# Assignment3_B_1_1

*Sangamesh*

*25 September 2018*

Reading data and making data into training and test data.

```
## [1] "head of training data"
```

```
##           Y1         X1        X2
## 1 -1.0565192  -6.236444 0.9615355
## 2 -0.5754127  -3.873848 0.5050130
## 3  5.0910630   5.640287 0.7175317
## 4  2.9475637   1.191125 0.3074231
## 5  2.9519538 -10.849769 0.5960600
## 6  3.1685278   2.603705 0.3109550
```

```
## [1] "head of testing data"
```

```
##            Y1         X1         X2
## 10 -0.7316911 -2.2906586 0.43611757
## 20  1.1998000  7.7123714 0.47222562
## 30  0.6124209 -1.4169026 0.01161898
## 40 -0.1139879  0.6901132 0.48192669
## 50  3.5655124  1.3302962 0.49526489
## 60  4.3900710  3.1920603 0.98896327
```

Fitting data for linear regression model

```
##
## Call:
## lm(formula = Y1 ~ X1 + X2, data = trainData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1226 -1.3189 -0.0519  1.1825  3.3815
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.77301    0.33680   8.233 1.66e-12 ***
## X1           0.19480    0.03465   5.623 2.24e-07 ***
## X2          -0.15753    0.61320  -0.257    0.798
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.711 on 87 degrees of freedom
## Multiple R-squared:  0.2666, Adjusted R-squared:  0.2498
## F-statistic: 15.81 on 2 and 87 DF,  p-value: 1.387e-06
```

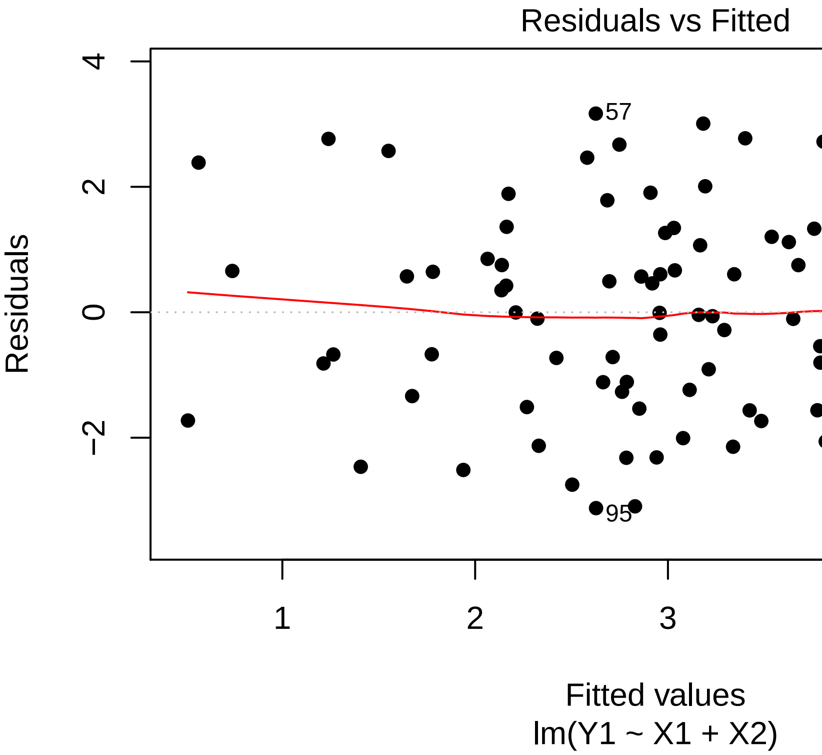
Several metrics useful for regression diagnostics : model.diag.metrics

```
## # A tibble: 6 x 11
##   .rownames     Y1     X1    X2 .fitted .se.fit   .resid   .hat .sigma
##   <chr>      <dbl>  <dbl> <dbl>   <dbl>   <dbl>    <dbl>  <dbl>  <dbl>
## 1 1         -1.06   -6.24 0.962    1.41   0.460 -2.46    0.0724   1.70
## 2 2         -0.575  -3.87 0.505    1.94   0.265 -2.51    0.0240   1.70
## 3 3          5.09    5.64 0.718   3.76    0.272  1.33    0.0252   1.71
## 4 4          2.95    1.19 0.307   2.96    0.203 -0.00905 0.0141   1.72
## 5 5          2.95  -10.8  0.596   0.566   0.480  2.39    0.0788   1.70
## 6 6          3.17    2.60 0.311   3.23    0.206 -0.0627  0.0145   1.72
## # ... with 2 more variables: .cooksd <dbl>, .std.resid <dbl>
```

