# MTcars regression project

c zingler

16/06/2020

## Motor Trends Cars Linear Regression Analysis

### Executive Summary

We will be analyzing the Motor Trend Car Road Tests data, it comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models).
This data was compiled in 1974. Simply put we're asking the question, which types of cars: manual or automatic transmission cars goes the furthest, using (mpg) as an indication. The findings are: Manual cars are more fuel efficient. For this Data set they go on average 7.245 miles further per gallon.

### Project requirement

We are to approach this task as a typical business initiated investigation, thus we can not assume a through understanding of statistic by the audience

#### Data description

Data Table name: mtcars

Format A data frame 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

#### Preparing the data for analysis

```
##load the graphics pakage in case of need
library(ggplot2)
## Load the data of interest
data(mtcars)
## Now we need to ensure some of the columns are factors for easier linear model fitting
mtcars$cyl <- factor(mtcars$cyl)
mtcars$vs <- factor(mtcars$vs)
mtcars$am <- factor(mtcars$am)
mtcars$gear <- factor(mtcars$gear)
```

```r
mtcars$carb <- factor(mtcars$carb)

## show the data
head(mtcars,10)
```

```
##                    mpg cyl  disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8 360.0 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6 225.0 105 2.76 3.460 20.22  1  0    3    1
## Duster 360        14.3   8 360.0 245 3.21 3.570 15.84  0  0    3    4
## Merc 240D         24.4   4 146.7  62 3.69 3.190 20.00  1  0    4    2
## Merc 230          22.8   4 140.8  95 3.92 3.150 22.90  1  0    4    2
## Merc 280          19.2   6 167.6 123 3.92 3.440 18.30  1  0    4    4
```

## Exploritory analysis

From the Scatter plot we can see the is a strong relationship between MPG and Cyl, Disp, Hp, Drat, Wt and am. This needs some quantification, please read below for subsequent discussion. In particular the relationship between MPG and Automatic and Manual transmissions.

So lets look at MPG under a single regressor of interest.

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

```
##
## 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 ***
## am1            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
```

```r
##and
basefit1 <- lm(mpg~am-1,data = mtcars)
summary(basefit1)
```

```
##
## Call:
## lm(formula = mpg ~ 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|)
## am0   17.147      1.125   15.25 1.13e-15 ***
## am1   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
```

So there is a 7.245 decrease in MPG for Automatic transmissions, and the second linear model details the MPG mean for each type of transmission. This can be verified by inspecting Appendix Fig.2

Now lets investigate a linear model of factors of interest. But note disp and wt are related, as well as drat and hp being related. So run the model without them.

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

```
##
## Call:
## lm(formula = mpg ~ cyl + hp + wt + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9387 -1.2560 -0.4013  1.1253  5.0513
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.70832    2.60489  12.940 7.73e-13 ***
## cyl6        -3.03134    1.40728  -2.154  0.04068 *
## cyl8        -2.16368    2.28425  -0.947  0.35225
## hp          -0.03211    0.01369  -2.345  0.02693 *
## wt          -2.49683    0.88559  -2.819  0.00908 **
## am1          1.80921    1.39630   1.296  0.20646
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.41 on 26 degrees of freedom
## Multiple R-squared:  0.8659, Adjusted R-squared:  0.8401
## F-statistic: 33.57 on 5 and 26 DF,  p-value: 1.506e-10
```

```
## and now for there means
bestfit1 <- lm(mpg~cyl+hp+wt+am-1,data = mtcars)
summary(bestfit1)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + hp + wt + am - 1, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -3.9387 -1.2560 -0.4013  1.1253  5.0513
##
## Coefficients:
##      Estimate Std. Error t value Pr(>|t|)
## cyl4 33.70832    2.60489  12.940 7.73e-13 ***
## cyl6 30.67698    3.10835   9.869 2.79e-10 ***
## cyl8 31.54465    3.88461   8.120 1.34e-08 ***
## hp   -0.03211    0.01369  -2.345  0.02693 *
## wt   -2.49683    0.88559  -2.819  0.00908 **
## am1   1.80921    1.39630   1.296  0.20646
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.41 on 26 degrees of freedom
## Multiple R-squared:  0.9892, Adjusted R-squared:  0.9868
## F-statistic: 398.6 on 6 and 26 DF,  p-value: < 2.2e-16
```

So all components are quite significant as the model has an R-squared of 86.59%. A very high degree of correlation

Now lets perform an ANOVA to compare the models

```
anova(basefit,bestfit)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ cyl + hp + wt + am
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     26 151.03  4    569.87 24.527 1.688e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

So the P-value is very small and quite significant, therefore We reject the Null hypothesis that WT, cyl, and hp do not contribute to the model accuracy.

### Inference on effect of Transmition on MPG

lets perform a T Test.

