Regression Model

Theo Livingwell

October 17, 2017

Executive summary

This edition of Motor Trend magazine explores the relationship between a set of variables and miles per gallon (MPG) (outcome). The two questions we shall try to answer are:

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

Perhaps this will help you decide on the type of transmission on your next car purchase.

Load required package, data and explore the data

```
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
data("mtcars")
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 ...
```

```
names(mtcars)
```

```
##  [1] "mpg"  "cyl"  "disp" "hp"   "drat" "wt"   "qsec" "vs"   "am"   "gear"
## [11] "carb"
```

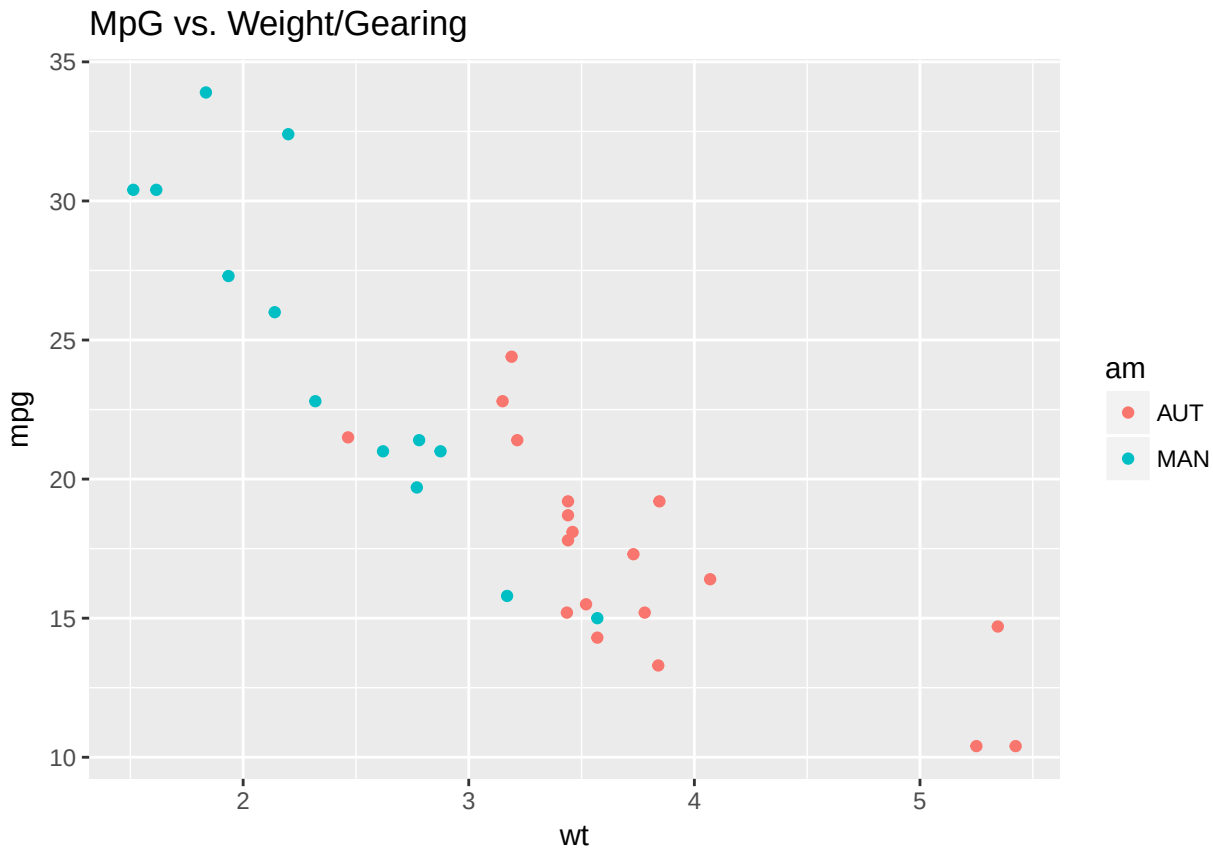
```
summary(mtcars)
```

```
##       mpg             cyl             disp             hp
##  Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0
##  1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5
```

```
##  Median :19.20   Median :6.000   Median :196.3   Median :123.0
##  Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7
##  3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0
##  Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0
##       drat             wt             qsec             vs
##  Min.   :2.760   Min.   :1.513   Min.   :14.50   Min.   :0.0000
##  1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000
##  Median :3.695   Median :3.325   Median :17.71   Median :0.0000
##  Mean   :3.597   Mean   :3.217   Mean   :17.85   Mean   :0.4375
##  3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000
##  Max.   :4.930   Max.   :5.424   Max.   :22.90   Max.   :1.0000
##        am              gear            carb
##  Min.   :0.0000   Min.   :3.000   Min.   :1.000
##  1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000
##  Median :0.0000   Median :4.000   Median :2.000
##  Mean   :0.4062   Mean   :3.688   Mean   :2.812
##  3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000
##  Max.   :1.0000   Max.   :5.000   Max.   :8.000
```

