# Exploratory data analysis of mtcars data set using regression models

*Mykyta Zharov*

*11/18/2019*

## Executive summary

In this project we will look at a data set mtcars, which was extracted from the 1974 Motor Trend US Magazine and comprises fuel consumption and 10 aspects of automobile design and performance for 32 autos. We want to understand the relationship between a set of variables and miles per gallon (MPG) (outcome). We are particulary interested in answering two questions:

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

To answer the above questions we are going to use exploratory analysis and regression models. Using Akaike Information Criteria (AIC) model selection procedure, the result is that variables weight and 1/4 mile time have significant impact in quantifying the difference of mpg between cars with automatic and manual transmissions. Obtained result is that on average, cars with manual transmission have 2.94 mpg more than cars with automatic transmission.

## Exploratory data analysis

Let us start by loading the data and printing out a few first rows.

```
# load data
data(mtcars)
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
```

Let us explore the structure of the data set.

```
dim(mtcars)
```

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

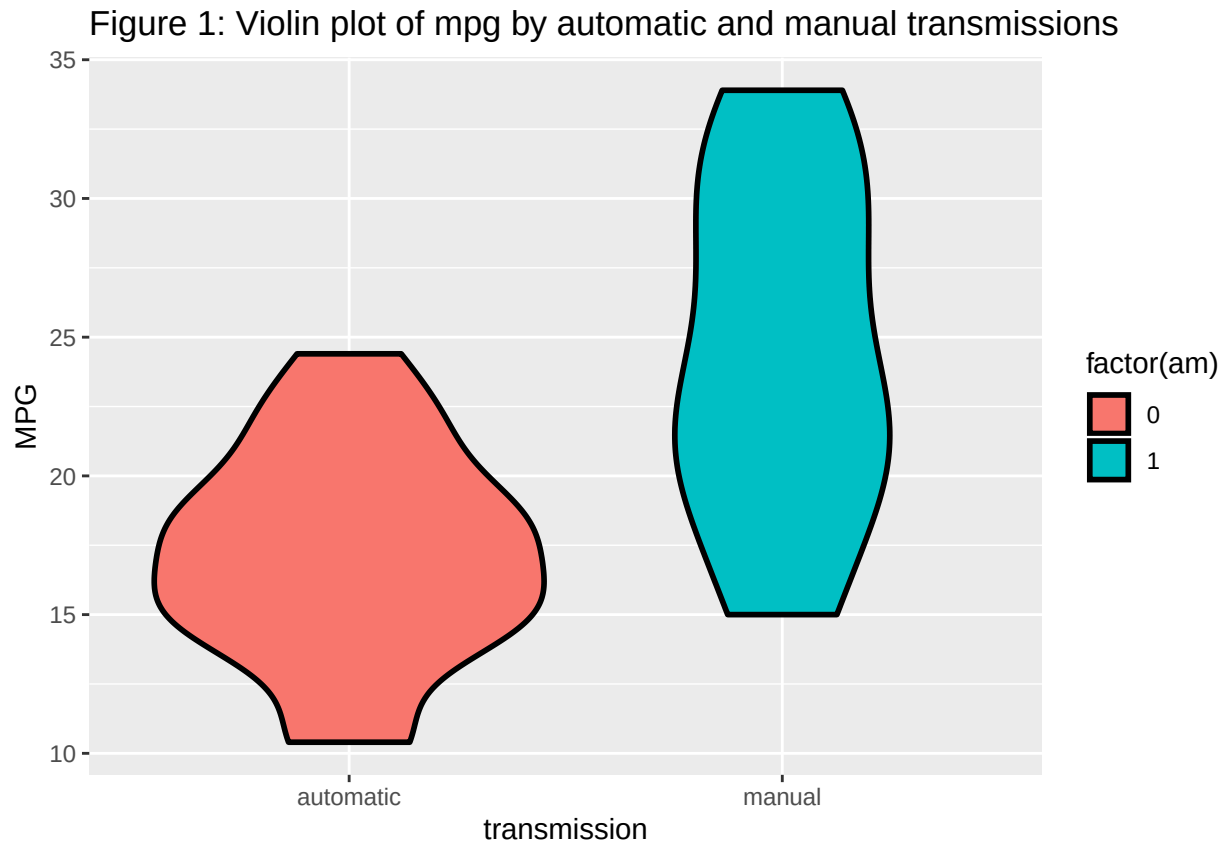
In total we have 32 observations. The variables have the following description:

- mpg - Miles/(US) gallon
- cyl - Number of cylinders
- disp - Displacement (cu.in.)
- hp - Gross horsepower
- drat - Rear axle ratio
- wt - Weight (1000 lbs)
- qsec - 1/4 mile time
- vs - Engine (0 = V-shaped, 1 = straight)

