**AIR UNIVERSITY, ISLAMABAD**



Department of Computer

     Science

FACULTYOF COMPUTING AND ARTIFICIAL INTELLIGENCE

**Data Science Lab: 09**

**Topic: Linear Regression**

**Lab Instructor: Mr. Muhammad Imran**

**Instructions:**

#### Plagiarism: Plagiarism cases will be dealt with strictly. If found plagiarized, both the involved parties will be awarded zero marks in this assignment, all of the remaining assignments, or even an F grade in the course. Copying from the internet is the easiest way to get caught!

**Deadline:** Late submission with marks deduction will be accepted according to the course policy shared earlier. Correct and timely submission of the assignment is the responsibility of every student; hence no relaxation will be given to anyone.

**Tip:** For timely completion of the assignment, start as early as possible. Furthermore, work smartly - as some of the problems can be solved using smarter logic.

1. Note: Follow the given instructions to the letter, failing to do so will result in a zero.

**Objectives:**

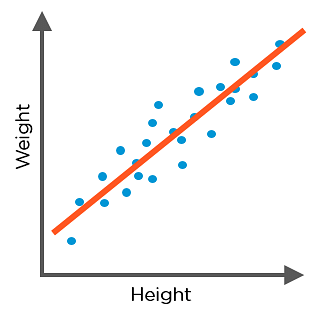
#### In this lab, you will:

Understand and apply the linear regression on a sample dataset in R studio.

**What is Linear Regression?**

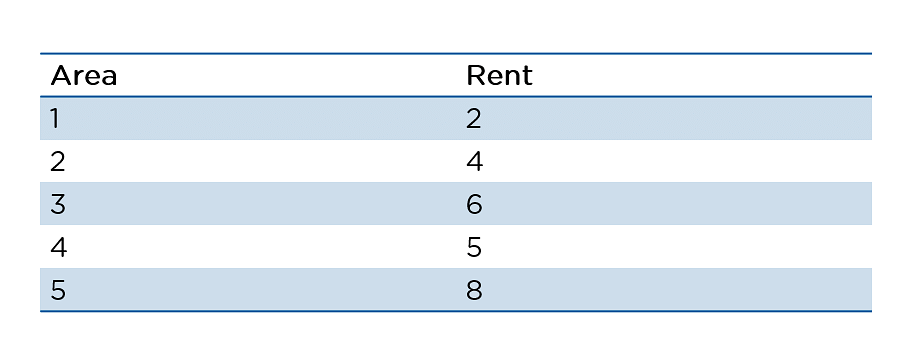
Linear regression is a form of statistical analysis that shows the relationship between two or more continuous variables. It creates a predictive model using relevant data to show trends.

Below is a graph that depicts the relationship between the heights and weights of a sample of people. The red line is the linear regression that shows the height of a person is positively related to its weight.



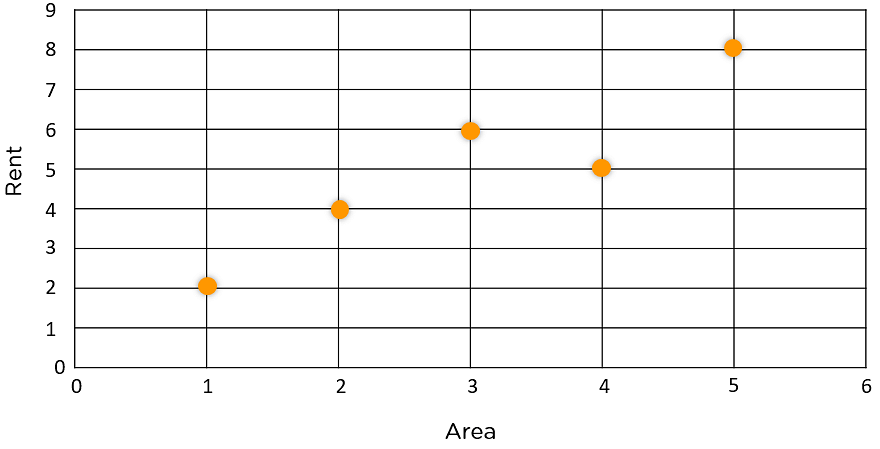
## How Does Linear Regression Work?

