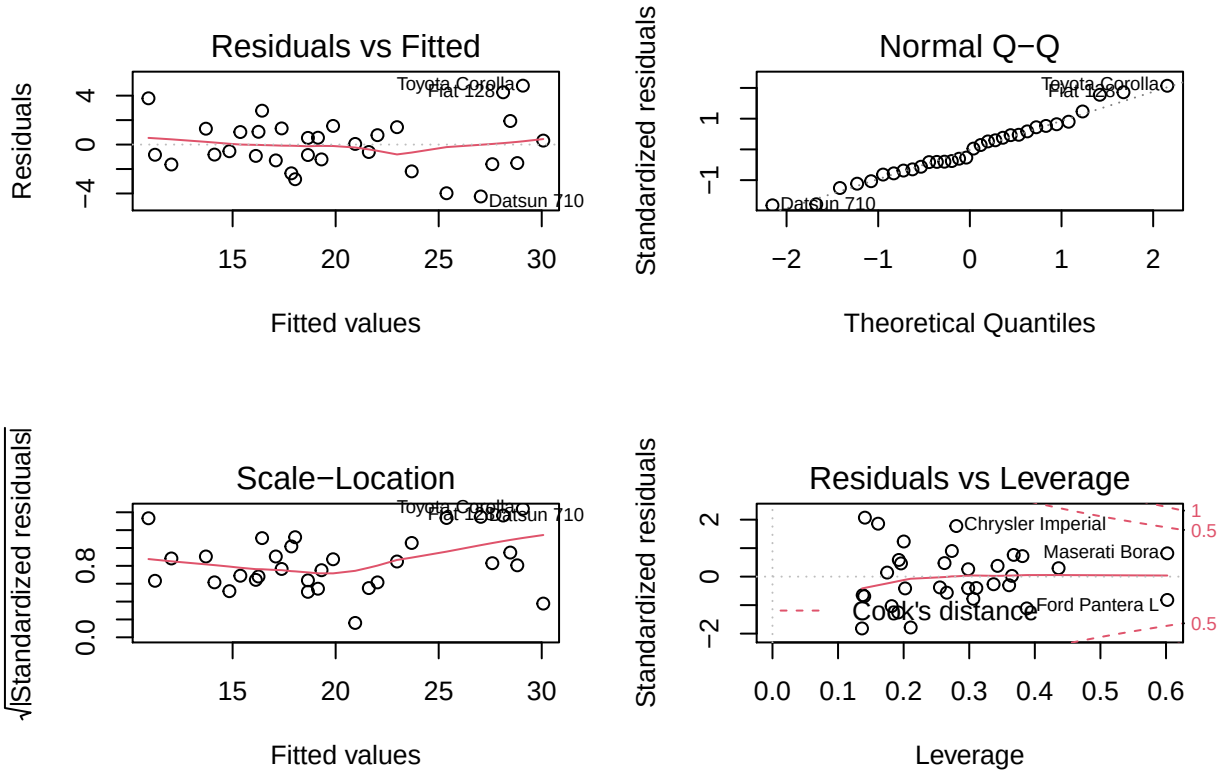
Plot n°1

```r
plot1 <- ggplot(data=mtcars,
                aes(x=am, y=mpg))
plot1 <- plot1 + geom_boxplot(aes(fill=am))
plot1 <- plot1 + labs(title="Miles per Galon relative to Transmission Type",
                      x="Transmission",
                      y="Miles per Galon")
plot1 <- plot1 + theme(legend.position = "none")
plot1
```



Plot n°2

```r
par(mfrow = c(2, 2))
plot(lm2)
```



The previous plots are are diagnostic of the residuals of the `lm2` model.

- **Residuals vs Fitted:** supports assumption of independence (Homoscedasticity)
- **Normal Q-Q:** The points follow closely the line concluding that residuals are normally distributed.
- **Scale-Location:** the random distribution confirms the constant variance assumption
- **Residuals vs Leverage:** Since all points are within the 0.5 bands, the conclusion is that there are no outliers.