title: "Regression Models Course Project: Motor data analisys"
author: "Andrey ABRAMOV"
date: "10th of October 2016"
output: pdf_document

Executive summary

In my report I will analyze data set name mtcars and try to find relationship between miles per galon ("mpg"" variable) and all other variables. Data set was taken from the 1974 Motor Trend US magazine and futured of 32 description for 1973?74 years models. I will apply regression models in order to explain what is a different in miles per gallon (mpg) for car with automatic (am=0) and manual (am=1) transmission. I will show the process of finding the best model. I will use logarithm of mpg due to heteroscedasticity in my model.I will show what is different in mpg for two cars with the same parameters and different transmission. Also will be shown dependence between the transmission type and a car horsepower. This result shows that cars with manual transmittion add 1.37*wt+0.356*carbˆ2 more MPG and subtracts -0.212*cyl-1.76*carb-1.72*wtˆ2 of MPG in average than cars with automatic tranmission. According to the model if you are choosing a car less than 2.2 tn of weight (and 4 cyl, 150 hp, 200 cu.in. engine, 4 forward gears and 4 carb) is better to take a car with automatic transmission.

Course project goal

I should explore the relationship between a set of variables and miles per gallon (MPG) (outcome). And interested in the following two questions: - Is an automatic or manual transmission better for MPG - Quantify the MPG difference between automatic and manual transmissions Instruction for Course project could be found at https://www.coursera.org/learn/regression-models/peer/nxntd/regression-models-course-project.

Data source

The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973?74 models).

```r
# library preparing
library(ggplot2)
library(broom)
library(grid)
library(gridExtra)
data(mtcars)
```

Data quick view

```r
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
```

```r
dim(mtcars)
```

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

So we have a data frame with 32 observations on 11 variables. The names of the variables correspond to the following data: [, 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 V/S [, 9] am Transmission (0 = automatic, 1 = manual) [,10] gear Number of forward gears [,11] carb Number of carburetors On my opinion #7, #8 and possible #5 are not descriptive characteristics of MPG of a car. I will not include them to analisys.

```r
mtcars <- subset(mtcars, select = -c(vs, qsec, drat))
mtcars$ttype[which(mtcars$am==1)] <- "manual"
mtcars$ttype[which(mtcars$am==0)] <- "auto"
mtcars$ttype <- as.factor(mtcars$ttype)
```

Exploratory data analyses

Let have a look on manual and automatic transmissions in a common plot - Enclose #1. According to this plot manual transmission has higher MPG values versus atomatic on the lower part of weight scale. Based on pair plot analisys (ENCLOSE #1) some correlation could be found between mpg and disp, hp, wt.

Inference analisys

First of all I need to check does mpg of different types of transmittion are from different groups? Let's use for that goal t-test:

```r
mtcars_inf_check <- t.test(mtcars$mpg ~ mtcars$am)
mtcars_inf_check$p.value
```

```
## [1] 0.001373638
```

So small value (0.13%) say that automatic and manual transmittion cars are from different groups.

Correlation analysis

Before I will try to find the best model for mpg let's have a look for correlation mpg to all other variables. Full correlation plot enclosed as Enclose #1 at the end of this document. It's seems like mpg hase a good relationship to numder of cylinders (cyl), engine volume (disp), horsepower (hp) an weight of the car (wt). On my opinion the better way is to check them by anova function below.

