Linear Regression study

Andres Mauricio Castro

Executive Summary

The following data analysis, is performed for the Motor Trend magazine, whom are interested in exploring the relationship between a set of variables and miles per gallon (MPG).

By using a dataset from the 1974 Motor Trend US magazine to answer the following questions, we are going to answer two main questions, they are interested in:

```
- Is an automatic or manual transmission better for miles per gallon (MPG)?
- How different is the MPG between automatic and manual transmissions?
```

By performing a set o hyphotesis tests, and constructing an optimized linear regression model, finding out if there are confounders which can help to explain better the relationship betweeng the given variables; after validating the final model, we can answer the two questions:

¿Is an automatic or manual transmission better for MPG? As per our analysis, we can conclude a manual transmission is better for MPG.

¿Quantify the MPG difference between automatic and manual transmissions? A manual transmission has 2.9 MPGs more than an automatic transmission.

See below all the statistical procedure followed, in order to answer the two questions.

Data Processing

We load the data, and see the class of variables in the dataset. The predictor variable is numeric, so, let's convert it into a factor, for better interpretability.

```
data(mtcars)
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

mtcars$am <- factor(mtcars$am,labels=c('Automatic','Manual'))
str(mtcars)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : Factor w/ 2 levels "Automatic","Manual": 2 2 2 1 1 1 1 1 1 1
...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

Exploratory Data Analysis

We look at the help of the dataset, to get a feel of the variables meaning, and how they can be related to the outcome, to construct and validate our model.

```
?mtcars
```

Also, we create a boxplot of the mpg by transmission type;in the figure 1, we can see there is a differente in the mpg by transmission type; the manual transmissions seems to get a better mpg than automatic tranmissions.

In the figure 2, we plot the relationships between the variables in the dataset; we can see the variables cyl, disp, hp, drat, wt, vs and am may have a strong correlation with mpg. In the next section, we will construct the best model, to see which variables are the confounders.

Finally, by performing a hypothesis test, we found the p-value is 0.001374, so, we can reject the null hypothesis, and claim that there is a signficiant difference in the mean MPG between manual and automatic transmissions.

```
aggregate(mpg~am, data = mtcars, mean)

##          am      mpg
## 1 Automatic 17.14737
## 2    Manual 24.39231

automaticData <- mtcars[mtcars$am == "Automatic",]
manualData <- mtcars[mtcars$am == "Manual",]
t.test(automaticData$mpg, manualData$mpg)

## 
##  Welch Two Sample t-test
## 
## data:  automaticData$mpg and manualData$mpg
## 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 of x mean of y
##  17.14737  24.39231
```

Model Building And Selection

Firstly, we create a model with all the variables as predictors, and by using an Stepwise Algorithm (function step), we can look for the best model, to explain the variability, including the confounders and the independent variable.

```
initialModel <- lm(mpg ~ ., data = mtcars)
optimizedModel <- step(initialModel, direction = "both")

summary(optimizedModel)

##
## 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
```

The best model calculated was **lm(formula = mpg ~ wt + qsec + am, data = mtcars)**

By looking at the optimized model, we see the variables wt and qsec are confounders, and the am is the independent variable. Also, the R2 value is 0.83, being the maximum obtained from the stepwise function. So, the optimized model can explain of the 83% of the variability.

Now, we can compare the model containing only the am variable as predictor, with the optimized model, and draw some conclusions.

```
baseModel <- lm(mpg ~ am, data = mtcars)
anova(baseModel, optimizedModel)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ wt + qsec + am
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     30 720.90
## 2     28 169.29  2    551.61 45.618 1.55e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value obtained is highly significant, and we reject the null hypothesis that the confounders don't contribute to the accuracy of the model.

Conclusion

In the figure 3, we plot the residuals for our best model, and draw some conclusions:

- The residuals are normally distributed (QQPlot).
- The variance is constant, as indicated in the Scale-Location plot.
- The residuals are homoskedastic, as the error variance is the same across all the values.

Now, let's look at other conclusions from our best model:

```
summary(optimizedModel)
```

```
##
## 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
```

- On average, manual transmissions have 2.9 MPGs more than automatic transmissions.
- mpg increases with increase of qsec.

- mpg will decrease approximately by 3.9, for every 1000 lb increase in wt (as per the definition of wt variable in the mtcars).

