**Simple Linear Regression**

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

**Ans**:

> calories\_consumed <- read.csv("calories\_consumed.csv")

> summary(calories\_consumed)

Weight gained 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

> weight\_gain\_model <- lm(formula = calories\_consumed$`Weight gained (grams)` ~ calories\_consumed$`Calories Consumed`, calories\_consumed)

> summary(weight\_gain\_model)

Call:

lm(formula = calories\_consumed$`Weight gained (grams)` ~ calories\_consumed$`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$`Calories Consumed` 0.42016 0.04115 10.211 2.86e-07

(Intercept) \*\*\*

calories\_consumed$`Calories Consumed` \*\*\*

---

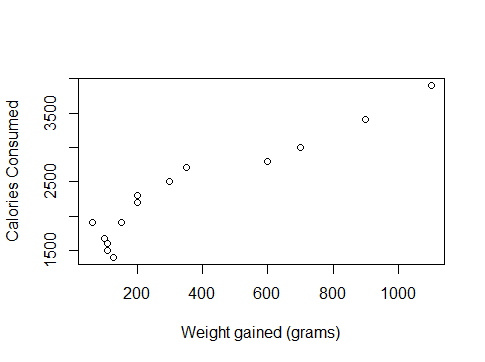
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

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)



**Conclusion**: As P value is 2.856e-07, which is <0.05, x variable is significant, also the R-squared value is 0.8968, which is >0.85 it shows strong linear relationship. Which means model predicts 89.6% output correct.

1. **Delivery\_time -> Predict delivery time using sorting time**

> delivery\_time <- read.csv("delivery\_time.csv")

> 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

> delivery\_time\_model <- lm(delivery\_time$`Delivery Time` ~ delivery\_time$`Sorting Time`, delivery\_time)

> summary(delivery\_time\_model)

Call:

lm(formula = delivery\_time$`Delivery Time` ~ delivery\_time$`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 \*\*

delivery\_time$`Sorting Time` 1.6490 0.2582 6.387 3.98e-06 \*\*\*

---

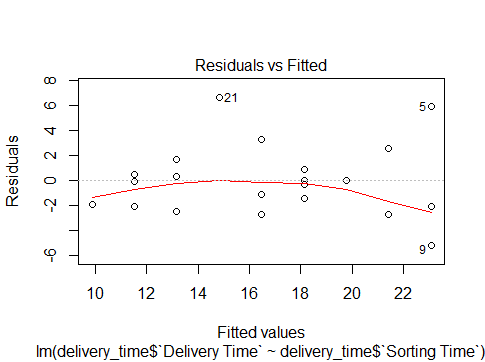
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

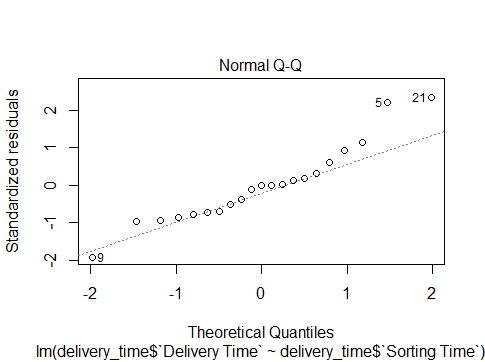
Residual standard error: 2.935 on 19 degrees of freedom

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

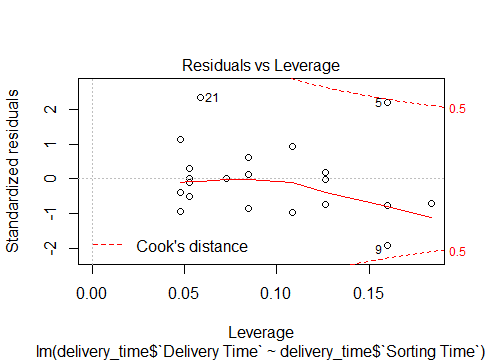
F-statistic: 40.8 on 1 and 19 DF, p-value: 3.983e-06

> plot(delivery\_time\_model)









p-value: 3.983e-06(significant), Multiple R-squared: 0.6823(Insignificant)

**Observation**:

By observing above model plots, the data points 5,9,21 are outliers. So further improve the model predictions capability we will discard those data points from the model and test it again.

