# Exploring the relationships between a set of variables and miles per gallon for a collection of cars

## Synopsis

Looking at a data set of a collection of cars, we are interested in exploring the relationship between a set of variables and miles per gallon (MPG). We are particularly interested in the following two questions:

- Is an automatic or manual transmission better for MPG
- Quantify the MPG difference between automatic and manual transmissions

## Exploring the data

Let's load dataset and look at the data primary features:

```
library(datasets)

data(mtcars)

dim(mtcars)
```

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

```
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
```

```
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
```

Dataset contains following 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 (lb/1000)
7. qsec - ¼ mile time
8. vs - V/S
9. am - Transmission (0 = automatic, 1 = manual)
10. gear - Number of forward gears
11. carb - Number of carburetors

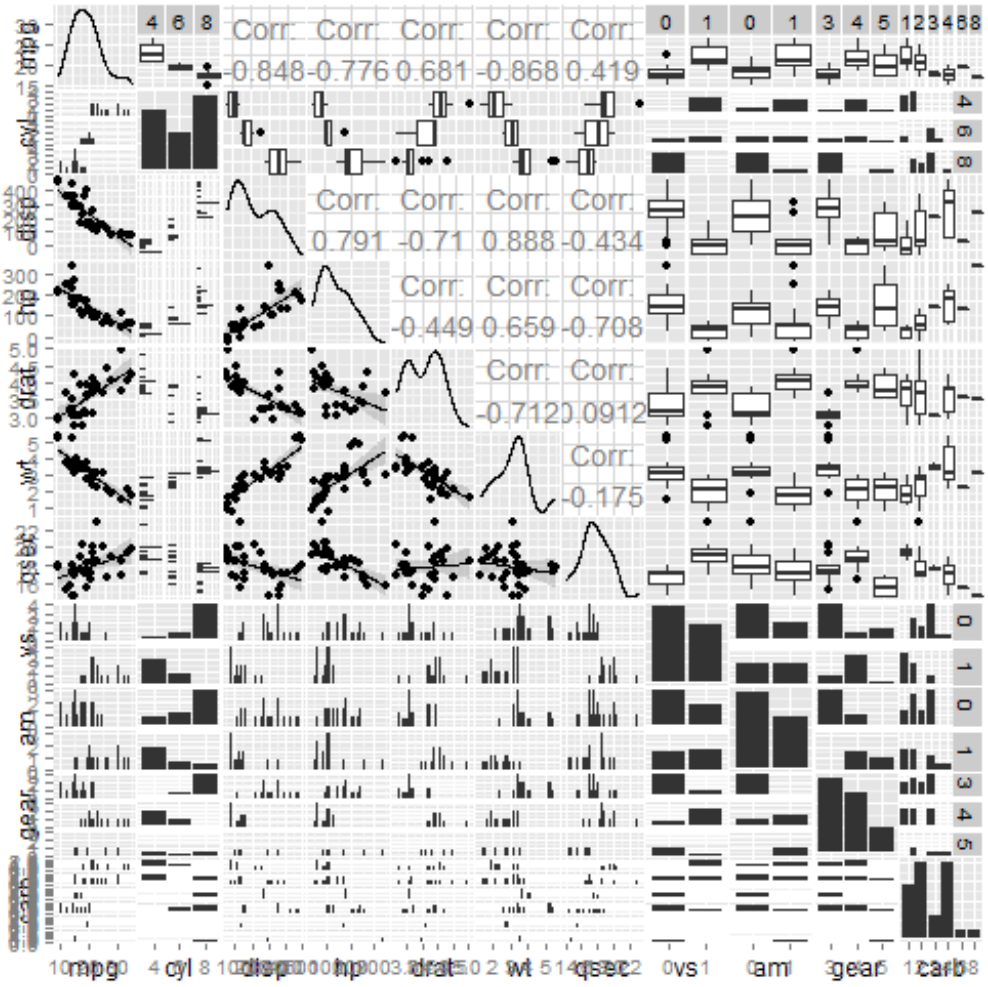
As we can see, there are bunch of variables that should be factors rather than numeric. Let's transofrm them appropriately:

```
mtcars$cyl = factor(mtcars$cyl)
mtcars$vs = factor(mtcars$vs)
mtcars$am = factor(mtcars$am)
mtcars$gear = factor(mtcars$gear)
mtcars$carb = factor(mtcars$carb)

summary(mtcars)
```

```
##       mpg        cyl         disp             hp             drat
##  Min.   :10.40   4:11   Min.   : 71.1   Min.   : 52.0   Min.   :2.760
##  1st Qu.:15.43   6: 7   1st Qu.:120.8   1st Qu.: 96.5   1st Qu.:3.080
##  Median :19.20   8:14   Median :196.3   Median :123.0   Median :3.695
##  Mean   :20.09          Mean   :230.7   Mean   :146.7   Mean   :3.597
##  3rd Qu.:22.80          3rd Qu.:326.0   3rd Qu.:180.0   3rd Qu.:3.920
##  Max.   :33.90          Max.   :472.0   Max.   :335.0   Max.   :4.930
##        wt             qsec       vs     am     gear   carb
##  Min.   :1.513   Min.   :14.50   0:18   0:19   3:15   1: 7
##  1st Qu.:2.581   1st Qu.:16.89   1:14   1:13   4:12   2:10
##  Median :3.325   Median :17.71                 5: 5   3: 3
##  Mean   :3.217   Mean   :17.85                        4:10
##  3rd Qu.:3.610   3rd Qu.:18.90                        6: 1
##  Max.   :5.424   Max.   :22.90                        8: 1
```