Regression analysis & Model selection

So, first of all I will prepare "base" model with weigth (wt) for comparing:

```r
auto_model2 <- lm(data = mtcars, mpg~disp+hp+gear+carb+wt+cyl+
                    I(disp*am)+I(hp*am)+I(gear*am)+I(carb*am)+I(wt*am)+I(cyl*am))
summary(auto_model2)
stepModel <- step(auto_model2, k=log(nrow(mtcars)))
summary(stepModel)
```

As a result 89.25% of multiple R-squared and 87.18% as adjusted R-Squared are not bad result. Gear, carb and wt are really significant in the model (p-value less than 1%). Plus two more variables for manual transmission - disp and cyl are significant. Let's find better model.

```
auto_model2 <- lm(data = mtcars, mpg~disp+hp+gear+carb+wt+cyl+
  I(disp*am)+I(hp*am)+I(gear*am)+I(carb*am)+I(wt*am)+I(cyl*am)+
  I(disp^2)+I(hp^2)+I(gear^2)+I(carb^2)+I(wt^2)+I(cyl^2)+
  I(disp^2*am)+I(hp^2*am)+I(gear^2*am)+I(carb^2*am)+I(wt^2*am)+I(cyl^2*am))
summary(auto_model2)
stepModel <- step(auto_model2, k=log(nrow(mtcars)))
summary(stepModel)
```

Result gave us 97% of multiple R-squared. But only some of variables are significant. Let's clean the list and try to find good model. I will exlude some insignificant variables from the model.

```
mtcars$log_mpg <- log(mtcars$mpg)
auto_model2 <- lm(data = mtcars, log_mpg~wt+hp+disp+cyl+carb+
  I(wt*am)+I(disp*am)+I(cyl*am)+I(carb*am)+
  I(wt^2)+I(hp^2)+I(disp^2)+I(cyl^2)+I(carb^2)+
  I(wt^2*am)+I(hp^2*am)+I(disp^2*am)+I(cyl^2*am)+I(carb^2*am))
summary(auto_model2)
base_model <- step(auto_model2, k=log(nrow(mtcars)))
summary(base_model)
confint(base_model)
```

I will stop the investigation and keep the next result as final: Multiple R-squared: 0.9556, Adjusted R-squared: 0.9139 F-statistic: 22.94 on 15 and 16 DF, p-value: 5.188e-08

Multiple R-squared means that we can explain about 95.56% of the variance of the MPG value. Have a look for our coefficients:

```
summary(base_model)
```

```
##
## Call:
## lm(formula = log_mpg ~ wt + hp + disp + cyl + carb + I(wt * am) +
##     I(cyl * am) + I(carb * am) + I(wt^2) + I(hp^2) + I(disp^2) +
##     I(cyl^2) + I(carb^2) + I(wt^2 * am) + I(carb^2 * am), data = mtcars)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.11861 -0.04152  0.00007  0.04280  0.11959
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -1.811e+00  1.325e+00  -1.366 0.190731
## wt             -1.458e+00  5.435e-01  -2.682 0.016364 *
## hp             -6.267e-03  2.329e-03  -2.691 0.016055 *
## disp           -7.781e-03  3.375e-03  -2.305 0.034880 *
## cyl             2.642e+00  6.479e-01   4.078 0.000877 ***
## carb            1.839e+00  4.357e-01   4.222 0.000648 ***
## I(wt * am)      1.379e+00  3.268e-01   4.219 0.000651 ***
## I(cyl * am)    -2.120e-01  8.804e-02  -2.408 0.028468 *
```

```
## I(carb * am)   -1.761e+00  4.304e-01  -4.091 0.000852 ***
## I(wt^2)         1.593e-01  6.373e-02   2.500 0.023654 *
## I(hp^2)         4.203e-05  1.067e-05   3.940 0.001171 **
## I(disp^2)       1.189e-05  4.987e-06   2.384 0.029867 *
## I(cyl^2)       -2.233e-01  5.392e-02  -4.142 0.000766 ***
## I(carb^2)      -3.900e-01  8.881e-02  -4.392 0.000455 ***
## I(wt^2 * am)   -1.719e-01  7.236e-02  -2.376 0.030320 *
## I(carb^2 * am)  3.561e-01  8.146e-02   4.372 0.000474 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08737 on 16 degrees of freedom
## Multiple R-squared:  0.9556, Adjusted R-squared:  0.9139
## F-statistic: 22.94 on 15 and 16 DF,  p-value: 5.188e-08
```

