Course Project - Regression models

Rafael Lavagna

30 de abril de 2018

Executive summary

The objective of this project is to analize a dataset of a collection of cars. In particular, we are going to explore the relationship between the miles per gallon consumed by the vehicles and its kind of transmission (manual or automatic) in order to infer whether one of them is better than the other or not, and, in that case, quantify that difference.

Exploratory Data Analysis

```
library(ggplot2)
data("mtcars")
```

In order to show how the miles per gallon varies between both different types of transmission we are going to transform the "am" variable into a factor one. Afterwards, we are going to build a boxplot showing the mentioned variation. *See figure 1 in the appendix* At first glance, it seems that cars with manual transmission have a better performance measured in miles per gallon than those with an automatic one. In order to verify this observation we are going to perform some regression models over the data in order to define if there is enough statistical evidence to support this claim.

Regression models

Simple Regression

As a start, we are going to fit a linear model using the variable "mpg" (miles per gallon) as the outcome and just the variable "am" (Transmission) as regressor. We are also going to consider an intercept for this model. It is important to mention that the level Automatic is considered the reference one.

```
fit <- lm(mpg~am,mtcars)
summary(fit)$coefficients
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## am           7.244939   1.764422  4.106127 2.850207e-04
```

Looking at the results obtained the linear model suggests that a car with automatic transmission will have a performance of 17.147 miles per gallon while a car with a manual one will have a performance 7.245 miles per gallon higher, that is 24.39 miles per gallon. It can be easily shown that these values are the arithmetic averages of each group.

We are going to build now a confidence interval for those values in order to check if the difference between both groups is statistically significant, in particular we are going to check whteher the interval of the "am" coefficient contains the value zero or not.

```
sumCoef <- summary(fit)$coefficients
sumCoef[1,1] + c(-1, 1) * qt(.975, df = fit$df) * sumCoef[1, 2]
```

```
## [1] 14.85062 19.44411
```

```
sumCoef[2,1] + c(-1, 1) * qt(.975, df = fit$df) * sumCoef[2, 2]
```

```
## [1]  3.64151 10.84837
```

As we can see, the interval for the "am" coeffcient is above zero, so, we can say with 95% confidence that cars with manual transmission have a higher performance than those with an automatic one and that this improvement is between 3.64 and 10.84 miles per gallon. Finally, we are going to take a look to the residuals of the regression in order to diagnose our model. *See figure 2 in the appendix* We observe no pattern in the residuals vs fitted plot. Furthermore, in the QQ plot we ovserve that residuals seems to follow normallity. In other words, there is no evidence in the residual plots against the fitness of our model. However, we observe a R-squared value of 0.3598. That means that just a 36% of the variation of the outcome "mpg" is explained by this model.

Multivariate Regression

We are now going to build a model with "mpg" as outcome but with all the other variables as regressors to see what happens with the "am" coeffcients. We are also going to run the anova test in order to check if some of the variables added are significant to the response "mpg".

```
fit2 <- lm(mpg~.,mtcars)
summary(fit2)$coefficients[9,]
```

```
##   Estimate Std. Error    t value   Pr(>|t|)
##  2.5202269  2.0566506  1.2254035  0.2339897
```

```
anova(fit,fit2)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     21 147.49  9     573.4 9.0711 1.779e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see from the summary that the variation explained by this new model increased (as expected) to a value of 0.869. Furthermore, the lower p-value of the anova test implies that at least one of the variables added to the model is significant to the outcome. Let's see now what happens with the confidence interval for the "am" coefficient.

```
sumCoef2 <- summary(fit2)$coefficients
sumCoef2[9,1] + c(-1, 1) * qt(.975, df = fit2$df) * sumCoef2[9, 2]
```

```
## [1] -1.756812  6.797266
```

This new interval for the am coefficient contains zero. That means that with 95% confidence we can't reject the null hypothesis that both groups are different, which is the same as saying that we can't conclude that cars with manual transmission have higher performance than those with an automatic one. As with the simple model, we are going to take a look at the residuals *See figure 3 in the appendix* Once again, we observe no pattern in the residuals vs fitted plot. Furthermore, in the QQ plot we ovserve that residuals seems to follow normallity. In other words, there is no evidence in the residual plots against the fitness of our model.

Appendix

Figure 1