Meta Data for model.diag.metrics Among the table columns, there are:

Y1: original values X1, X2: the observed values .fitted: the fitted values .resid: the residual errors

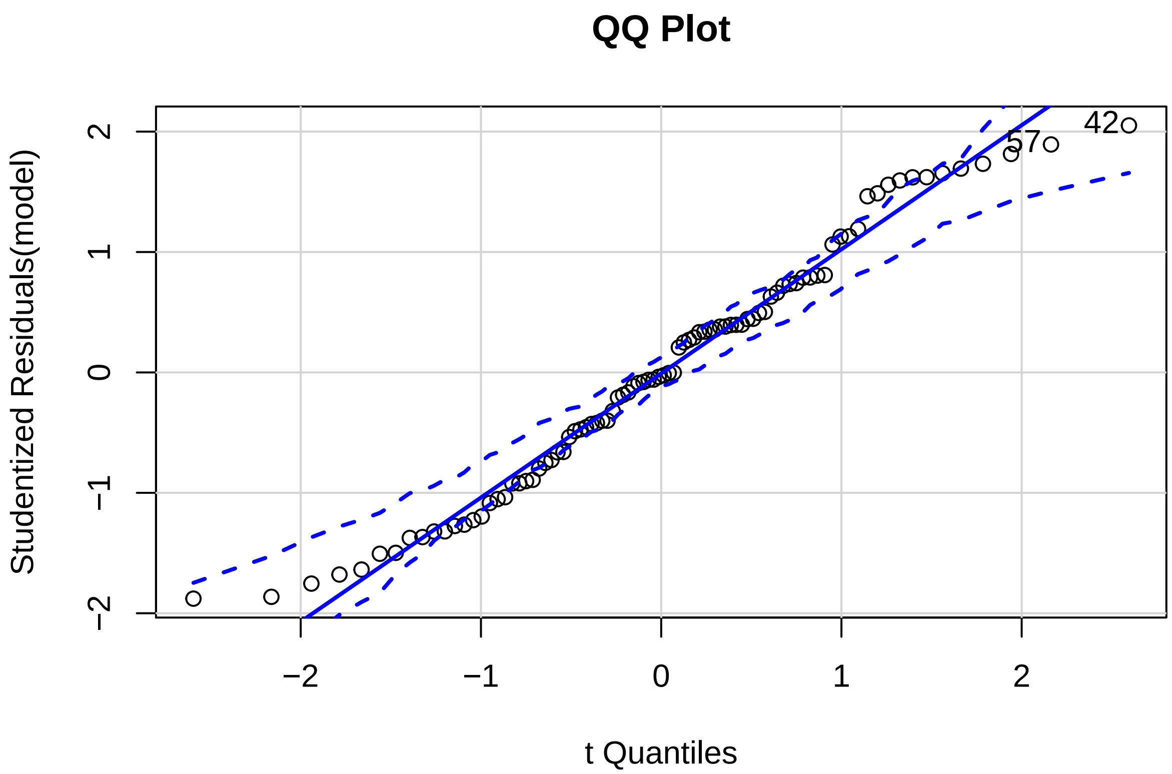
**Let's see correlation between the features:**



