PROJECT

ANEESH MAKKAR, 7902203

23/11/22

Question

Question. State the response variable you would like to model.

Ans. The response variable that I will model is the number of innings played by an player.

Question. State your proposed explanatory variables and explain why you think they would be accurate predictors.

Ans. The explanatory variables that I will model are "number of times a player scores 50" and "number of times a player scores 0". 0 means that the player gets out at 0 score.

Number of times a player scores 50 should help predict number of innings played by a player as if a player scores 50 'n' number of times then the player must had played at least n innings and it is rarely possible that in every innings a player score 50. So, a player must had played more than n matches if that players scores 50 in n matches.

Number of times a player scores 0 should help predict number of innings played by a player as if a player scores 0 'n' number of times then the player must had played at least n innings and it is rarely possible that in every innings a player gets out scoring 0. So, a player must had played more than n matches if that players scores 0 in n matches.

Data Set

Dataset in tabular form

```
cric <- read.csv("stats.csv")

knitr::kable(cric,"latex",align=c("l","c","c"))
```

Player	Inns	X50	X0
SR Tendulkar (INDIA)	452	96	20
V Kohli (INDIA)	245	62	13
RT Ponting (AUS/ICC)	365	82	20
RG Sharma (INDIA)	220	43	13
ST Jayasuriya (Asia/SL)	433	68	34
HM Amla (SA)	178	39	4
AB de Villiers (Afr/SA)	218	53	7
CH Gayle (ICC/WI)	294	54	25
KC Sangakkara (Asia/ICC/SL)	380	93	15
SC Ganguly (Asia/INDIA)	300	72	16
TM Dilshan (SL)	303	47	11
LRPL Taylor (NZ)	217	51	9
HH Gibbs (SA)	240	37	22
Saeed Anwar (PAK)	244	43	15
BC Lara (ICC/WI)	289	63	16
DPMD Jayawardene (Asia/SL)	418	77	28
DA Warner (AUS)	126	23	2
ME Waugh (AUS)	236	50	16
AJ Finch (AUS)	128	29	11
S Dhawan (INDIA)	142	33	5
DL Haynes (WI)	237	57	13
JH Kallis (Afr/ICC/SA)	314	86	17
Q de Kock (SA)	124	26	4
JE Root (ENG)	142	35	5
MJ Guptill (NZ)	183	37	15
NJ Astle (NZ)	217	41	19
AC Gilchrist (AUS/ICC)	279	55	19
WU Tharanga (Asia/SL)	223	37	17
V Sehwag (Asia/ICC/INDIA)	245	38	14
Mohammad Yousuf (Asia/PAK)	273	64	15
Babar Azam (PAK)	81	17	3
Tamim Iqbal (BAN)	217	51	19
EJG Morgan (ENG/IRE)	228	47	16
Yuvraj Singh (Asia/INDIA)	278	52	18
KS Williamson (NZ)	144	39	5
G Kirsten (SA)	185	45	11
ME Trescothick (ENG)	122	21	13
PR Stirling (IRE)	131	26	10
F du Plessis (SA)	136	35	3
R Dravid (Asia/ICC/INDIA)	318	83	13
JM Bairstow (ENG)	81	14	6
SPD Smith (AUS)	113	25	5
CG Greenidge (WI)	127	31	3
WTS Porterfield (IRE)	142	20	10
G Gambhir (INDIA)	143	34	11
IVA Richards (WI)	167	45	7
BRM Taylor (ZIM)	203	39	15
Mohammad Hafeez (PAK)	216	38	19
S Chanderpaul (WI)	251	59	6

The data represents the record of 49 cricketers till the year 2021 in all the three formats. (ODI, T20 and test)

