Regression Models - Motor Trend Project

Yadder Aceituno

March 24, 2019

Overview

This report shows the analysis in mtcars dataset. The objective is to answer the next two questions:

- Is an automatic or manual transmission better for MPG (Miles per gallon)?
- What is the MPG difference between automatic and manual transmissions?

To reach it, first we will examine what kind of information the dataset have. Then, we will create a first model to determine what is the relationship between the MPG and the type of transmission (automatic or manual). Having the model we will determine if we need to create another model using anothers columns from the dataset.

Loading data and taking a first look.

For this analysis I will use the "mtcars" data. Let's load it and view a summary of the data.

```
data("mtcars");
#Creating factor variable for am.
mtcars$am <- factor(mtcars$am, levels = c(0, 1), labels = c("Automatic", "Manual"))
```

The columns that we are interested are "mpg"" and "am". Let's look some rows of these columns.

Visual Exploratory Analysis

Let's create a boxplot to see how is the relation between the type of transmission and mpg (miles per gallon). With this plot we can analyze how will be the model that will be created. Look Appendix 1.

Taking a look with our plot we can see that there is a difference about 7 miles per gallon between the means from the two different transmissions. Also, we see that the manual transmission is better for mpg.

Let's confirm the difference between the means:

```
mean(mtcars[mtcars$am == "Manual",]$mpg) - mean(mtcars[mtcars$am == "Automatic",]$mpg)
```

```
## [1] 7.244939
```

We can conclude that manual transmission is better than automatic transmission, and the difference is 7.24 miles per gallon. But we won't because we don't know if there is another variable which could explain with a better way the mpg variable. We need to create regression models to check it.

Simple Regression Analysis

We create our model with:

```
mdl1 <- lm(mpg~am, data = mtcars);
```

The variables mdl1 has the model. Let's take a summary of the model.

```r
summary(mdl1)$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
```

As we know, our model is defined by the formula $mpg = \beta_0 + \beta_1 am$, where:

- mpg means miles per gallon.
- am means automatic or manual transmission (1 = Manual, 0 = Automatic)
- β_0 means the average for mpg when transmission is automatic
- β_1 means the increment of mpg compared with β_0 when transmission is manual.

Also, we can see that the value of R^2 is 0.3598, which means that the model cover 36% of the variance. As result, we need to create another model including more variables of the data set.

Identifying Variables

To indentify the variables to include in our new model, we need to analyze how much related are all the variables in the dataset with the variable called mpg. To see it, we will use the tool pairs. Look appendix 2

As we could see in the previous plot from appendix 2, the variables with a strongest relation with mpg are:

- cyl: Number of cylinders
- disp: Engine displacement
- hp: Horsepower
- wt: Weight

We can confirm it seeing the correlation valueS between the variables from the dataset.

```r
# Correlation values for mpg with another variables from the same dataset.
round(cor(mtcars[,names(mtcars) != "am"])[1,],2)
```

```
##   mpg   cyl  disp    hp  drat    wt  qsec    vs  gear  carb
##  1.00 -0.85 -0.85 -0.78  0.68 -0.87  0.42  0.66  0.48 -0.55
```

We could see that the mentioned variables (cyl, disp, hp and wt) have values above 80%. We can even include the variable drat(rear axle ratio) because have better correlation than am(automatic/manual transmision) but we won't because could cause overfitting in our model.

Multiple Regression Model

Due we already have the variables to use, let's create our model and see the summary from this new one.

```r
mdl2 <- lm(mpg ~ am + cyl + disp + hp + wt, data = mtcars)

summary(mdl2)$coef
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 38.20279869 3.66909647 10.412045 9.084987e-11
## amManual     1.55649163 1.44053603  1.080495 2.898430e-01
## cyl         -1.10637984 0.67635506 -1.635797 1.139322e-01
## disp         0.01225708 0.01170645  1.047036 3.047194e-01
## hp          -0.02796002 0.01392172 -2.008374 5.509659e-02
## wt          -3.30262301 1.13364263 -2.913284 7.256888e-03
```

We can see with this new model that the value of R^2 is 0.8551, which means that the model cover about 86% of the variance. Definitely the new model is better than the first one.