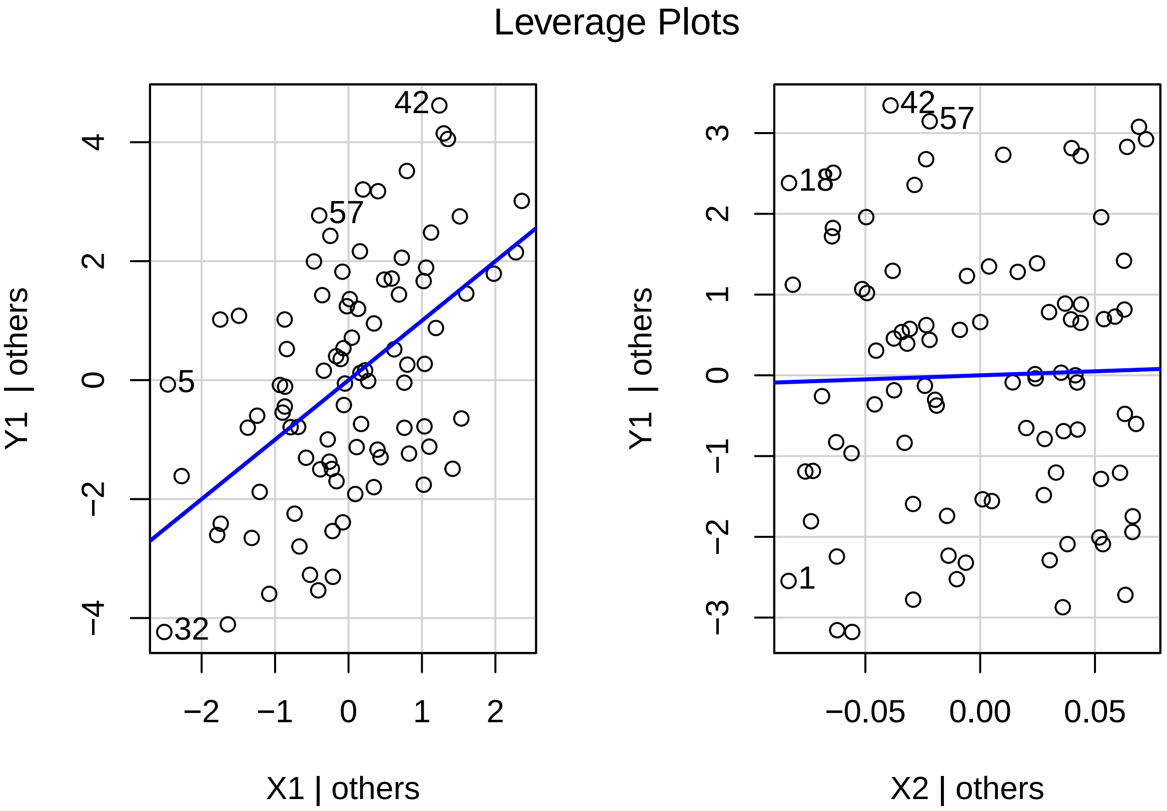
We can see the plot in residual and fitted plot here now:

Note how the residuals plot of this last model shows some important points still lying far away from the middle area of the graph. Since the behaviour is random in nature we were successful in this test.

```
## Loading required package: carData
```



```
## 42 57
## 38 52
```