We can better understand how linear regression works by using the example of a dataset that contains two fields, Area and Rent, and is used to predict the house’s rent based on the area where it is located. The dataset is:

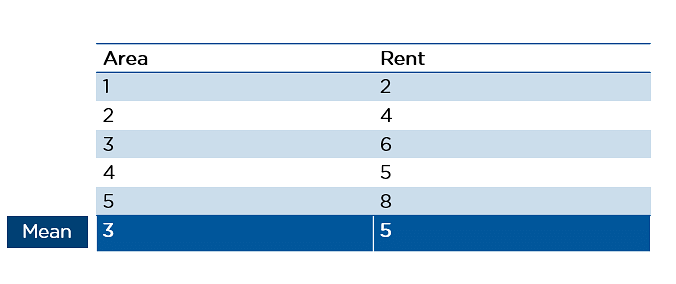


As you can see, we are using a simple dataset for our example. Using this uncomplicated data, let’s have a look at how linear regression works, step by step:

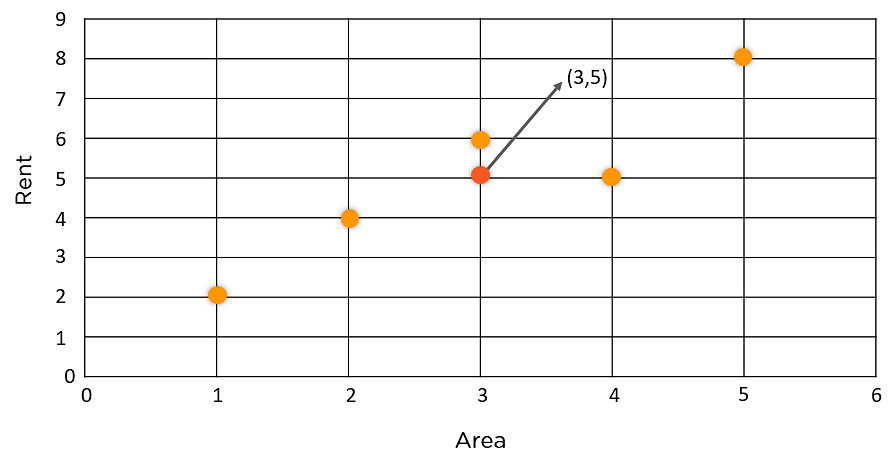
1. With the available data, we plot a graph with Area in the X-axis and Rent on Y-axis. The graph will look like the following. Notice that it is a linear pattern with a slight dip.



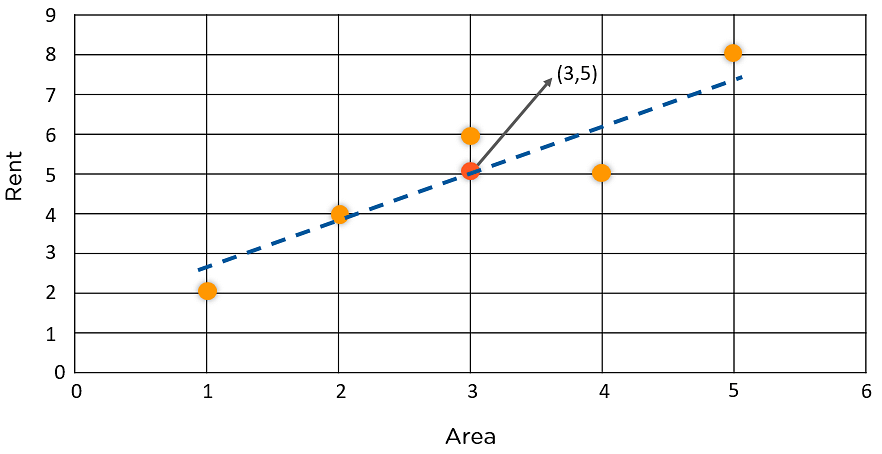
2. Next, we find the mean of Area and Rent.



