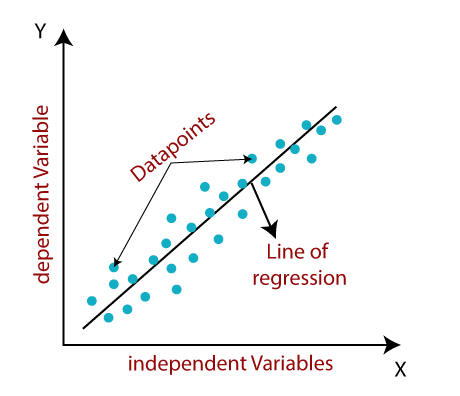
Galton was interested in understanding the relationship between various physical characteristics in humans, such as height, weight, and head size. To do this, he developed a method of measuring the strength of the relationship between two variables by calculating the correlation coefficient.

Galton also developed the method of least squares, which is used to find the line of best fit through a set of data points. This method forms the basis for linear regression, which is a statistical technique used to model the relationship between two variables by fitting a linear equation to the observed data.

While Galton is generally credited with the development of linear regression, it is worth noting that the method has evolved over time and has been refined by many other statisticians and mathematicians.



Independent variable = Cause

Dependant Variable = Effect

y=mx+c

y=input data

x= input data

m=slope

c=y intercept of line

m=∑(x-ẍ)(y-ẏ) / ∑(x- ẍ)2

x= 1,2,3,4,5

y=3,4,2,4,5

ẍ= mean of x=3

ẏ= mean of y=3.6

ẍ=3

=∑(x-ẍ)= (1-3)+(2-3)+(3-3)+(4-3)+(5-3)

| **x** | **y** | **x-ẍ(ẍ=3)** | **y-ẏ** | **(x-ẍ)(y-ẏ)** | **(x- ẍ)2** |
| --- | --- | --- | --- | --- | --- |
| 1 | 3 | -2 | -0.6 | 1.2 | 4 |
| 2 | 4 | -1 | 0.4 | -0.4 | 1 |
| 3 | 2 | 0 | -1.6 | 0 | 0 |
| 4 | 4 | 1 | 0.4 | 0.4 | 1 |
| 5 | 5 | 2 | 1.4 | 2.8 | 4 |
| ∑= |  |  |  | 4 | 10 |

m = 4/10 = 0.4 ; m=0.4

y= mx + c

3.6 = 0.4\*3+c c=y-mx

c=2.4

| x | y (predicted)( y= mx + c) |
| --- | --- |
| 1 | 2.8 (.4\*1+2.4) |
| 2 | 3.2 (.4\*2+2.4) |
| 3 | 3.6 (.4\*3+2.4) |
| 4 | 4 (.4\*4+2.4) |
| 5 | 4.4 (.4\*5+2.4) |
| 100 |  |

R^2= ∑(y p -ẏ)2 / ∑(y- ẏ) 2

ẏ=3.6

| x | y (predicted)( y= mx + c) | **y p -ẏ** | **(y p -ẏ)2** | **y- ẏ** | **(y- ẏ) 2** |
| --- | --- | --- | --- | --- | --- |
| 1 | 2.8 | -0.8 | 0.64 | -0.6 | 0.36 |
| 2 | 3.2 | -0.4 | 0.16 | 0.4 | 0.16 |
| 3 | 3.6 | 0 | 0 | -1.6 | 2.56 |
| 4 | 4 | 0.4 | 0.16 | 0.4 | 0.16 |
| 5 | 4.4 | 0.8 | 0.64 | 1.4 | 1.96 |
|  |  | ∑ | 1.6 |  | 5.2 |
|  |  |  |  |  |  |

R^2=1.6/5.2= 0.3076

R=0.5547

#plot

import matplotlib.pyplot as plt

from scipy import stats

x = [5,7,8,7,2,17,2,9,4,11,12,9,6]

y = [99,86,87,88,111,86,103,87,94,78,77,85,86]

slope, intercept, r, p, std\_err = stats.linregress(x, y)

def myfunc(x):

return slope \* x + intercept

mymodel = list(map(myfunc, x))

plt.scatter(x, y)

plt.plot(x, mymodel)

plt.show()

#pre

from scipy import stats

x = [5,7,8,7,2,17,2,9,4,11,12,9,6]

y = [99,86,87,88,111,86,103,87,94,78,77,85,86]

slope, intercept, r, p, std\_err = stats.linregress(x, y)

def myfunc(x):

return slope \* x + intercept

ypred = myfunc(20)

print(ypred)

