Regression Model Project

Naqeeb Asif

23 February 2018

Synopsis

In this experiment mtcars data is explored. First the relationship between miles per gallon and transmission type is examined. It is found out that the transmission type impacts the miles per gallon but after adding other features the transmission type becomes less significant. The residuals of the new model have almost normal distribution with few outliers.

Loading the data

In this section the data is loaded.

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

Data Processing

In this section the exploratoty analysis is performed on the data.

Converting Features to 'Factor' type

In the dataset that somes features should not be of numeric type so lets convert those features to 'factor'.

```
library(dplyr)

data_mtcars <- mtcars %>% mutate(cyl=factor(cyl), gear=factor(gear),carb=factor(carb),
                                 vs=factor(vs,labels = c("V-shape","Straight Line"),
                                           levels = c(0,1)),
                                 am=factor(am,labels = c("Automatic","Manual"),
                                           levels = c(0,1)))
```

Exploratory Analysis

In this section exploratory analysis is performed on the data.

Reltionship Between Miles per Gallon and Transmssion Type

As it can be seen from the box plots in appendix, the mean of mpg for automatic transmission type is less than that of manual transmission type. Now lets quantify the means.

```
library(dplyr)
mData <- data_mtcars %>% select(mpg,am) %>%group_by(am) %>%summarise(mean=mean(mpg))
```

The mean of mpg for automatic transmission type is 17.15 and the mean of mpg for manual transmission is 24.39. The difference between mean for the two transmission types are 7.24. So we can say that manual transision type gives better miles per gallon.

Result of exploratory analysis

Different relationships can be seen from the plots in appenices 1 and 2. We can see from the boxplots that mean of mpg varies for different values of the features. From the scatter plot it can be seen that some features (‘disp’,‘hp’.‘wt’) have inverse relation with mpg while other features (‘drat’,‘qsec’) have direct relation with mpg. Form the correlation plot in appendix 4 it can be seen that the features which impact ‘mpg’ the most are ‘cyl’,‘disp’,‘hp’ and ‘wt’

Regression Model

Now lets look at regression model between transmission type and miles per gallon

```r
fit1 <- lm(mpg~am,data=data_mtcars)
sum_model <-summary(fit1)
```

From the above results we can conclude the following:

- p-value for the variable is `3e-04` which is less than 0.05 so the model is acceptable.
- The coefficient beta0 is `17.15` which is equal to the mean of mpg for automatic transmission type.
- The coefficient beta1 `7.24` which is equal to the difference between mean of mpg for two transmission types as explained above.
- The coefficients also give the result that mpg for manual transmission type is larger than that automatic transmission type.
- r-squared is only `0.359798943425465` which means only 36% of the variance is covered by the model.

Better Model

From the above results it can be seen that the model is not a good model .Lets add the four features ,which we found in the exploratory analysis, in the model.

```r
better_fit <- update(fit1,mpg~am+cyl+disp+hp+wt)
comp_models<-anova(fit1,better_fit)
```

From the above results it can be seen the model performs better than the previous one as the p-value is `8.63680441696425e-08` very close to 0. Now lets look at the summary of the model.

```r
summary_model <- summary(better_fit)
```

Above model gives r-squared = `0.866` therefore it covers 86.6% variance and it is a better model.

Residual Plot analysis

The residual plot is plotted in apendex 4 . Following are the results which are obtained from the plots:

- No relation is found between residuals and fitted values showing that residuals have constant variance.