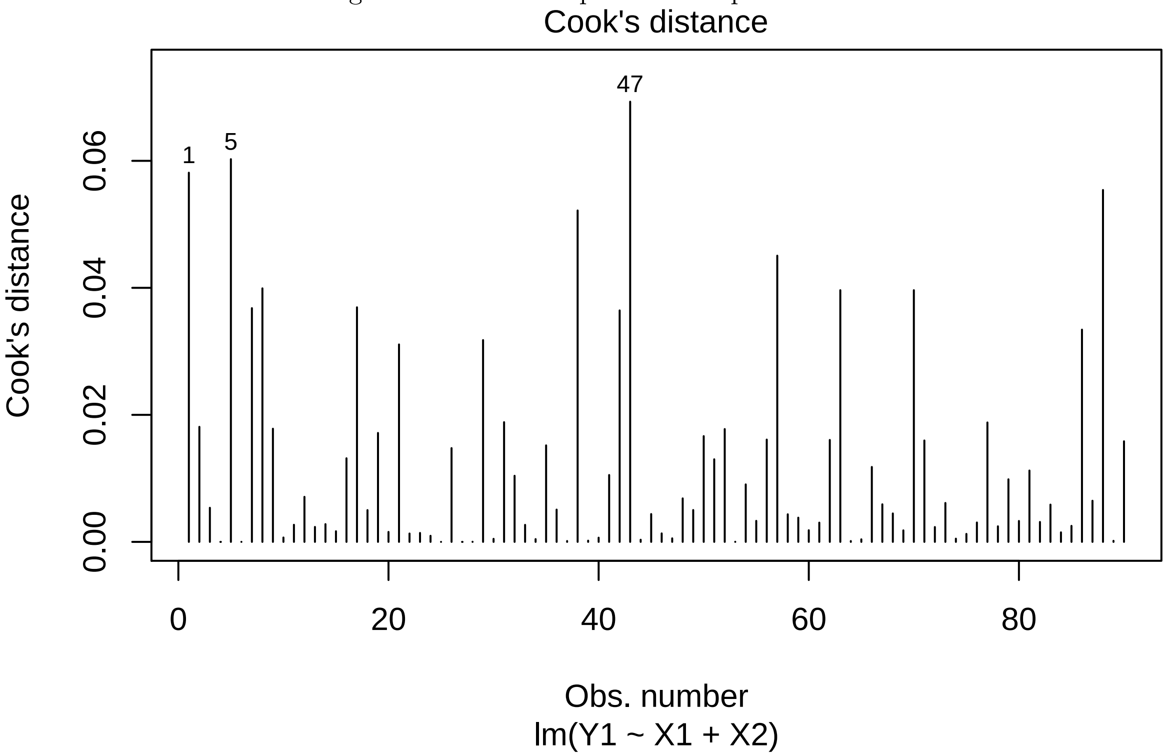
Check outliers:

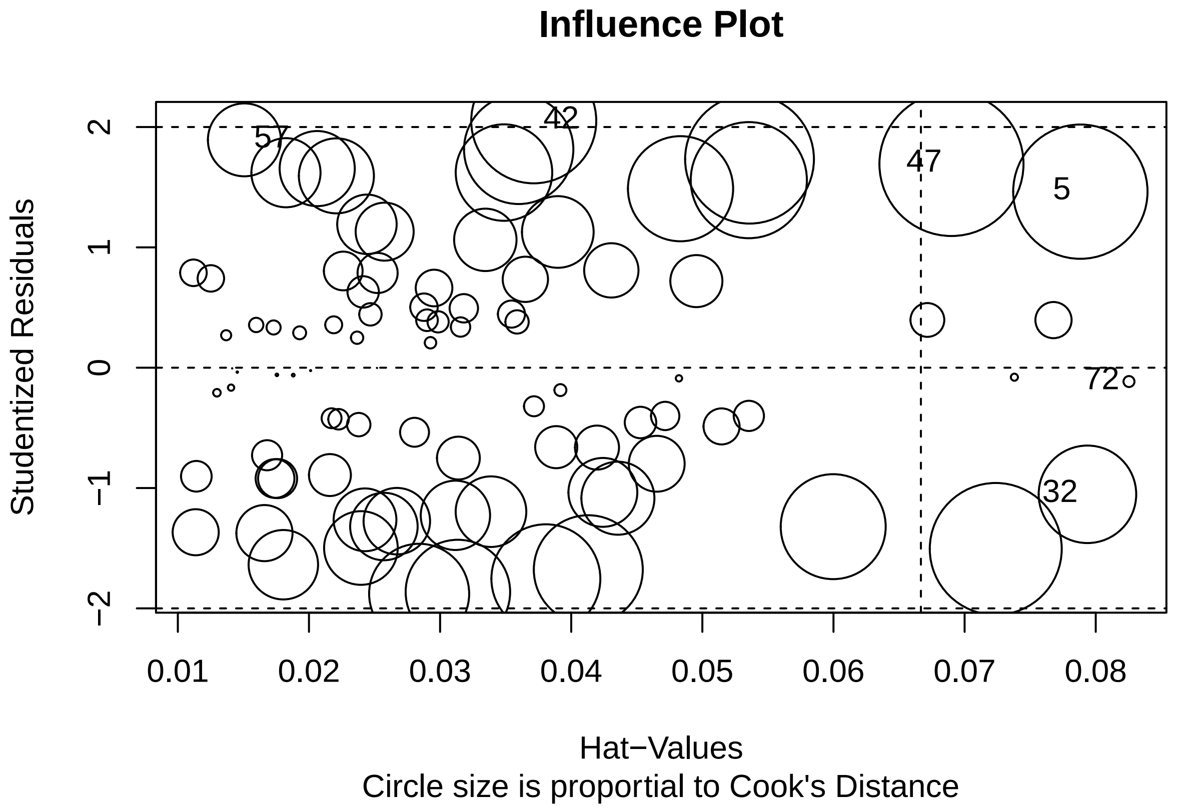
Check if errors are auto corelated

```
## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

##
##  Durbin-Watson test
##
## data:  model
## DW = 2.1766, p-value = 0.7891
## alternative hypothesis: true autocorrelation is greater than 0
```