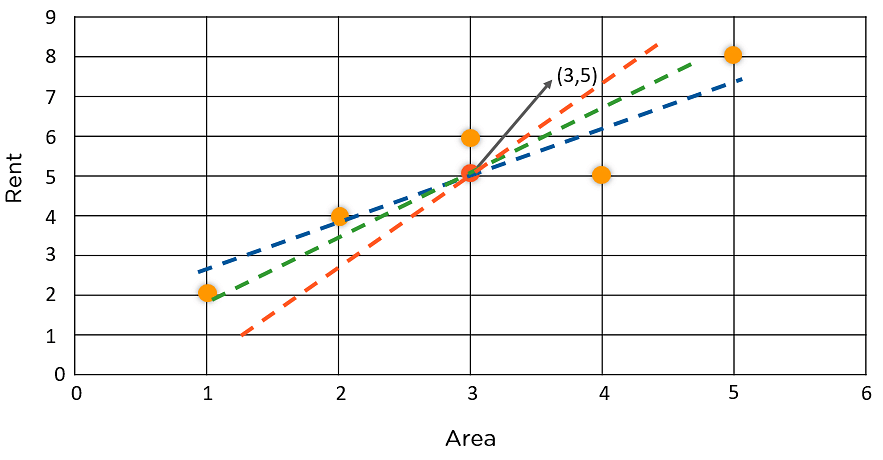
3. We then plot the mean on the graph.



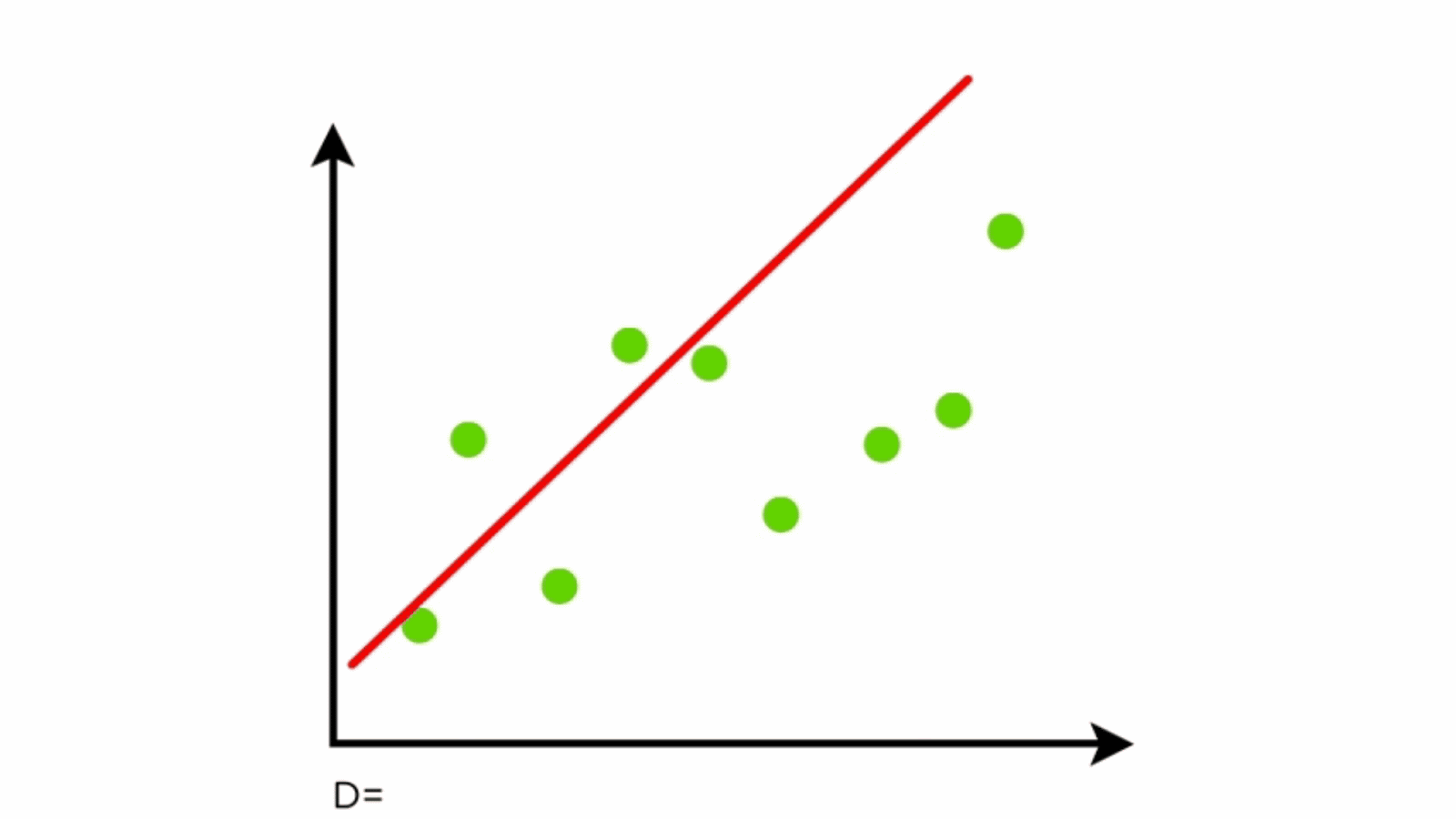
4. We draw a line of best fit that passes through the mean.



