8. Data Science – Machine Learning – R value

Contents

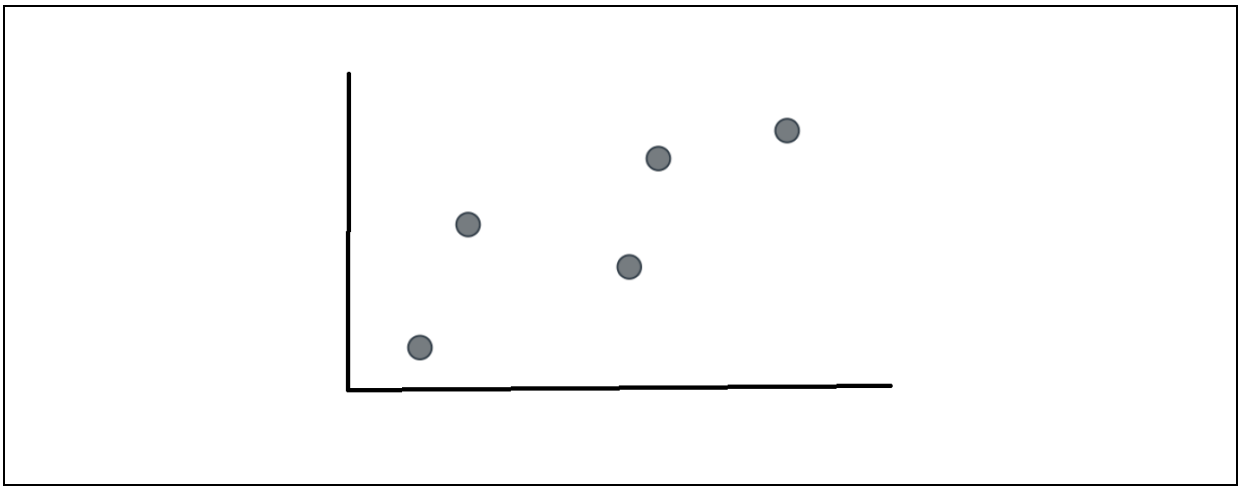
8. Data Science – Machine Learning – R value

1. Regression

- ✓ By using regression we can find the relationship between variables.
- ✓ Once we understand the relationship in between the variables then we can predict the future outcomes

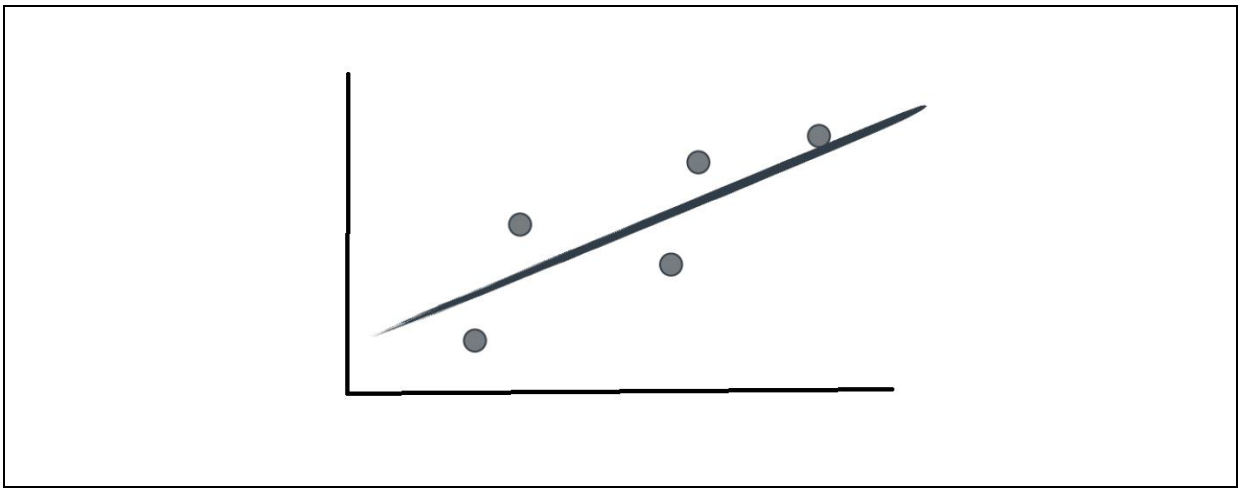
2. A line

- ✓ If two variables having relationship then if we draw this relationship in a two dimensional then we get a straight line.
- ✓ The picture of linear regression is simple.
- ✓ Let us say we have some points, a line will travel in between these points



3. The goal

- ✓ The goal of linear regression is to draw the best fitted line.
- ✓ Best fitted line means that the line which passes as close as possible to these points.



