Course Project - Regression Models

Venkat Ram Rao (11/22/2020)

Executive Summary

This report is for the final course project for the **Regression Models** course, part of the **John Hopkins Statistics and Machine Learning Specialization** on **Coursera**

The purpose of the report is to explore the relationship between a set of variables and miles per gallon (MPG) (outcome) and answer the following questions:

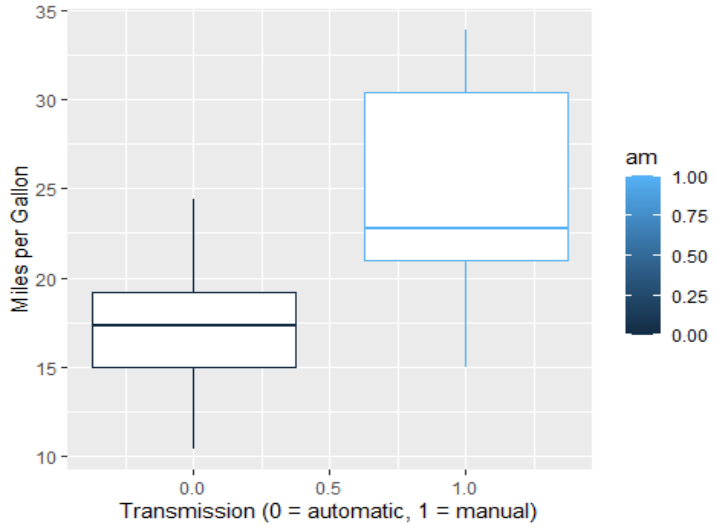
1. "Is an automatic or manual transmission better for MPG"
2. "Quantify the MPG difference between automatic and manual transmissions"

The analysis below has reached the conclusion that changing the transmission from Automatic to Manual causes an increase in Fuel economy of **2.9358** mpg when controlled for all other factors

Exploratory analysis

mtcars is a dataset provided by R and contains 32 sample records from 1974 comparing fuel comsumption(mpg) and 10 aspects of automobile design and performance for 32 automobiles. (**Appendix 1.** Has the R code)

Below is a simple plot (**Appendix 2.**) showing the miles/gallon for Manual and Automatic transmission. At first glance, there seems to be a clear difference in mpg based on the Transmission.



A better analysis can be done by running a T-test. **Appendix 3**. Has the code and the final output of the T-test. It showed that there is a mean difference of **7.24492** mpg between cars with Manual versus Automatic transmission. The 95% confidence interval **(-11.28,-3.21)** does not contain zero meaning we can say with a 95% confidence that Manual Transmission leads to better fuel economy .

However, there are other variables which effect the mpg so additional analysis will need to be done as below

Regression Models

Data Prep:

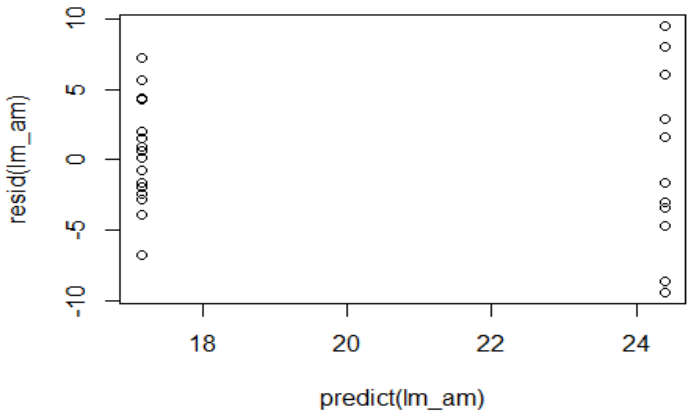
One note - it seems all the data in mtcars is defined as numeric. I refactored the 'am' column to better reflect the data:

```
mtcars$am<-relevel(as.factor(mtcars$am),
ref='0',labels=c('Automatic','Manual'))
```

Simple model:

Starting with a simple model of mpg vs Transmission type (**R-code in Appendix 4**) shows a model with a high R-squared of (**.9487**) and shows that the improvement in mpg when going from Automatic to Manual transmission is **7.24473**. However, plotting the residuals(below) shows that they are definitely not random. Hence, a simple regression model with just Transmission type is not reliable.

```
plot(predict(lm_am),resid(lm_am))
```