This result shows that cars with manual transmittion add 1.37*wt+0.356*carb^2 more MPG and subtracts -0.212*cyl-1.76*carb-1.72*wt^2 of MPG in average than cars with automatic tranmission.

```r
# manual tranmission car example
manual <- matrix(c(cyl=4, disp=200, hp=150, wt=3.0, am=1, gear=4, carb=4), nrow = 1, ncol = 7)
manual <- as.data.frame(manual)
names(manual) <- c("cyl", "disp", "hp", "wt", "am", "gear", "carb")
# automatic tranmission car example
auto <- manual # copy all parameters from manual transmission
auto$am <- 0 # make a cer with automatic transmission
# compare our cars
exp(predict(base_model, newdata = manual)) - exp(predict(base_model, newdata = auto))
```

```
##       1
## 4.78685
```

According to our model car with manual transmission, 4 cylinders, 200 cu.in. engine volume, 150 Gross horsepower, 3.0 tn of weight, 4 forward gears and 4 carburetors will have 4.79 miles per gallon more than an automatic transmission car with the same parametrs. Comparing the MPG of two cars with different transmissions from the number of horsepower is given in Enclose #4.

Residual analysis

According to Enclose #2 I could say that: 1) Residuals and Fitted does not shows dependence on each other 2) Scale-Location plot confirms the constant variance assumption 3) Normal Q-Q plot are strong and and shoes that residuals are normally distributed 4) Residuals vs. Leverage shows that no outliers are present, as all values fall well within the 0.5 bands.

```r
sum((abs(dfbetas(base_model)))>1) # >1 - due to small value of n (=32)
```

