Course Project: Linear Regression

S. Servaes

18/10/2019

Summary

This report will examine the mtcars data set and explore the relationship between miles per gallon (MPG) and transmission type. Specifically, this project will examine: 1) Is an automatic or manual transmission better for MPG; and 2) Quantify the difference between automatic and manual transmissions.

Results indicate that vehicles with automatic transmissions have fuel mileage significantly lower than vehicles with manual transmissions. Regression analysis demonstrates that the MPG of a vehicle can be predicted given the weight, displacement, number of cylinders and the transmission type. Based on the best-fit regression model it can be said that vehicles with manual transmissions do give 0.14 miles per gallon more than vehicles with automatic transmissions.

Exploratory Data Analysis

Dataset

First we take a look at the raw data available here by plotting the first lines.

```
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_num <- mtcars
```

Dataset visualisation

Now we modify the dataset changing cyl, vs, gear, carb and am into factor variables.

The plot (appendix: figure 1) shows that mpg tends to be lower in cars with an automatic gearshift.

Quantitative Analysis

Linear model

Next we test this difference for significance.

```
t.test(mpg~am, data=mtcars)
```

```
##
##  Welch Two Sample t-test
##
## data:  mpg by 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
```

With a p-value of 0.001374, we can constitute that the automatic gearshift have a significantly lower mpg compared to cars with a manual gearshift.

As the difference is significant we quantify this in the following step.

```
lmmodel1 <- lm(mpg~am, data = mtcars)
summary(lmmodel1)
```

```
##
## Call:
## lm(formula = mpg ~ am, 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|)
## (Intercept)   17.147      1.125  15.247 1.13e-15 ***
## amManual       7.245      1.764   4.106 0.000285 ***
## ---
## 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:  0.3598, Adjusted R-squared:  0.3385
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

From this we can see that the average of MPG for automatic is 17.15, while the average for manual is 7.25 higher. However, as the R-squared has a value of 0.3598, this model only describes ~36% of the variation. Therefore, multivariate linear regression is needed. In order to determine which variables would be useful to be included in the model, a correlation plot (appendix: figure 2) of the absolute correlation coefficients and an analysis of variance was done.

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## cyl          2  824.8   412.4  51.377 1.94e-07 ***
## disp         1   57.6    57.6   7.181   0.0171 *
## hp           1   18.5    18.5   2.305   0.1497
## drat         1   11.9    11.9   1.484   0.2419
## wt           1   55.8    55.8   6.950   0.0187 *
## qsec         1    1.5     1.5   0.190   0.6692
## vs           1    0.3     0.3   0.038   0.8488
## am           1   16.6    16.6   2.064   0.1714
## gear         2    5.0     2.5   0.313   0.7361
## carb         5   13.6     2.7   0.339   0.8814
## Residuals   15  120.4     8.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The significant variables in the analysis of variance are cyl, disp and wt. This is also clearly visible from the correlation plot.

Multivariate Linear Regression

Both cyl, disp and wt have p-values lower than 0.05 in the analysis of variance, suggesting these possibly influence mpg. Therefore these variables were chosen to be included in the model and are then compared to the previous model with anova.

```
lmmodel2 <- lm(mpg ~ am + cyl + disp + wt, data = mtcars)
summary (lmmodel2)
```

```
##
## Call:
## lm(formula = mpg ~ am + cyl + disp + wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5029 -1.2829 -0.4825  1.4954  5.7889
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.816067   2.914272  11.604 8.79e-12 ***
## amManual     0.141212   1.326751   0.106  0.91605
## cyl6        -4.304782   1.492355  -2.885  0.00777 **
## cyl8        -6.318406   2.647658  -2.386  0.02458 *
## disp         0.001632   0.013757   0.119  0.90647
## wt          -3.249176   1.249098  -2.601  0.01513 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.652 on 26 degrees of freedom
## Multiple R-squared:  0.8376, Adjusted R-squared:  0.8064
## F-statistic: 26.82 on 5 and 26 DF,  p-value: 1.73e-09
```

```
anova(lmmodel1, lmmodel2)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + cyl + disp + wt
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     26 182.87  4    538.03 19.124 1.927e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The R-squared of the new model, lmmodel2, is 0.8376. Therefore the new model manages to explain 83.76% of the variance. According to the Anova this was a significantly better fit than the initial model only including am as a variable. Therefore, cyl, disp and wt affect the correlation between mpg and am. Automatic cars have 0.14 MPG less then cars with manual transmission, when these confounding variables are taken into account.

Conclusion

Is an automatic or manual transmission better for MPG?

It initially appeared that manual transmission cars have a better MPG compared to automatic cars. However when including confounding variables such as the amount of cylinders, displacement and weight, this difference

is strongly reduced, though still in the favour of the cars with manual transmission. A large part of the difference is therefore explained by other variables.

Quantify the MPG difference between automatic and manual transmissions

Analysis has shown that when only the transmission was used in the model, manual cars have an MPG increase of 7.25. However, when variables wt and hp are included, the manual car advantage drops to only 0.14 MPG as the other variables (such as the amount of cylinders, the displacement and the weight) strongly contribute to this effect.

