correlation and Simple Regression Assignment

Pradeep Paladugula

6/11/2020

1. When do we want to use correlation rather than using ANOVA? Please provide an example.
Answer: Correlation and ANOVA either of the analysis does not assume dependence of any variable on other variable, neither does they tries to find out the relationship between the two. It simply estimates the degree of association between variables. For both the analysis the values of dependent and independant variables are random. But we use correaltion over ANOVA when we test interdependence of association between variables.

2. Please explain why partial correlation is useful and provide an example.
definition: A correlation between two variables in which the effects of other variables are held constant is knownas partial Correlation. I can conduct Partial Correlation with more than just 1 third-variable. I can include as many thrid-variables as I wish for the analysis. PCor explains how variables work together to explain patterns in the data. When multiple regression is taken in to consideration it explains how individual variable explain the relationship.

example: in The below data, the score relates to total happines score which is a constant variable and variance accounted by the Perceptions of corruption and variance accounted by the GDP per capita score. And also unique variance accounted by the both the variables for the Happiness score. What regions rank the highest in overall score of happiness and each of the two factors contributing to happiness.

```
library(ggm)
library(reprex)
happinessscore <- read.csv('datasets-894-813759-2019.csv')
#vgsalesNew <- na.omit(vgsales)
head(happinessscore)
```

```
##   Overall.rank Country.or.region Score GDP.per.capita Social.support
## 1            1           Finland 7.769          1.340          1.587
## 2            2           Denmark 7.600          1.383          1.573
## 3            3            Norway 7.554          1.488          1.582
## 4            4           Iceland 7.494          1.380          1.624
## 5            5       Netherlands 7.488          1.396          1.522
## 6            6       Switzerland 7.480          1.452          1.526
##   Healthy.life.expectancy Freedom.to.make.life.choices Generosity
## 1                   0.986                        0.596      0.153
## 2                   0.996                        0.592      0.252
```

```
## 3                   1.028                          0.603      0.271
## 4                   1.026                          0.591      0.354
## 5                   0.999                          0.557      0.322
## 6                   1.052                          0.572      0.263
##   Perceptions.of.corruption
## 1                     0.393
## 2                     0.410
## 3                     0.341
## 4                     0.118
## 5                     0.298
## 6                     0.343
```

```
pcor(c('GDP.per.capita', "Perceptions.of.corruption",'Score'),
var(happinessscore))
```

```
## Warning in var(happinessscore): NAs introduced by coercion
```

```
## [1] -0.01285421
```

```
3.  When should we use regression instead of ANOVA?
Answer:
  Differences:
  a) Regression: Regression modal is based on one more continuous predictor
variable
     ANOVA: ANOVA modal based on one or more categorical predictor variables
  b) Regression: Focuses on fixed or independent or continuous variables
     ANOVA: Focuses on random variables
  C) Regression: Only single error term
     ANOVA: Several error terms
  d) Regression: Used mainly for forecasting and predictions
     ANOVA: Used mainly to determine if data from various groups have a
common means or not
```

#e) regression: Francis Galton who coined the term 'regression' in the 19th century
#ANOVA: ANOVA got wide popularity when Sir Ronald Fisher included the term in his book 'statistical methods for research workers'

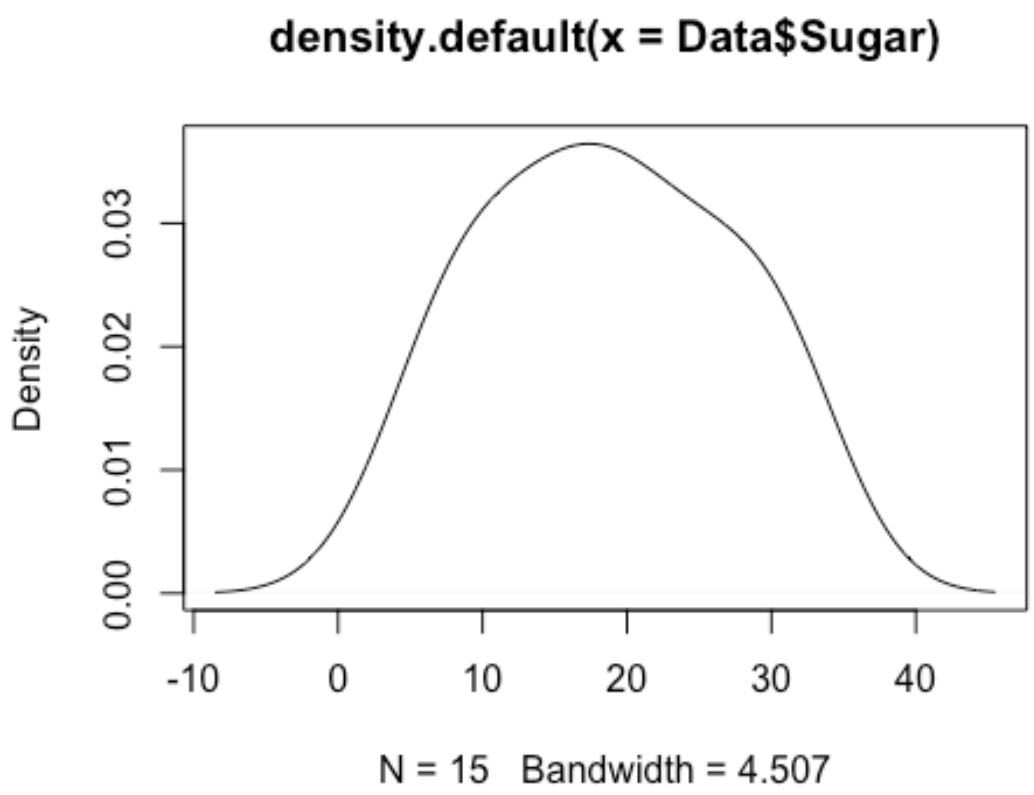
```
4. Please explain the relationship between SStotal, SSregression and SSerror.
Answer:
  SStotal: The Sum of squares total is denoted as SST, It is the sqaured
difference between the observed variable and its mean (dispersion of the
obsered variables around mean, it is a mesure of the total variability of the
dataset)

  SSregression: The Sum of Squares due to regression or SSR. It is the sum of
difference betweent he predicted value and mean of the dipendent variable(the
measure taht decsribes how well the line fits the data). If the value of SSR
is equal to the SST, It means our regression model captures the observed
variability and is perfect.

  SSerror: The Sum of Squares error or SSE. The error is the difference
```

between the observed variable and the predicted value. we usually want to minimize the error. smaller the error, the better the estimation power of regression. it is also called as residual sum of squares.

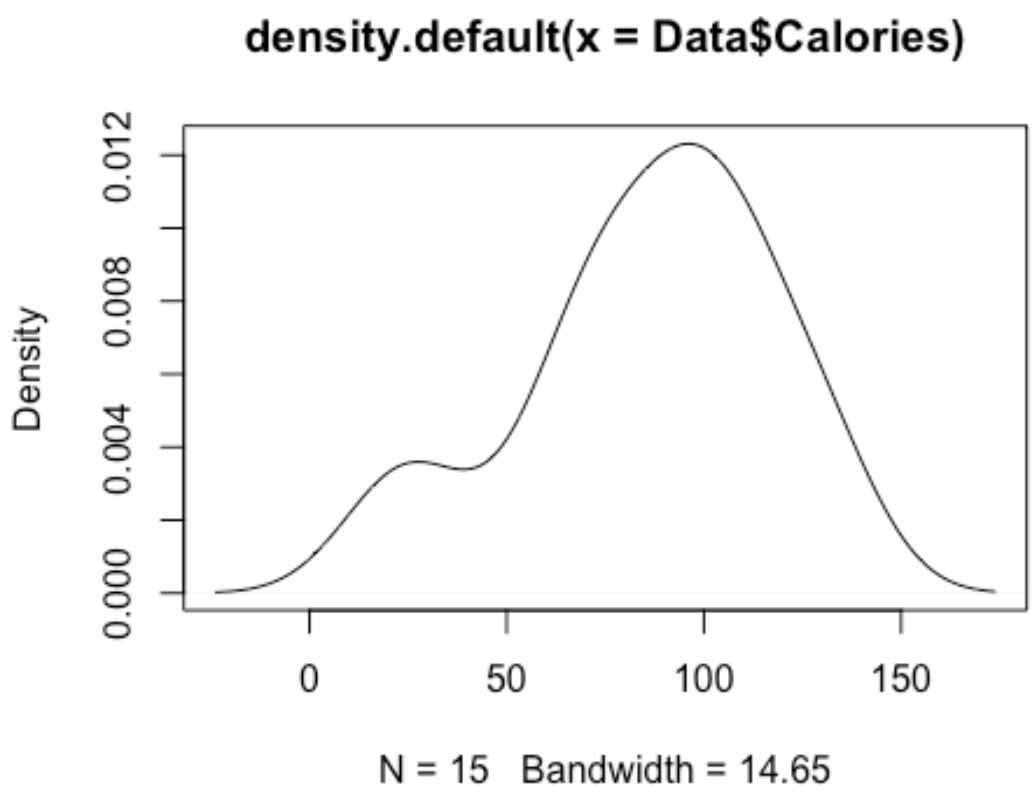
The above three are mathematically related: SST = SSR + SSE

5. Please use the following data to build a regression model and write a summary. IV is sugar and DV is calories.
Sugar: 5, 8, 9, 10, 15, 18, 14, 17, 20, 22, 24, 26, 30 ,30, 32
Calories: 20, 30, 60, 70, 100, 95, 70, 83, 103, 112, 130, 80, 95, 130, 112

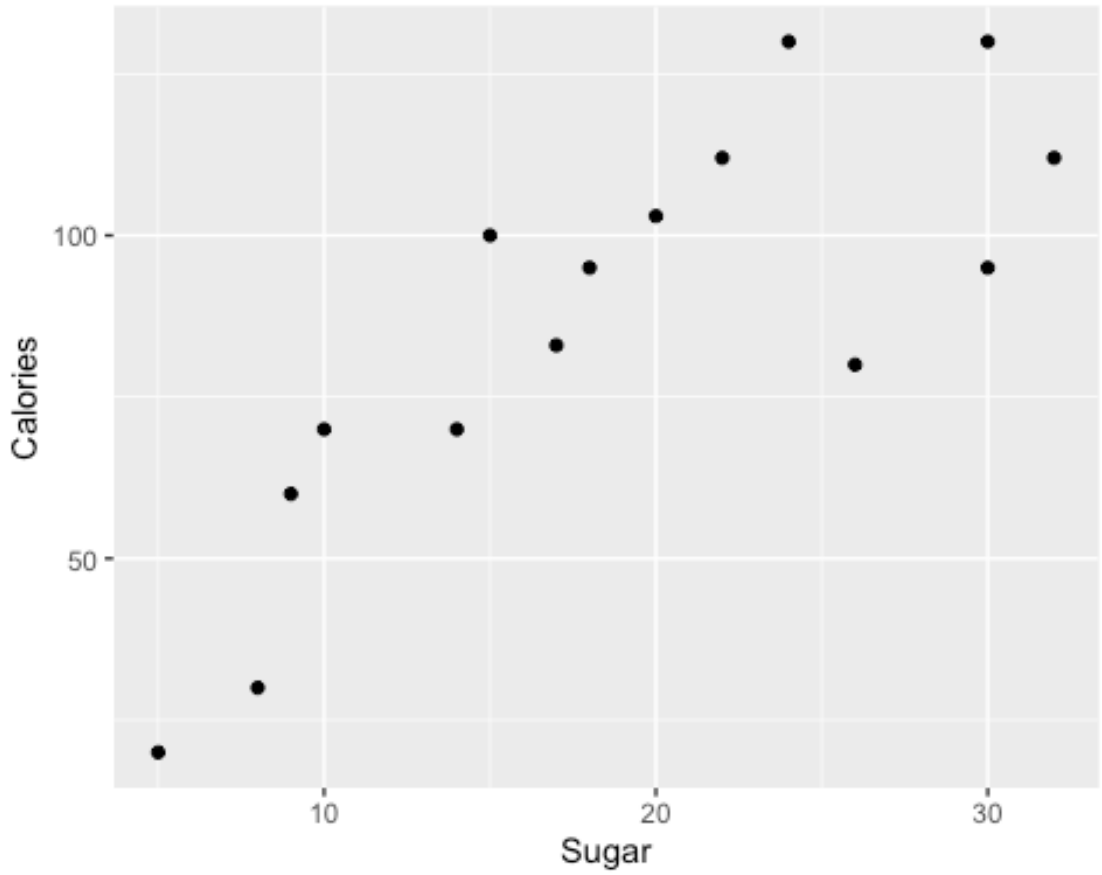
```
Sugar <- c(5, 8, 9, 10, 15, 18, 14, 17, 20, 22, 24, 26, 30 ,30, 32)
Calories <- c(20, 30, 60, 70, 100, 95, 70, 83, 103, 112, 130, 80, 95, 130, 112)
Data <- data.frame(Sugar, Calories)

#density plot to check non zero variaance
plot(density(Data$Sugar))
```



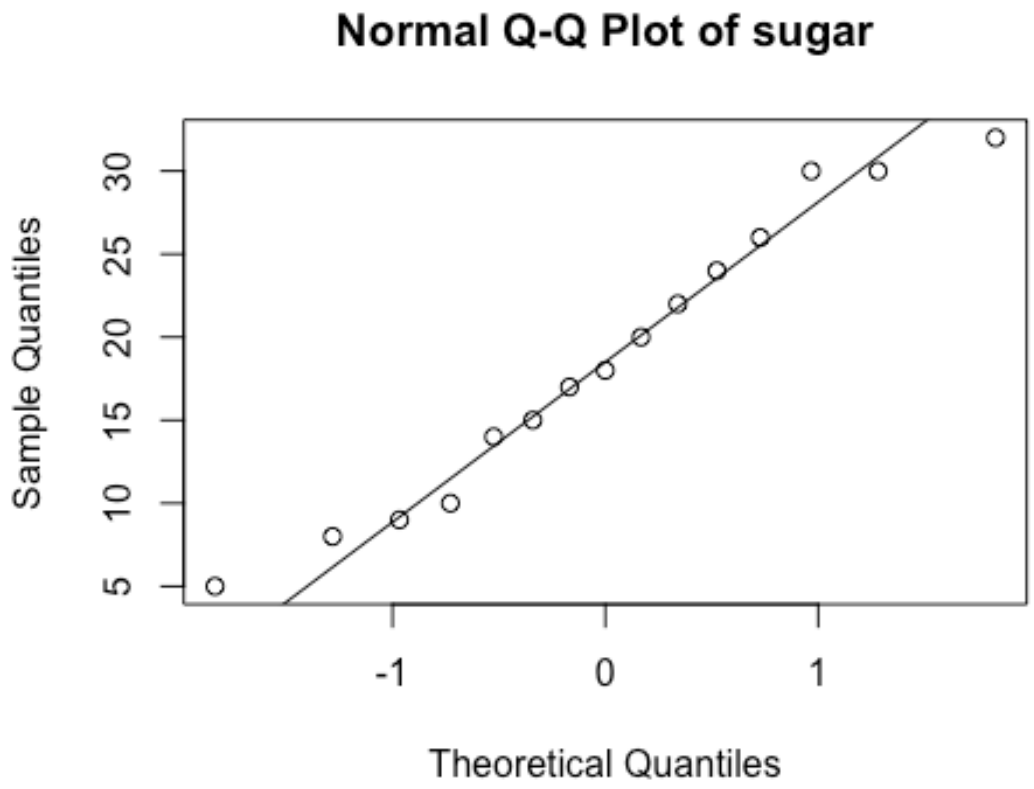

```r
plot((density(Data$Calories)))

#Homoscedasticity/linearity test: sccatter plot
library(ggplot2)
```



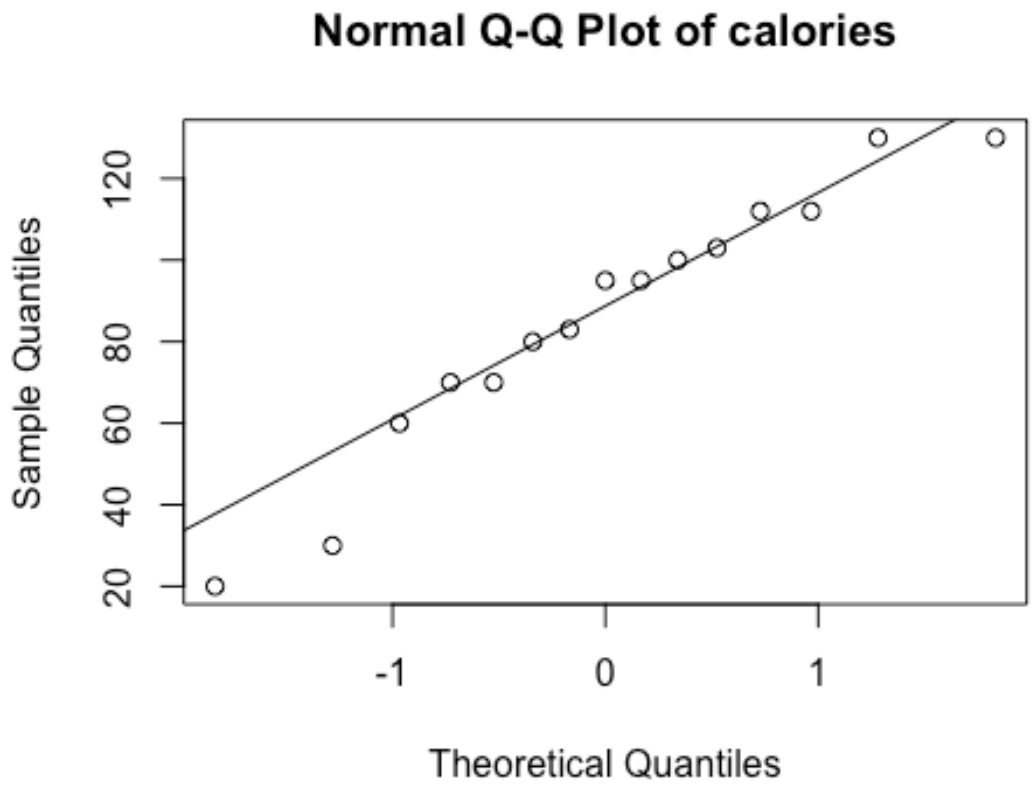

```
scatterPlot <- ggplot(data = Data, aes(Sugar, Calories))
scatterPlot + geom_point() + labs(main = 'Heteroscedasticity or linearity
test between Sugar and Calories')
```




```r
##Normally Distributed Error
qqnorm(Data$Sugar, main = "Normal Q-Q Plot of sugar")
qqline(Data$Sugar)
```

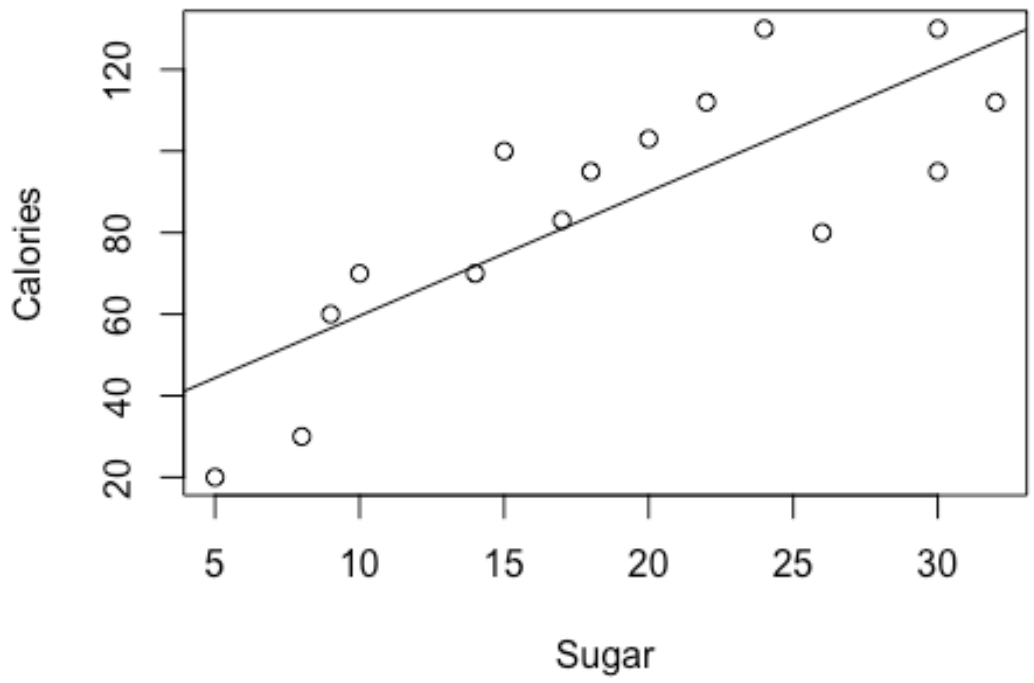



```r
qqnorm(Data$Calories, main = "Normal Q-Q Plot of calories")
qqline(Data$Calories)
```

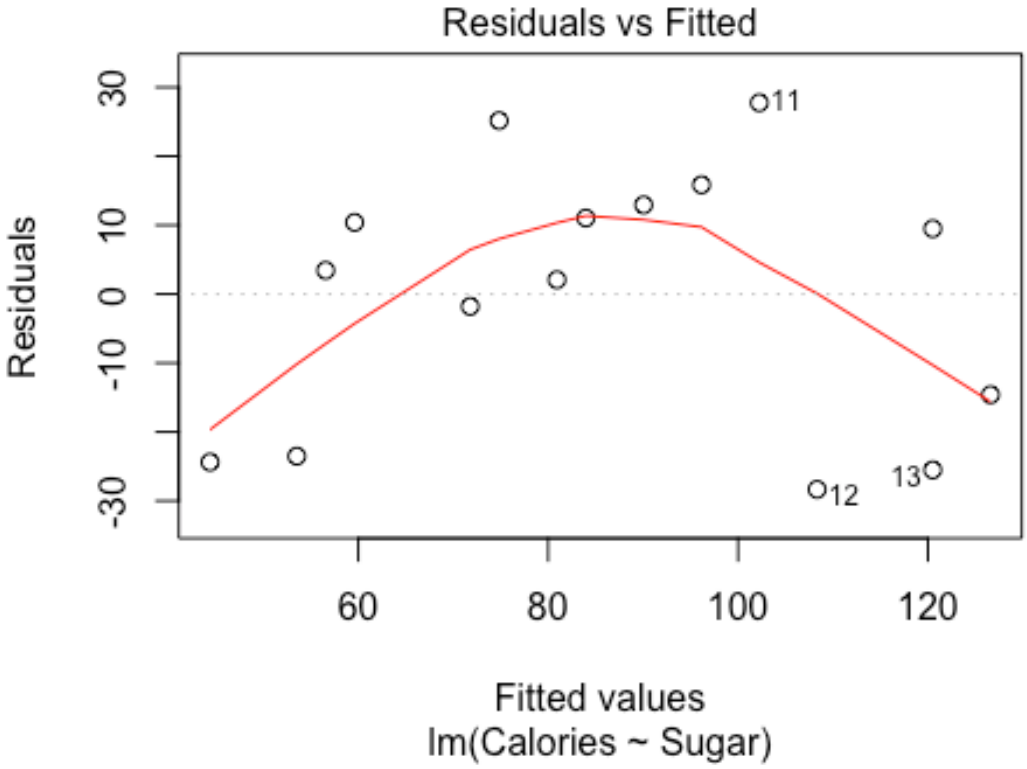


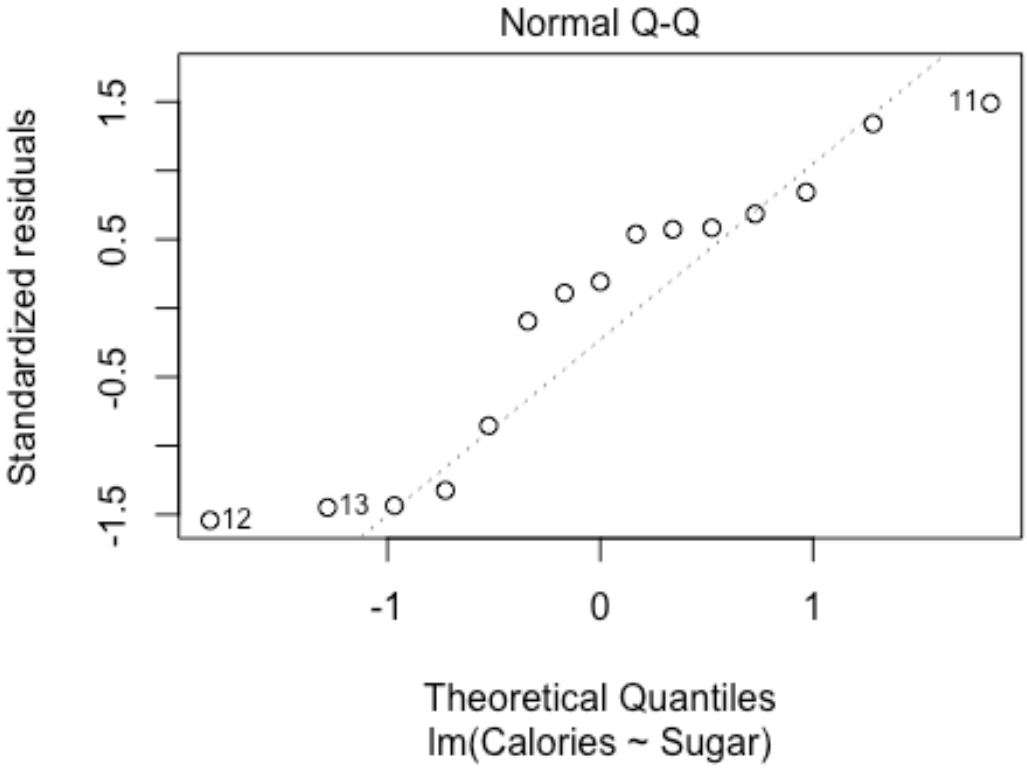

```
##Regression Model
model <- lm(Calories ~ Sugar, data = Data)
summary(model)

##
## Call:
## lm(formula = Calories ~ Sugar, data = Data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.332 -19.060   3.438  11.985  27.758
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  29.1542    12.4132   2.349 0.035315 *
## Sugar         3.0453     0.6074   5.013 0.000237 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.56 on 13 degrees of freedom
## Multiple R-squared:  0.6591, Adjusted R-squared:  0.6329
## F-statistic: 25.13 on 1 and 13 DF,  p-value: 0.0002373
```

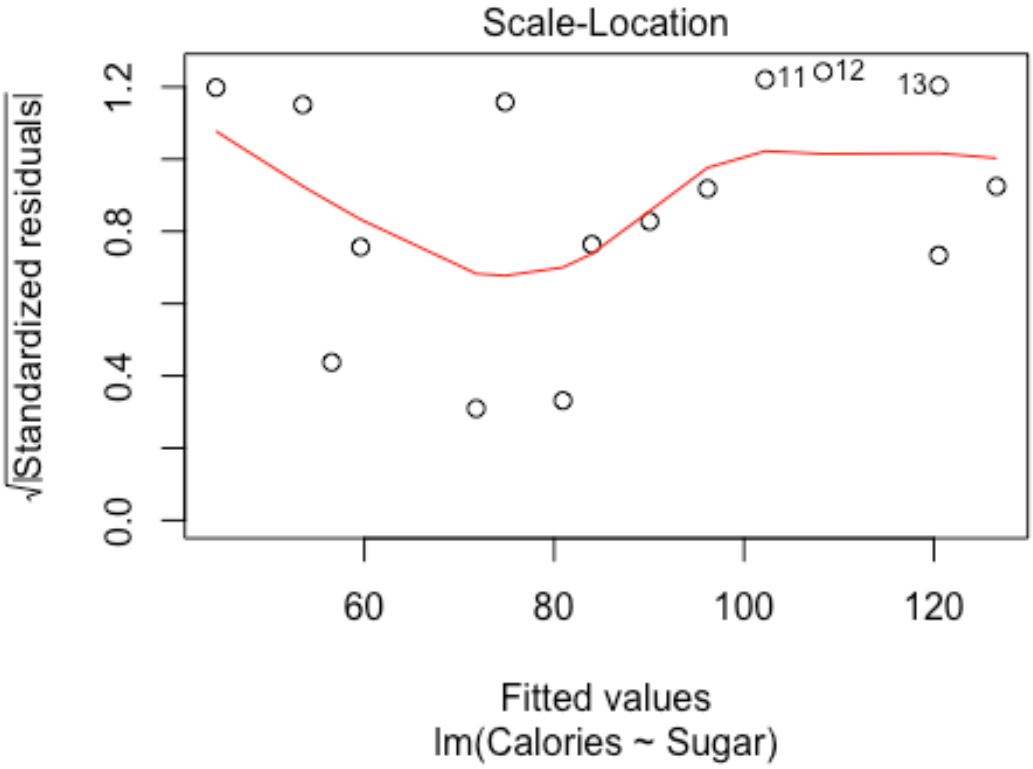
```r
##Model Plot
plot(Calories ~ Sugar, data = Data)
abline(model)
```

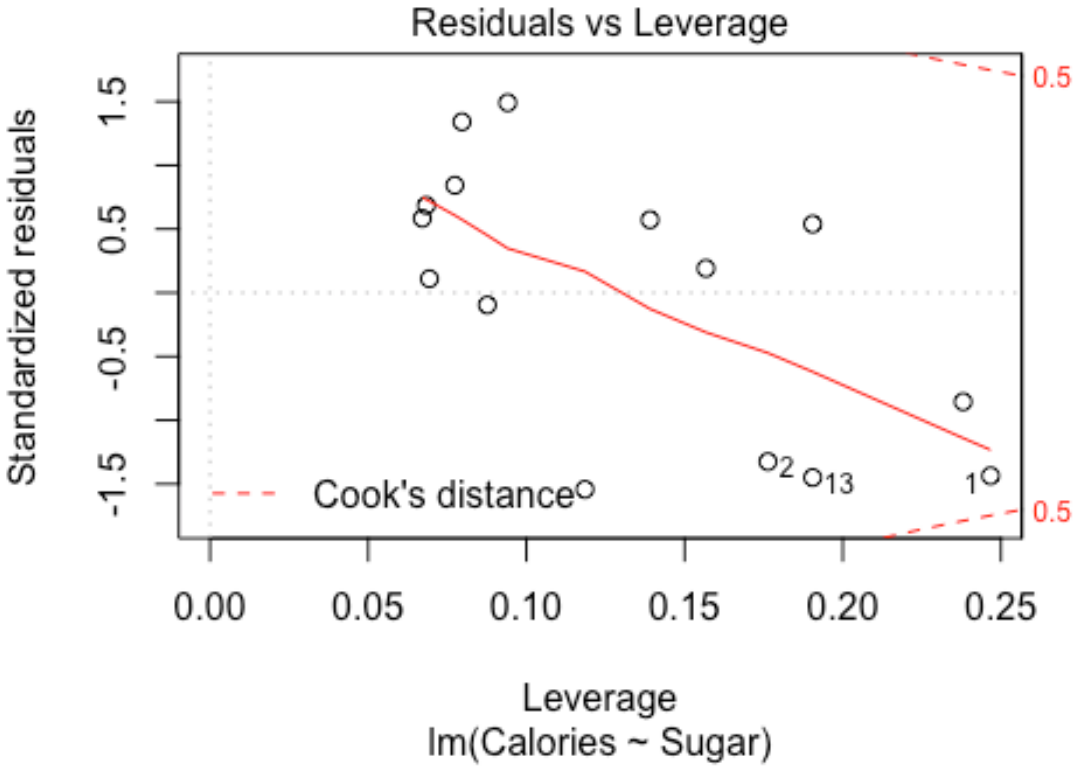



```r
plot(model)
```


Residuals vs Fitted
Residuals
30
10
0
-10
-30
11
12
13
60
80
100
120
Fitted values
lm(Calories ~ Sugar)


Normal Q-Q
Standardized residuals
1.5
0.5
-0.5
-1.5
11
12
13
-1
0
1
Theoretical Quantiles
lm(Calories ~ Sugar)


Scale-Location
√|Standardized residuals|
0.0
0.4
0.8
1.2
11
12
13
60
80
100
120
Fitted values
lm(Calories ~ Sugar)




```
#regression Equation
#Y = 29.1542 + 3.0435X
```

I have conducted simple regression modal to predict the calories to the amount of sugar intake. Their is no further adjustment made to the data. A significant regression equation was found ($F(1, 13) = 25.13$, $p = 0.0002373 < 0.001$), with and Rsquare = 0.66). Both intercept($p = 0.04 < 0.05$) and predictor($p < 0.001$) were statiscally significant, which falls in 99.9% area under the curve. I observed from the residual vs fitted plot this linear model is not appropriate and Normal QQ plot of the theoritical quantiles and standardized residuals are linear or they are nearly normal to my predictions because the ouliers in the cooks distance graph are inside the average cooks distnace values.