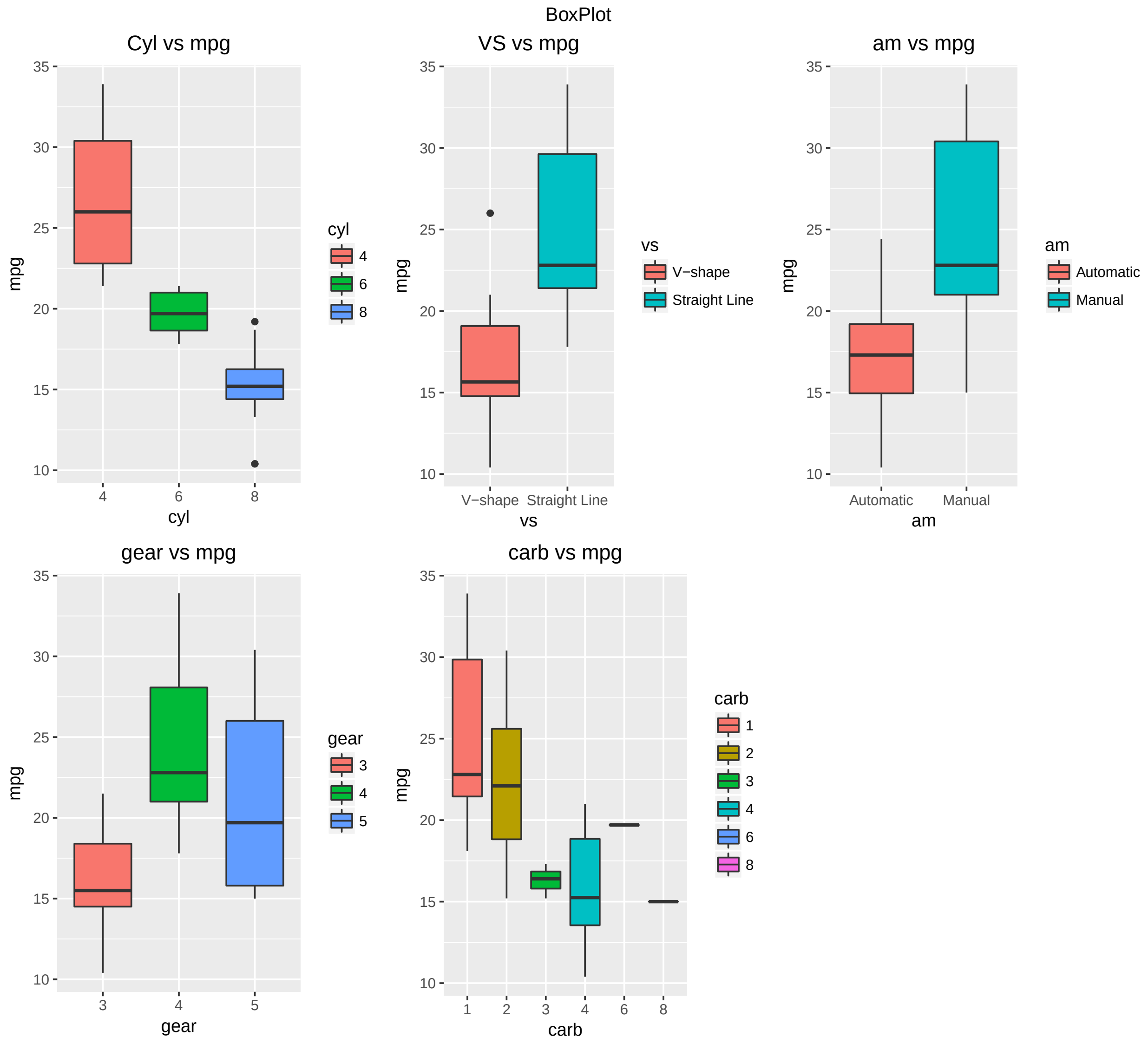
- Normal Q-Q plot indicates the normal distribution of the residuals.
- Leverage plot indicates that there are some outliers in the plot causing the leverage in the plot.

