Regresion Models Course Project

Hernan Solano

19/5/2018

Executive Summary

This is the final project for the Regression Models course of the Data Science Specialization. We are going to use the mtcars dataset from the default data of the R software in order to determing the relationship between the nature of the miles per galon (MPG) and the transmission or am variable (0 = automatic, 1 = manual) of a series of vehicles. We perform some Exploratory Data Analisis and then chose an appropiate regression model in order to quantify the difference of MPG of each transmission type.

Exploratory Data Analysis

First, lets identify the variable types and how the table looks like:

```
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
```

Since our focus is on the mpg and am variables the following code shows their key measures:

```
round(summary(mtcars$mpg),2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.40   15.43   19.20   20.09   22.80   33.90
```

```
round(summary(mtcars$am),2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00    0.00    0.41    1.00    1.00
```

In this case, the range of the mpg variable is (10.4 - 33.9) and its standard deviation is: 6.03. This is consistent with the figure 1 that shows the histogram of the mpg variable. On the other, the am variable only can take as values a 0 or a 1 and 41% of the cars have a manual transmission.

Furthermore, if we see figure 2, we can see that cars with a manual transmission might have a higher mean mpg value than the vehicules with an automatic transmision.

Regression Model

We are going to use a linear regression model in order to identify the relationship between these two variables. Then, our outcome will be the mpg variable and the predictor will be the transmission variable as a factor. We add -1 to the model to avoid confussions with the coefficients.

```
    fit <- lm(mpg ~as.factor(am)-1,data = mtcars)
    summary(fit)
```

```
##
## Call:
## lm(formula = mpg ~ as.factor(am) - 1, 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|)
## as.factor(am)0   17.147      1.125   15.25 1.13e-15 ***
## as.factor(am)1   24.392      1.360   17.94  < 2e-16 ***
## ---
## 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:  0.9487, Adjusted R-squared:  0.9452
## F-statistic: 277.2 on 2 and 30 DF,  p-value: < 2.2e-16
```

```
    confint(fit)
```

```
##                   2.5 %   97.5 %
## as.factor(am)0 14.85062 19.44411
## as.factor(am)1 21.61568 27.16894
```

According to the summary, none of the estimated coefficients is equal to 0 and the confidence intervals confirm it. We ploted the line on figure 3 so we can appreciate the increasing trend between automatic and manual transmisions. Additionally, if we look at the diagnosis plots in figure 4 there are 3 cars with high leverage in comparison with the others, these are: the Toyota Corolla, the Maserati Bora and the Ford Pantera L.

We can also calculate the difference between the coefficients, that is:

Manual transmition coefficient - automatic transmition coefficient

24.39 - 17.15 = 7.24

We can also confirm this difference by running a t-test like this:

```
    test <- t.test(mpg ~as.factor(am),data = mtcars)
    test$conf.int
```

```
## [1] -11.280194  -3.209684
## attr(,"conf.level")
## [1] 0.95
```

```
    test$estimate
```

```
## mean in group 0 mean in group 1
##        17.14737        24.39231
```

This means that with a p-value of 0.0013736 we can reject the null hypothesis and say that the mpg values of the 2 compared groups are different. We can conclude by saying that cars with manual transmissions have 7.24 mpg more than cars with automatic transmissions.

Appendix

Figure 1

This is the histogram of the mpg variable:

```
    ggplot(data = mtcars,aes(x = mpg)) + geom_histogram(fill = "#014d64") +
    theme_economist() + labs(title = "Figure 1: Histogram of mpg")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

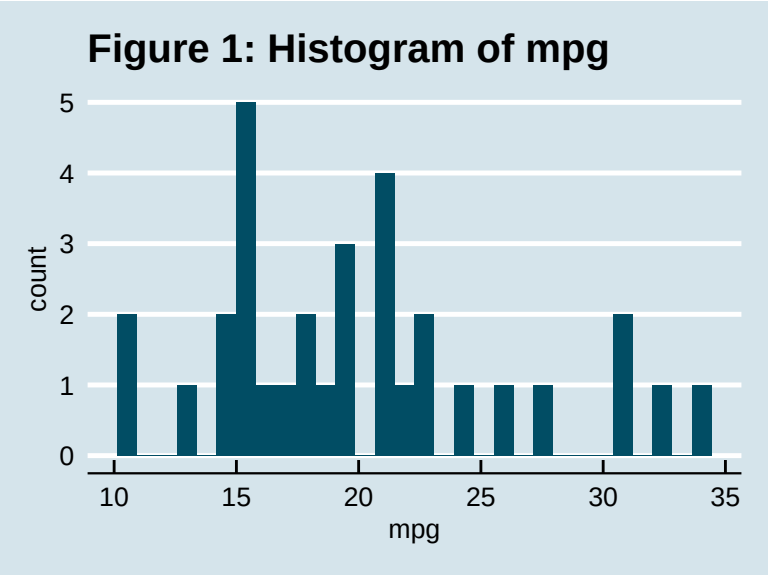

Figure 2

This is a generalized Pairs Plot using ggpairs:

```
    ggpairs(data = mtcars,columns = c("wt","am","mpg")) +
    labs( title = "Figure 2: Pairs Plot") +theme_economist()
```

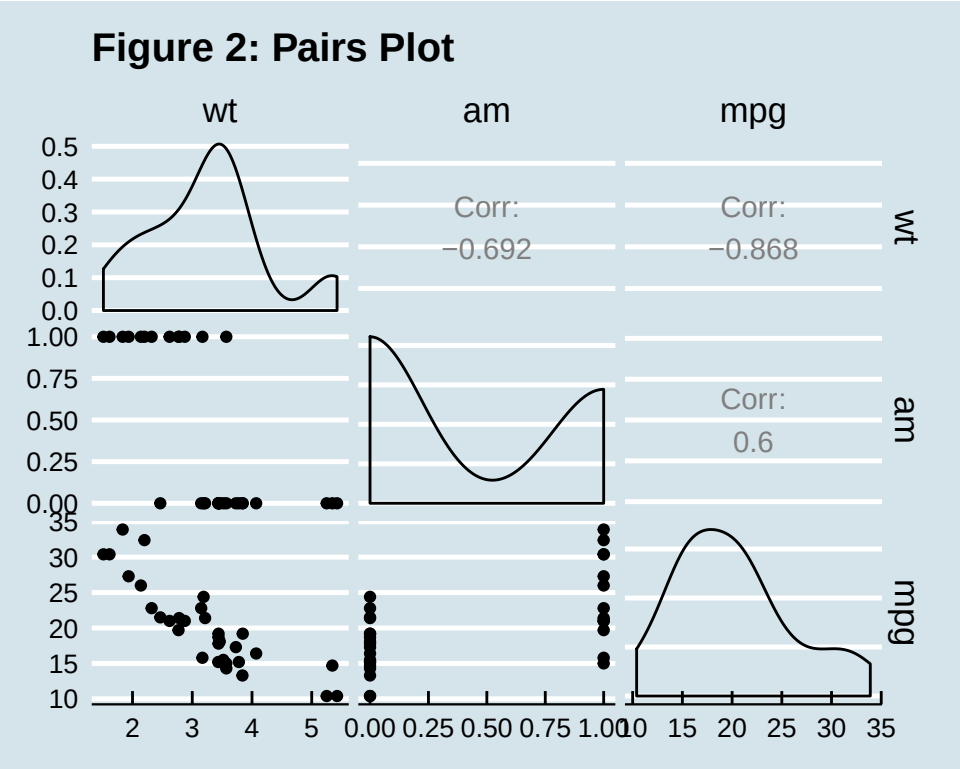

Figure 3

This is a plot of the regression line modeled:

```
ggplot(data = mtcars,aes(x = am,y = mpg)) + geom_point()+geom_smooth(method = lm) +
labs( title = "Figure 3: Regression line") + theme_economist()
```

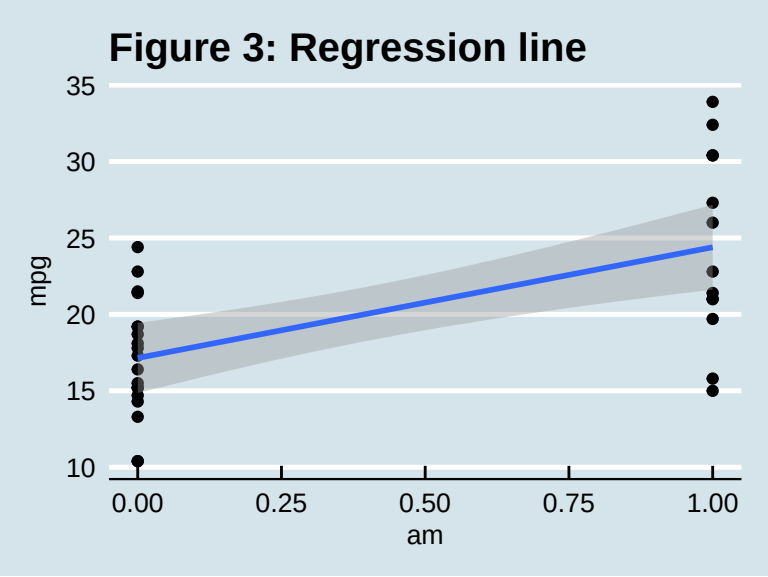

Figure 4

Diagnosis plots:

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