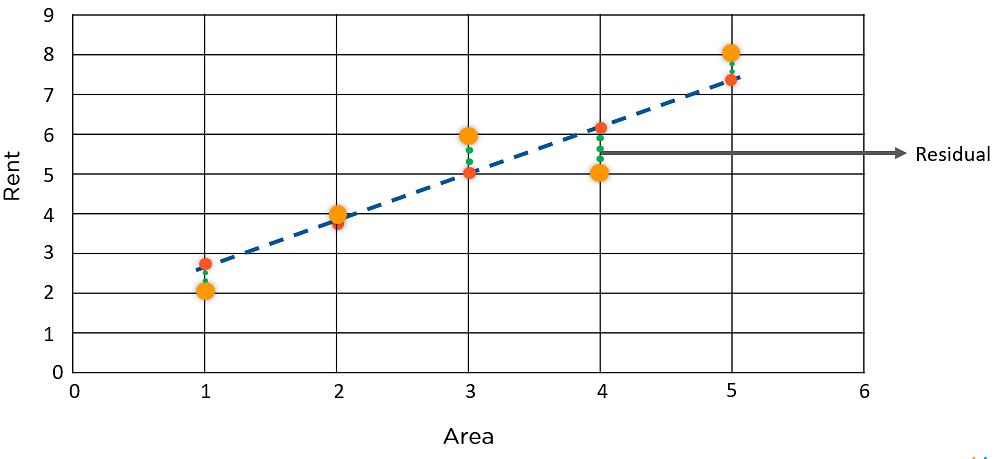
5. But we encounter a problem. As you can see below, multiple lines can be drawn through the mean:



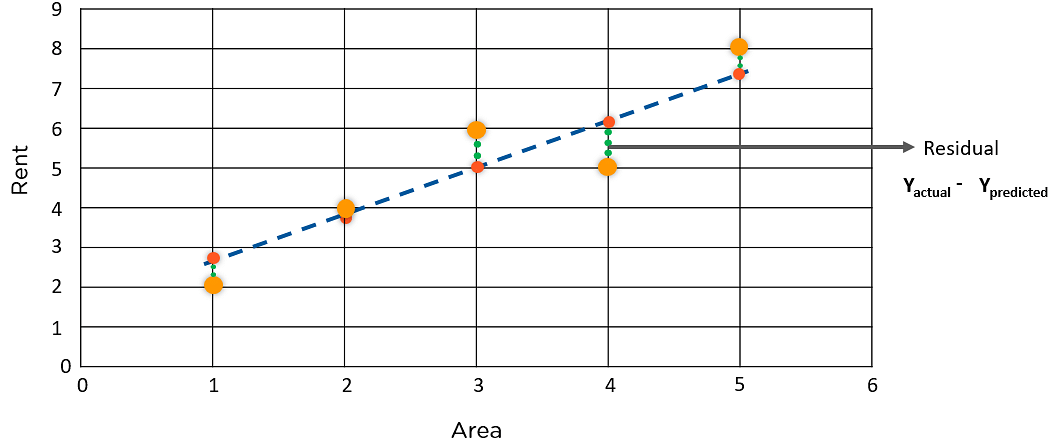
6. To overcome this problem, we keep moving the line to make sure the best fit line has the least square distance from the data points



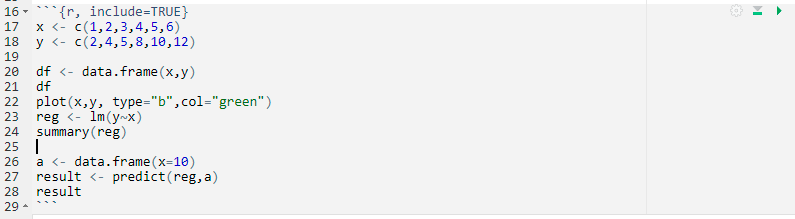
7. The least-square distance is found by adding the square of the residuals

