title: "Regression Models - Course Project, 2015" author: "Albion Dervishi" date: "April 25, 2015"

In this project we will analyze the data in mtcars dataset, specifically relationship between many variables recorded and the miles per gallon (MPG) of the cars.

Types of variables in the dataset: mpg = Miles/(US) gallon; cyl = Number of cylinders; wt = Weight (lb/1000); vs = V/S; am = Transmission (0 = automatic, 1 = manual); gear = Number of forward gears; carb = Number of carburetors

Dataset- "mtcars"

Creation of correlation model includes all variables as predicour initial model in correlation of mpg. This model includes all variables as predictors of mpg. In base our results in all cells for regression a model seems that are compatible in linear model. This analysis also gives us information, when we will do a linear model later on. (Fig 1)

```
library(dplyr);library(ggplot2); library(grid); library(gridExtra);
library(corrplot)data(mtcars)
wt<-factor(mtcars$wt); cyl <- factor(mtcars$cyl); vs <- factor(mtcars$vs); gear <-
factor(mtcars$gear)
carb <- factor(mtcars$carb); am <- factor(mtcars$am,labels=c("Automatic","Manual"))
Ccars<- cor(mtcars); corrplot(Ccars)
```

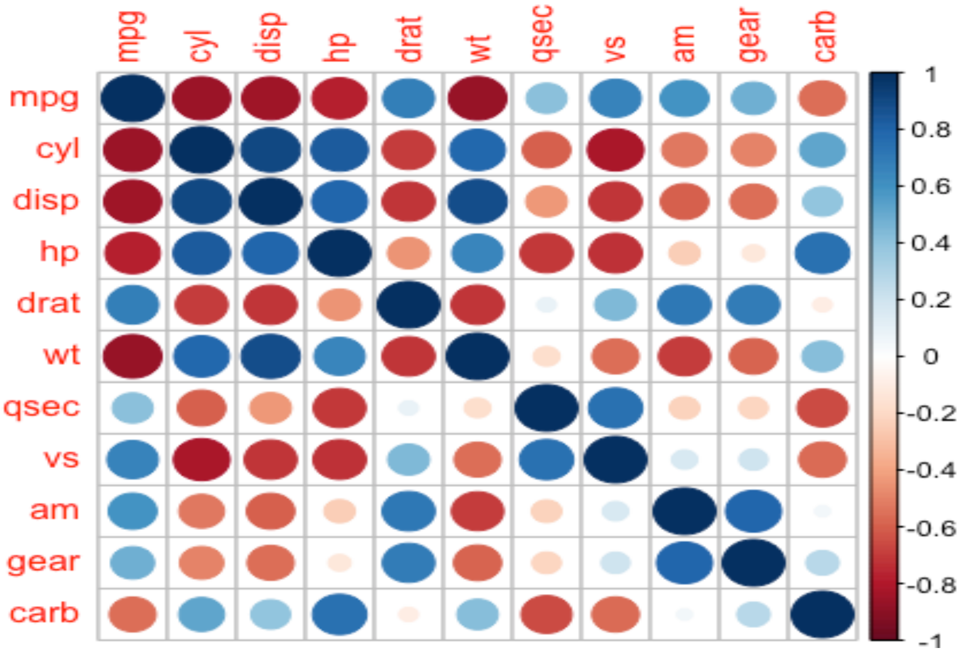

Fig.1

Multivariate Model

Creation of the models that reports fuel consumption on the basis of observed variables in the dataset mtcars. (Fig.2)

I1 - Value of MPG and type of transmission have linear relationship and from the model, we can infer that transmission mode has an impact on MPG. Model suggests that average MPG for automatic transmission is 17 and average MPG for manual transmission is 24. Manual transmission get 7.2 MPG improvement when compared to automatic.

l2-Value of MPG and weight of car mpg have a clear correlation, will decrease by 5 for every 1000 lb increase in wt.

l3 - Shows number of cylinders, cyl increases from 4 to 6 and 8, mpg will decrease by a factor of 2.8 respectively.

l4- Value of MPG and VS have clear positive correlation by increasing factor 7.9

l5- Value of MPG and number of cylinders have positive correlation, cyl increases from 4 to 6 and 8, mpg will increase by a factor of 3.9