> delivery\_time\_model1 <- lm(delivery\_time$`Delivery Time` ~ delivery\_time$`Sorting Time` , data = delivery\_time[c(-5,-9,-21),])

> summary(delivery\_time\_model1)

Call:

lm(formula = delivery\_time$`Delivery Time` ~ delivery\_time$`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 \*\*\*

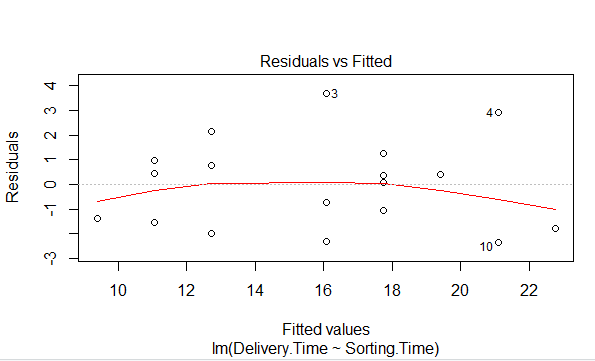
---

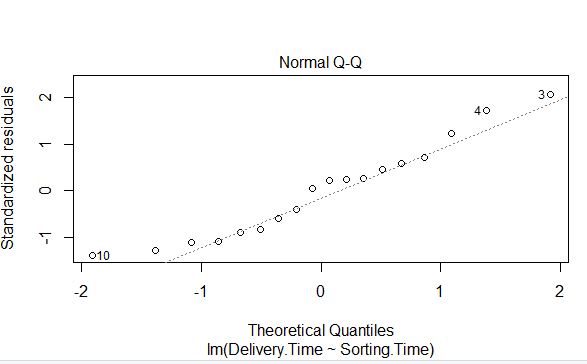
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

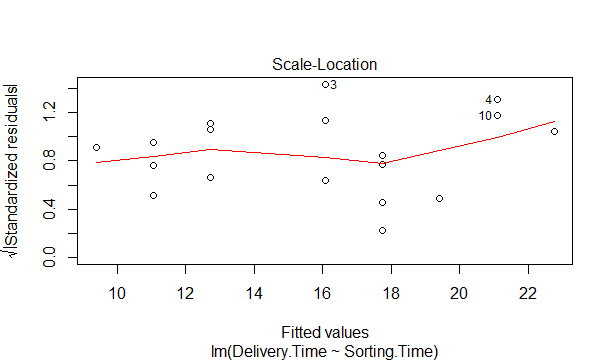
Residual standard error: 1.839 on 16 degrees of freedom

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

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







**Conclusion**: As P value is 1.273e-07, which is <0.05, x variable is significant, also the R-squared value is 0.8332, which is >0.85 it shows strong linear relationship. Which means model predicts 83.32% output correct.

1. **Emp\_data -> Build a prediction model for Churn\_out\_rate**

>emp\_data <- read\_csv("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

> emp\_data\_model <- lm(emp\_data$Churn\_out\_rate ~ emp\_data$Salary\_hike)

> summary(emp\_data\_model)

Call:

lm(formula = emp\_data$Churn\_out\_rate ~ emp\_data$Salary\_hike)

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 \*\*\*

emp\_data$Salary\_hike -0.10154 0.01618 -6.277 0.000239 \*\*\*

---

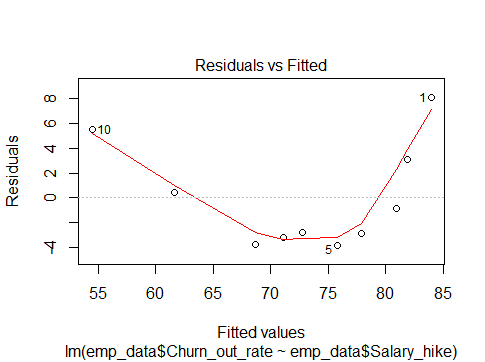
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

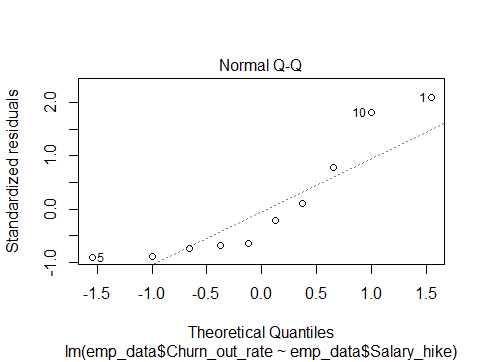
Residual standard error: 4.469 on 8 degrees of freedom

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

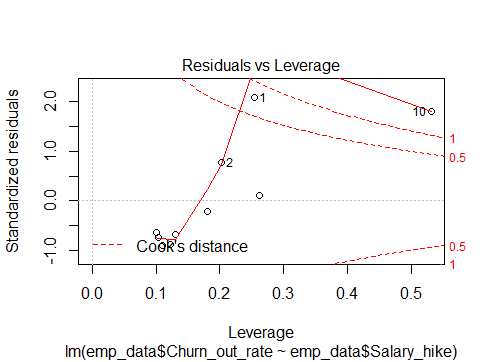
F-statistic: 39.4 on 1 and 8 DF, p-value: 0.0002386

> plot(emp\_data\_model)









**Conclusion**: As P value is 0.0002386, which is <0.05, x variable is significant, also the R-squared value is 0.8312, which is >0.85 it shows strong linear relationship. Which means model predicts 83.12% output correct.

1. **Salary\_hike -> Build a prediction model for Salary\_hike**

>Salary\_Data <- read\_csv("Salary\_Data.csv")

> summary(Salary\_Data)

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

> Salary\_Data\_model <- lm(Salary\_Data$Salary ~ Salary\_Data$YearsExperience, Salary\_Data)

> summary(Salary\_Data\_model)

Call:

lm(formula = Salary\_Data$Salary ~ Salary\_Data$YearsExperience,

data = Salary\_Data)

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 \*\*\*

Salary\_Data$YearsExperience 9450.0 378.8 24.95 < 2e-16 \*\*\*

---

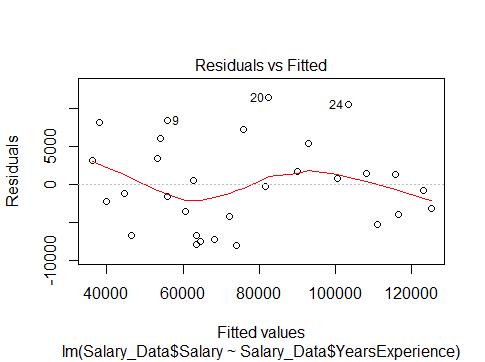
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

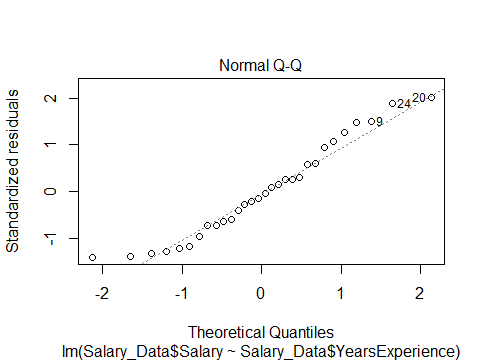
Residual standard error: 5788 on 28 degrees of freedom

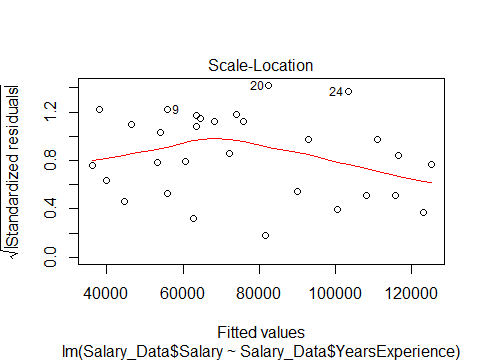
Multiple R-squared: 0.957, Adjusted R-squared: 0.9554

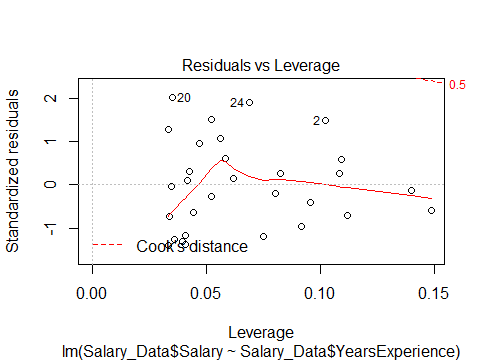
F-statistic: 622.5 on 1 and 28 DF, p-value: < 2.2e-16

> plot(Salary\_Data\_model)









**Conclusion**: As P value is 2.2e-16, which is <0.05, x variable is significant, also the R-squared value is 0.957, which is >0.85 it shows strong linear relationship. Which means model predicts 95.7% output correct.