Identifying Residual Plot and Regression Model

As a data scientist we always have to work on data which may or may not be always in support of the problem we are trying to solve. In regression analysis while working on multiple models we have to focus on many datapoint and their impacts on the predictions. One of such datapoint is Residual plots and how we can read it and correct or improve our regression model by visually understanding the Residual plots.

I am going to user Cars dataset from R, which has following two fields:

- speed numeric Speed (mph)
- dist numeric Stopping distance (ft)

Summary of the data can be seen as below :

```
summary(cars)
```

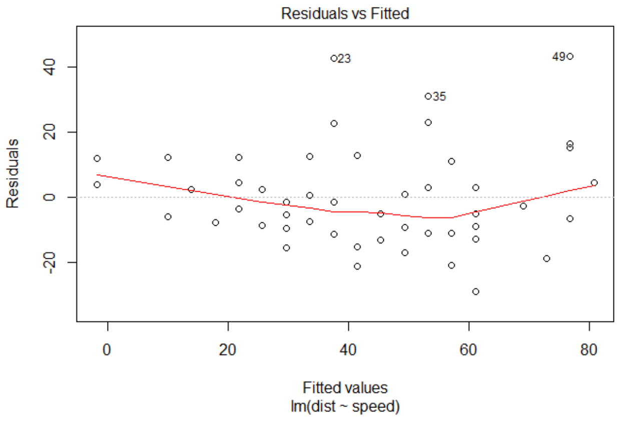
```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00  
```

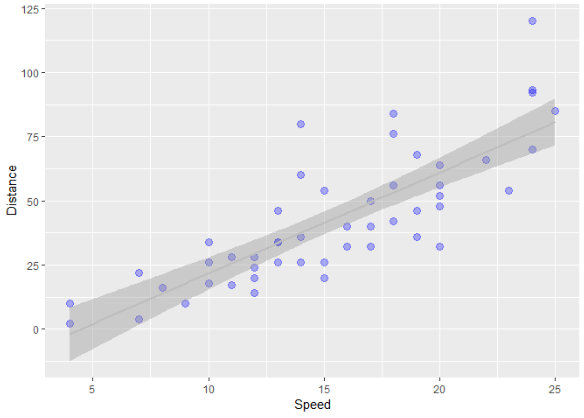
Regression

I am going to use speed of the car in mph to identify the stopping distance covered in ft.

- Predictor: Speed
- response : Stopping Distance
- Model1 : Distance = intercept + slope * speed

Below Residuals vs Fitted plots shows how the residuals are seen close to the regression line . Points below the regression line have positive error and points above the regression line have negative error. Positive errors indicates points lies below the regression line and have predicted value greater than the actual value. Negative Error indicates points lies above the regression line and have predicted value less than the actual value.





Residual plots if it doesn't show any pattern by looking at the data points (as shown in above plot) we can say that this model may be able to explain some of the variation of the data. This doesn't suggest that it's a best model but it suggest that it's not a bad model. Since there is no visible pattern of the data that can be seen can say that model is on the right track.

Moving steps ahead

Now we will generate some data points that are randomly generated and normally distributed.

X <- Generate 100 data points between 3 and -3

y <- x + sin(x)

y1 <- x + sin(x) + rnorm(100,sd=.2)

y2 <- x + rnorm(100,sd=.2)

My objective is to see how residual plots would come when we try to build linear regression model lm(Y~X). Let's dive into it and see the power of residual plots visually.

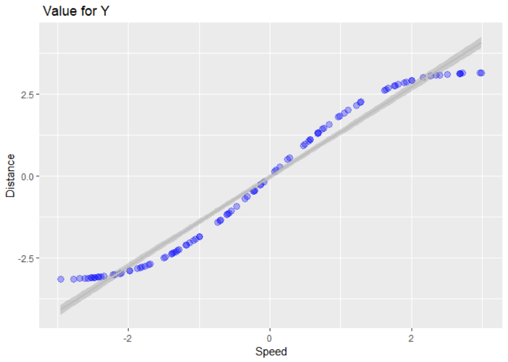
You can see how y,y1,y2 are distributed , by histogram and scatterplot.

```
x <- runif(100,-3,3) # Generate 100 data points between 3 and -3


y <-  x + sin(x)

y1 <- x + sin(x) + rnorm(100,sd=.2)

y2 <-  x + rnorm(100,sd=.2)

dt_chk <- data.frame('x'= x,'y'=y ,'y1'=y1, 'y2'=y2)

head(dt_chk )
```