l6- Value of MPG and number of carburetors have clear negative correlation by decreasing factor 2

```
grid.arrange(l1, l2, l3, l4, l5, l6, ncol=3, main ="mtcars")

l1<-qplot(am, mpg, data = mtcars)+ stat_smooth(method="lm", se=FALSE)
coef(lm(mpg ~ am, data = mtcars))

## (Intercept)          wt
##   37.285126   7.244939

l2<-qplot(wt, mpg, data = mtcars)+ geom_abline(intercept = 37, slope = -5)
coef(lm(mpg ~ wt, data = mtcars))

## (Intercept)          wt
##   37.285126  -5.344472

l3<-qplot(cyl, mpg, data = mtcars)+ geom_abline(intercept = 37.8, slope = -2.8)
coef(lm(mpg ~ cyl, data = mtcars))

## (Intercept)         cyl
##    37.88458   -2.87579

l4<-qplot(vs, mpg, data = mtcars)+ geom_abline(intercept = 16.6, slope = 7.9)
coef(lm(mpg ~ vs, data = mtcars))

## (Intercept)          vs
##   16.616667    7.940476

l5<-qplot(gear, mpg, data = mtcars)+ geom_abline(intercept = 5.6 , slope = 3.9)
coef(lm(mpg ~ gear, data = mtcars))

## (Intercept)        gear
##    5.623333    3.923333

l6<-qplot(carb, mpg, data = mtcars)+ geom_abline(intercept = 25.8 , slope = -2)
coef(lm(mpg ~ carb, data = mtcars))

## (Intercept)        carb
##   25.872334  -2.055719
```

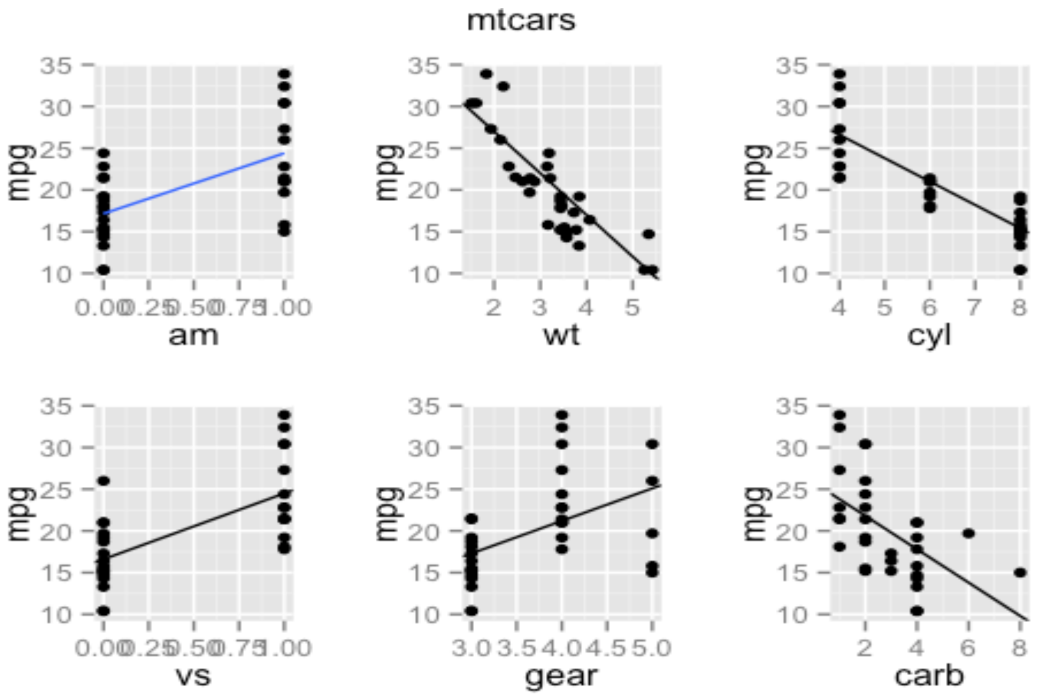

Fig. 2

Inference of output

The analysis suggests that selected model has statistically importance for variables in the “mtcars” dataset in relationship to the variation of mpg. The steps function will perform this selection by calling lm repeatedly. It selects the best variables to use in predicting mpg with other variables "cyl","hp", "am", “disp”, "gear" and "carb".

```
lm<- lm(mpg ~ ., data = mtcars); summary(lm)
## lm(formula = mpg ~ ., 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|)
## (Intercept) 12.30337   18.71788   0.657   0.5181
## 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
## am           2.52023    2.05665   1.225   0.2340
## 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:  0.869,  Adjusted R-squared:  0.8066
## F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```

```
anova(lm)

## Analysis of Variance Table
##
## Response: mpg
##           Df Sum Sq Mean Sq  F value    Pr(>F)
## cyl        1 817.71  817.71 116.4245 5.034e-10 ***
## disp       1  37.59   37.59   5.3526  0.030911 *
## hp         1   9.37    9.37   1.3342  0.261031
## drat       1  16.47   16.47   2.3446  0.140644
## wt         1  77.48   77.48  11.0309  0.003244 **
## qsec       1   3.95    3.95   0.5623  0.461656
## vs         1   0.13    0.13   0.0185  0.893173
## am         1  14.47   14.47   2.0608  0.165858
## gear       1   0.97    0.97   0.1384  0.713653
## carb       1   0.41    0.41   0.0579  0.812179
## Residuals 21 147.49    7.02
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The adjusted R-squared value of 0.80 tells us that 80% of the variability is explained by this model. From these results we can conclude that more than 80% of the variability is explained by the above model. The p-value obtained with anova test is significant (p<0.05) and we reject the null hypothesis in the variables wt and disp.