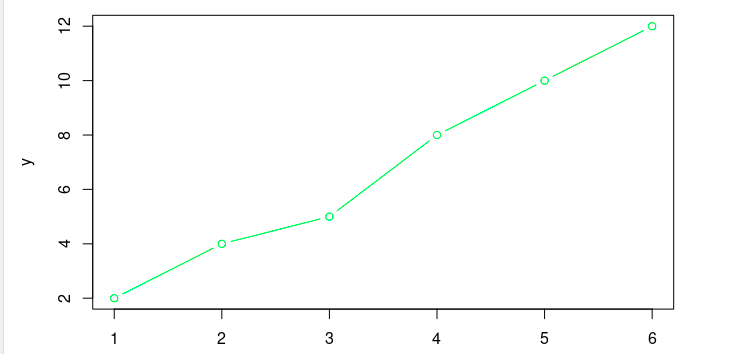


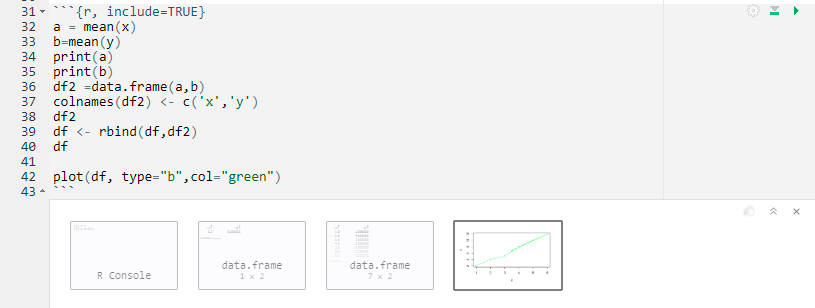
8. We now arrive at the relation that, Residual is the distance between Y-actual and Y-pred.

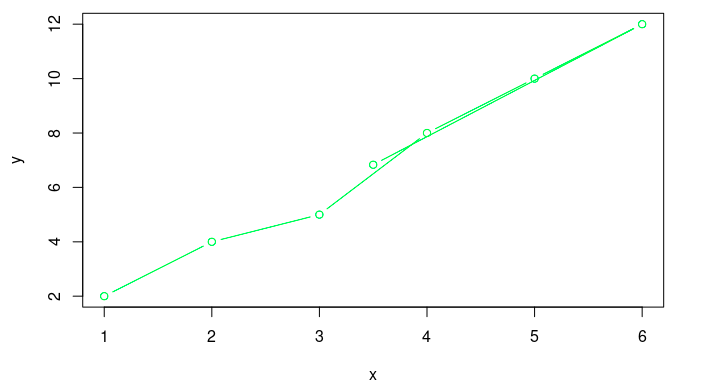


**Example 01:**









**Example 02:**

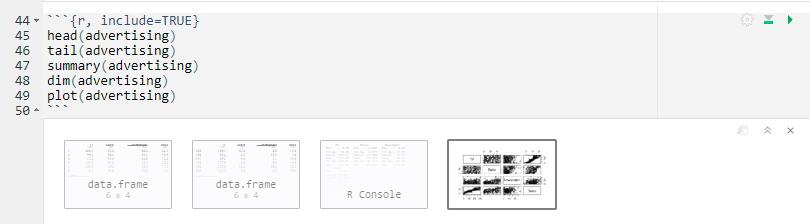
Now we will now look at a real-life scenario where we will predict the sales by using regression analysis in R. The sample dataset is given on GCR.

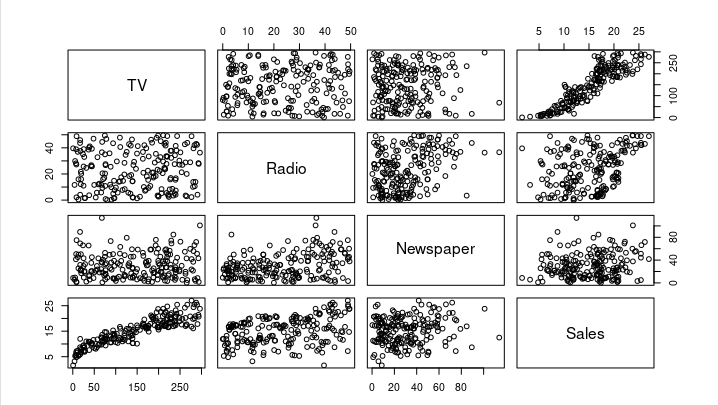
So let’s start linear regression step-by-step. Since we will perform linear regression in RStudio, we will open that first. We type the following code in R:

**Step 01:**

* Import the dataset. Displays the top 6 rows of a dataset
* Gives certain statistical information about the data.
* Plot the variables to see their trends

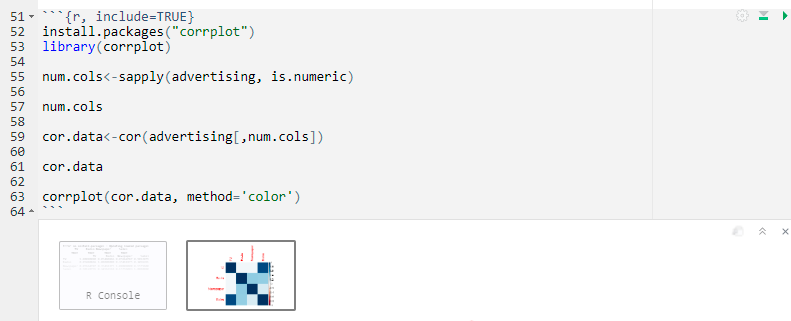
The output will look like below:

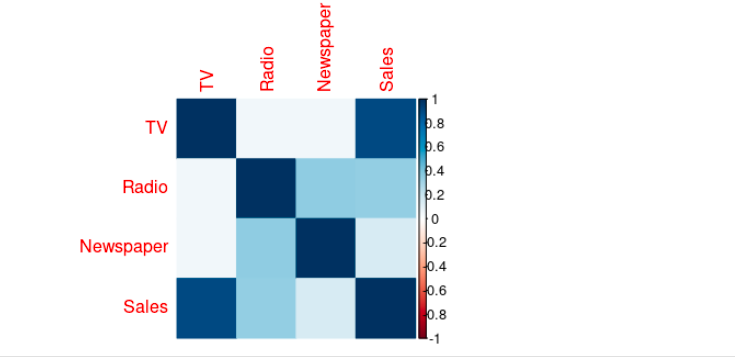




**Step 02:**

Let’s now see how the variables are correlated to each other. For that, we’ll take only the numeric column values.

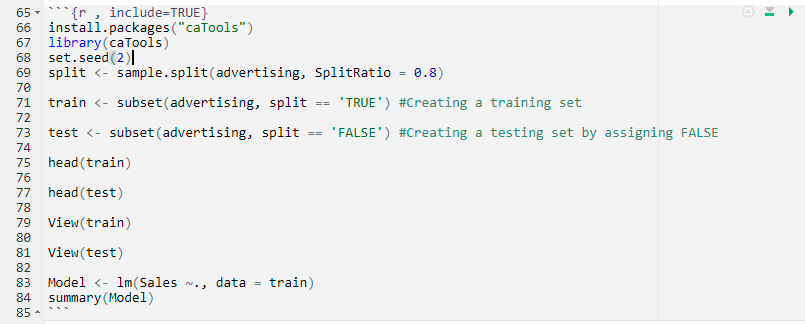




As you can see from the above correlation matrix, the variables have a high degree of correlation between each other and with the sales variable.

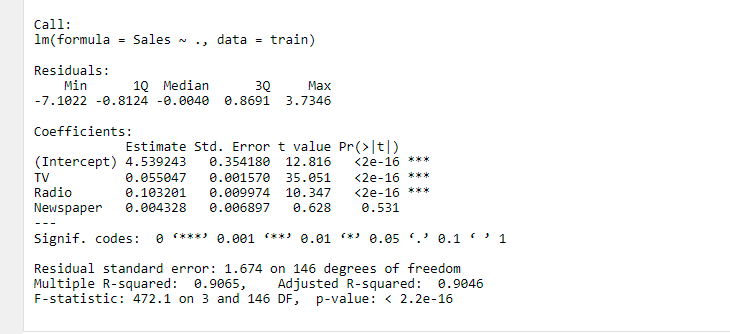
**Step 03:**

Let’s now split the data from training and testing sets.