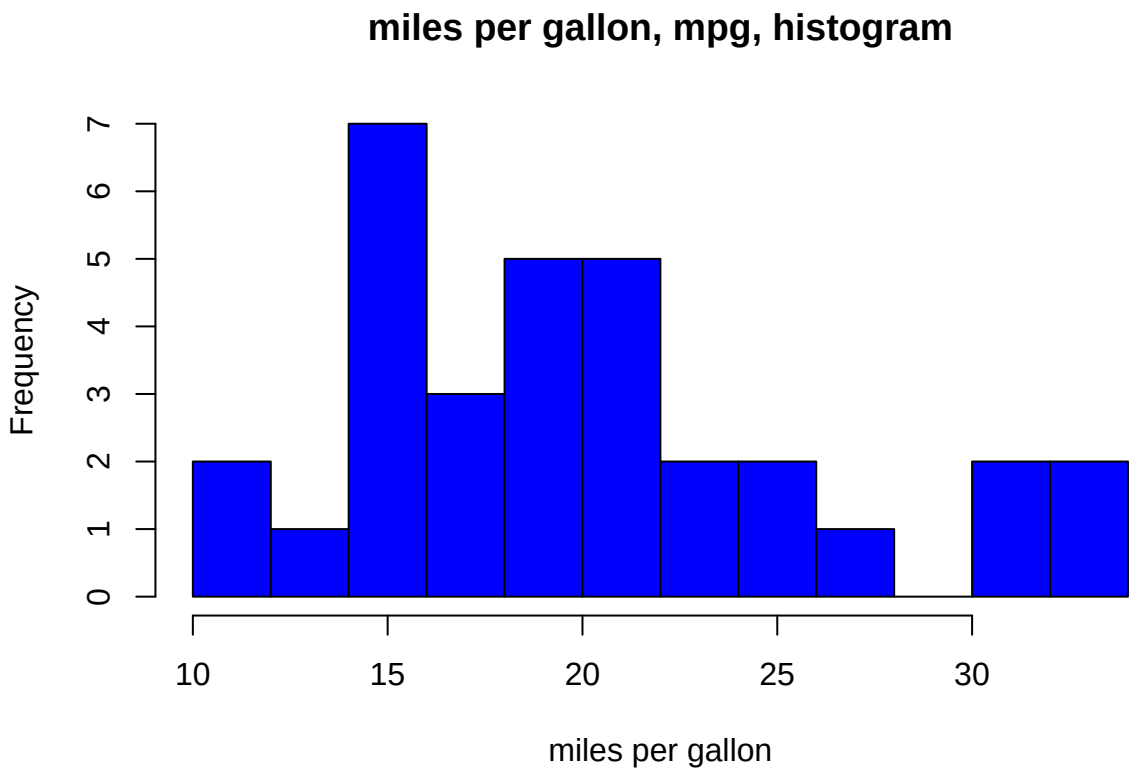
The data frame has 32 obervations of 11 variables. The variables are of the type numeric. To manual and automatic transmission cars successfully, lets change the variable am from the type numeric to factor. Herre AUT = automatic transmission and MAN = manual transmission

```
mtcars$am <- as.factor(mtcars$am)
levels(mtcars$am) <- c("AUT", "MAN")
print(qplot(x=wt, y=mpg, colour=am, data=mtcars,main="MpG vs. Weight/Gearing"))
```