#Bad Fit

import matplotlib.pyplot as plt

from scipy import stats

x = [89,43,36,36,95,10,66,34,38,20,26,29,48,64,6,5,36,66,72,40]

y = [21,46,3,35,67,95,53,72,58,10,26,34,90,33,38,20,56,2,47,15]

slope, intercept, r, p, std\_err = stats.linregress(x, y)

def myfunc(x):

return slope \* x + intercept

mymodel = list(map(myfunc, x))

plt.scatter(x, y)

plt.plot(x, mymodel)

plt.show()

**Task**

x= 5,7,8,7,2,17

y=99,86,87,88,111,86

Program

import matplotlib.pyplot as plt  
from scipy import stats  
  
x = [5,7,8,7,2,17,2,9,4,11,12,9,6]  
y = [99,86,87,88,111,86,103,87,94,78,77,85,86]  
  
slope, intercept, r, p, std\_err = stats.linregress(x, y)  
  
def myfunc(x):  
  return slope \* x + intercept  
  
mymodel = list(map(myfunc, x))  
  
plt.scatter(x, y)  
plt.plot(x, mymodel)  
plt.show()

**Predict speed of car**

from scipy import stats

x = [5,7,8,7,2,17,2,9,4,11,12,9,6]

y = [99,86,87,88,111,86,103,87,94,78,77,85,86]

slope, intercept, r, p, std\_err = stats.linregress(x, y)

def myfunc(x):

return slope \* x + intercept

speed = myfunc(10)

print(speed)

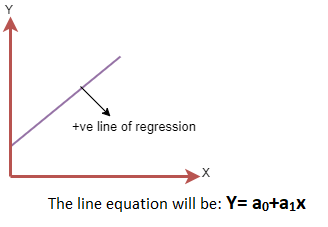
## Bad Fit?

Let us create an example where linear regression would not be the best method to predict future values.

Example

These values for the x- and y-axis should result in a very bad fit for linear regression:

import matplotlib.pyplot as plt  
from scipy import stats  
  
x = [89,43,36,36,95,10,66,34,38,20,26,29,48,64,6,5,36,66,72,40]  
y = [21,46,3,35,67,95,53,72,58,10,26,34,90,33,38,20,56,2,47,15]  
  
slope, intercept, r, p, std\_err = stats.linregress(x, y)  
  
def myfunc(x):  
  return slope \* x + intercept  
  
mymodel = list(map(myfunc, x))  
  
plt.scatter(x, y)  
plt.plot(x, mymodel)  
plt.show()



slope, intercept, r, p, std\_err = stats.linregress(x, y)

The return value is an object with the following attributes:

**slope**

Slope of the regression line.

**intercept**

Intercept of the regression line.

**Rvalue**

The Pearson correlation coefficient. The square of rvalue is equal to the coefficient of determination.

**pvalue**

A p-value **measures the probability of obtaining the observed results** The p-value for a hypothesis test whose null hypothesis is that the slope is zero, using Wald Test with t-distribution of the test statistic. See *alternative* above for alternative hypotheses.

**stderr**

Standard error of the estimated slope (gradient), under the assumption of residual normality.

**intercept\_stderr**

Standard error of the estimated intercept, under the assumption of residual normality.

the slope of a linear regression line can be negative. The slope represents the rate at which the output variable (dependent variable) changes with respect to a change in the input variable (independent variable).

If the slope is positive, it means that the output variable increases as the input variable increases. If the slope is negative, it means that the output variable decreases as the input variable increases.

In the case of a simple linear regression, where there is only one input variable, the slope is negative when the relationship between the input and output variables is inverse or negative, meaning that as the input variable increases, the output variable decreases. In this case, the slope of the regression line is negative.

In statistics, the p-value is a measure of the evidence against a null hypothesis. Specifically, it i**s the probability of obtaining a test statistic at least as extreme as the one observed, assuming the null hypothesis is true.**

In simpler terms, the p-value represents the probability that the results obtained in a statistical analysis occurred by chance, rather than as a result of the relationship being tested.

**To calculate the p-value, you first need to choose a statistical test that is appropriate for the type of data you are working with and the research question you are trying to answer.** The specific test used will determine how the p-value is calculated.

Once you have performed the test, you will obtain a test statistic (e.g., t-statistic, F-statistic, chi-square statistic, etc.) that measures the strength of the relationship between the variables being tested. The p-value is then calculated based on this test statistic and the null hypothesis being tested.

If the p-value is less than a predetermined significance level (often 0.05), it is typically interpreted as evidence against the null hypothesis, and the relationship being tested is considered statistically significant. If the p-value is greater than the significance level, the null hypothesis cannot be rejected, and the relationship being tested is not considered statistically significant.

It's important to note that the p-value is just one piece of evidence to consider when interpreting statistical results, and it should always be interpreted in the context of the research question and the available evidence