```
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 0 mean in group 1
##        17.14737        24.39231
```

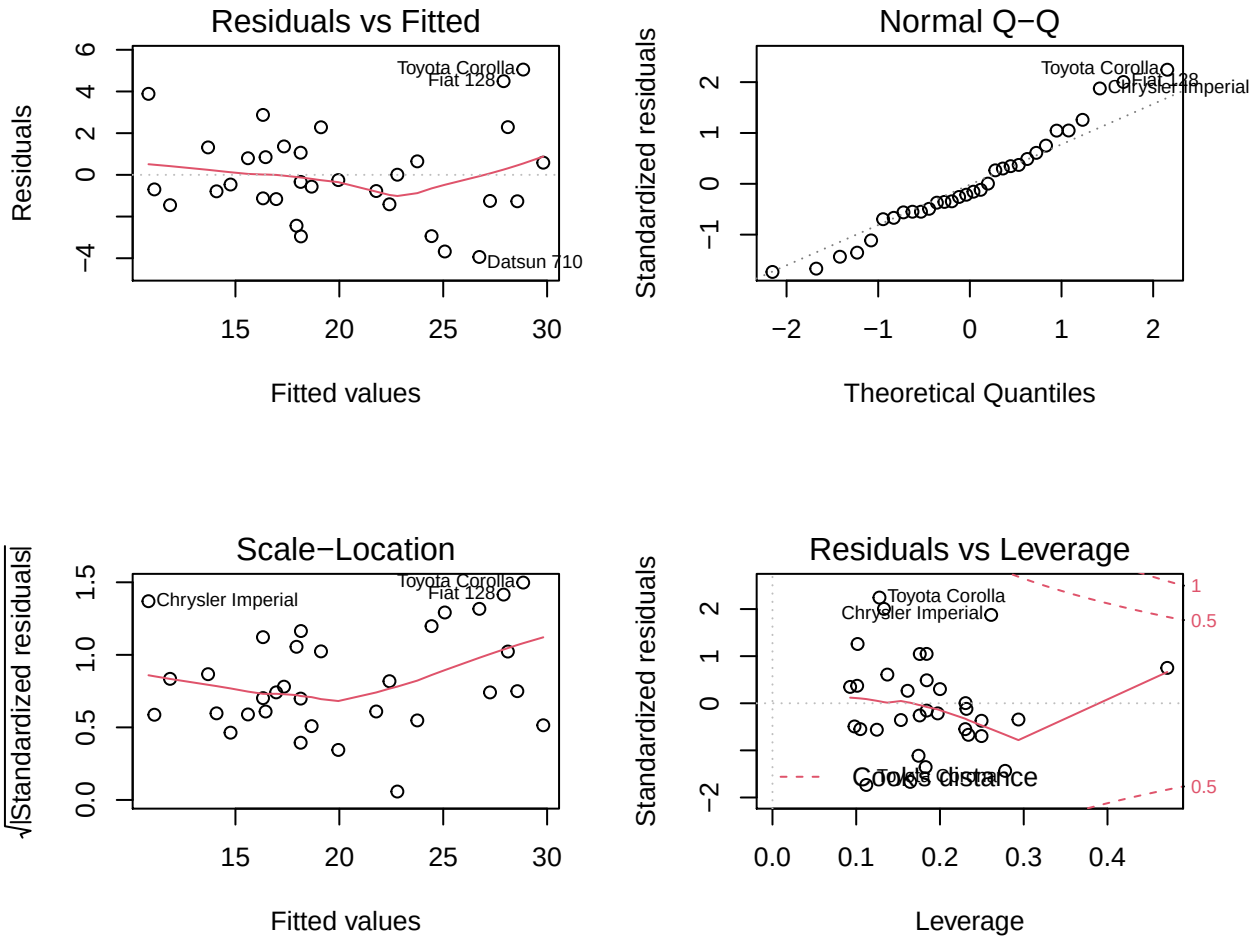
And we see that the Type of transmission has a significant impact on the level of Miles pre Gallon achieved. that is greater than a 95% confidence, as seen in the T Test.

## Appendix

### Residual, Diagnostics and Plots

We compute some regression diagnostics on the bestfit model residuals

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




From the above plots, we can make the following observations,

- The points in the Residuals vs. Fitted plot seem to be randomly scattered on the plot and verify the independence condition.
- The Normal Q-Q plot consists of the points which mostly fall on the line indicating that the residuals are normally distributed.
- The Scale-Location plot consists of points scattered in a constant band pattern, indicating constant variance.
- There are some distinct points of interest (outliers or leverage points) in the top right of the plots.

We now compute some regression diagnostics of our model to find out these interesting leverage points as shown in the following section.

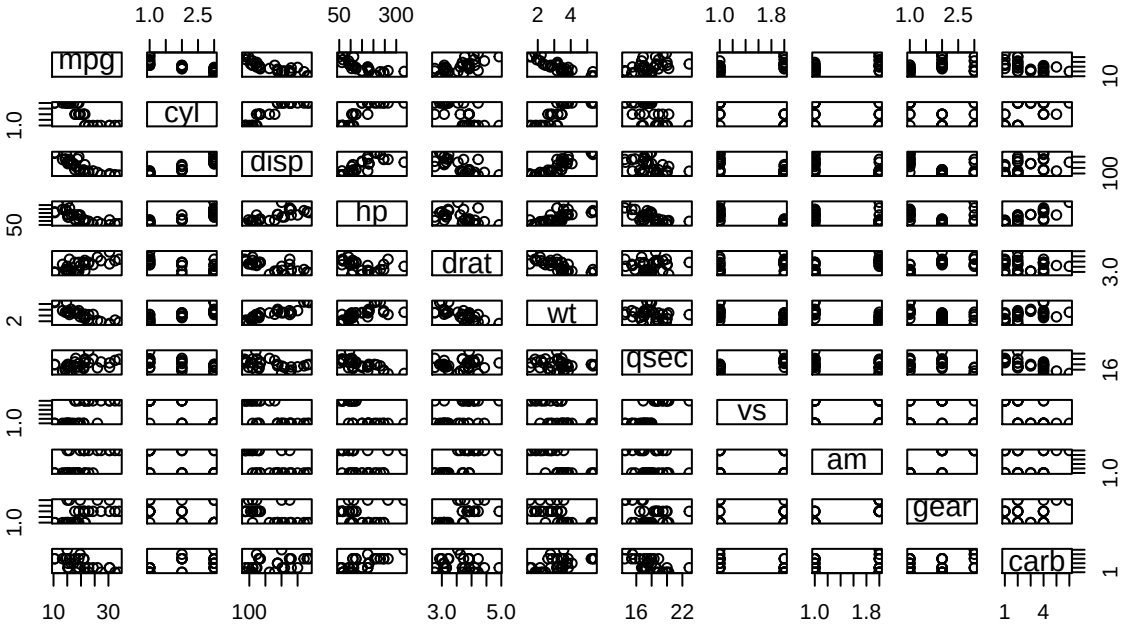
```
Lev <- hatvalues(bestfit)
tail(sort(Lev),3)
```