```
mtcars$am <- as.factor(mtcars$am)
g <- ggplot(mtcars, aes(x=factor(0),y=mpg)) + geom_boxplot(aes(fill=am)) + facet_grid(.~am)
g <- g + theme(axis.title.x=element_blank(), axis.text.x=element_blank(), axis.ticks.x=element_blank())
g <- g + ggtitle("Miles per gallon by Transmission kind") + ylab("Miles per gallon") +
theme(plot.title = element_text(hjust = 0.5))
g <- g + scale_fill_discrete(name="Transmission", breaks=c("0","1"),labels=c("Automatic", "Manual"))
g
```

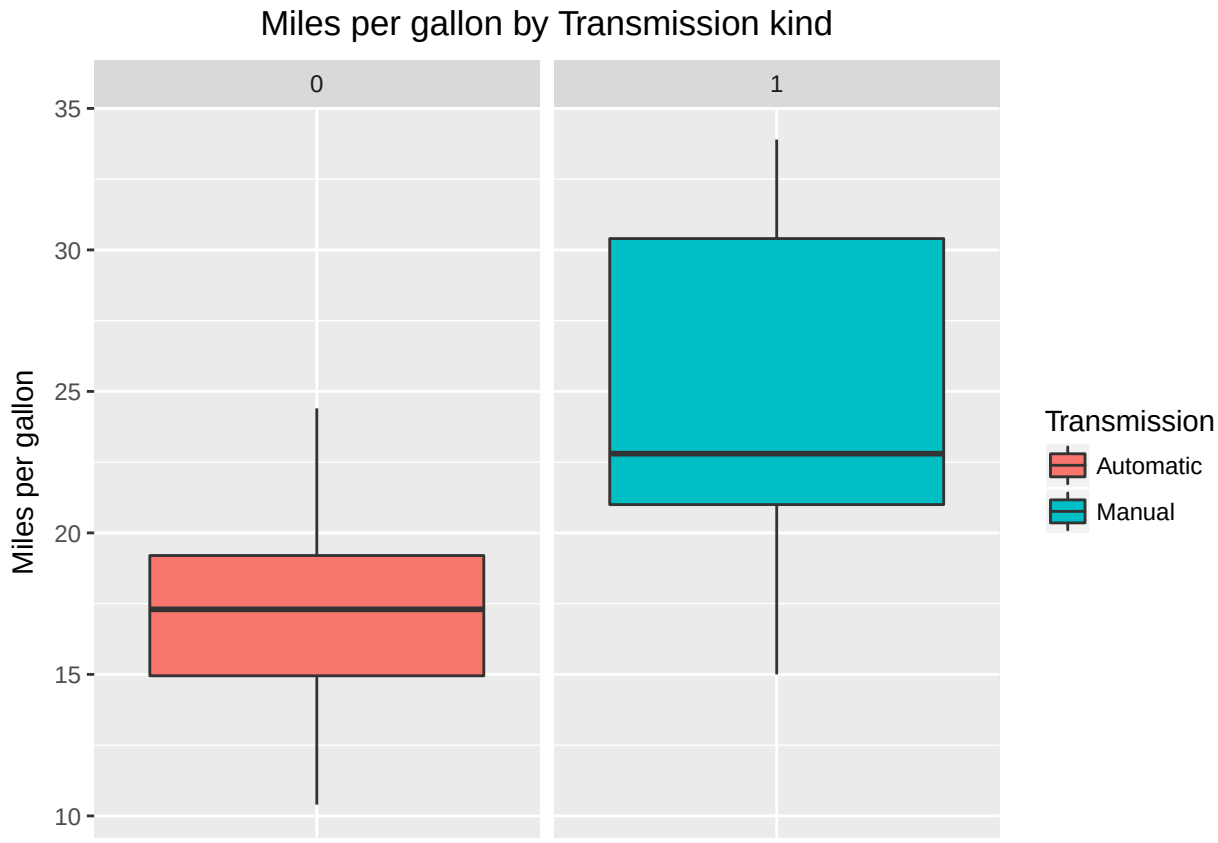


Figure 2

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

