Regression Model Course Project

Sebastian Jaroszewicz

12/10/2020

Introduction

The objective of this study is to ooking at a data set of a collection of cars and to explore the relationship between a set of variables and miles per gallon (MPG) (outcome). In particular it is of interest to answer the following questions:

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

The data set that will be used to carry out the study will be the mtcars.

Exploratory Data Analysis

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

```
a <- dim(mtcars)
```

The data set is composed of with 32 observations on 11 (numeric) variables.

1. mpg Miles/(US) 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

```
# Transform same varialbes into factor
mtcars$am <- factor(mtcars$am,labels=c("Automatic","Manual"))
mtcars$cyl <- as.factor(mtcars$cyl)
mtcars$vs <- as.factor(mtcars$vs)
mtcars$gear <- as.factor(mtcars$gear)
mtcars$carb <- as.factor(mtcars$carb)
```

In order to better understand the data, we made a box plot graph mpg by tansmission type (see appendix).

The plot shows a significant difference between manual and automatic transmissions. To quantitatively analyze this difference we are going to perform a t-test

```r
testResults <- t.test(mtcars$mpg ~ mtcars$am)
testResults
```

```
##
##  Welch Two Sample t-test
##
## data:  mtcars$mpg by mtcars$am
## t = -3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.280194  -3.209684
## sample estimates:
## mean in group Automatic    mean in group Manual
##                17.14737                24.39231
```

The T-Test rejects the null hypothesis that the difference between transmission types is 0. The difference estimate between the 2 transmissions is about 7 mpg in favor of the manual.

Regression Model

To study the relationship between mpg and the other variables, we performed a regression model for this dataset. We use multiple linear regression and the R step function, which chooses the best model.

```r
fit <- lm(mpg~.,mtcars)
summary(fit)
fit_Step <- step(fit)
```

