**EXPERIMENT – 8**

**Aim:** Exercises to draw a scatter diagram, residual plots, outliers leverage and influential data points in R

**Theory and Technique:**

**Scatter Plot** - A scatter plot is a set of dotted points representing individual data pieces on the horizontal and vertical axis. In a graph in which the values of two variables are plotted along the X-axis and Y-axis, the pattern of the resulting points reveals a correlation between them.

**Scatter plot in R Programming Language using the plot() function.**

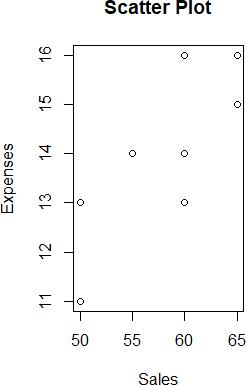
Syntax: plot(x, y, main, xlab, ylab, xlim, ylim, axes)

Code:

x = c(50,50,55,60,65,65,65,60,60,50)

y = c(11,13,14,16,16,15,15,14,13,13)

plot(x, y, main="Scatter Plot", xlab="Sales", ylab="Expenses")



**Residual plots** are often used to assess whether or not the residuals in regression analysis are normally distributed and whether or not they exhibit heteroscedasticity.

Code:

x = c(6,7,7,8,10,10,11,12,14,15,16)

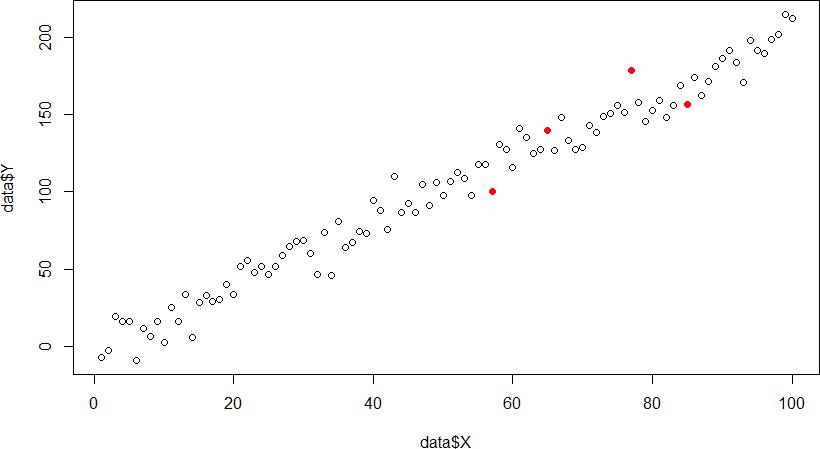
y = c(55,40,50,41,35,28,38,32,28,18,13)

mod = lm(y~x)

summary(mod)

plot(x, y, main="Size of Data Vs Requests", xlab="Gigabytes", ylab="Processed Requests", pch=16, col="blue")

abline(a=70.16, -3.39, col="red");



**Outliers**: Outliers are the points that are distinct and deviant from the bulk of the dataset. In general, the outliers have **high residual** values means that the difference is greater than the b/w observed and predicted value.

Code:

data <- data.frame(x,y)

plot(data$x, data$y)

* Example: Detecting outliers
* Identify observations with high residuals

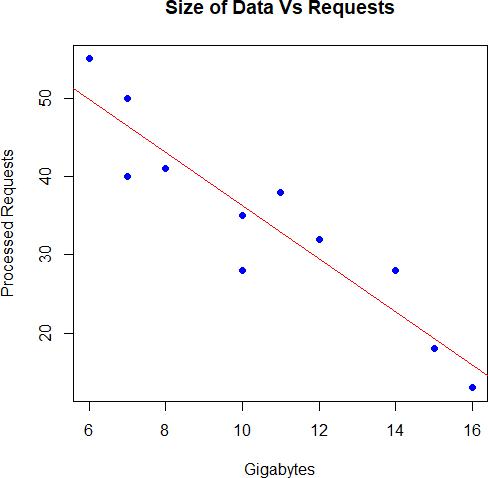
outliers <- which(abs(resid(mod)) > 2 \* sd(resid(mod))) X <- 1:100

Y <- 2 \* X + rnorm(100, mean = 0, sd = 10) model <- lm(Y ~ X, data = data)

data <- data.frame(X = 1:100, Y = 2 \* X + rnorm(100, mean = 0, sd = 10)) outliers <-which(abs(resid(model)) > 2 \* sd(resid(model)))

plot(data$X, data$Y)

points(data$X[outliers], data$Y[outliers], col = "red", pch = 19)



**Influential Points:**

An influential point is a point that has a large impact on the regression. Surprisingly, these are not the same thing. A point can be an outlier without being influential. A point can be influential without being an outlier. A point can be both or neither

Code:

influential <- cooks.distance(mode threshold <- 3 / length(data$X)

influential\_obs <- which(influential > threshold)

* Highlight influential observations in the scatterplot plot(data$X, data$Y) points(data$X[influential\_obs], data$Y[influential\_obs], col = "orange", pch = 19)