4. Can we use regression in everywhere?

- ✓ If there is a relationship in between the variables then only we can use linear regression algorithm.
- ✓ If there is no relationship in between the variables then we cannot use linear regression algorithm.

5. r value

- ✓ r value explains about how variables are related each other.
- ✓ This is very important step to recognise relationship in between values of x-axis and y-axis.

6. r value range

- ✓ The r values range, from -1 to 1.
- ✓ While calculating if we get **r** value as near to **-1** or **1** then those variables are **strongly related** each other.
- ✓ While calculating if we get **r** value as **0** then it confirms that there is **no relationship** in between the variables.

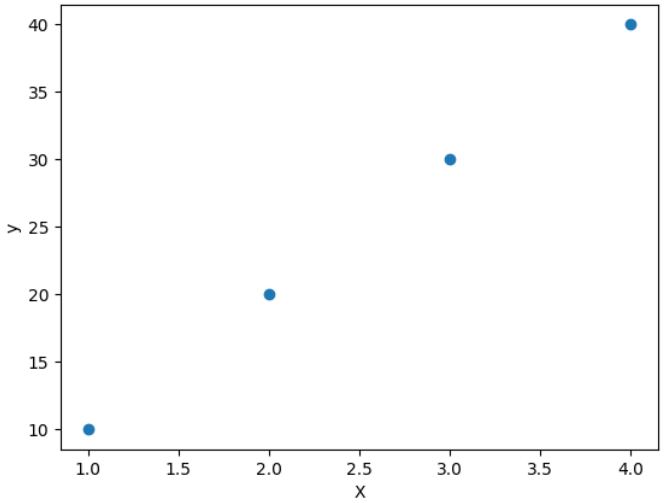
7. Calculate r value

- ✓ By using python scipy module we can calculate r value easily

Program Name: Plotting x and y values
demo1.py

```
import pandas as pd
from scipy import stats
import matplotlib.pyplot as plt

d = {
        "area": [1, 2, 3, 4],
        "rice_yield": [10, 20, 30, 40]
}

df = pd.DataFrame(d)

X = df.area.values
y = df.rice_yield.values

plt.xlabel("X")
plt.ylabel("y")

plt.scatter(X, y)
plt.show()
```

Output


40
35
30
25
20
15
10
y
1.0
1.5
2.0
2.5
3.0
3.5
4.0
X

Program Name: Calculating r value
demo2.py

```
import pandas as pd
from scipy import stats

d = {
        "area": [1, 2, 3, 4],
        "rice_yield": [10, 20, 30, 40]
}

df = pd.DataFrame(d)

X = df.area.values
y = df.rice_yield.values

slope, intercept, r_value, p_value, std_err = stats.linregress(X, y)

print(r_value)
```

Output

```
1.0
```

Note:

- ✓ The result 1.0
- ✓ So we can confirm that there is strong a relationship.
- ✓ So we can apply linear regression algorithm on top of these variables for future prediction

Program Name: Calculating r value
demo3.py

```
import pandas as pd
from scipy import stats

d = {
        "a": [600, 3000, 2, 3600, 4],
        "b": [550000, 565000, 610000, 680000, 725000]
}

df = pd.DataFrame(d)

X = df.a.values
y = df.b.values

slope, intercept, r_value, p_value, std_err = stats.linregress(X, y)

print(r_value)
```

Output

```
-0.06533879637370224
```

Note:

- ✓ The result -0.06533879637370224 this value is very near to 0
- ✓ So we can confirm that there is no relationship.
- ✓ So we cannot apply linear regression algorithm on top of these variables, even if we apply then we will get wrong prediction results