Comparing Models

We can compare our models using ANOVA function.

```
anova(mdl1, mdl2)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + cyl + disp + hp + wt
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     26 163.12  4    557.78 22.226 4.507e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see that RSS (Residual Sum of Squares) value for model 2 is better than RSS vsalue for model 1. And we can see that p-value is 4.504 * 10-8. We conclude that the other variables help to explain the model.

Now let's check the normality condition from residuals and some outliers from our model. Look appendix 3.

Checking the "Residuals vs Fitted" plot we see that our residuals are randomly distributed so that our independence condition is correct.

The "Normal QQ" plot show us the normality of the residuals. That's correct.

We can see the constant variance conditios looking the "Scale Location" plot because they are scattered but they are following a pattern.

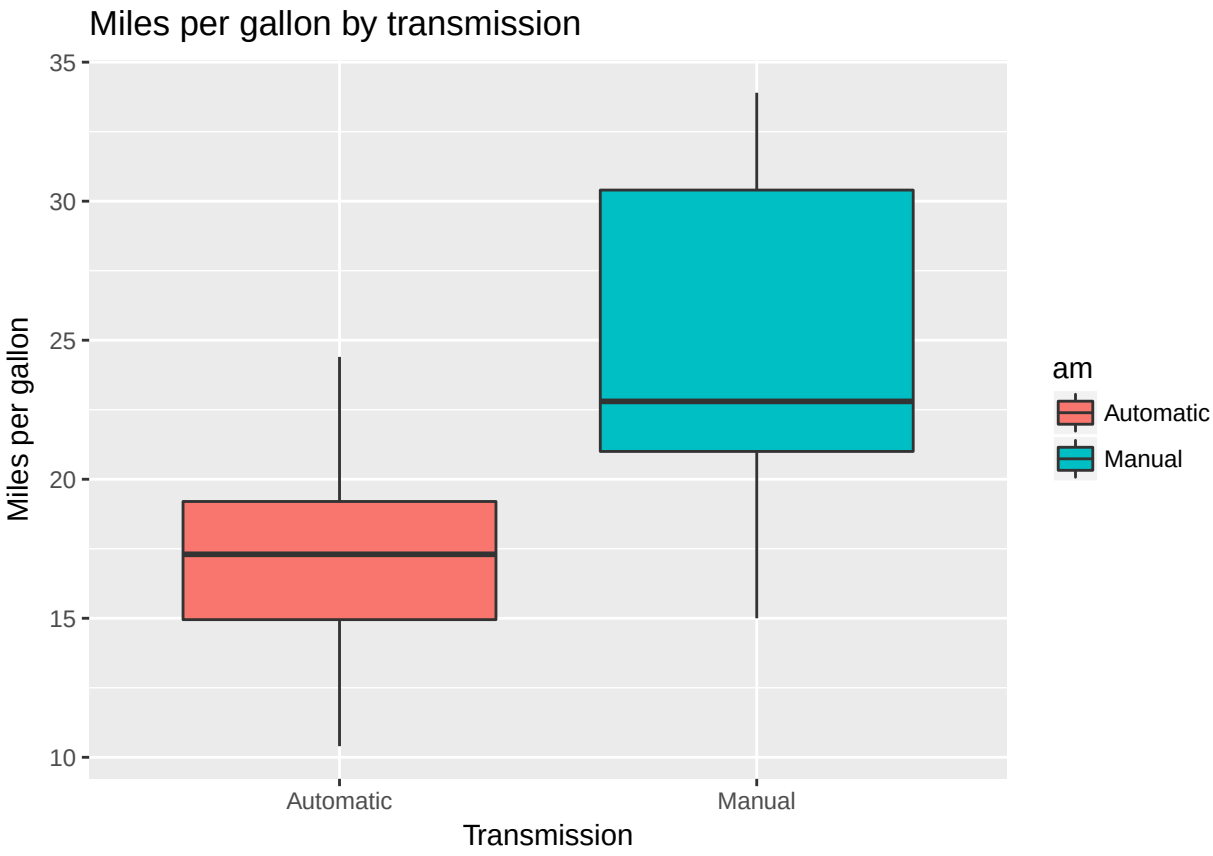
Conclusions:

- Manual transmission is better than automatic transmission for miles per gallon.
- The type of transmission doesn't explain the miles per gallon variable according the dataset.
- Including another variables (cylinders, horse power, engine displacement and weight), manual transmission increase 1.56 miles per gallon than automatic transmission.

Appendix:

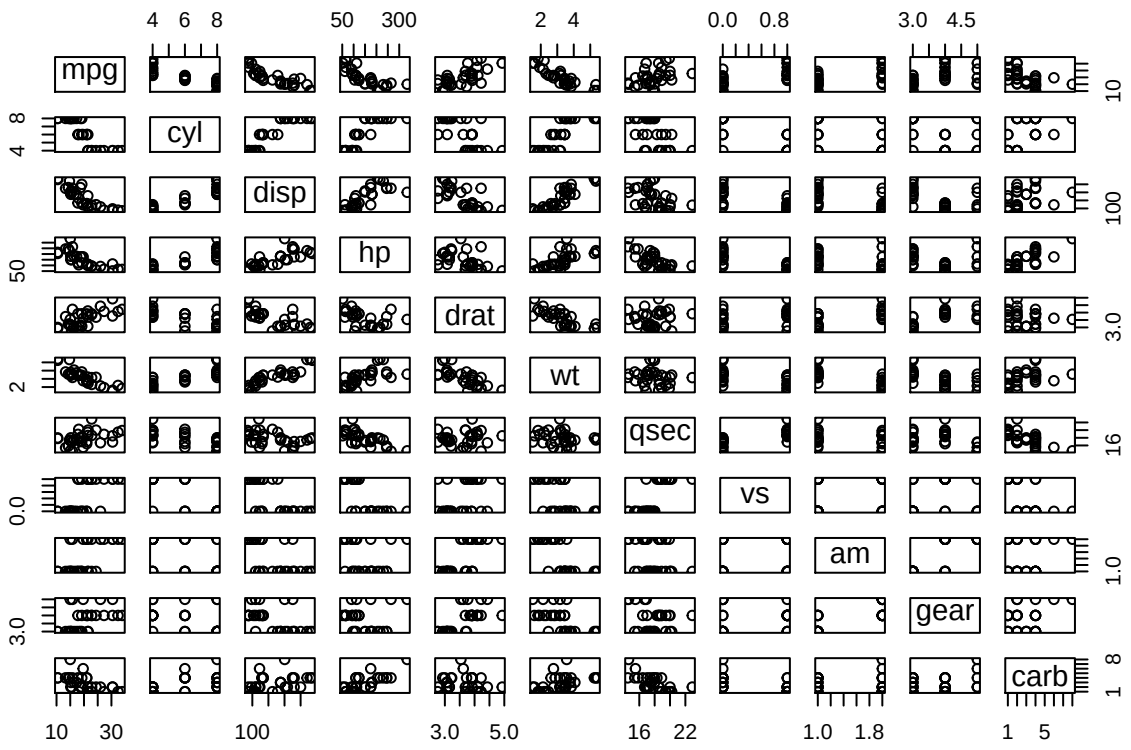
1. Plot: Miles per gallon by transsmission