APPENDIX

Figure 1 - Data points:

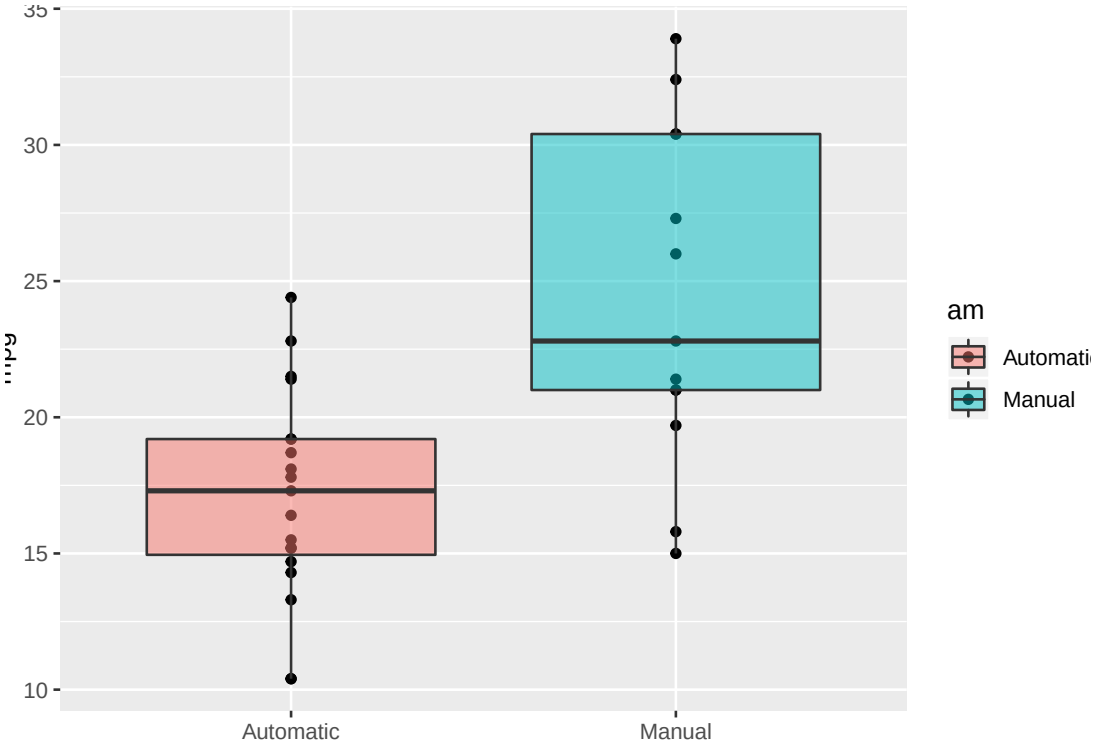

Figure 2 - Correlation plot (abs):

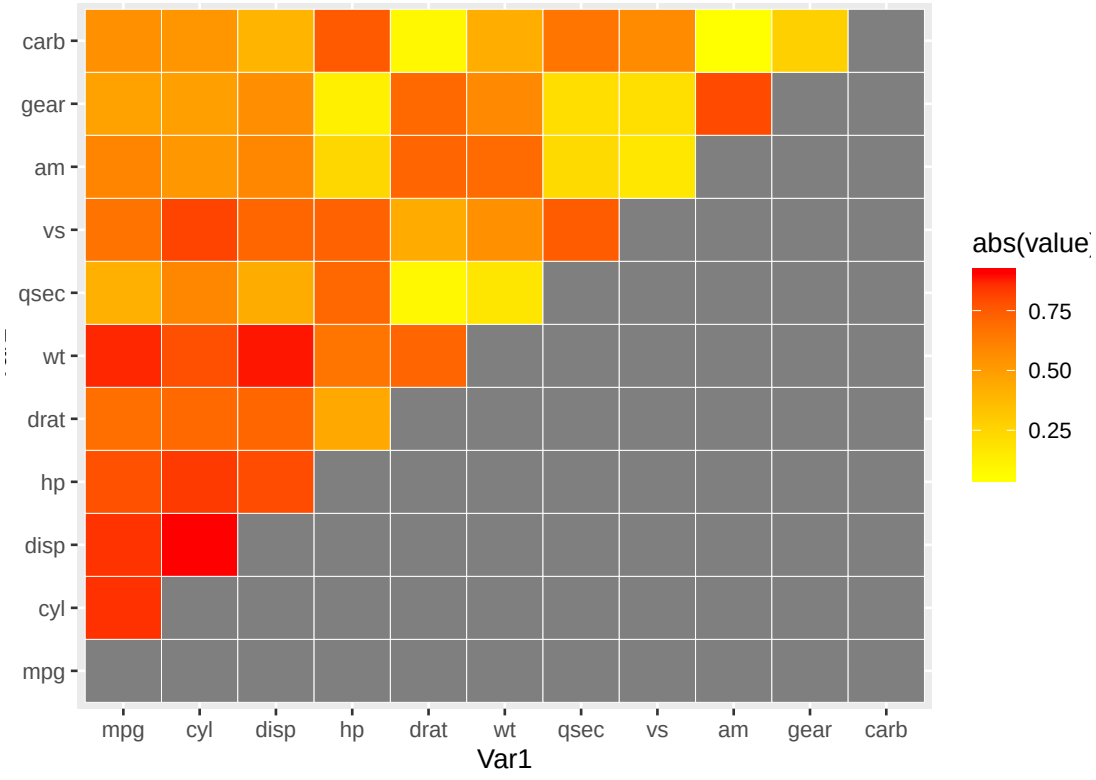

Figure 3 - Residuals: