**Simple Linear Regression Assignment**

**1 - Calories\_consumed-> predict weight gained using calories consumed**

> Calories\_consumed <- read.csv("E:/Data Science Asignments/Simple regression/calories\_consumed.csv")

> summary(Calories\_consumed)

Weight.gained..grams. Calories.Consumed

Min. : 62.0 Min. :1400

1st Qu.: 114.5 1st Qu.:1728

Median : 200.0 Median :2250

Mean : 357.7 Mean :2341

3rd Qu.: 537.5 3rd Qu.:2775

Max. :1100.0 Max. :3900

> var(Calories\_consumed$Calories.Consumed)

[1] 565668.7

> sd(Calories\_consumed$Calories.Consumed)

[1] 752.1095

> var(Calories\_consumed$Weight.gained..grams.)

[1] 111350.7

> sd(Calories\_consumed$Weight.gained..grams.)

[1] 333.6925

> #Creating Linear Model for weight gain

> WeightGainModel <- lm(Weight.gained..grams. ~ Calories.Consumed, data = Calories\_consumed)

> summary(WeightGainModel)

Call:

lm(formula = Weight.gained..grams. ~ Calories.Consumed, data = Calories\_consumed)

Residuals:

Min 1Q Median 3Q Max

-158.67 -107.56 36.70 81.68 165.53

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -625.75236 100.82293 -6.206 4.54e-05 \*\*\*

Calories.Consumed 0.42016 0.04115 10.211 2.86e-07 \*\*\*

---

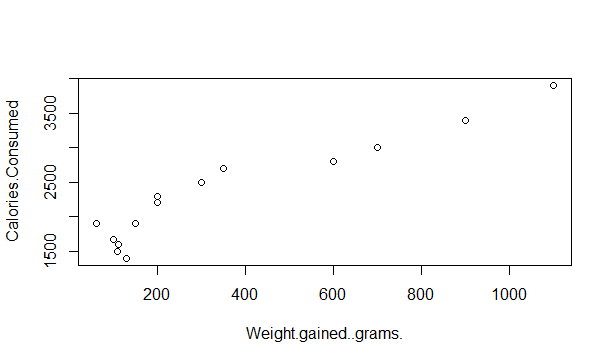
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 111.6 on 12 degrees of freedom

Multiple R-squared: 0.8968, Adjusted R-squared: 0.8882

F-statistic: 104.3 on 1 and 12 DF, p-value: 2.856e-07

> plot(Calories\_consumed)



**Here the P-value is less than 0.05.**

**So X varibale is significance and also Multiple R-Square value is 0.8968.**

**That’s mean this model will predict the output 89.68% time correct**

### 2 - Delivery\_time -> Predict delivery time using sorting time

**Reding the data from data file and saving into a variable**

|  |
| --- |
|  |
|  |

> delivery\_time <- read.csv("E:/Data Science Asignments/Simple regression/delivery\_time.csv")

> var(delivery\_time$Sorting.Time)

> summary(delivery\_time)

Delivery.Time Sorting.Time

Min. : 8.00 Min. : 2.00

1st Qu.:13.50 1st Qu.: 4.00

Median :17.83 Median : 6.00

Mean :16.79 Mean : 6.19

3rd Qu.:19.75 3rd Qu.: 8.00

Max. :29.00 Max. :10.00

> # Variance and Standard deviation of Delivery.Time column

> var(delivery\_time$Delivery.Time)

[1] 25.75462

> sd(delivery\_time$Delivery.Time)

[1] 5.074901

[1] 6.461905

> sd(delivery\_time$Sorting.Time)

[1] 2.542028

> #Creating Linear Model for delivery time

> deliverTimeModel <- lm(Delivery.Time ~ Sorting.Time, data = delivery\_time)

> summary(deliverTimeModel)

Call:

lm(formula = Delivery.Time ~ Sorting.Time, data = delivery\_time)

Residuals:

Min 1Q Median 3Q Max

-5.1729 -2.0298 -0.0298 0.8741 6.6722

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 6.5827 1.7217 3.823 0.00115 \*\*