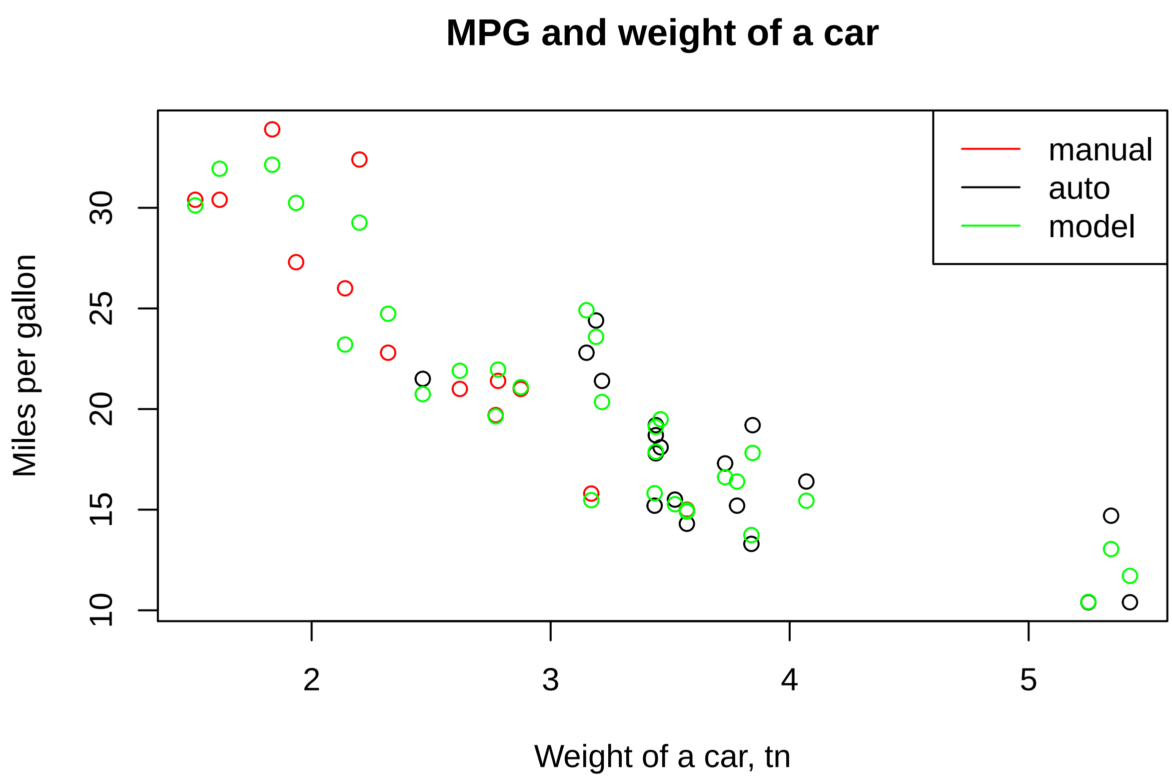
```
## [1] 4
```

The dfbetas value is not so huge. So the analysis meet our assumptions. No residues of heteroscedasticity in our model (Enclose #3).

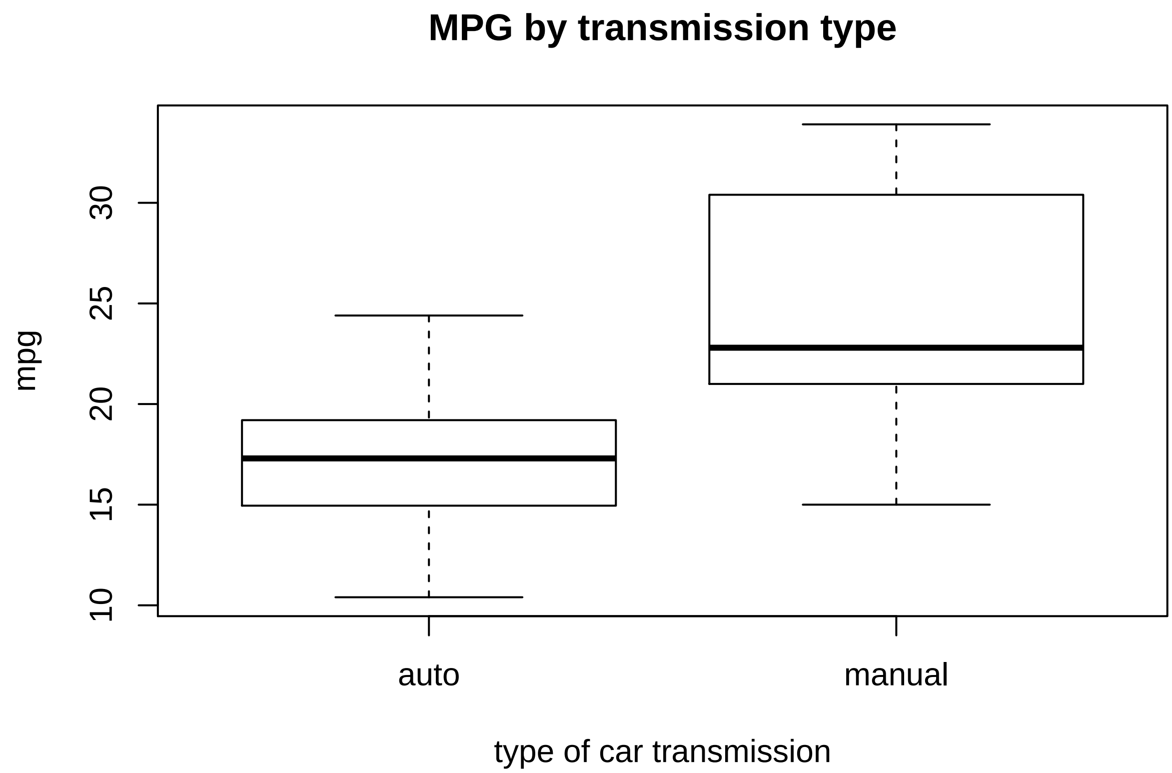
ENCLOSE 1 A

Income car data

```
mtcars$mpg_predict <- exp(predict(base_model, newdata = mtcars))
plot(y=mtcars$mpg, x=mtcars$wt, type='p', ylab ="Miles per gallon",
     xlab="Weight of a car, tn", main = "MPG and weight of a car", col=mtcars$ttype)
points(y=mtcars$mpg_predict, x=mtcars$wt, type='p', col='green')
legend('topright', c("manual","auto","model"), lty=c(1,1,1), col=c("red","black","green"))
```