Appendix

Figure 1

MPGs by Transmission Type.

```
boxplot(mpg~am, data = mtcars,
        col = c("blue", "red"),
        xlab = "Transmission",
        ylab = "Miles per Gallon",
        main = "MPG by Transmission Type")
```

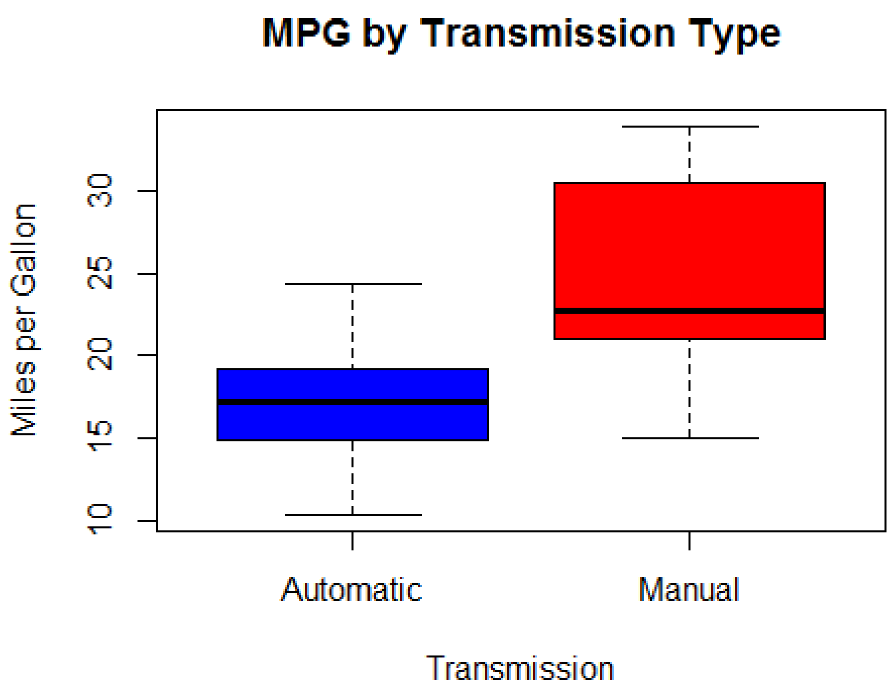

Figure 2

Pairs plot for the mtcars dataset.

```
pairs(mtcars)
```

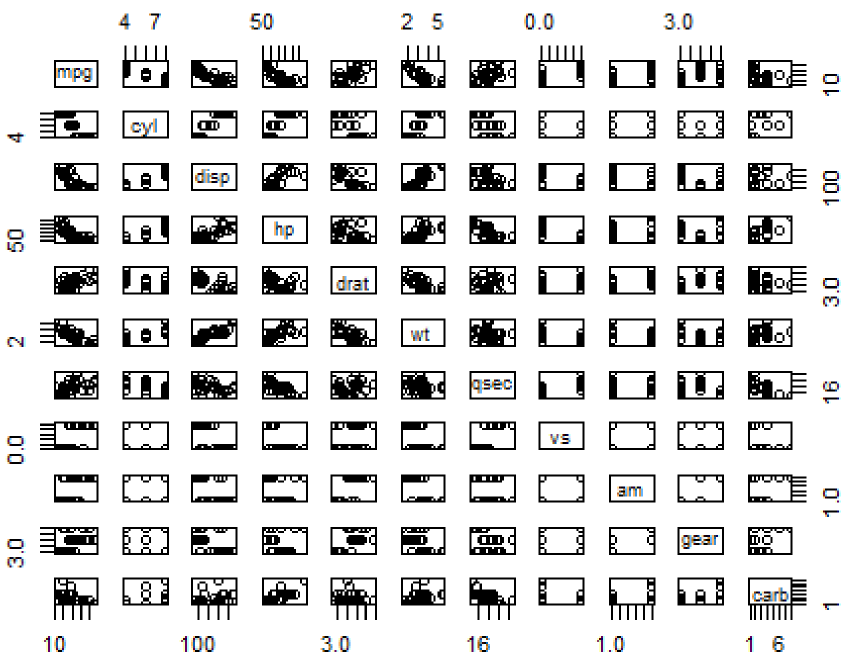

Figure 3

Residuals Plot.

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