Linear Regression and Correlation using RStudio.

These instructions assume that you've already installed the necessary package and that you've imported and attached your data file.

To perform linear regression in Rstudio, you will use the following R command:

plot(x, y, main="title", xlab="x label", ylab="y label")

- x = Variable that you want on the x axis. This is the independent variable.
- y = Variable that you want on the y axis. This is the dependent variable.
- main = Title for the Scatterplot
- xlab = Label for the x axis.
- ylab = Label for the y axis.

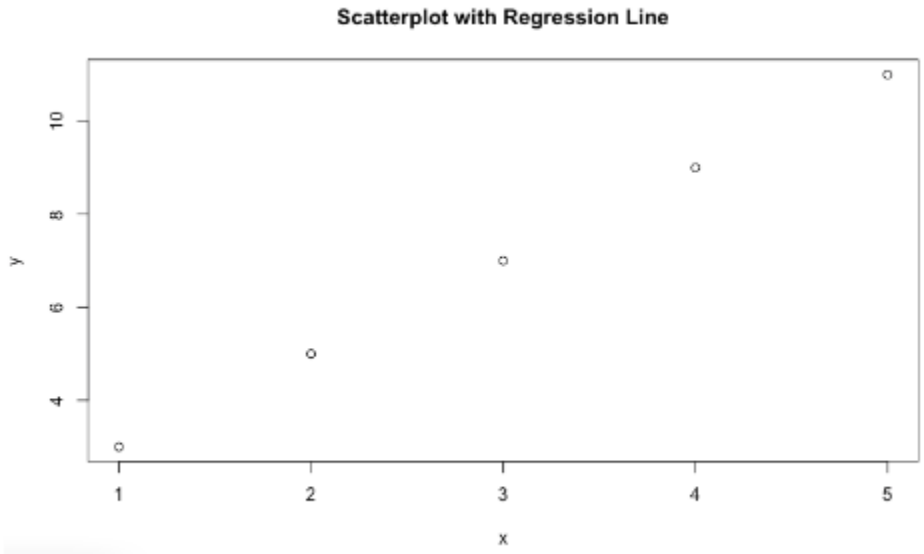
Let's do a simple example. Input the following arrays of data into R:

x <- c(1, 2, 3, 4, 5)
y <- c(3, 5, 7, 9, 11)

Create the scatterplot for this array using the following command:

plot(x, y, main = "Scatterplot with Regression Line", xlab = "x", ylab = "y")

The following plot will be created.

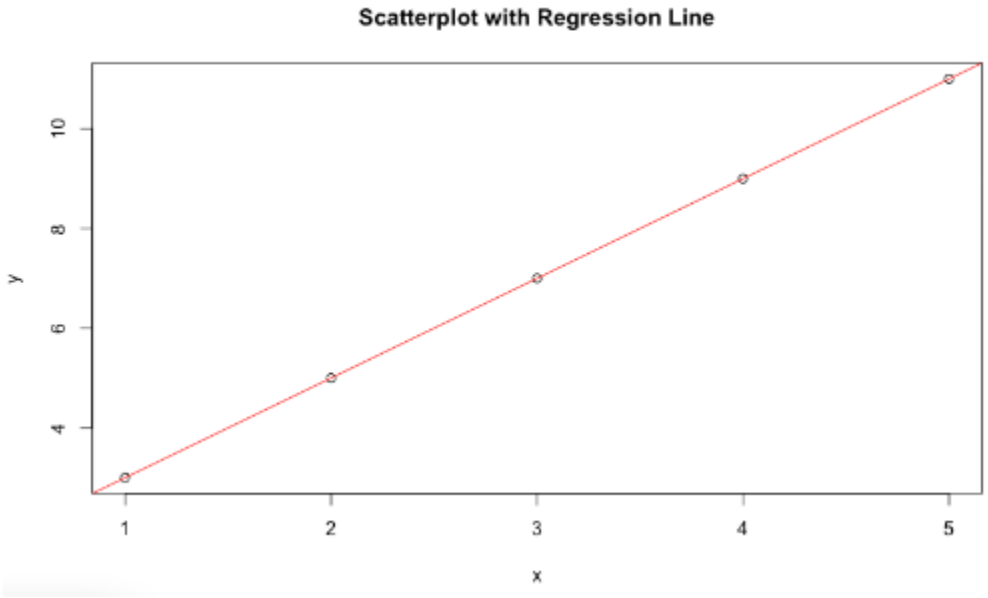


Now, let apply a linear regression model to this plot using the following command:

model <- lm(y ~ x)

The following command will draw a red regression line through the data points.

abline(model, col = "red")



To find the linear regression equation, you will need to extract the coefficients from R using the following commands:

coefficients<-coef(model)
print(coefficients)

Note - This print command simply means that the output will be in the terminal window.

Here's what the output will look like:

(Intercept) x
1 2

Recall that the equation of a line is in the form of $y = mx + b$. From the output, the slope is $m=2$ and the y-intercept is $b=1$. That means our equation will be $y = 2x + 1$.

Finally, we are always interested in the goodness of fit information such as the correlation coefficient, Pearson's r, and the coefficient of determination r^2.

The following commands are used to find correlation coefficient which ranges from -1 to 1 and is a representation of the strength of the relationship between the variables:

```
cor_coef <- cor(x, y)
print(cor_coef)
```

The commands below are used to find the coefficient of determination r^2 which can be interpreted as the percentage to which the variable x explains the variability in variable y.

```
summary_data <- summary(model)
r_squared <- summary_data$r.squared
print(r_squared)
```

Finally, to find Pearson's r, use the following commands:

```
cor_coef <- cor(x, y, method = "pearson")
print(cor_coef)
```