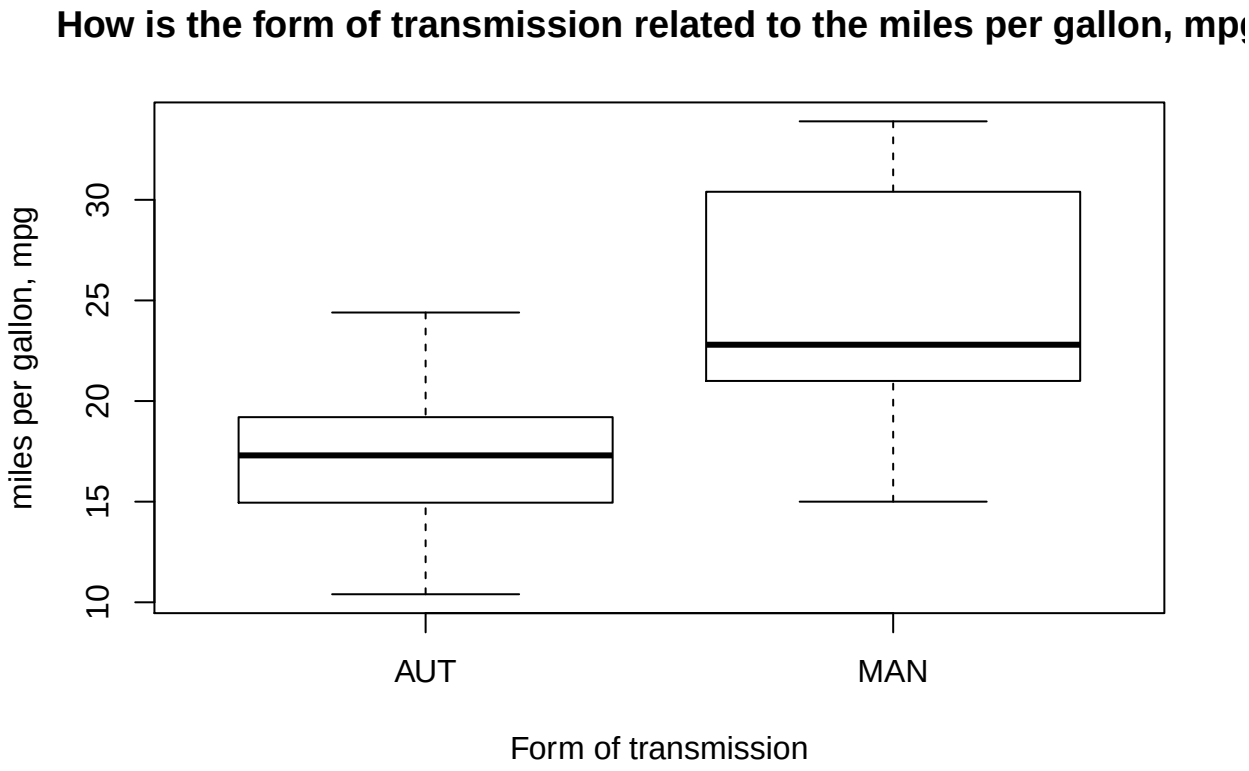
Lets look at how the distribution of mpg approximates Gausian (normal) distribution with a histogram

```
hist(mtcars$mpg, breaks=10, col = "blue", xlab = "miles per gallon", main = "miles per gallon, mpg, his
```



what's the relationship between transmission type and mpg?

```
boxplot(mpg~am, data=mtcars, xlab="Form of transmission", ylab="miles per gallon, mpg", main="How is th
```



The boxplot shows the mean of AUT to be 17.5 which is lower than the mean of MAN which is 22.5. Manual transmission cars are higher on miles per pallon, mpg, compared to automatic transmission

```
mpg_aut_trans <- mtcars[mtcars$am =="AUT", ]$mpg
mpg_man_trans <- mtcars[mtcars$am == "MAN", ]$mpg
t.test(mpg_aut_trans, mpg_man_trans)
```

```
##
##  Welch Two Sample t-test
##
## data:  mpg_aut_trans and mpg_man_trans
## 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
```

A p-value of 0.001374, suggest that we accept the alternative hypothesis that automatic cars have less mpg compared to manual cars. This woulf be the case if we assum that all features of manual and automatic cars are the same.

let's take a look at a linear model

```
lm_fit <- lm(mpg~am, mtcars)
summary(lm_fit)
```

```
##
```

```
## 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 ***
## amMAN          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
```

the altenative hypothesis is accepted by a p-value of 0.000285. The R squared value is 0.3598 our model explains 35.98% of variance

let's use the step function to look at a a multivariate regression model

```
new_model1 <- step(lm(data=mtcars, mpg~ .,), trace = 0, steps=10000)
summary(new_model1)
```

```
##
## 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 ***
## amMAN         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
```

An R squared value of 0.85 indicates the model explains 84% of variance in mpg. Weight of cars and acceleration speed have the highest relation in explaining the variance in mpg

lets look at a model with 3 variables- wt, qsec, and am

```
three_fit_model <- lm(mpg~am + wt+qsec, data=mtcars)
anova(lm_fit,three_fit_model)
```

```
## Analysis of Variance Table
##
```

```
## Model 1: mpg ~ am
## Model 2: mpg ~ am + wt + qsec
##   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
```

this model shhow 84% of variation in mpg, and p-value of 3.745e-09. Again we accept the alternative hypothesis that our multivariate model is marked difference from our simple linear model