Conclusion

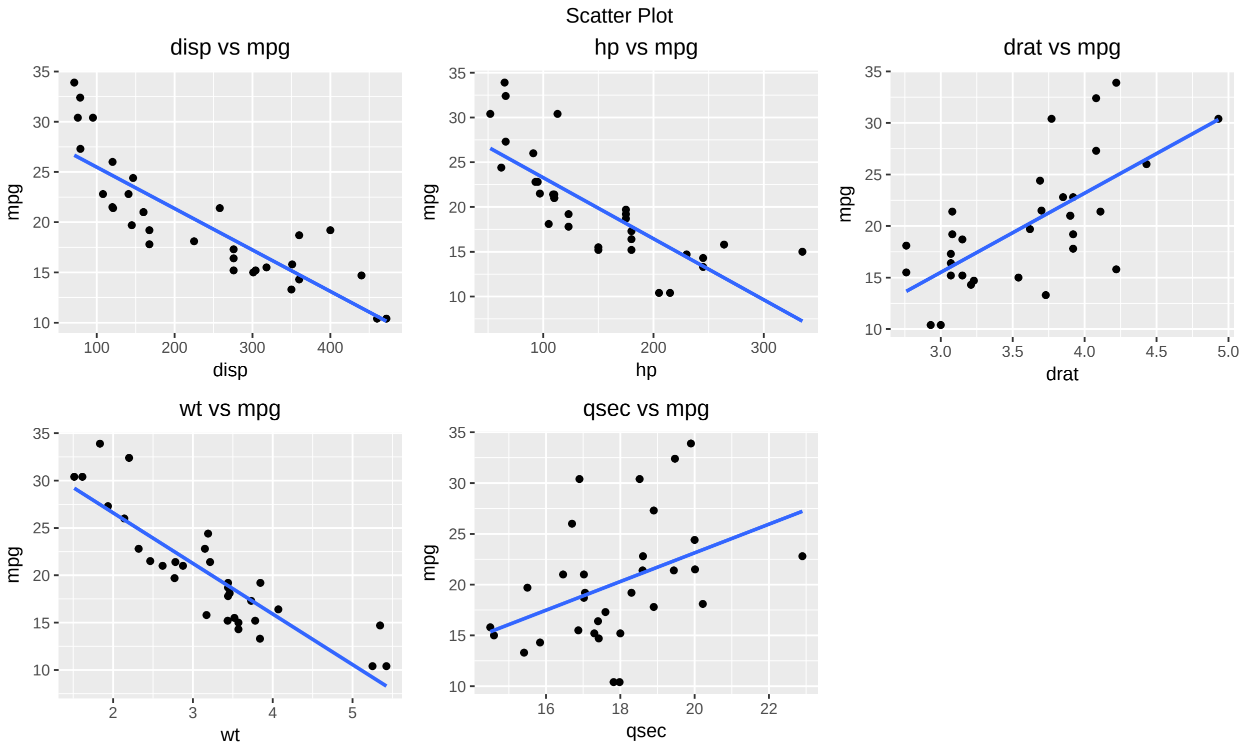
We can conclude that manual transmission type car has more miles per gallon (average of 7) than the automatic transmission type. Moreover when we inculde other features then transmission type has less significant impact on the miles per gallon of the car. The residual obtained after fitting the most correlated features have constant variance and almost normal distribution.

Appendix

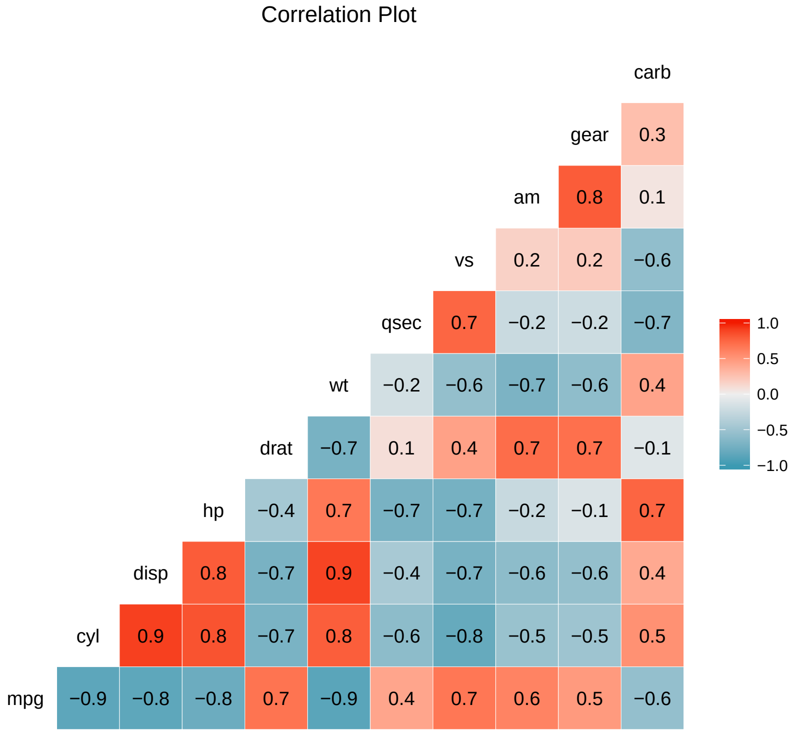
A.1 BoxPlots between mpg and diffrent factor features.