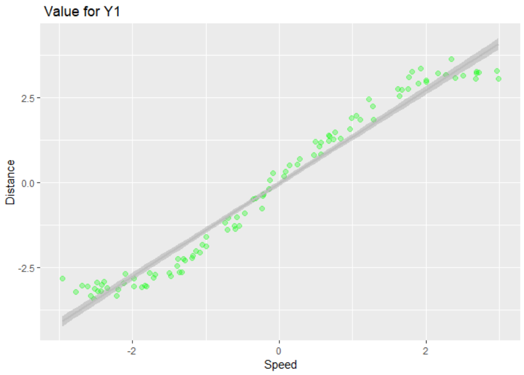
```
##            x          y        y1         y2
## 1  1.9309903  2.8668188  3.352629  2.0231600
## 2  0.0890511  0.1779846  0.320571  0.4316150
## 3 -1.0534957 -1.9226530 -1.818391 -1.2771628
## 4  0.5725210  1.1142737  1.169030  0.6309607
## 5 -1.0802123 -1.9622702 -2.048347 -1.0824503
## 6  2.5016344  3.0987964  3.146961  2.8058144
```

Now if we plot regression line and actual data points , and try to understand how good the model is, most of the time we may feel our model is good since datapoints are close to regression line. In 1st plot , it looks very clearly that our data is making a sine waves along with regression line, whereas in 2nd plot it looks very close that our model might be good choice for this data. Third plot is also showing the same information.