- am - Transmission (0 = automatic, 1 = manual)
- gear - Number of forward gears
- carb - Number of carburetors

Let us start by looking at a violin plot of mpg by automatic and manual transmissions:

```
library(stats);
library(ggplot2);
ggplot(mtcars, aes(y=mpg, x=factor(am,
                labels = c("automatic", "manual")), fill=factor(am)))+
    geom_violin(colour="black", size=1)+
    xlab("transmission") + ylab("MPG") +
    ggtitle("Figure 1: Violin plot of mpg by automatic and manual transmissions")
```




From Figure 1 we see that cars with automatic transmission have a lower mpg value than cars with manual transmission. To quantify this observation we would like to build a multivariate regression model, wich takes into account other predictor variables. We need to understand which variables need to be taken as a predictors to better describe our data.

## Multivariate linear regression

Let us start by looking at a regression model, where we take all variables as predictors.

```
model=lm(mpg~ cyl+disp+hp+drat+wt+qsec+factor(vs)+factor(am)+gear+carb , data=mtcars)
summary(model)$coef
```

```
##                Estimate  Std. Error    t value   Pr(>|t|)
## (Intercept) 12.30337416 18.71788443  0.6573058 0.51812440
```

```
## cyl         -0.11144048  1.04502336 -0.1066392 0.91608738
## disp         0.01333524  0.01785750  0.7467585 0.46348865
## hp          -0.02148212  0.02176858 -0.9868407 0.33495531
## drat         0.78711097  1.63537307  0.4813036 0.63527790
## wt          -3.71530393  1.89441430 -1.9611887 0.06325215
## qsec         0.82104075  0.73084480  1.1234133 0.27394127
## factor(vs)1  0.31776281  2.10450861  0.1509915 0.88142347
## factor(am)1  2.52022689  2.05665055  1.2254035 0.23398971
## gear         0.65541302  1.49325996  0.4389142 0.66520643
## carb        -0.19941925  0.82875250 -0.2406258 0.81217871
```

From the output above p-values show that if we include all the variables, none of them will be a significant predictor of mpg with 95% confidence level. Let us apply stepwise selection procedure by AIC to help us select a subset of predictor variables that best explain the mpg variable.

```
library(MASS);
fit <- lm(mpg ~ cyl+disp+hp+drat+wt+qsec+factor(vs)+factor(am)+gear+carb, data = mtcars)
step <- stepAIC(fit, direction="both", trace=FALSE)
summary(step)
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + factor(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 ***
## factor(am)1   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
```

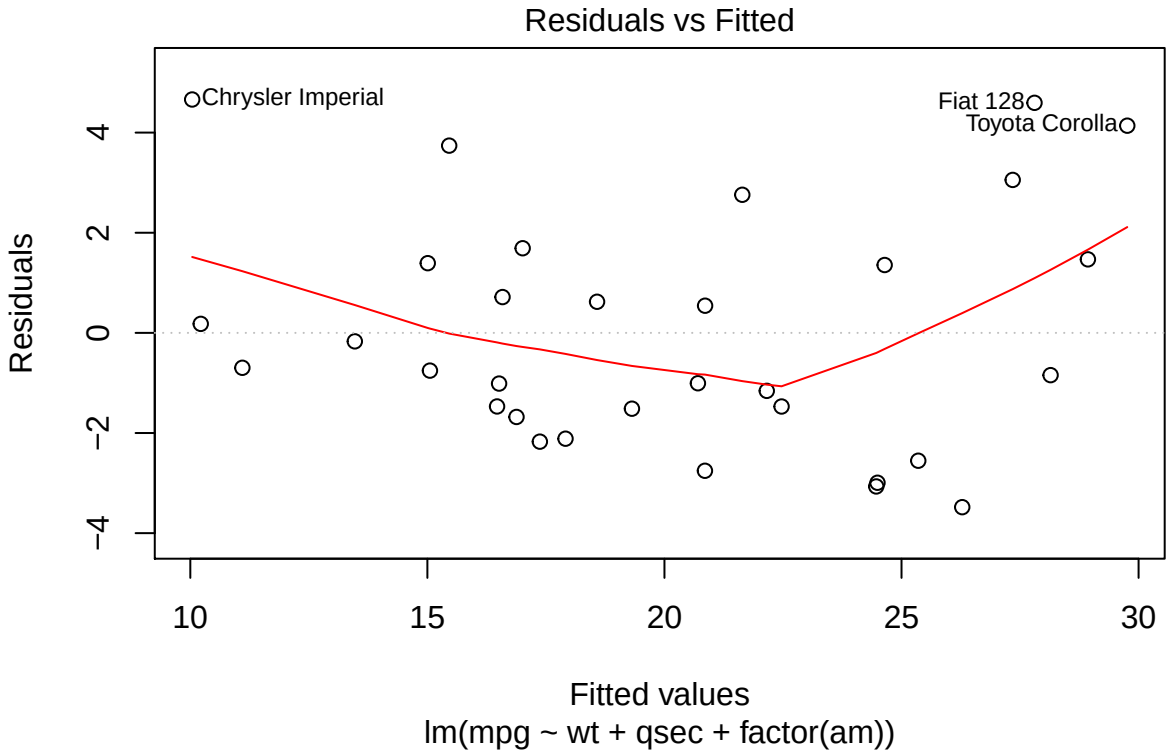
From the output above we see that by stepwise selection procedure by AIC only three variables are left that describe the mpg variable in the best way: wt (Weight), qsec (1/4 mile time), and am (Transmission). The adjusted R^2 value from the output above is almost 84%, which implies that over model describes almost 84% of variation in mpg. This value is satisfactory. The coefficient for am variable, implies that on average, cars with manual transmission have 2.94 mpg more than cars with automatic transmission.