Reference

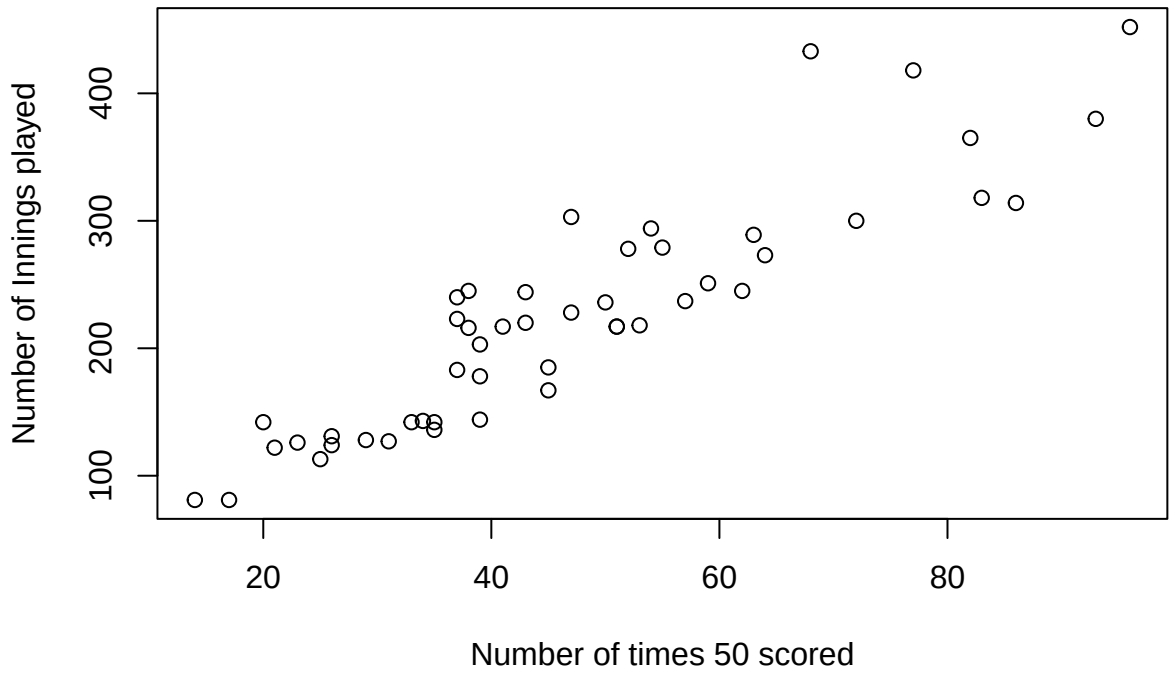
KRISHNA. (2021). Cricket Dataset. Retrieved from Cricket Dataset: https://www.kaggle.com/datasets/notkrishna/cricket-statistics-for-all-formats?select=odb.csv

Explanation of variable and the units of measurement.

The variable Inns is the number of innings a cricketer played.The variable X50 is the number of times a player scores 50.The variable X0 is the number of times a player scores 0.

Scatter plot between Inns and X50.

```
plot(cric$X50,cric$Inns, xlab = "Number of times 50 scored",ylab = "Number of Innings played")
```



Calculating degree of variability between Inns and x50(r^2)

```
cor(cric$Inns,cric$X50)^(2)
```

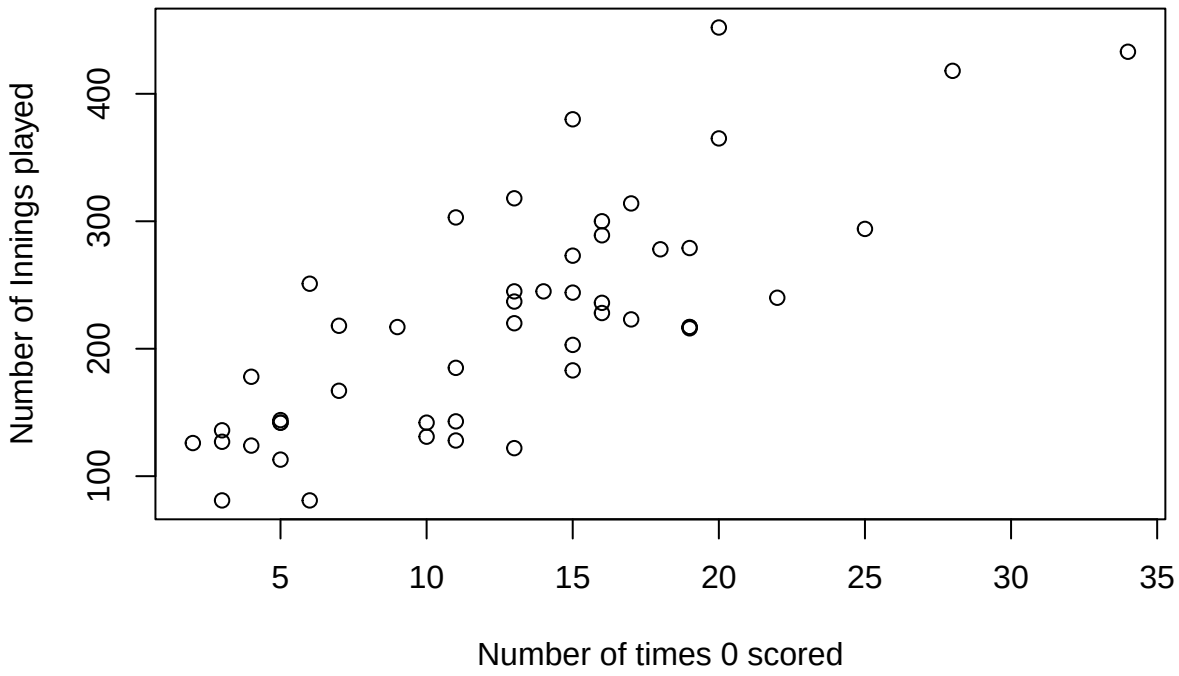
```
## [1] 0.8207977
```

$r^2 = 0.8207977$

So, there exists a positive relationship between number of times 50 scored and number of innings played.

Scatter plot between Inns and X0.

```
plot(cric$X0,cric$Inns, xlab = "Number of times 0 scored",ylab = "Number of Innings played")
```



Calculating degree of variability between Inns and x0(r^2)

```
cor(cric$Inns,cric$X0)^(2)
```

```
## [1] 0.5943858
```

$r^2 = 0.5943858$

So, there exists a postive relationship between number of times 0 scored and number of innings played.

Preliminary Model

```
cric.lm <- lm(Inns ~ X0, data = cric)
summary(cric.lm)
```

```
##
## Call:
## lm(formula = Inns ~ X0, data = cric)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -102.244  -43.340   -1.044   32.718  157.823
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   94.368     17.608   5.359 2.47e-06 ***
## X0             9.990      1.204   8.299 9.15e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 57.8 on 47 degrees of freedom
## Multiple R-squared:  0.5944, Adjusted R-squared:  0.5858
## F-statistic: 68.87 on 1 and 47 DF,  p-value: 9.148e-11
```

Regression Line:

$\hat{y} = 94.368 + 9.99X1$

```
cric2.lm <- lm(Inns ~ X50, data = cric)
summary(cric2.lm)
```

```
##
## Call:
## lm(formula = Inns ~ X50, data = cric)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -67.96 -26.52  -6.90  23.65 124.55
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  30.7219    14.2349   2.158   0.0361 *
## X50           4.0842     0.2784  14.672   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 38.42 on 47 degrees of freedom
## Multiple R-squared:  0.8208, Adjusted R-squared:  0.817
## F-statistic: 215.3 on 1 and 47 DF,  p-value: < 2.2e-16
```

Regression Line:

$\hat{y} = 30.7219 + 4.0842X2$

```
cric3.lm <- lm(Inns ~ X0 + X50, data = cric)
summary(cric3.lm)
```

```
##
## Call:
## lm(formula = Inns ~ X0 + X50, data = cric)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -50.861 -16.225  -1.501  10.758  89.794
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  11.4820     9.5984   1.196    0.238
## X0            5.0304     0.6275   8.017 2.79e-10 ***
## X50           3.1147     0.2183  14.268  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25.08 on 46 degrees of freedom
## Multiple R-squared:  0.9252, Adjusted R-squared:  0.922
## F-statistic: 284.7 on 2 and 46 DF,  p-value: < 2.2e-16
```

$RegressionLine : \hat{y} = 5.0304X1 + 3.1147X2 + 11.4820$

For Inns ~ X0, $R^2_{adj} = 0.5858$

For Inns ~ X50, $R^2_{adj} = 0.817$

For Inns ~ X0 + X50, $R^2_{adj} = 0.922$

So, R^2_{adj} has increased from both Inns ~ X0 and Inns ~ X50.

Second Order Model

```
cric.full <- lm(Inns ~ X0 + X50 + I(X0^2) + I(X50^2) + X0 * X50 , data = cric)

summary(cric.full)
```

```
##
## Call:
## lm(formula = Inns ~ X0 + X50 + I(X0^2) + I(X50^2) + X0 * X50,
##     data = cric)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.139 -11.953  -4.397   7.728  93.778
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 32.725784  22.070708   1.483  0.14543
## X0           1.133768   1.916151   0.592  0.55716
## X50          3.122231   0.885361   3.527  0.00102 **
## I(X0^2)      0.099216   0.092283   1.075  0.28831
## I(X50^2)    -0.001708   0.011398  -0.150  0.88161
## X0:X50       0.017499   0.063021   0.278  0.78260
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.41 on 43 degrees of freedom
```

```
## Multiple R-squared:  0.9338, Adjusted R-squared:  0.9261
## F-statistic: 121.4 on 5 and 43 DF,  p-value: < 2.2e-16
```

Regression Line:

$\hat{y} = 32.725784 + 1.133768X1 + 3.122231X2 + 0.099216X1^2 - 0.001708X2^2 + 0.017499X1X2$

Conducting ANOVA Test

LEVEL OF SIGNIFICANCE: 0.05

HYPOTHESES: H_o: $\beta_1 = \beta_2 = 0$ H_a: At least one β_i != 0 where i = 1,2

DECISION RULE: if p-value $< \alpha$ then reject H_0. if P-value $> \alpha$ then fail to reject H_0.

TEST STATISTIC: 121.4

P-VALUE: ≈ 0

CONCLUSION: p-value $< \alpha$. So, we reject H_0. So, we have sufficient evidence to conclude that at least one of the model terms is significant.

Model Refinement

```
summary(cric.full)
```

```
##
## Call:
## lm(formula = Inns ~ X0 + X50 + I(X0^2) + I(X50^2) + X0 * X50,
##     data = cric)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.139 -11.953  -4.397   7.728  93.778
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 32.725784  22.070708   1.483  0.14543
## X0           1.133768   1.916151   0.592  0.55716
## X50          3.122231   0.885361   3.527  0.00102 **
## I(X0^2)      0.099216   0.092283   1.075  0.28831
## I(X50^2)    -0.001708   0.011398  -0.150  0.88161
## X0:X50       0.017499   0.063021   0.278  0.78260
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.41 on 43 degrees of freedom
## Multiple R-squared:  0.9338, Adjusted R-squared:  0.9261
## F-statistic: 121.4 on 5 and 43 DF,  p-value: < 2.2e-16
```

Only the coefficient of X50 has significant p-value as it is less than 0.05. X0 has p-value greater than 0.05 but we will keep it in model as it is first order term.

p-value of X50: 0.00102 p-value of X0: 0.55716

Proposing Reduced Model

```
cric.reduced <- lm(Inns ~ X0 + X50, data= cric)
```

Performing nested F-Test

```
anova(cric.reduced,cric.full)
```

```
## Analysis of Variance Table
##
## Model 1: Inns ~ X0 + X50
## Model 2: Inns ~ X0 + X50 + I(X0^2) + I(X50^2) + X0 * X50
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1     46 28944
## 2     43 25622  3    3321.6 1.8582  0.151
```

LEVEL OF SIGNIFICANCE: 0.05

HYPOTHESES: H_o: $\beta_3 = \beta_4 = \beta_5 = 0$ H_a: At least one β_i != 0 where i = 3,4,5

DECISION RULE: if p-value $< \alpha$ then reject H_0. if P-value $> \alpha$ then fail to reject H_0.

TEST STATISTIC: 1.8582

P-VALUE: 0.151

CONCLUSION: p-value $> \alpha$. So, we fail to reject H_0. So, we have insufficient evidence to conclude that the coefficients of the terms $X50^2$, $X0^2$ and X50 * X0 are significant.

Final Model and Assessment

Conducting ANOVA test on reduced model

```
summary(cric.reduced)
```

```
##
## Call:
## lm(formula = Inns ~ X0 + X50, data = cric)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -50.861 -16.225  -1.501  10.758  89.794
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  11.4820     9.5984   1.196    0.238
## X0            5.0304     0.6275   8.017 2.79e-10 ***
## X50           3.1147     0.2183  14.268  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 25.08 on 46 degrees of freedom
## Multiple R-squared:  0.9252, Adjusted R-squared:  0.922
## F-statistic: 284.7 on 2 and 46 DF,  p-value: < 2.2e-16
```

LEVEL OF SIGNIFICANCE: 0.05

HYPOTHESES: H_o: $\beta_1 = \beta_2 = 0$ H_a: At least one β_i != 0 where i = 1,2

DECISION RULE: if p-value $< \alpha$ then reject H_0. if P-value $> \alpha$ then fail to reject H_0.

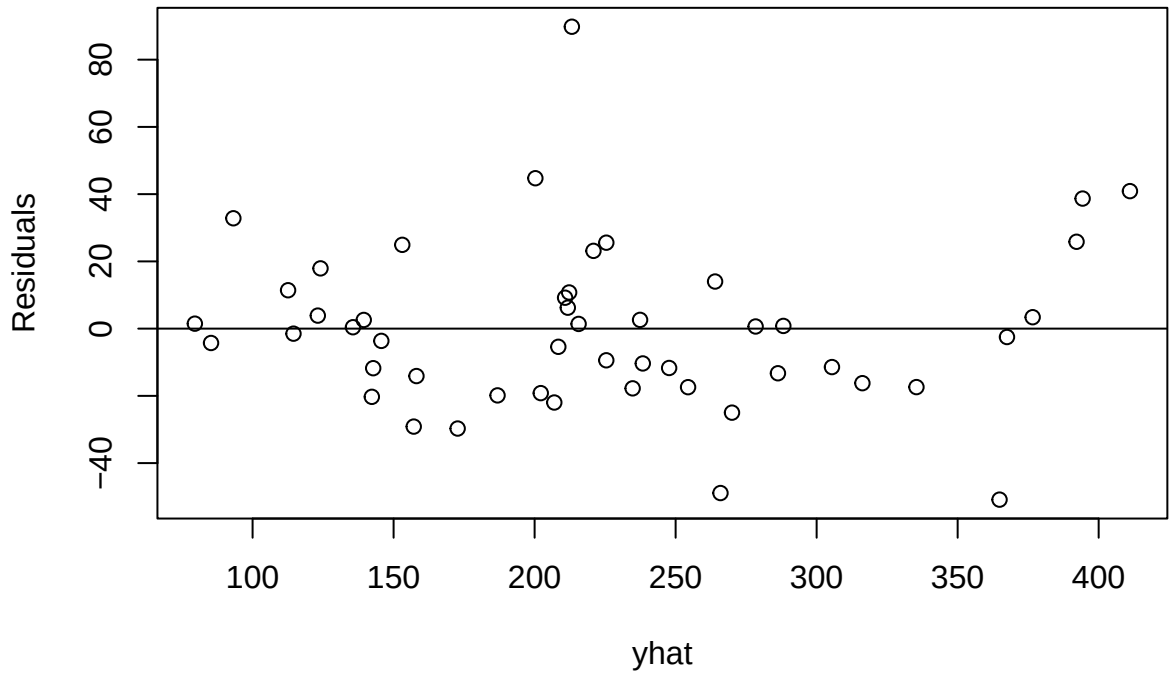
TEST STATISTIC: 284.7

P-VALUE: ≈ 0

CONCLUSION: p-value $< \alpha$. So, we reject H_0. So, we have sufficient evidence to conclude that at least one of the model terms is significant.

Residual Plot:

```
cric.res <- resid(cric.reduced)
cric.fitted <- fitted.values(cric.reduced)
plot(cric.fitted,cric.res,ylab = "Residuals", xlab="yhat")
abline(h = 0)
```

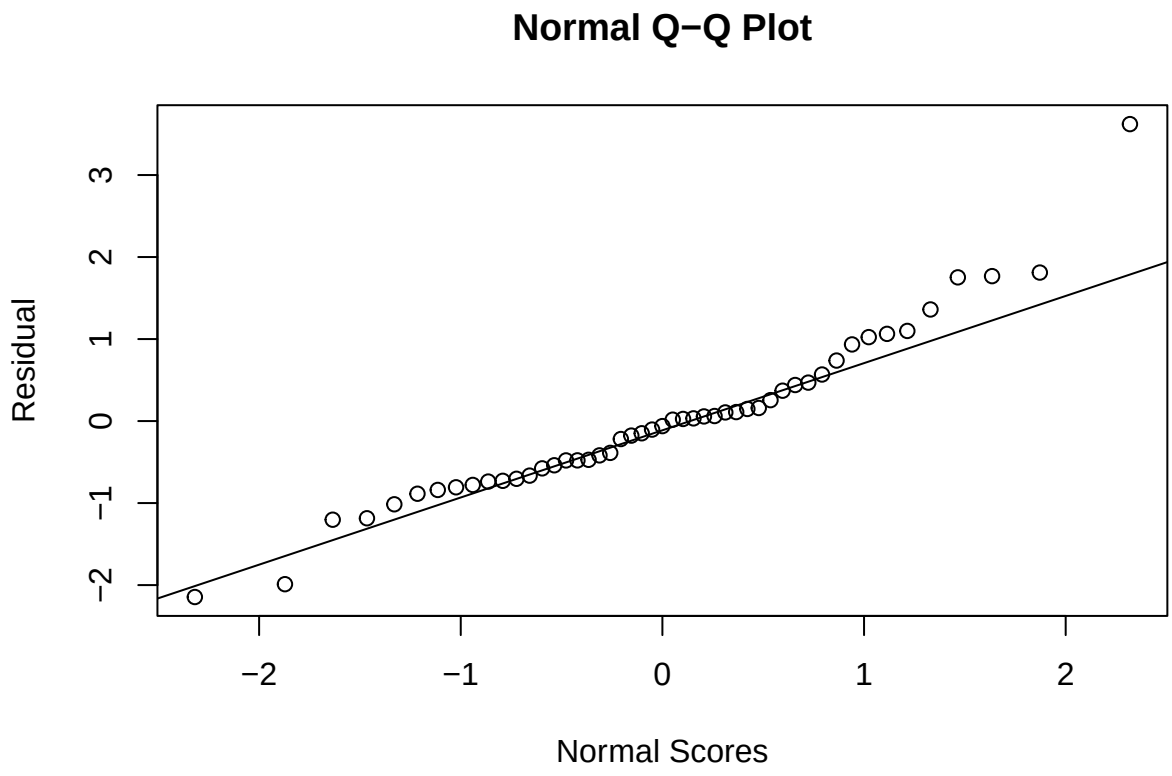


Linearity: There is no discernible pattern in the points so the data is linear.

Not Constant Variance: The width of the error terms is not similar throughout the plot and be the same above and below the axis.

Normal Quantile Plot:

```
cric.stdres = rstandard(cric.reduced)
qqnorm(cric.stdres, ylab = "Residual", xlab = "Normal Scores")
qqline(cric.stdres)
```



Normality of Error terms: error terms follow a normal distribution as all observations are near the qunatile-quantile line.

Conclusion

According to my opinion, the explanatory variables were able to predict the response variable as the final reduced model has those variables. Moreover, the model passed the nested f test. Another thing that I want to add is that the data follows normal distribution according to the normal quantile quantile plot and residual plot.

Final Regression Line

$\hat{y} = 5.0304X1 + 3.1147X2 + 11.4820$