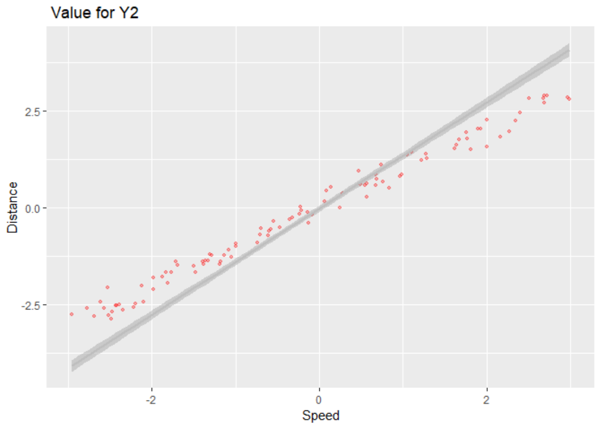
Value for Y



Value For Y1



Value for Y2

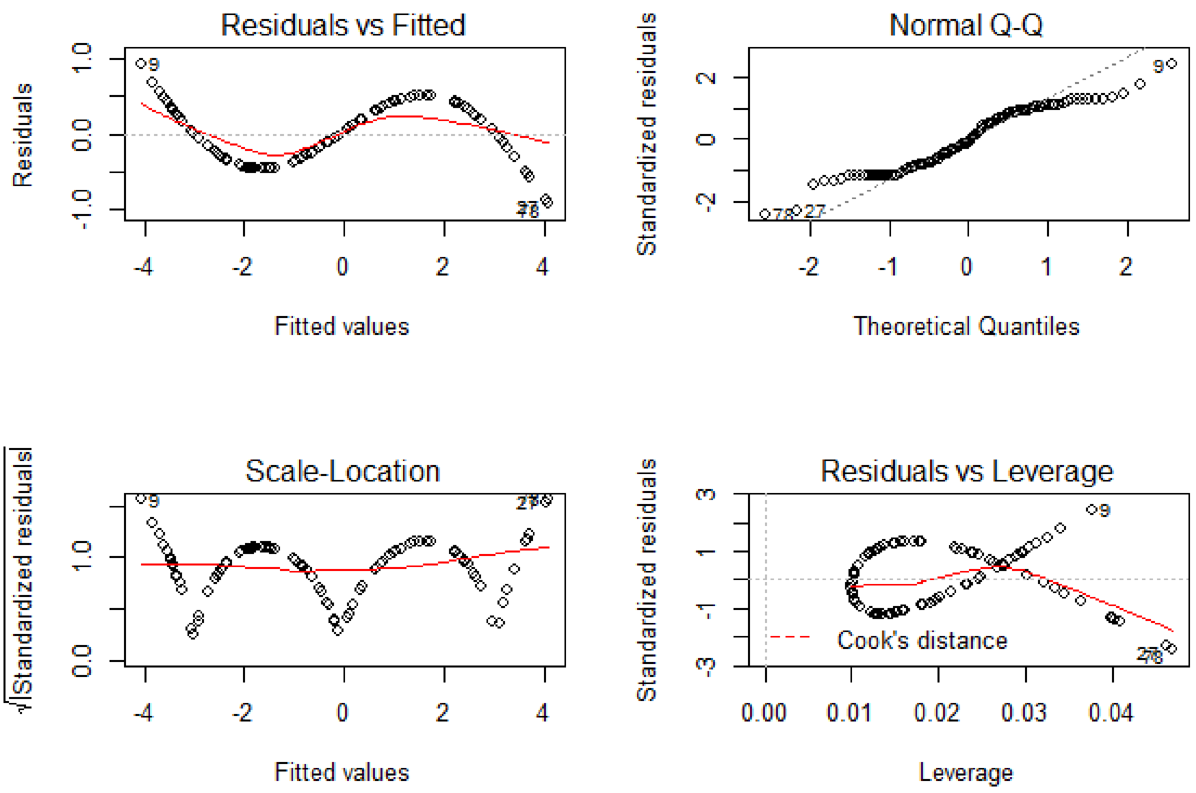


Data points near the regression line.

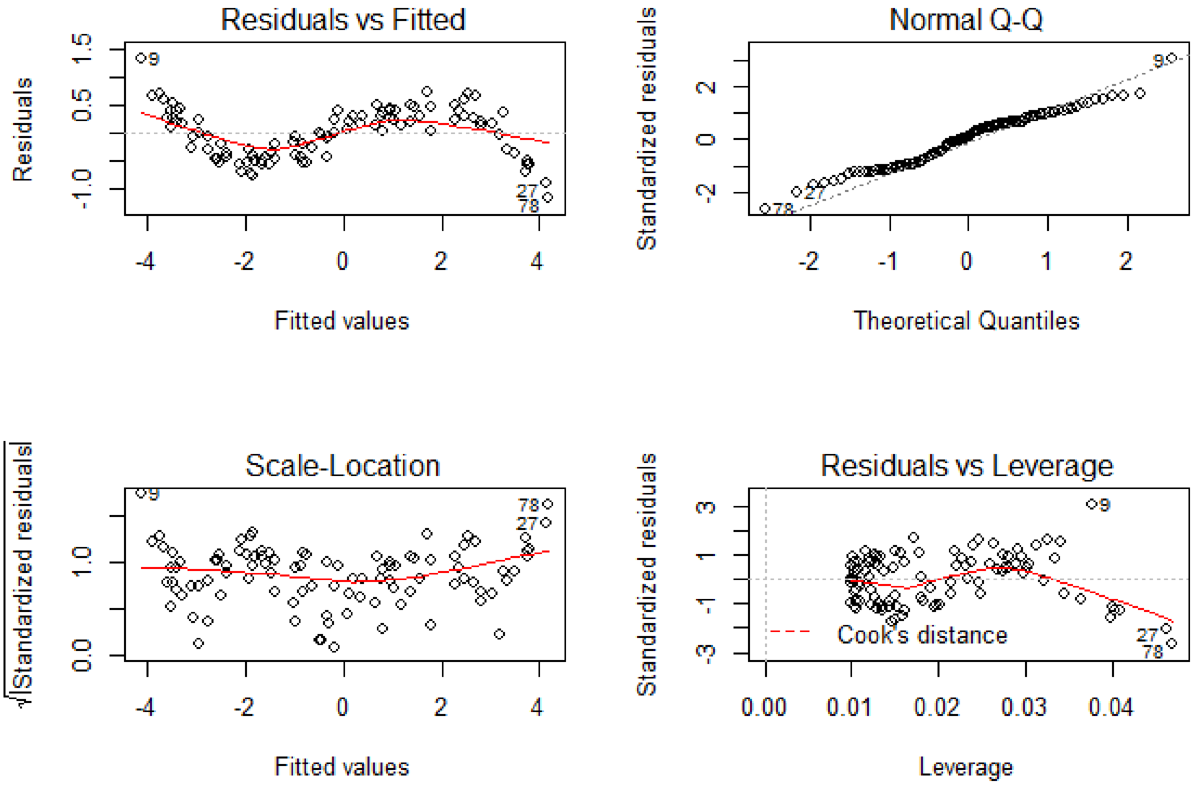
How do we validate it now ?