The model suggest to use the formula **mpg ~ cyl + hp + wt + am**

```r
model = lm(mpg~ wt + qsec + am, data=mtcars)
summary(model)
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.6178     6.9596   1.382 0.177915
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## amManual      2.9358     1.4109   2.081 0.046716 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.459 on 28 degrees of freedom
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8336
## F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

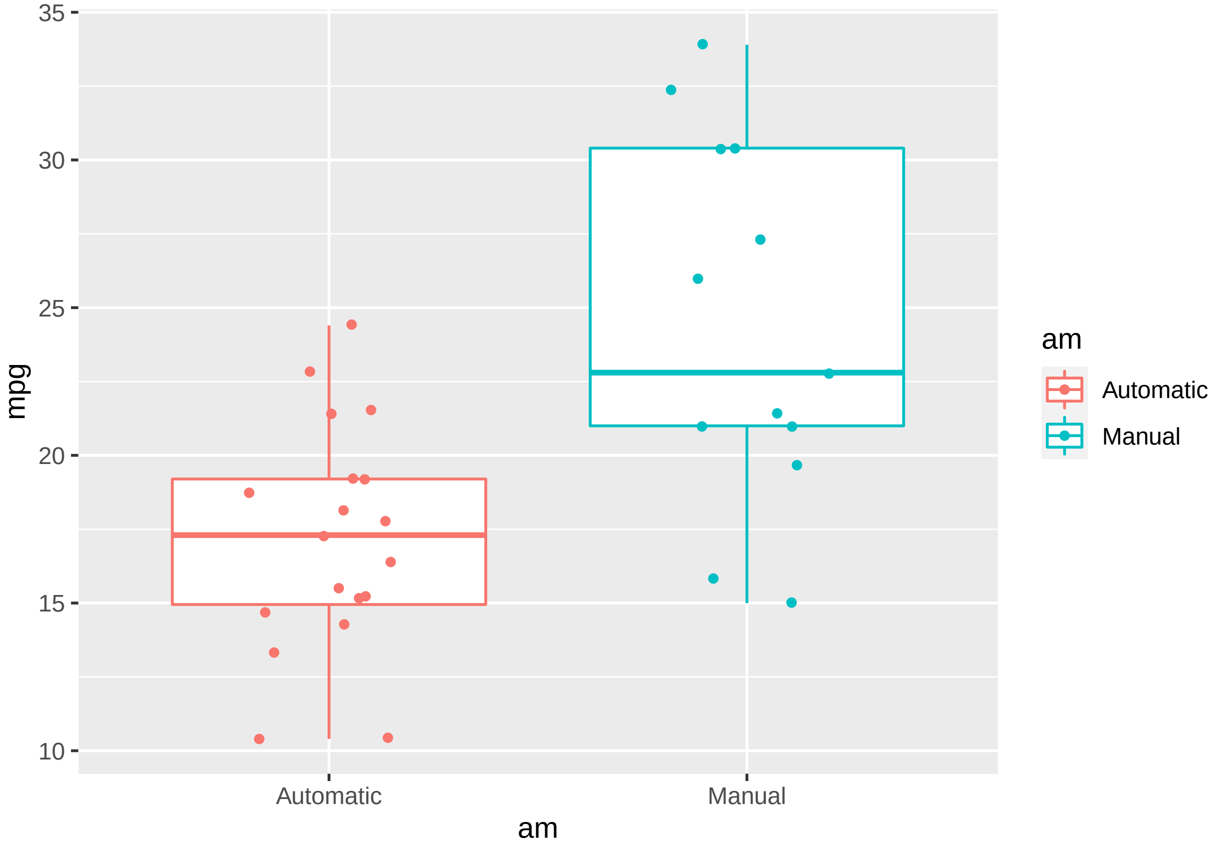
In th appendix we plot the residuals of the model.

Conclusions

From our study we can determine that there is a difference in mpg in relation to transmission type in favor of manual. But a better explanation adjust with weight and qsec.

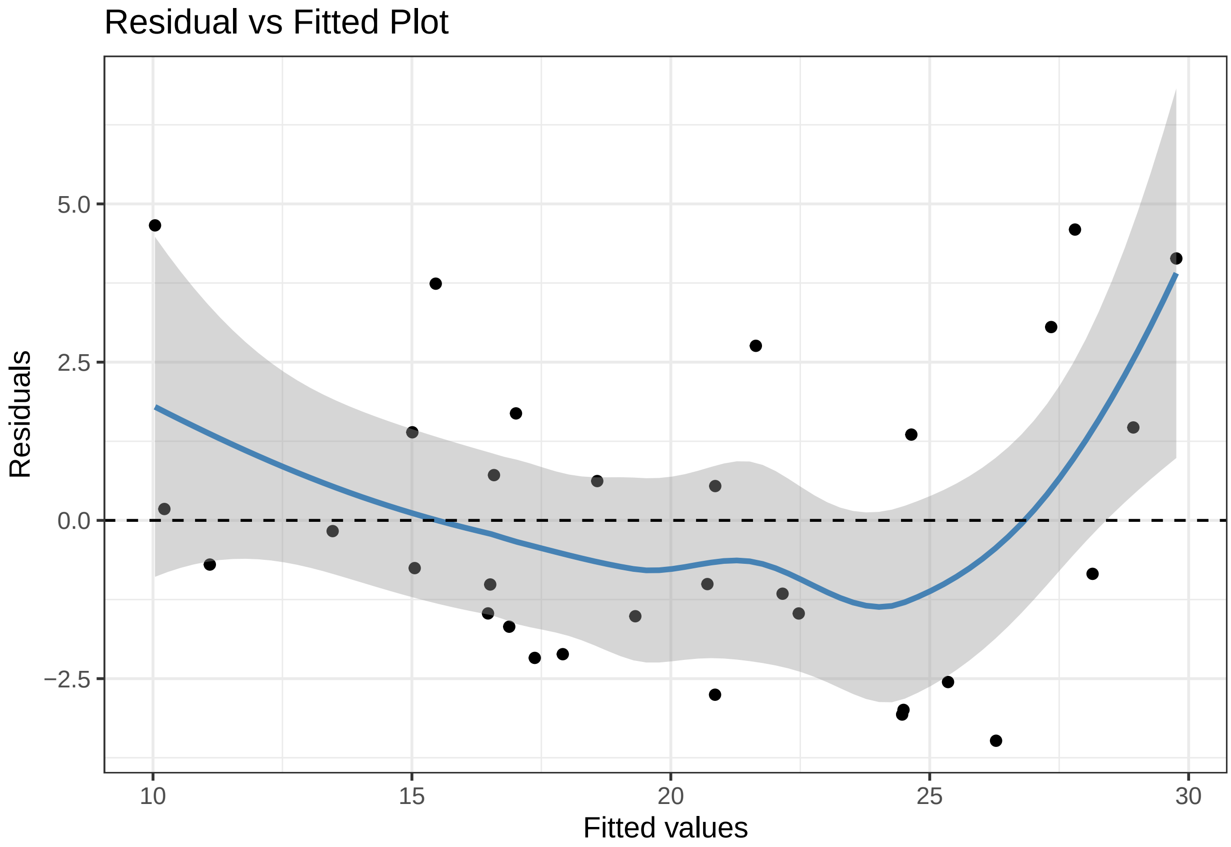
Appendix

```
#boxplot(mpg ~ am, data = mtcars, col = (c("red","blue")), ylab = "Miles Per Gallon", xlab = "Transmiss
g <- ggplot(mtcars, aes(x=am, y=mpg, color = am)) +
        geom_boxplot()
g <- g + geom_jitter(shape=16, position=position_jitter(0.2))
g
```



```
ggplot(model, aes(.fitted, .resid)) + geom_point() +
  stat_smooth(method="loess", col="steelblue") + geom_hline(yintercept=0, linetype="dashed") +
  xlab("Fitted values")+ylab("Residuals") +
  ggtitle("Residual vs Fitted Plot") + theme_bw()
```

```
## `geom_smooth()` using formula 'y ~ x'
```


Residual vs Fitted Plot
Residuals
5.0
2.5
0.0
−2.5
10
15
20
25
30
Fitted values