We can see that there are outliers in this dataset mainly row 5,42,57,32 from X1, and 18,42,57,1 from X2. So, overall number 42 and 57 are outliers.

Influential Variables : Fiding influential data points is required. So let us look at variable plots:





```
##      StudRes        Hat        CookD
## 5   1.463036 0.07882556 0.0602640116
## 32 -1.052152 0.07936606 0.0317723820
## 42  2.051376 0.03714193 0.0521849091
## 47  1.692974 0.06899694 0.0693171600
## 57  1.893423 0.01507526 0.0177631223
## 72 -0.115048 0.08253501 0.0004014576
```

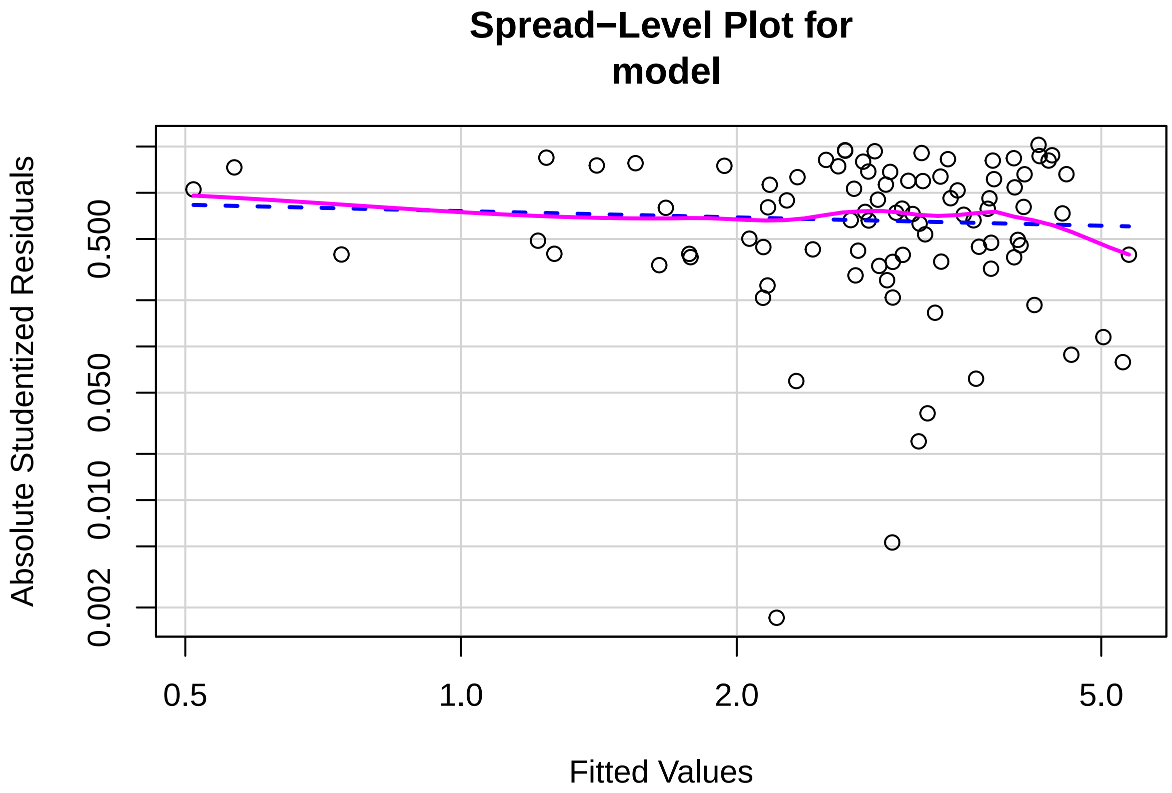
Checking for Multi-collinearity:

```
##       X1       X2
## 1.004315 1.004315
```