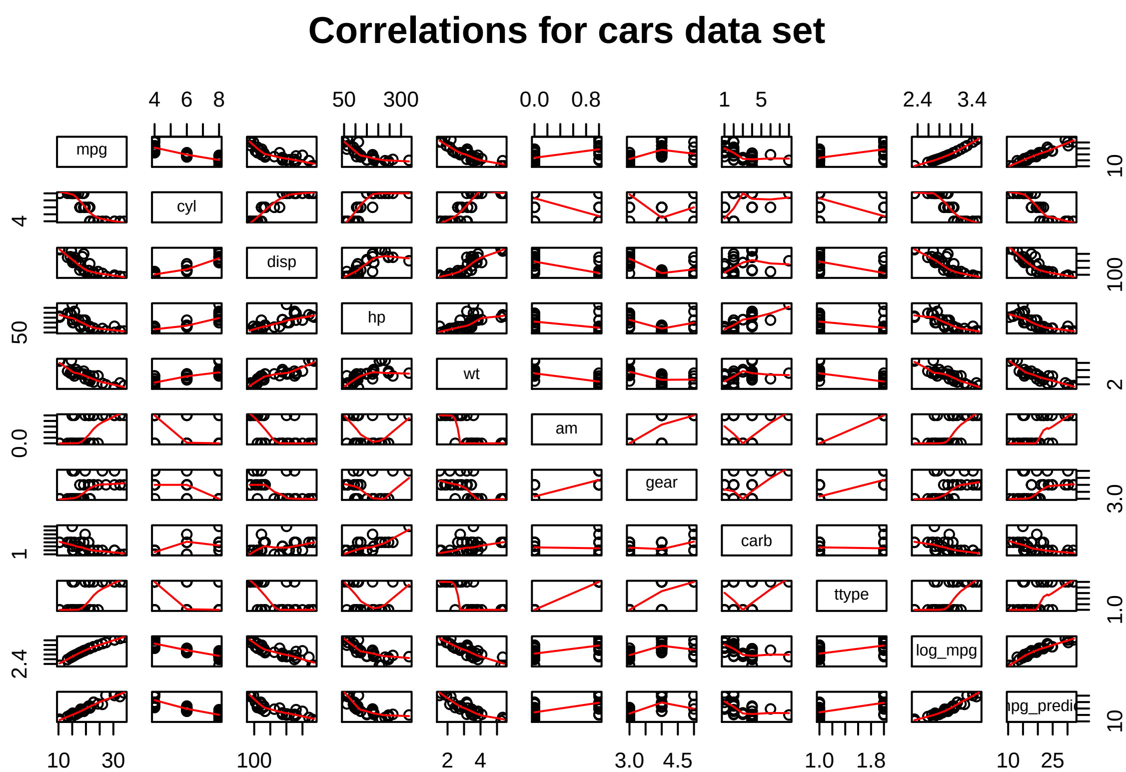

```
boxplot(data = mtcars,mpg ~ ttype, xlab="type of car transmission", ylab="mpg",
        main="MPG by transmission type")
```



ENCLOSE 1 B

Correlation analysis plot

```
pairs(mtcars, panel=panel.smooth, main="Correlations for cars data set")
```