**Step 04:**

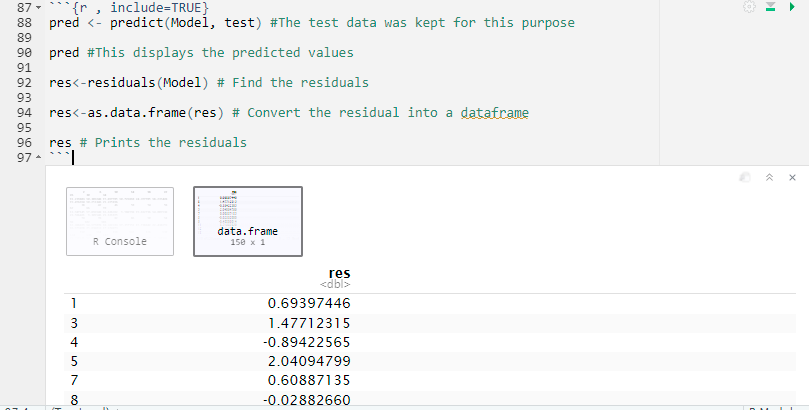
Now that we have the test and train variables, let’s go ahead and create the model:



**Step 05:**

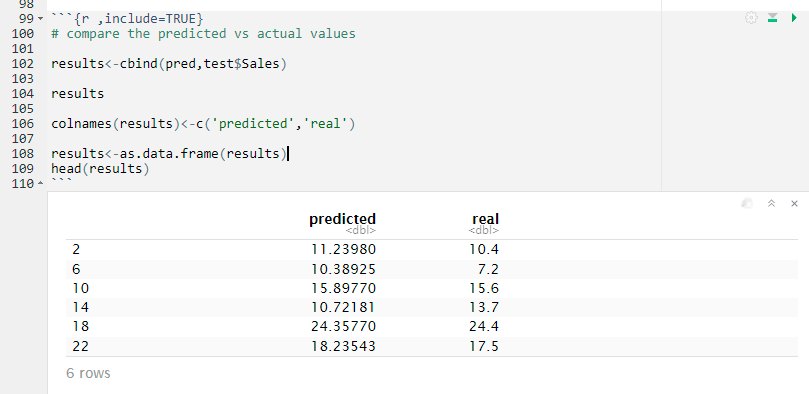
Prediction on Test Data

* Displays the predicted values by predicting values
* Find the residuals
* Convert the residual into a data frame



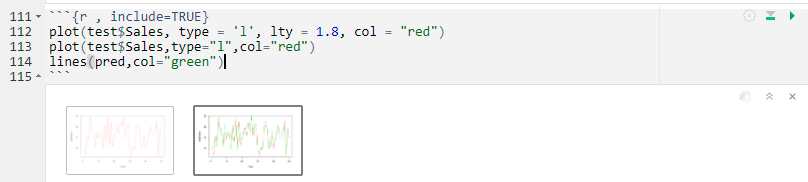
**Step 06:**

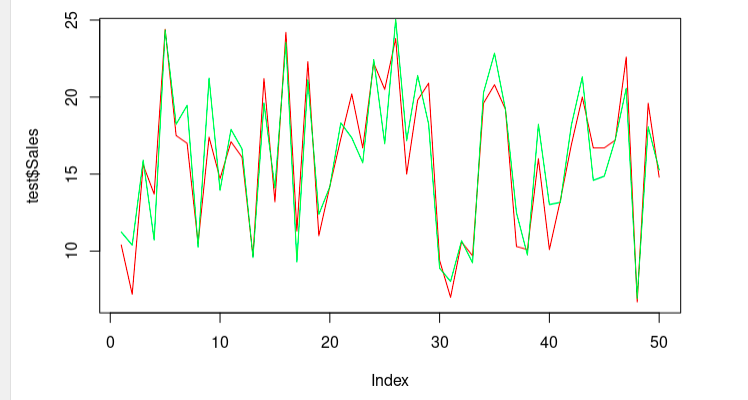
Compare the predicted vs actual values



**Step 07:**

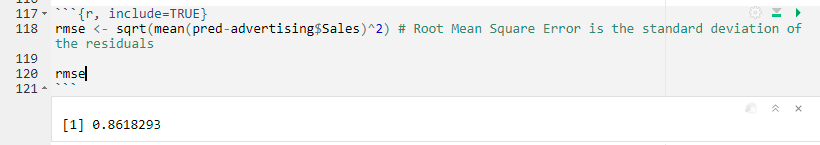
 Let’s now, compare the predicted vs actual values





**Step 08:**

From the above output, we can see that the graphs of the predicted revenue and expected revenue are very close. Let’s check out the accuracy so we can validate the comparison.



You learned about the various commands, packages and saw how to plot a graph in RStudio. Although this is a good start, there is still so much more to discover about linear regression.