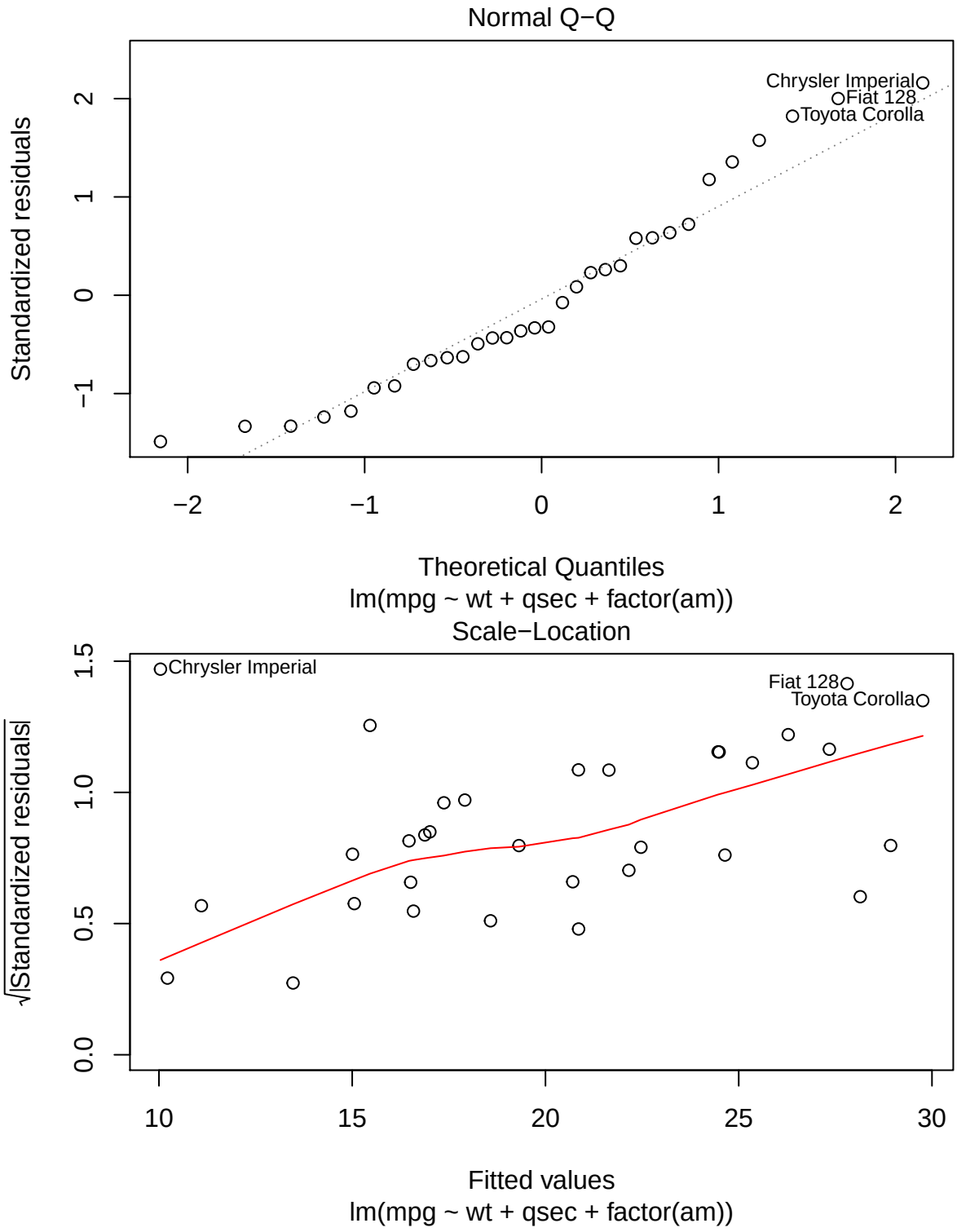
In the appendix section one can find plots for the diagnostics of the model. It is shown that no specific pattern exists in the residuals and the residuals are normally distributed.