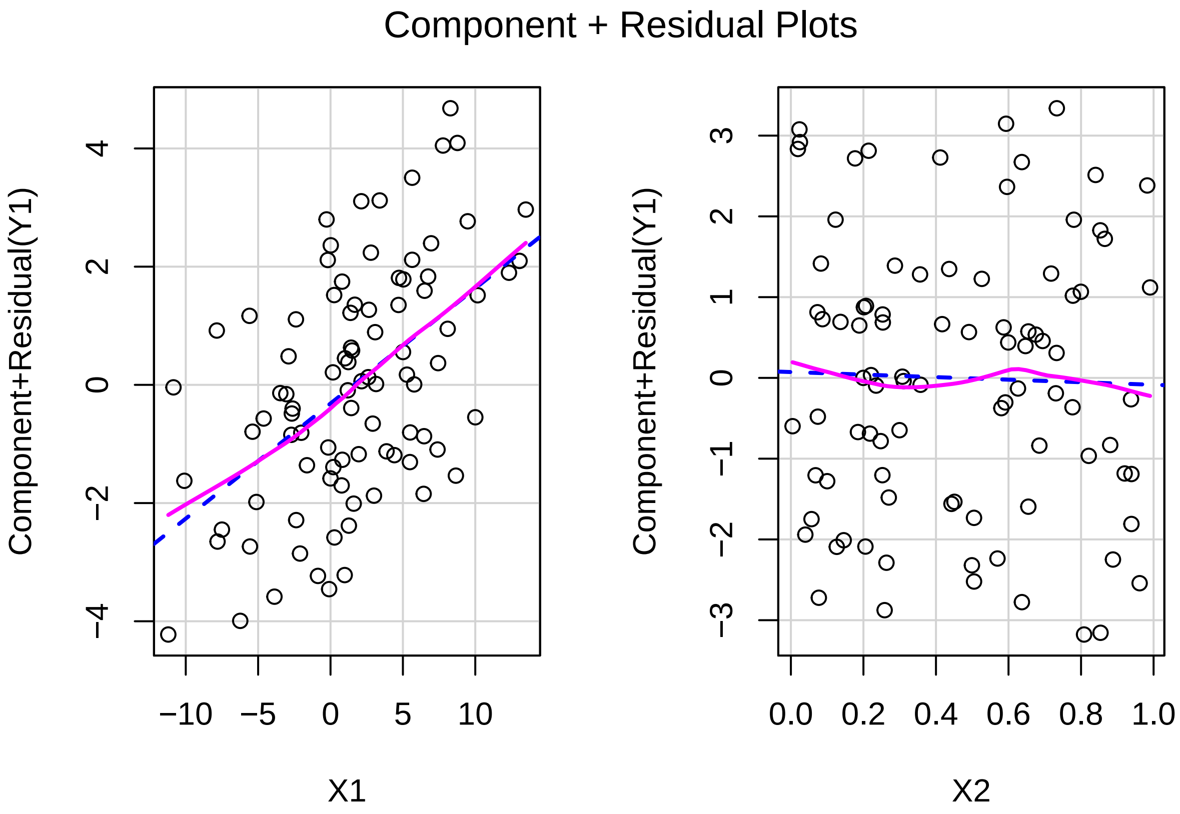
```
##    X1    X2
## FALSE FALSE
```

Non-constant Error Variance:

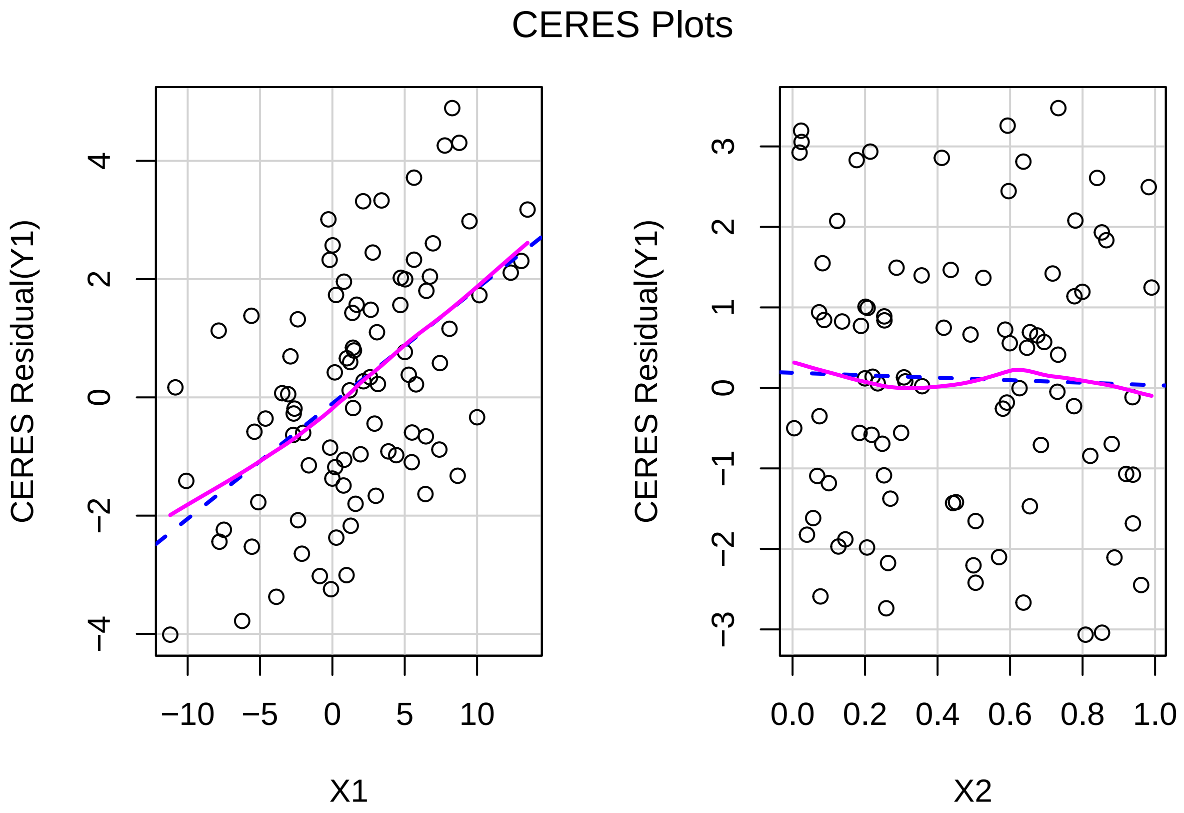
```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.0003181395, Df = 1, p = 0.98577
```



```
##
## Suggested power transformation:  1.136489
```



Nonlinearity test:



So all the factors are linear which is required.

Non-independence of Errors:

```
##  lag Autocorrelation D-W Statistic p-value
##    1      -0.1061408       2.17664   0.482
##  Alternative hypothesis: rho != 0
```