Sorting.Time 1.6490 0.2582 6.387 3.98e-06 \*\*\*

---

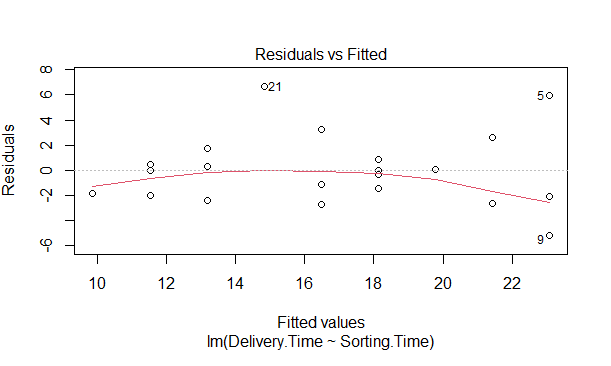
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

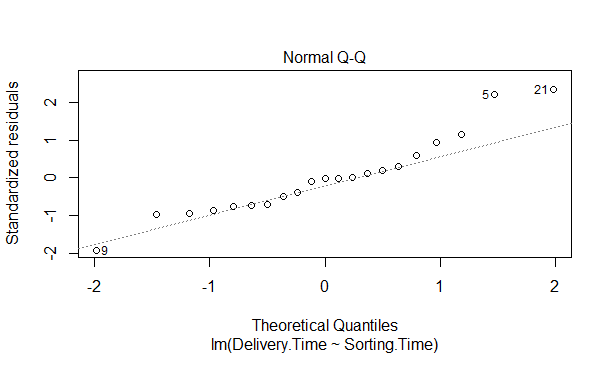
Residual standard error: 2.935 on 19 degrees of freedom

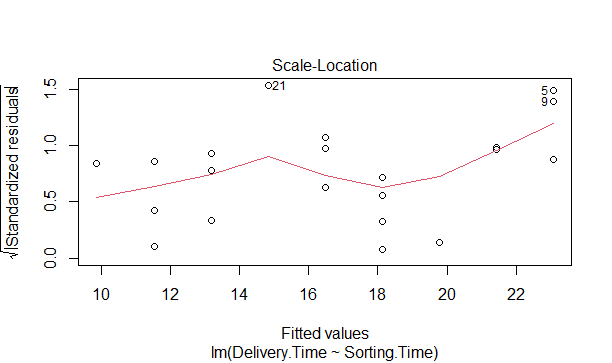
Multiple R-squared: 0.6823, Adjusted R-squared: 0.6655

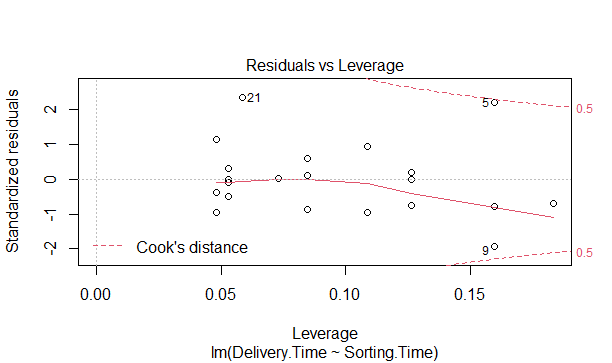
F-statistic: 40.8 on 1 and 19 DF, p-value: 3.983e-06

> plot(deliverTimeModel)









**Hence the P-value is less than 0.05.**

**So X varibale is significance and also Multiple R-Square value is 0.6823.**

**That’s mean this model will predict the output 68.23% time correct**

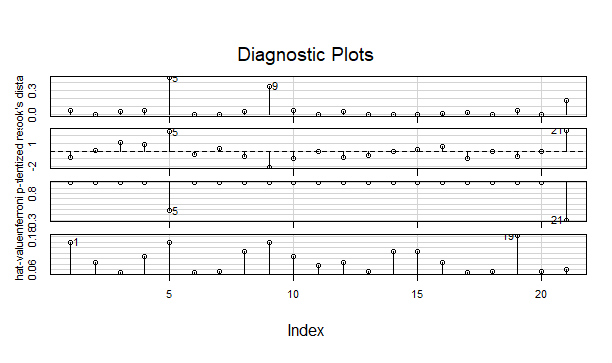
> library(mvinfluence)

Loading required package: car

Loading required package: carData

Loading required package: heplots

> influenceIndexPlot(deliverTimeModel)



> deliverTimeModel <- lm(Delivery.Time ~ Sorting.Time, data = delivery\_time[c(-5,-9,-21),])

> summary(deliverTimeModel)

Call:

lm(formula = Delivery.Time ~ Sorting.Time, data = delivery\_time[c(-5,

-9, -21), ])

Residuals:

Min 1Q Median 3Q Max

-2.3407 -1.5027 0.2275 0.9328 3.6815

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 6.0240 1.1751 5.126 0.000102 \*\*\*

Sorting.Time 1.6741 0.1872 8.941 1.27e-07 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

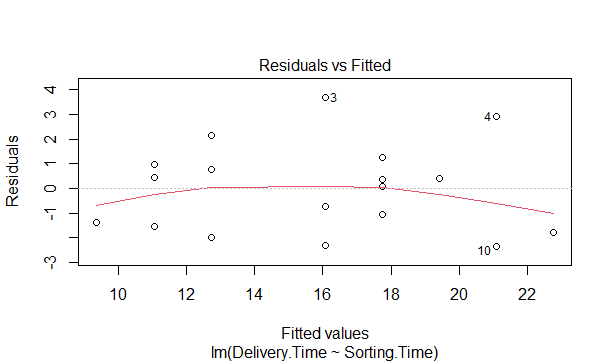
Residual standard error: 1.839 on 16 degrees of freedom

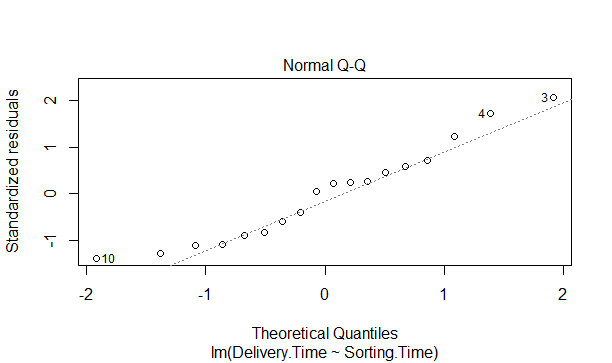
Multiple R-squared: 0.8332, Adjusted R-squared: 0.8228

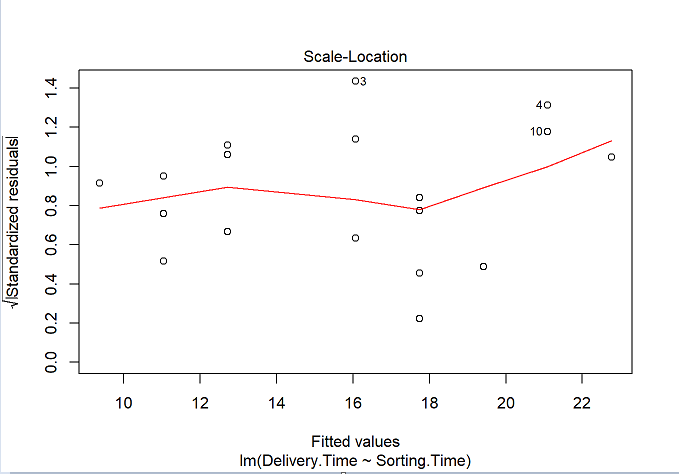
F-statistic: 79.94 on 1 and 16 DF, p-value: 1.273e-07

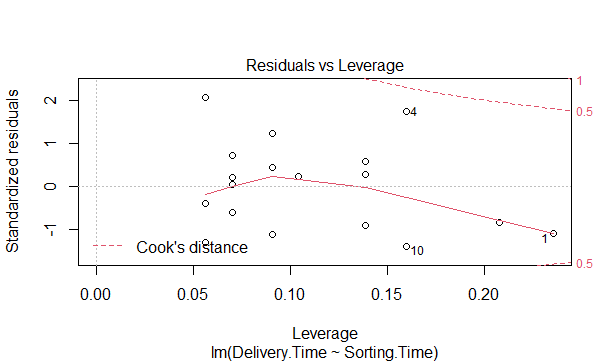
> plot(deliverTimeModel)

Hit <Return> to see next plot:









**After removing 3 points Multiple R-Square value is increased to 0.8332.**

**That’s mean this model will predict the output 83.32% time correct.**

### 3 - Emp\_data -> Build a prediction model for Churn\_out\_rate

**Reding the data from data file and saving into a variable**

#\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

> #\*3 - Emp\_data -> Build a prediction model for Churn\_out\_rate

> Emp\_data <- read.csv("E:/Data Science Asignments/Simple regression/emp\_data.csv")

> summary(Emp\_data)

Salary\_hike Churn\_out\_rate

Min. :1580 Min. :60.00

1st Qu.:1618 1st Qu.:65.75

Median :1675 Median :71.00

Mean :1689 Mean :72.90

3rd Qu.:1724 3rd Qu.:78.75

Max. :1870 Max. :92.00

> # Variance and Standard deviation of Salary\_hike column

> var(Emp\_data$Salary\_hike)

[1] 8481.822

> sd(Emp\_data$Salary\_hike)

[1] 92.09681

> # Variance and Standard deviation of Churn\_out\_rate column

> var(Emp\_data$Churn\_out\_rate)

[1] 105.2111

> sd(Emp\_data$Churn\_out\_rate)

[1] 10.25725

> #Creating Linear Model for Churn\_out\_rate

> Churn\_out\_rate\_Model <- lm(Churn\_out\_rate ~ Salary\_hike, data = Emp\_data)

> summary(Churn\_out\_rate\_Model)

Call:

lm(formula = Churn\_out\_rate ~ Salary\_hike, data = Emp\_data)

Residuals:

Min 1Q Median 3Q Max

-3.804 -3.059 -1.819 2.430 8.072

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 244.36491 27.35194 8.934 1.96e-05 \*\*\*

Salary\_hike -0.10154 0.01618 -6.277 0.000239 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.469 on 8 degrees of freedom

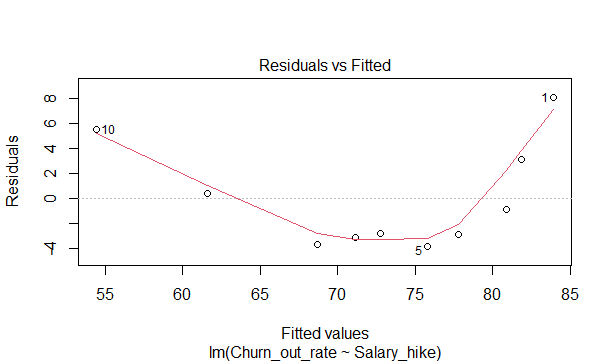
Multiple R-squared: 0.8312, Adjusted R-squared: 0.8101

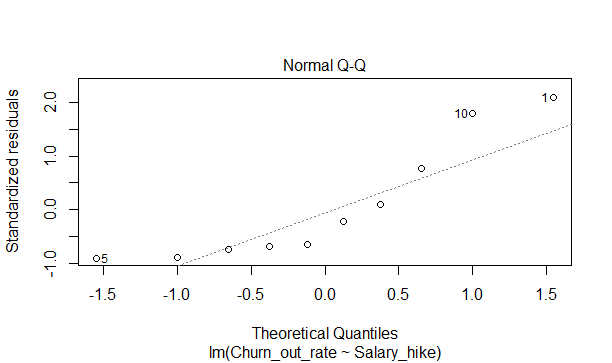
F-statistic: 39.4 on 1 and 8 DF, p-value: 0.0002386

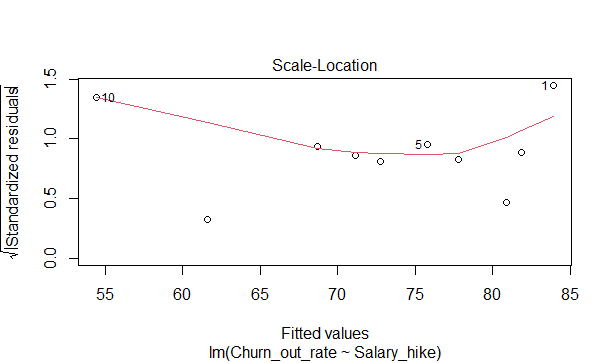
> plot(Churn\_out\_rate\_Model)

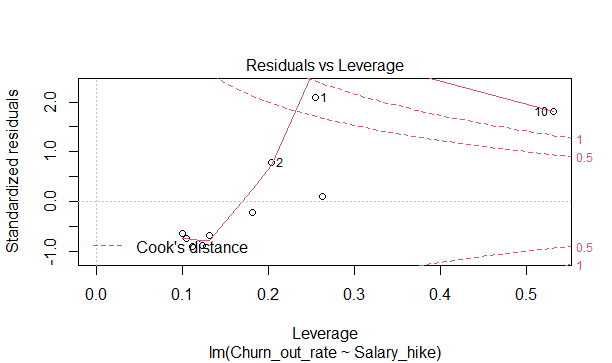
Hit <Return> to see next plot: plot(Churn\_out\_rate\_Model)

Hit <Return> to see next plot

****

****

****

****

**Hence the P-value is less than 0.05.**

**So X varibale is significance and also Multiple R-Square value is 0.8312**

**That’s mean this model will predict the output 83.12% time correct**

### 4 - Salary\_hike -> Build a prediction model for Salary\_hike

**Reding the data from data file and saving into a variable**

> Salary\_hike <- read.csv("E:/Data Science Asignments/Simple regression/Salary\_Data.csv")

> summary(Salary\_hike)

YearsExperience Salary

Min. : 1.100 Min. : 37731

1st Qu.: 3.200 1st Qu.: 56721

Median : 4.700 Median : 65237

Mean : 5.313 Mean : 76003

3rd Qu.: 7.700 3rd Qu.:100545

Max. :10.500 Max. :122391

> # Variance and Standard deviation of Salary\_hike column

> var(Salary\_hike$YearsExperience)

[1] 8.053609

> sd(Salary\_hike$YearsExperience)

[1] 2.837888

> # Variance and Standard deviation of Churn\_out\_rate column

> var(Salary\_hike$Salary)

[1] 751550960

> sd(Salary\_hike$Salary)

[1] 27414.43

> #Creating Linear Model for Salary\_hike

> Salary\_hike\_Model <- lm(Salary ~ YearsExperience, data = Salary\_hike)

> summary(Salary\_hike\_Model)

Call:

lm(formula = Salary ~ YearsExperience, data = Salary\_hike)

Residuals:

Min 1Q Median 3Q Max

-7958.0 -4088.5 -459.9 3372.6 11448.0

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 25792.2 2273.1 11.35 5.51e-12 \*\*\*

YearsExperience 9450.0 378.8 24.95 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5788 on 28 degrees of freedom

Multiple R-squared: 0.957, Adjusted R-squared: 0.9554

F-statistic: 622.5 on 1 and 28 DF, p-value: < 2.2e-16

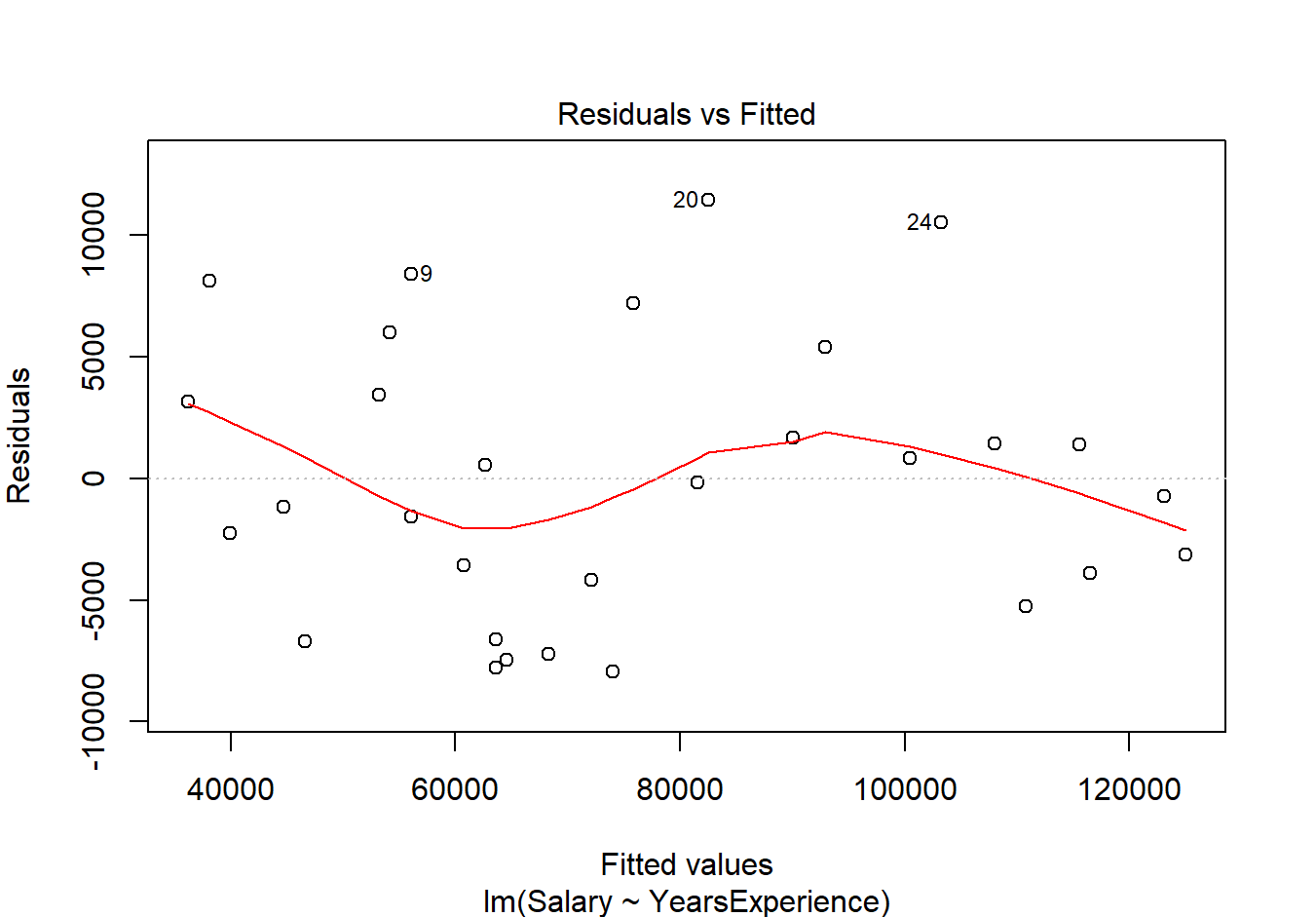
> plot(Salary\_hike\_Model)

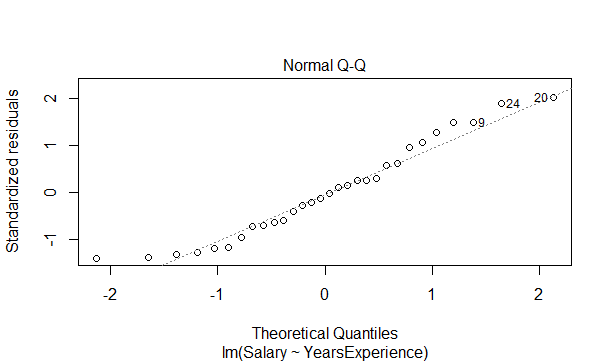
Hit <Return> to see next plot: plot(Salary\_hike\_Model)

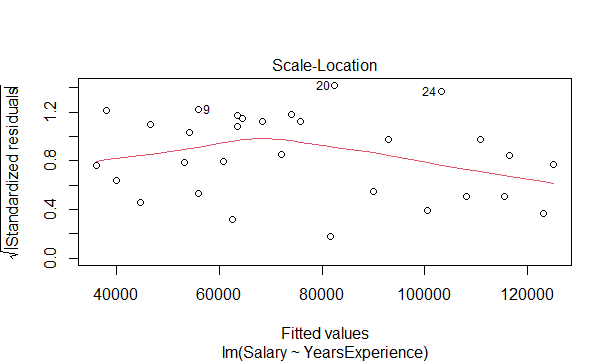
Error: Unable to establish connection with R session

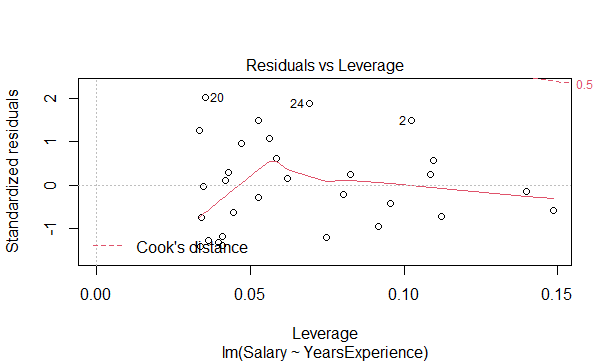
Hit <Return> to see next plot:

Hit <Return> to see next plot:









**Hence the P-value is less than 0.05.**

**So X varibale is significance and also Multiple R-Square value is 0.957**

**That’s mean this model will predict the output 95.7% time correct**