A.2 Scatter Plots between mpg and numerical features

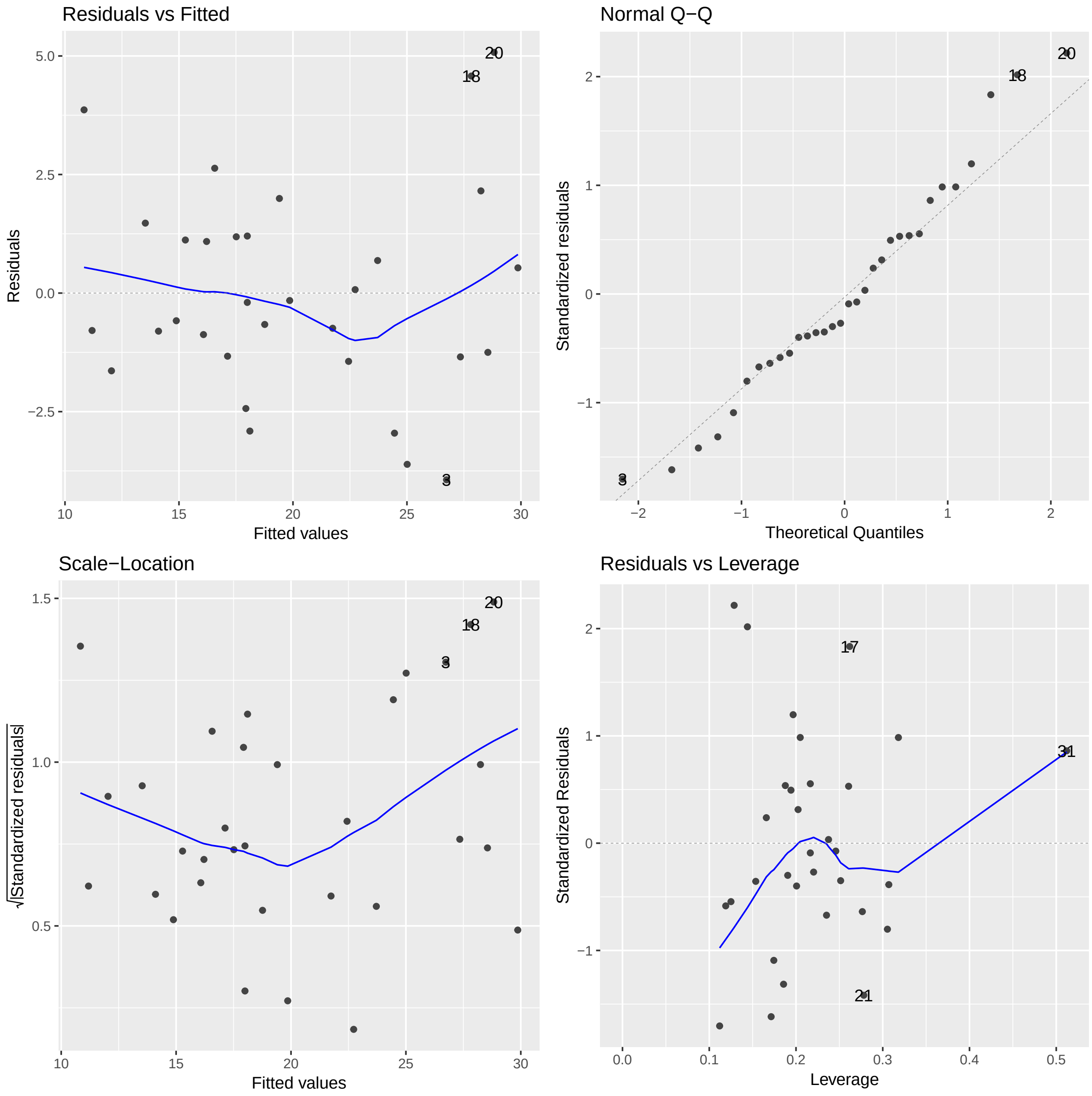


A.3 Correlation Plot between different variables



	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
gear										0.3
am									0.8	0.1
vs								0.2	0.2	−0.6
qsec							0.7	−0.2	−0.2	−0.7
wt						−0.2	−0.6	−0.7	−0.6	0.4
drat					−0.7	0.1	0.4	0.7	0.7	−0.1
hp				−0.4	0.7	−0.7	−0.7	−0.2	−0.1	0.7
disp			0.8	−0.7	0.9	−0.4	−0.7	−0.6	−0.6	0.4
cyl		0.9	0.8	−0.7	0.8	−0.6	−0.8	−0.5	−0.5	0.5
mpg	−0.9	−0.8	−0.8	0.7	−0.9	0.4	0.7	0.6	0.5	−0.6

A.4 Residual Plot of the model


Residuals vs Fitted
Residuals
5.0
2.5
0.0
−2.5
10
15
20
25
30
Fitted values
20
18
3
Normal Q−Q
Standardized residuals
2
1
0
−1
−2
−1
0
1
2
Theoretical Quantiles
20
18
3
Scale−Location
√|Standardized residuals|
1.5
1.0
0.5
10
15
20
25
30
Fitted values
20
18
3
Residuals vs Leverage
Standardized Residuals
2
1
0
−1
0.0
0.1
0.2
0.3
0.4
0.5
Leverage
17
31
21