Regression through the origin

This analysis and graphic using the origin as a pivot point picking the line that minimizes the sum of the squared vertical distances of the points to the line for MPG and mode of transmission(in two levels "Automatic" and "Manual"). Subtract the means so that the origin is the mean of the MPG and mode of transmission as factor. (Fig.3)

```
data(mtcars)
mtcars<-data.frame(mtcars)
a<-factor(mtcars$mpg)
x <-mtcars$mpg - mean(mtcars$mpg)
y<-factor(mtcars$am,labels=c("Automatic","Manual"))
car <- as.data.frame(table(a,x, y))
names(car) <- c("MPG","mpg", "am", "freq")

g <- ggplot(filter(car, freq > -1), aes(am, MPG), + facet_grid(.~am))
g <- g + xlab("Type of car transmission")
g <- g + ylab("Miles per gallon")
g <- g  + scale_size(range = c(4, 20), guide = "none" )
g <- g + geom_point(colour="grey90",  aes(size = freq+ 0.2, show_guide = FALSE))
g <- g + geom_point(aes(colour=freq, size = freq))
g <- g + geom_point(colour="grey80")
g <- g + scale_colour_gradient(low = "yellow", high="red")
g <- g + geom_smooth(aes(group=freq), method="lm", fullrange=TRUE)
g
```

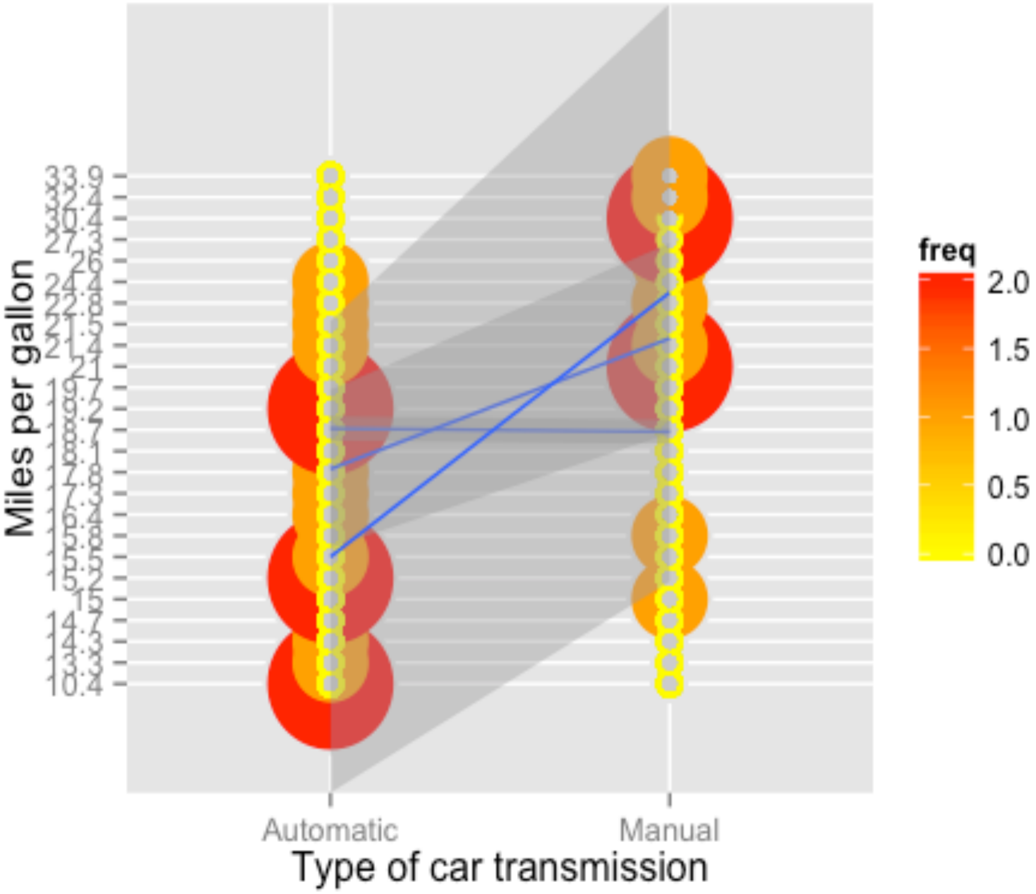
Miles per gallon
33.9
32.4
30.4
27.3
26
24.4
22.8
21.5
21.4
21
19.7
19.2
18.7
18.1
17.8
17.3
16.4
15.8
15.5
15.2
15
14.7
14.3
13.3
10.4
Automatic
Manual
Type of car transmission
freq
2.0
1.5
1.0
0.5
0.0

Fig.3