Let’s plot residual plot and see how does that looks for each y,y1,y2.

```
# lets build the model
model3 <- lm(y~x,dt_chk)
model4 <- lm(y1~x,dt_chk)
model5 <- lm(y2~x,dt_chk)
par(mfrow = c(2, 2))
plot(model3)
```



```
plot(model4)
```


Residuals vs Fitted
Residuals
1.5
0.5
-1.0
-4
-2
0
2
4
Fitted values
9
27
78
Normal Q-Q
Standardized residuals
2
0
-2
-2
-1
0
1
2
Theoretical Quantiles
9
27
78
Scale-Location
√|Standardized residuals|
1.0
0.0
-4
-2
0
2
4
Fitted values
9
78
27
Residuals vs Leverage
Standardized residuals
3
1
-1
-3
0.00
0.01
0.02
0.03
0.04
Leverage
Cook's distance
9
27
78

```
plot(model5)
```

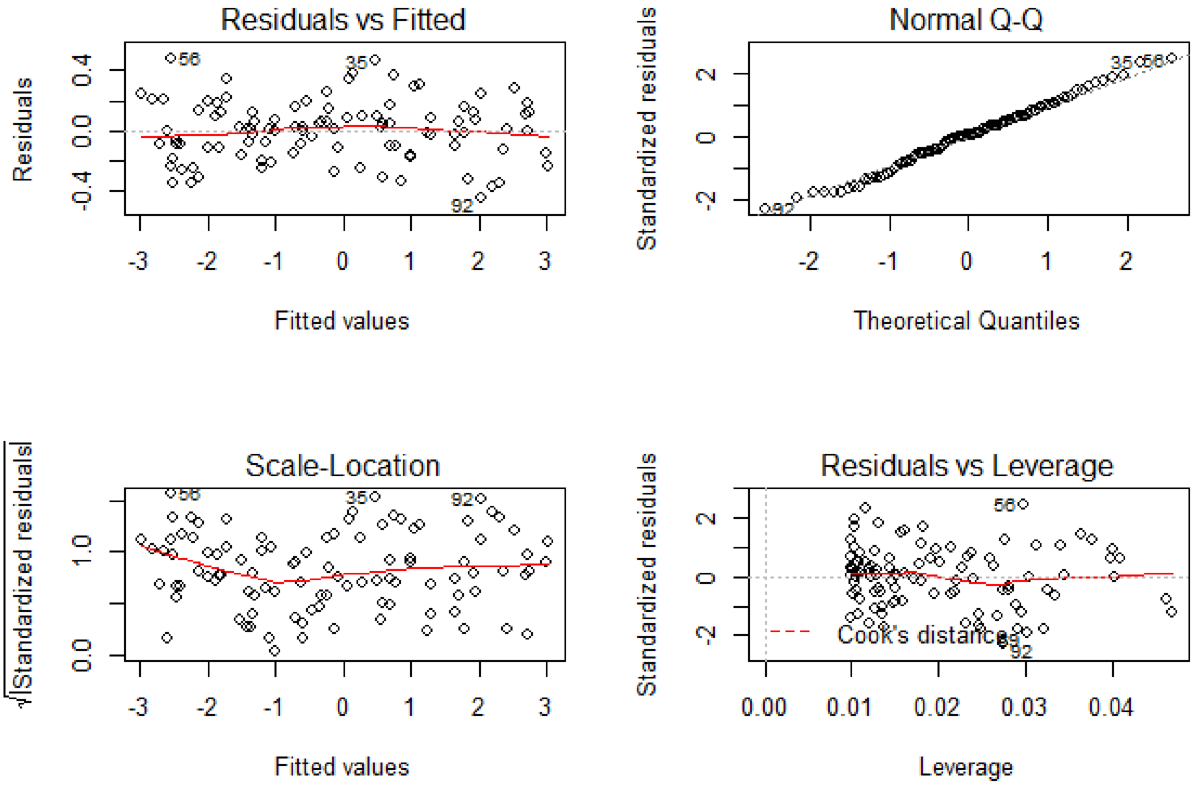


Figure 1: *(Plot for model 13)* In contrast to last plot *(model 15)* here we see that our residuals plots gives very clear pattern which seems to follow a clear sine wave path in the residual plot. Except third plot first two plots seems to follow sine wave and hence it’s clear that we are missing something in the regression model for these equations. We can try explaining these patterns in the model to improve our regression model (Model 133 and Model 14).

This blog suggest how residual plots are important and how they can give us hidden information about the data and help us build the model that better explain the response variable .