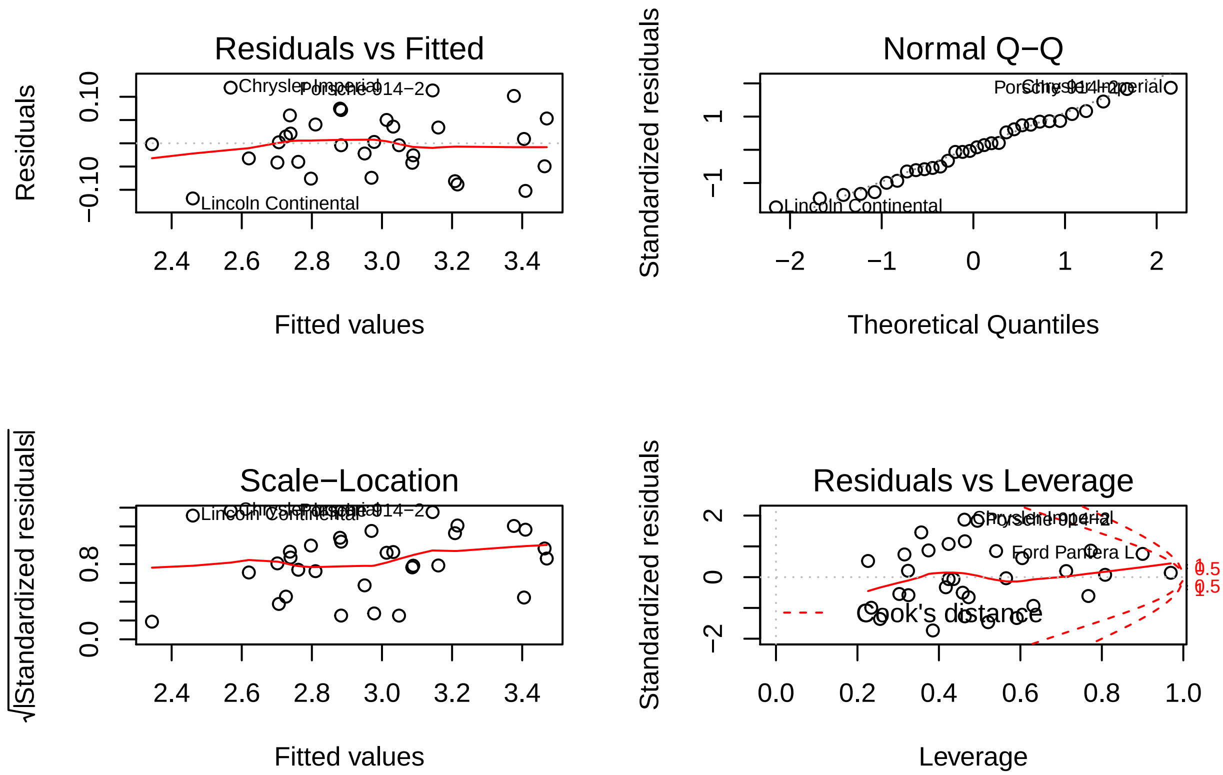


ENCLOSE 2

Regression analysis plot

```r
par(mfrow = c(2, 2))
plot(base_model)
```

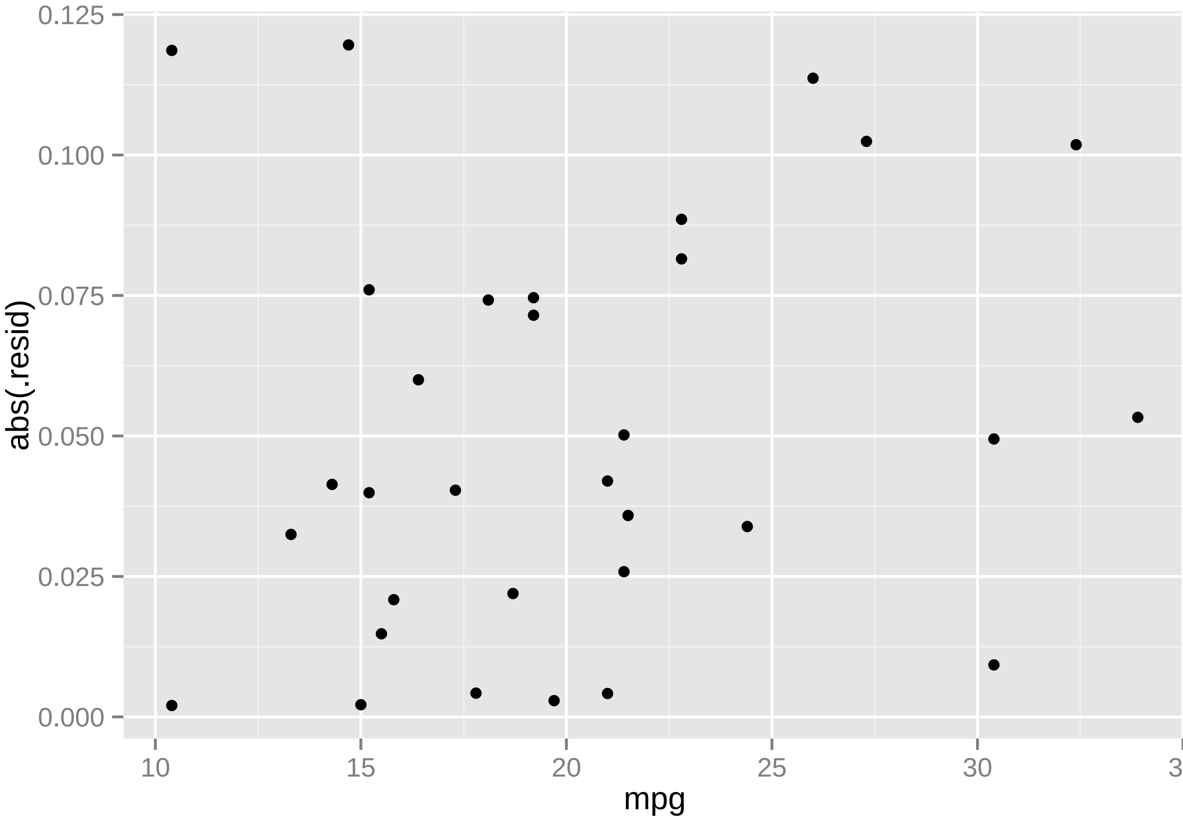
```
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced

## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
```



ENCLOSE 3

Residual analisys plot

```r
mtcars <- augment(base_model, mtcars) # include residual into dataset
qplot(data=mtcars, mpg, abs(.resid)) # have a look for residuals
```