```
##       Toyota Corona Lincoln Continental       Maserati Bora
##           0.2777872           0.2936819           0.4713671
```

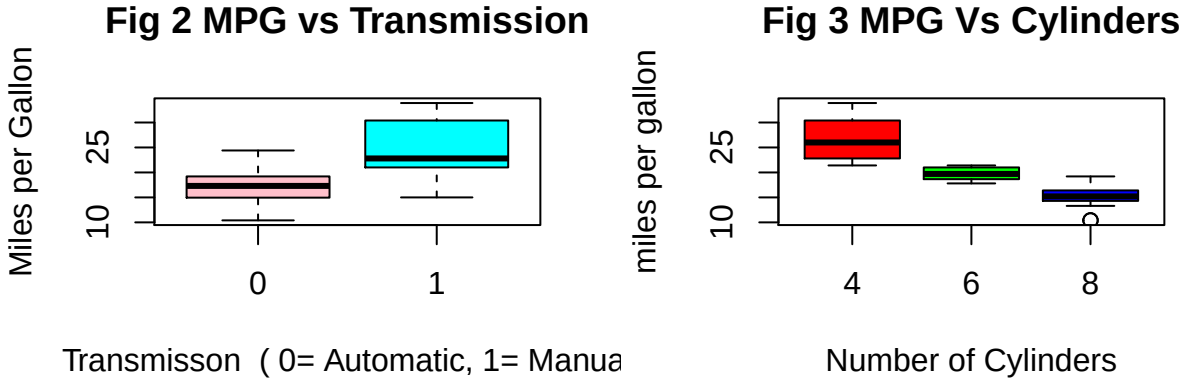
```
resinf <- dfbetas(bestfit)
tail(sort(resinf[,6]),3)
```

```
## Chrysler Imperial          Fiat 128     Toyota Corona
##         0.3507458         0.4292043         0.7305402
```

The same cars are mentioned in the residual plots, our analysis looks correct.




Scatter plots matrix of MTCARS dataset




e