```
p <- ggplot(mtcars, aes(x = am, y = mpg, fill = am))
p + geom_boxplot() + labs(title = "Miles per gallon by transmission", x = "Transmission", y = "Miles pe
```



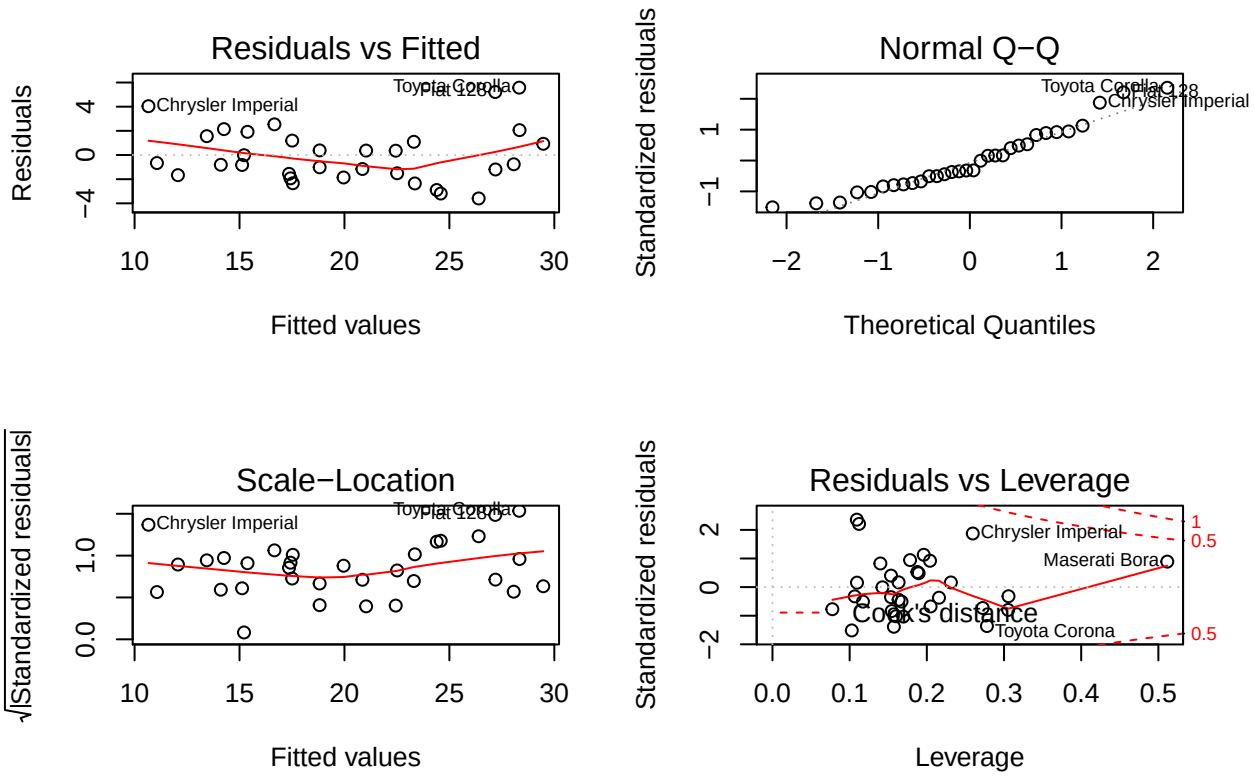
2. Plot: Pairs plot

```
pairs(mpg ~ ., data = mtcars)
```



3. Plot: Model 2

```
par(mfrow=c(2, 2))
plot(md12)
```


Residuals vs Fitted
Residuals
Fitted values
Chrysler Imperial
Toyota Corolla
Fiat 128
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Toyota Corolla
Fiat 128
Chrysler Imperial
Scale−Location
√|Standardized residuals|
Fitted values
Chrysler Imperial
Toyota Corolla
Fiat 128
Residuals vs Leverage
Standardized residuals
Leverage
Chrysler Imperial
Maserati Bora
Toyota Corona
Cook's distance
1
0.5
0.5