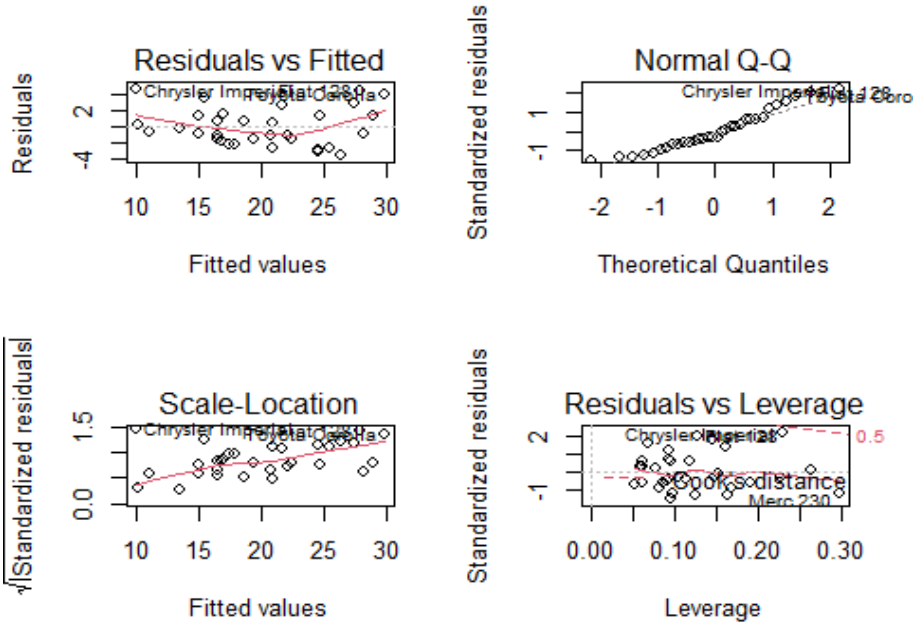
Best model: Running a regression with all the variables (**Appendix 5.)** shows multiple variables with a strong correlation with mpg. In particular, weight(wt) has a strong negative coefficient**(-3.71530**). 1/4 mile time (qsec), Number of forward gears(gear) and Rear axle ratio(drat) seem to have a strong positive coefficients as well : **0.82104, 0.65541** & **0.78711** respectively.

The best model can be found by using the step function – (**R-Code in Appendix 6.**)

This gives us a model(lm_best) based on **wt**, **qsec** and **transmission** with an R squared of **0.9879** and the coefficients below**:**

```
##      Estimate Std. Error t value Pr(>|t|)
## wt    -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec   1.2259     0.2887   4.247 0.000216 ***
## am0    9.6178     6.9596   1.382 0.177915
## am1   12.5536     6.0573   2.072 0.047543 *
```

Below is the residual plot which is reasonably random. (**R – code in Appendix 7.**) giving a higher confidence on the model accuracy



Conclusion

Looking at the coefficients, it looks like:

1. Changing the transmission from Automatic to Manual causes an increase in Fuel economy of **2.9358** mpg when controlled for all other factors.

2. Weight and 1/4 mile time are also influential. Each 1000 lb increase in weight reduces fuel economy by **3.9165** mpg,

3. Increasing qsec also increases Fuel Economy. This is interesting as there is no logical reason for this to be the case. This needs to be investigated further. There is a possibility that some bias crept in and qsec has strong correlations with other factors (e.g. hp)

Assumptions:

The big issue with this analysis is that there are only 32 records in the data set. This is small and, if possible, additional data needs to be identified.

APPENDIX

1. Exploratory Analysis

```
data(mtcars)
dim(mtcars)
```

```
## [1] 32 11
```

```
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  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

2. Creating a BoxPlot

```
library(ggplot2)
my_boxplot <- ggplot(mtcars,aes(x=am,y=mpg, group=am, col=am)) +
geom_boxplot() + xlab('Transmission (0 = automatic, 1 = manual)') +
ylab('Miles per Gallon')
my_boxplot
```

3. T-test for Manual and Automatic Transmission

```
mtcars_am0 <- mtcars[which(mtcars$am == 0), names(mtcars) %in%
c("mpg")]
mtcars_am1 <- mtcars[which(mtcars$am == 1), names(mtcars) %in%
c("mpg")]
t.test(mtcars_am0, mtcars_am1, paired = FALSE, alternative="two.sided",
var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  mtcars_am0 and mtcars_am1
```

```
## 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
```

4. Regression of MPG against Trasmission

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

##
## Call:
## lm(formula = mpg ~ 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|)
## factor(am)0   17.147      1.125   15.25 1.13e-15 ***
## 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
```

5. Regression with all variables.

```
lm_all <- lm(mpg~. -1,data = mtcars)
summary(lm_all)

##
## Call:
## lm(formula = mpg ~ . - 1, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4506 -1.6044 -0.1196  1.2193  4.6271
##
## Coefficients:
##      Estimate Std. Error t value Pr(>|t|)
## cyl  -0.11144    1.04502  -0.107   0.9161
```

```
## disp  0.01334    0.01786   0.747   0.4635
## hp   -0.02148    0.02177  -0.987   0.3350
## drat  0.78711    1.63537   0.481   0.6353
## wt   -3.71530    1.89441  -1.961   0.0633 .
## qsec  0.82104    0.73084   1.123   0.2739
## vs    0.31776    2.10451   0.151   0.8814
## am0  12.30337   18.71788   0.657   0.5181
## am1  14.82360   18.35265   0.808   0.4283
## gear  0.65541    1.49326   0.439   0.6652
## carb -0.19942    0.82875  -0.241   0.8122
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.65 on 21 degrees of freedom
## Multiple R-squared:  0.9895, Adjusted R-squared:  0.984
## F-statistic: 179.8 on 11 and 21 DF,  p-value: < 2.2e-16
```

6. Finding the best model

```
lm_best<-step(lm_all,direction ="both")

summary(lm_best)

##
## Call:
## lm(formula = mpg ~ wt + qsec + am - 1, 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|)
## wt    -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec   1.2259     0.2887   4.247 0.000216 ***
## am0    9.6178     6.9596   1.382 0.177915
## am1   12.5536     6.0573   2.072 0.047543 *
## ---
## 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:  0.9879, Adjusted R-squared:  0.9862
## F-statistic: 573.7 on 4 and 28 DF,  p-value: < 2.2e-16
```

7. Residual Plot for the best model(lm_best)

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