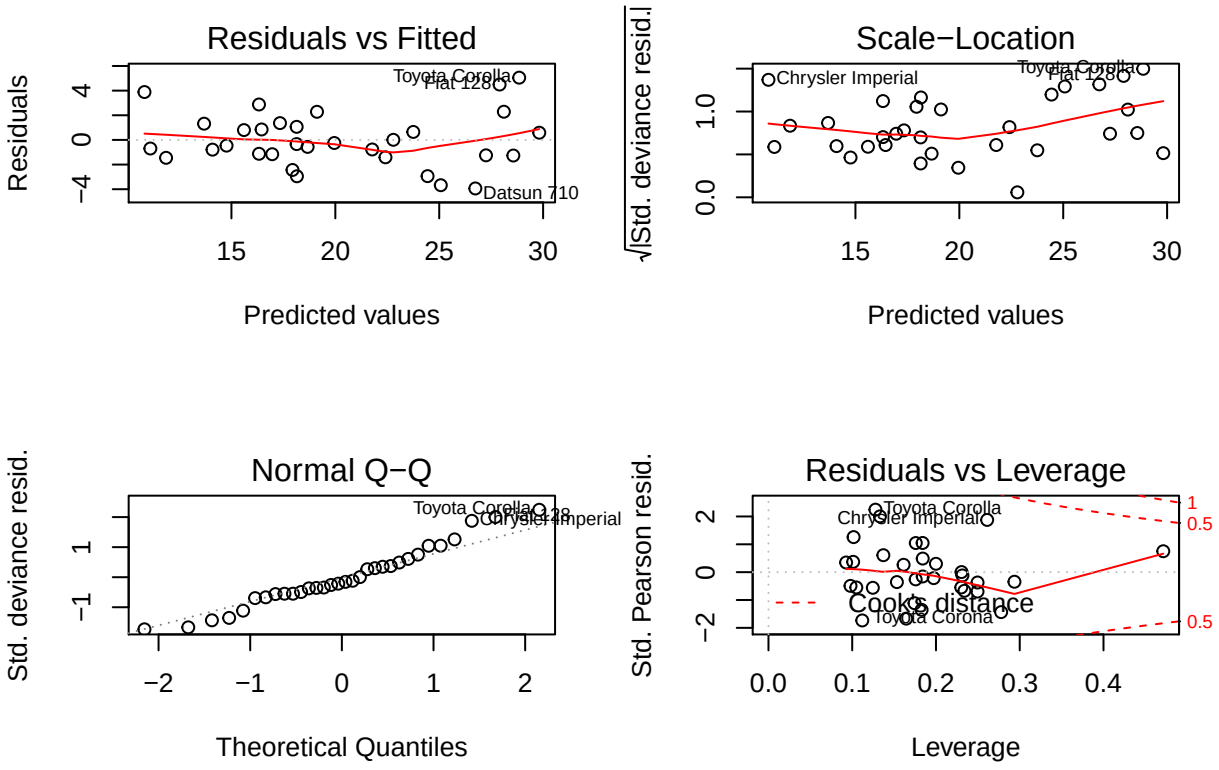
```
summary(three_fit_model)
```

```
##
## Call:
## lm(formula = mpg ~ am + wt + qsec, 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
## amMAN         2.9358     1.4109   2.081 0.046716 *
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## ---
## 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
```

our model explains 84% of variancein mpg. It indicates manual transmission has 2.94 mpg more compared to automatic transmission cars. Manual transmission is therefore better on mpg compared to automatic transmission.

Here is another model

```
model_final <-glm(mpg ~ as.factor(cyl) + as.factor(am) + hp + wt, data=mtcars)
layout(matrix(c(1,2,3,4),2,2))
plot(model_final)
```




```
summary(model_final)
```

```
##
## Call:
## glm(formula = mpg ~ as.factor(cyl) + as.factor(am) + hp + wt,
##     data = mtcars)
##
## Deviance 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 ***
## as.factor(cyl)6  -3.03134    1.40728  -2.154  0.04068 *
## as.factor(cyl)8  -2.16368    2.28425  -0.947  0.35225
## as.factor(am)MAN  1.80921    1.39630   1.296  0.20646
## hp               -0.03211    0.01369  -2.345  0.02693 *
## wt               -2.49683    0.88559  -2.819  0.00908 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 5.808677)
##
##     Null deviance: 1126.05  on 31  degrees of freedom
## Residual deviance:  151.03  on 26  degrees of freedom
```

```
## AIC: 154.47
##
## Number of Fisher Scoring iterations: 2
```