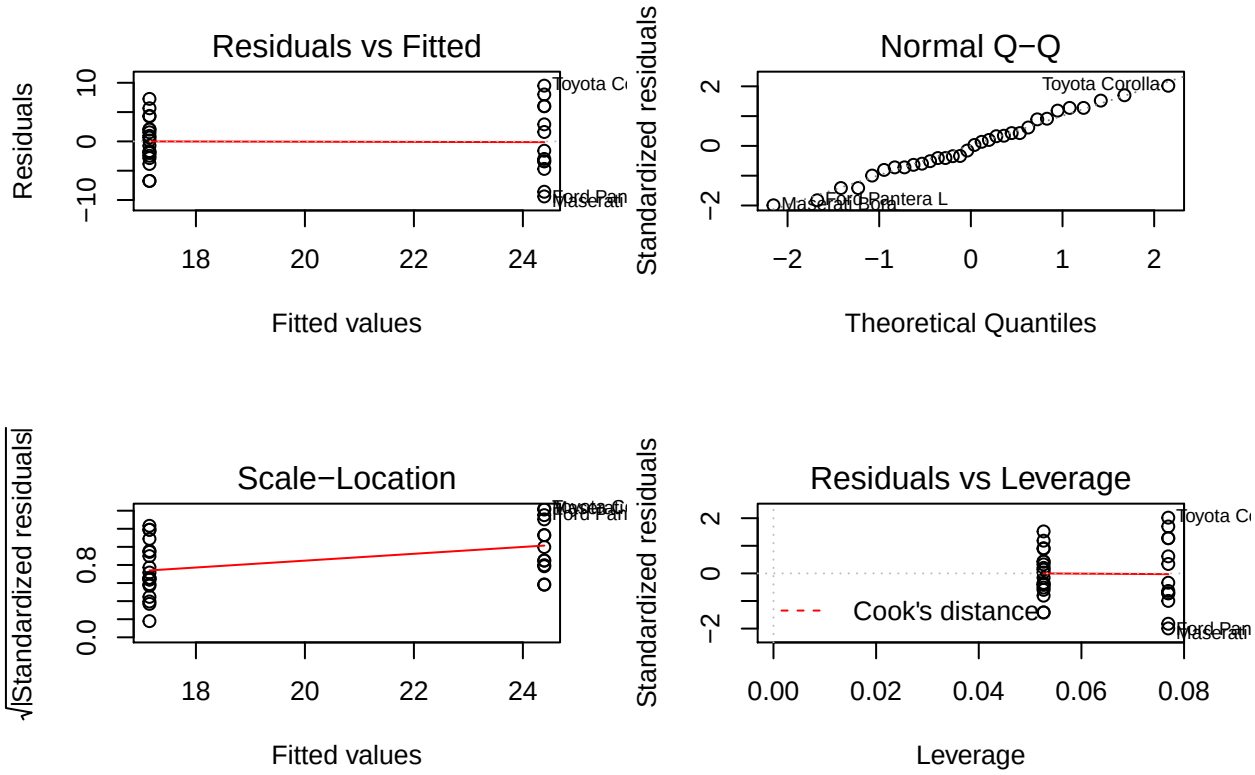
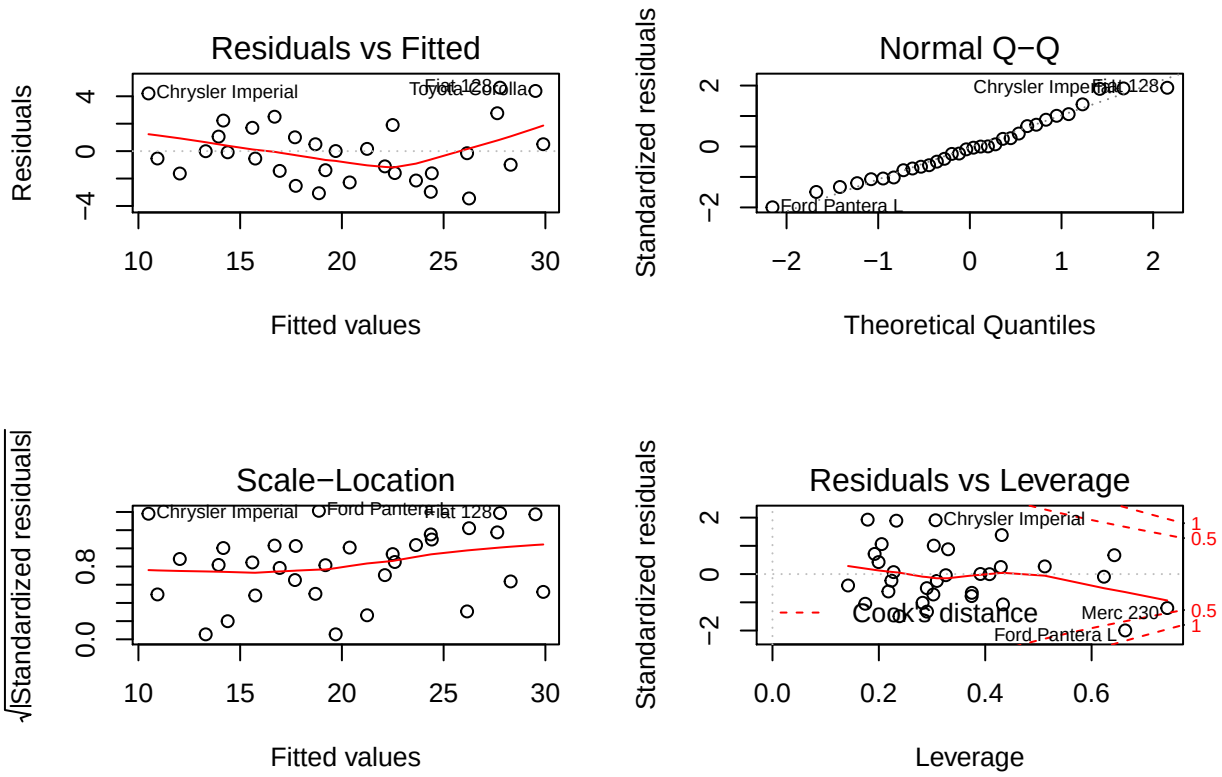


Figure 3

```
par(mfrow=c(2,2))
plot(fit2)
```


Residuals vs Fitted
Residuals
Fitted values
Chrysler Imperial
Toyota Corolla
Fiat 128
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Chrysler Imperial
Fiat 128
Ford Pantera L
Scale−Location
√|Standardized residuals|
Fitted values
Chrysler Imperial
Ford Pantera L
Fiat 128
Residuals vs Leverage
Standardized residuals
Leverage
Chrysler Imperial
Cook's distance
Merc 230
Ford Pantera L
1
0.5
0.5
1