ENCLOSE 4

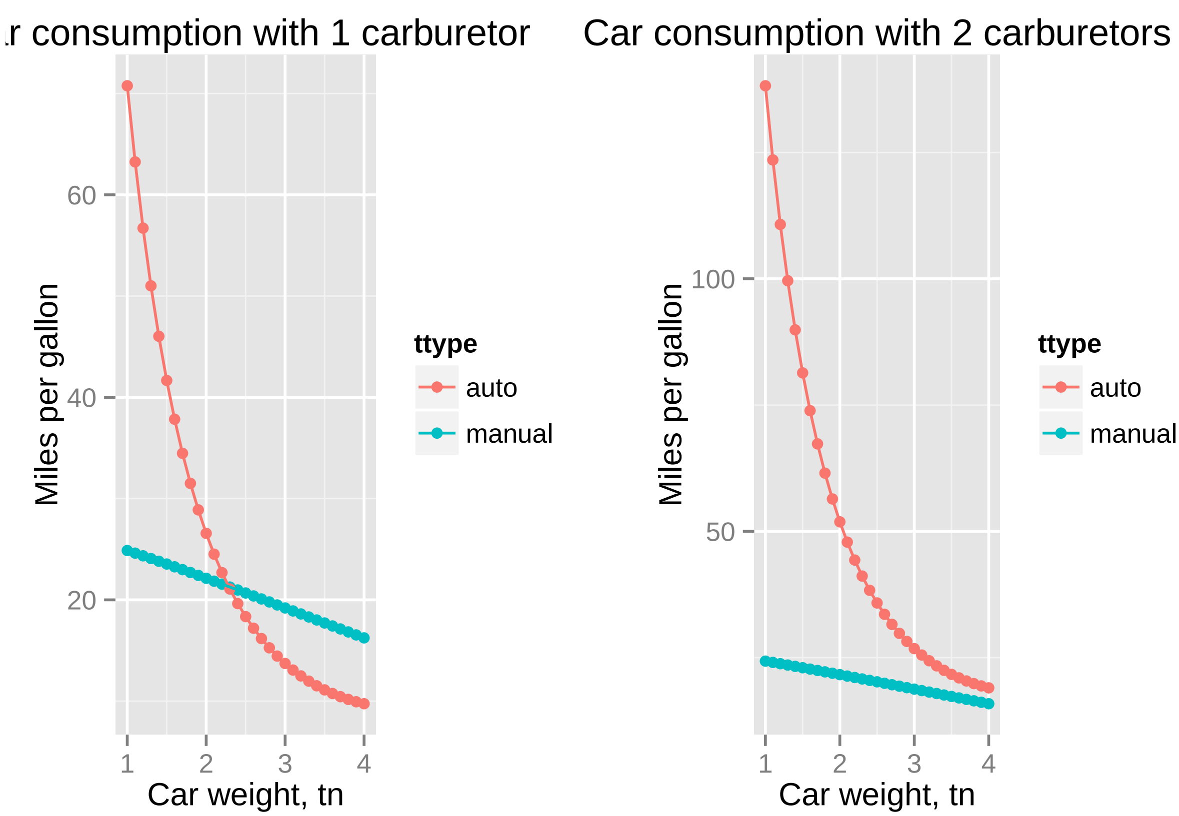
Manual transmission versus automatic transmission by a weigth of car

```
# manual transmission car example
manual <- matrix(nrow = 31, ncol = 7)
manual <- as.data.frame(manual)
names(manual) <- c("cyl", "disp", "hp", "wt", "am", "gear", "carb")
manual$hp <- 150
manual$cyl <- 4
manual$disp <- 200
manual$wt <- seq(1.0, 4.0, by=0.1)
manual$am <- 1
manual$gear <- 4
manual$carb <- 1
# prepare from 1 to 8 range of carburetors
B <- manual # make a copy of data set
for (i in 2:8) {
  manual$carb <- i
  B <- rbind(B, manual)
}
manual <- B
manual$mpg <- exp(predict(base_model, newdata = manual))
```

```r
# automatic tranmission car example
auto <- B
auto$am <- 0
auto$mpg <- exp(predict(base_model, newdata = auto))

# comparing dataset
A <- rbind(manual, auto)
A$carb <- as.factor(A$carb)
A$am <- as.factor(A$am)
A$ttype <- "auto"
A$ttype[which(A$am==1)] <- "manual"
```

Comparing of two cars witn 4 cyl, 150 horsepower, 200 cu.in. engine vol, 150 horsepower, 3.0 tn, 4 forward gears and 4 carb and different weight:

```r
B <- A[which(A$carb==1),]
p1 <- qplot(y=mpg,  x=wt,  data =   B, color  = ttype,  geom    =   c("point", "line"), xlab = "Car wei
B <- A[which(A$carb==2),]
p2 <- qplot(y=mpg,  x=wt,  data =   B, color  = ttype,  geom    =   c("point", "line"), xlab = "Car wei
grid.arrange(p1, p2, ncol = 2)
```




```r
B <- A[which(A$carb==4),]
p1 <- qplot(y=mpg,  x=wt,  data =   B, color  = ttype,  geom    =   c("point", "line"), xlab = "Car wei
B <- A[which(A$carb==6),]
```

```
p2 <- qplot(y=mpg,  x=wt,  data =   B, color  = ttype,  geom    =   c("point", "line"), xlab = "Car wei
grid.arrange(p1, p2, ncol = 2)
```