Let's look at relationships between different variables:

```
library(GGally)
library(ggplot2)

ggpairs(mtcars, lower = list(continuous = "smooth"), params = c(method = "loess"))
```




# Building data model

Let's try to build data model based on all possible variables:

```
fit1 <- lm(mpg ~ am, mtcars)
fit2 <- lm(mpg ~ am + cyl, mtcars)
fit3 <- lm(mpg ~ am + cyl + disp, mtcars)
fit4 <- lm(mpg ~ am + cyl + disp + hp, mtcars)
fit5 <- lm(mpg ~ am + cyl + disp + hp + drat, mtcars)
fit6 <- lm(mpg ~ am + cyl + disp + hp + drat + wt, mtcars)
fit7 <- lm(mpg ~ am + cyl + disp + hp + drat + wt + qsec, mtcars)
fit8 <- lm(mpg ~ am + cyl + disp + hp + drat + wt + qsec + vs, mtcars)
fit8 <- lm(mpg ~ am + cyl + disp + hp + drat + wt + qsec + vs + gear, mtcars)
fit9 <- lm(mpg ~ am + cyl + disp + hp + drat + wt + qsec + vs + gear + carb, mtcars)

anova(fit1, fit2, fit3, fit4, fit5, fit6, fit7, fit8, fit9)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + cyl
## Model 3: mpg ~ am + cyl + disp
## Model 4: mpg ~ am + cyl + disp + hp
## Model 5: mpg ~ am + cyl + disp + hp + drat
## Model 6: mpg ~ am + cyl + disp + hp + drat + wt
## Model 7: mpg ~ am + cyl + disp + hp + drat + wt + qsec
## Model 8: mpg ~ am + cyl + disp + hp + drat + wt + qsec + vs + gear
## Model 9: mpg ~ am + cyl + disp + hp + drat + wt + qsec + vs + gear + carb
##   Res.Df    RSS Df Sum of Sq       F   Pr(>F)
## 1     30 720.90
## 2     28 264.50  2    456.40 28.4297 7.89e-06 ***
## 3     27 230.46  1     34.04  4.2402  0.05728 .
## 4     26 183.04  1     47.42  5.9078  0.02809 *
## 5     25 182.38  1      0.66  0.0820  0.77855
## 6     24 150.10  1     32.28  4.0216  0.06331 .
## 7     23 141.21  1      8.89  1.1081  0.30916
## 8     20 134.00  3      7.20  0.2992  0.82547
## 9     15 120.40  5     13.60  0.3388  0.88144
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As we can see, besides transmission information, adding number of cylinders and gross horsepower pass sygnificance test. Now, let's check if interactions between these values are significant:

```
fit1 <- lm(mpg ~ am + cyl + hp, mtcars)
fit2 <- lm(mpg ~ am + cyl + hp + am*cyl, mtcars)
fit3 <- lm(mpg ~ am + cyl + hp + am*cyl + am*hp, mtcars)
fit4 <- lm(mpg ~ am + cyl + hp + am*cyl + am*hp + cyl*hp, mtcars)
fit5 <- lm(mpg ~ am + cyl + hp + am*cyl + am*hp + cyl*hp + am*cyl*hp, mtcars)
anova(fit1, fit2, fit3, fit4, fit5)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am + cyl + hp
## Model 2: mpg ~ am + cyl + hp + am * cyl
## Model 3: mpg ~ am + cyl + hp + am * cyl + am * hp
## Model 4: mpg ~ am + cyl + hp + am * cyl + am * hp + cyl * hp
## Model 5: mpg ~ am + cyl + hp + am * cyl + am * hp + cyl * hp + am * cyl *
##     hp
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     27 197.20
## 2     25 187.04  2   10.1614 0.6160 0.5500
## 3     24 186.30  1    0.7400 0.0897 0.7676
## 4     22 167.90  2   18.3990 1.1154 0.3473
## 5     20 164.95  2    2.9499 0.1788 0.8376
```

As we can see, interactions have no significance. So let's take original model:

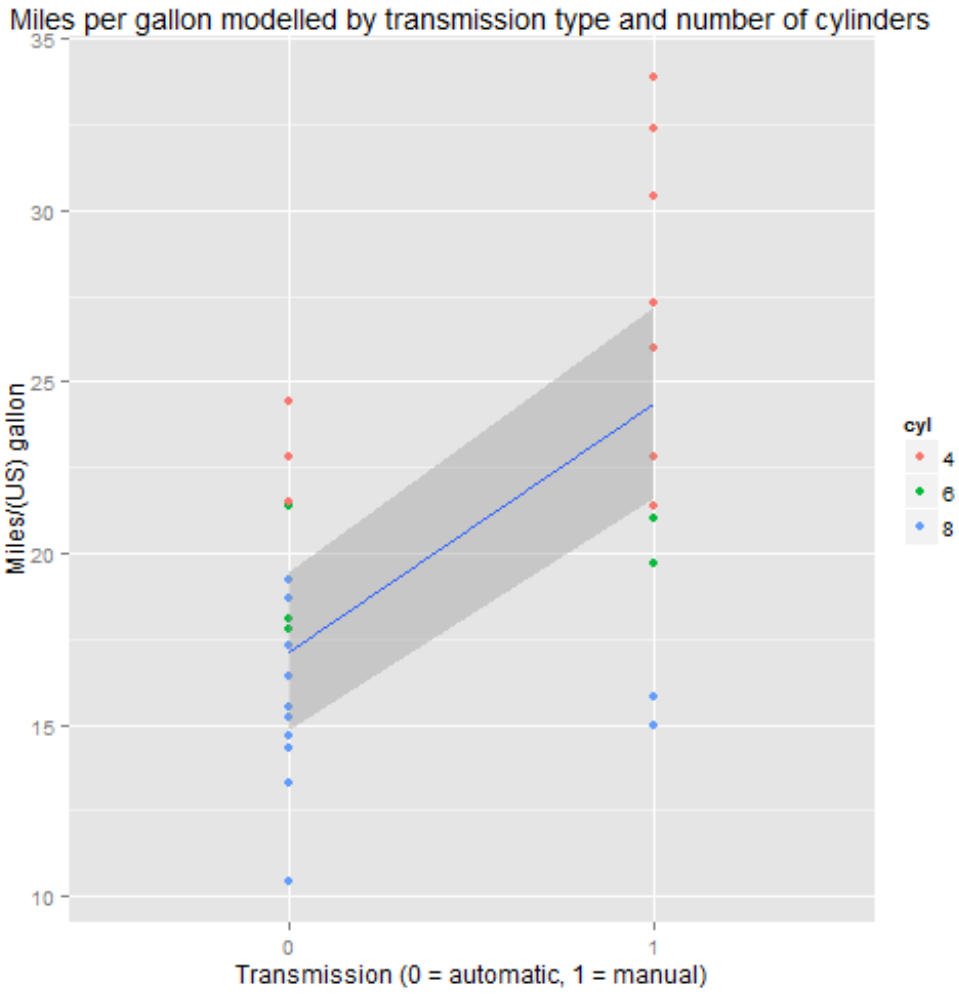
```
fit <- lm(mpg ~ am + cyl + hp, mtcars)
summary(fit)
```