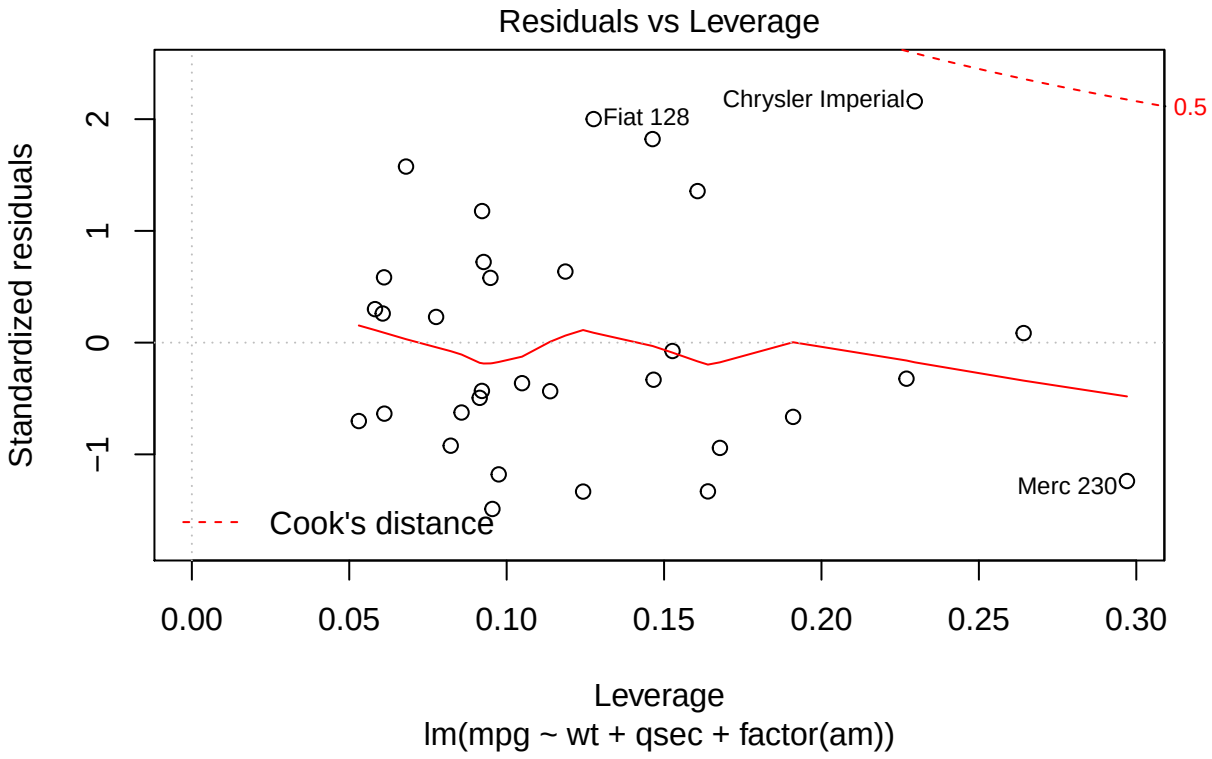
## Appendix

### Diagnostics for obtained model

```
model=lm(mpg~ wt+qsec+factor(am) , data=mtcars)
plot(model)
```




Normal Q−Q
Chrysler Imperial
Fiat 128
Toyota Corolla
Standardized residuals
−2
−1
0
1
2
Theoretical Quantiles
lm(mpg ~ wt + qsec + factor(am))
Scale−Location
Chrysler Imperial
Fiat 128
Toyota Corolla
√|Standardized residuals|
0.0
0.5
1.0
1.5
10
15
20
25
30
Fitted values
lm(mpg ~ wt + qsec + factor(am))


Residuals vs Leverage
Standardized residuals
Leverage
lm(mpg ~ wt + qsec + factor(am))
Fiat 128
Chrysler Imperial
Merc 230
0.5
Cook's distance
2
1
0
−1
0.00
0.05
0.10
0.15
0.20
0.25
0.30