```
##
## Call:
## lm(formula = mpg ~ am + cyl + hp, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.231 -1.535 -0.141  1.408  5.322
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 27.29590    1.42394  19.169  < 2e-16 ***
## am1          4.15786    1.25655   3.309  0.00266 **
## cyl6        -3.92458    1.53751  -2.553  0.01666 *
## cyl8        -3.53341    2.50279  -1.412  0.16943
## hp          -0.04424    0.01458  -3.035  0.00527 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.703 on 27 degrees of freedom
## Multiple R-squared:  0.8249, Adjusted R-squared:  0.7989
## F-statistic: 31.79 on 4 and 27 DF,  p-value: 7.401e-10
```

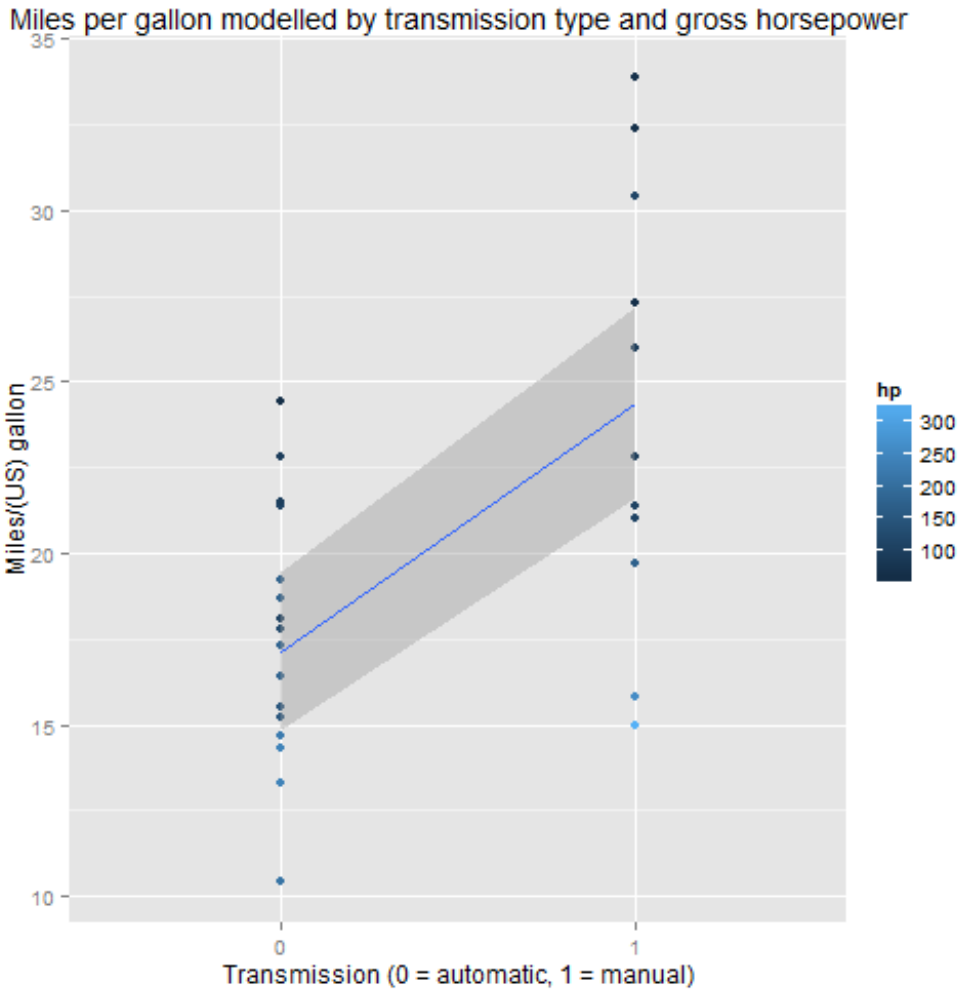
As we can see, manual transmission brings more miles per gallon with mean value 4.1578565 and standard error 1.25655. So, this value is significant based on t-test, for >95% confidence interval.

Let's plot given model:

```
ggplot(mtcars, aes(x = am, y = mpg, group = 1)) +
geom_point(aes(color = cyl)) +
geom_smooth(method = "lm", formula = y ~ x) +
labs(x = 'Transmission (0 = automatic, 1 = manual)') +
labs(y = 'Miles/(US) gallon') +
labs(title = 'Miles per gallon modelled by transmission type and number of cylinders')
```

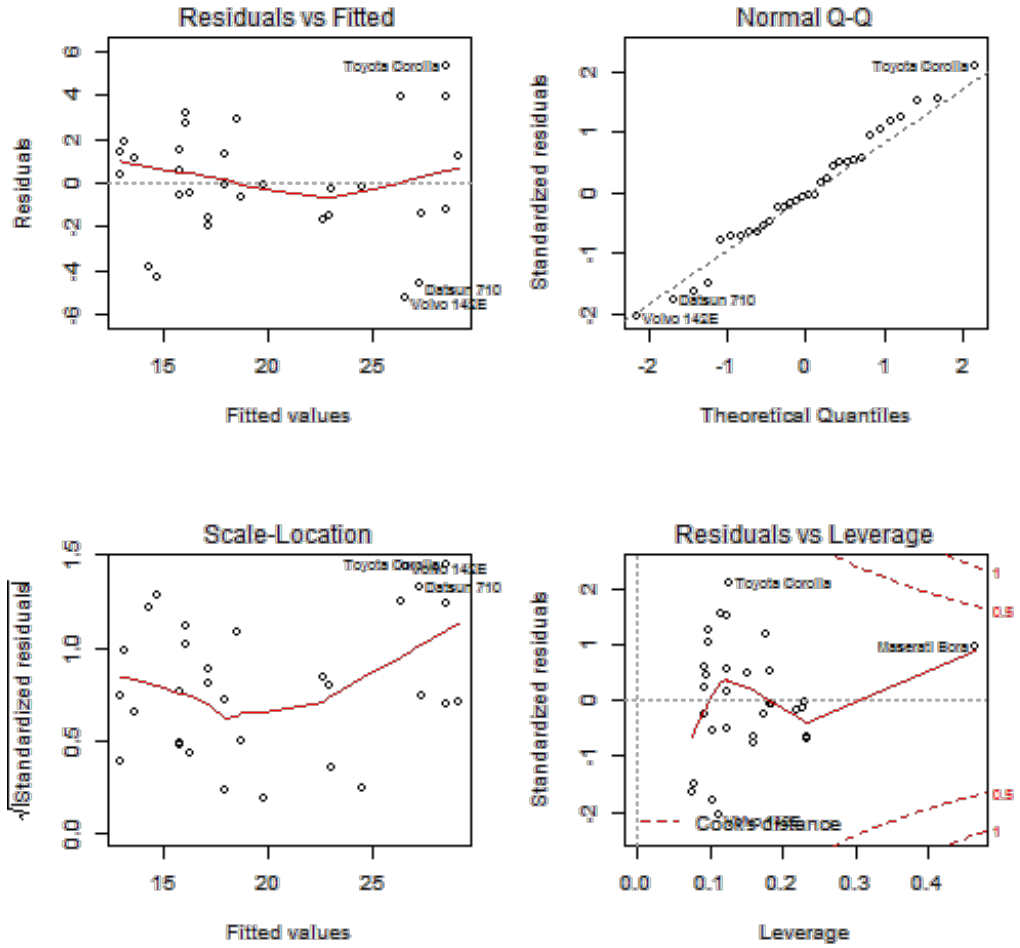



```
ggplot(mtcars, aes(x = am, y = mpg, group = 1)) +
geom_point(aes(color = hp)) +
geom_smooth(method = "lm", formula = y ~ x) +
labs(x = 'Transmission (0 = automatic, 1 = manual)') +
labs(y = 'Miles/(US) gallon') +
labs(title = 'Miles per gallon modelled by transmission type and gross horsepower')
```




And finally let's look at some diagnostics:

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




As we can see, our model is acceptable and can be considered significant for investigations.

# Summary

So, final conclusion is that:

- we can build acceptable model of relation between miles per gallon value and transmission type, number of cylinders and gross horsepower
- manual transmission brings more miles per gallon with mean value 4.1578565 and standard error 1.25655
- manual transmission is significantly better than automatic for miles per gallon