DECODING :

Simple Linear Regression

Formulae and Calculations

Nishigandha Sharma · Follow

Published in Analytics Vidhya

4 min read · Aug 10, 2020

We are all aware of the most simple equation in Statistics and Machine Learning model; the Linear Regression Equation. With this article, I aim to bring in clarity on

how the formula can be calculated by hand for the line equation. Here is the formula:

$y = mx + c$, where m is the slope and c is the y-intercept.

First let's look at the calculation of the simple linear equation with 1 variable with the following age and weight example of school children. Here Age is the predictor *(X)* and Weight is *(y)* which is to be predicted based on Age.

Note: For simple linear regression, you X and y variable needs to be numeric in nature.

	Age	Weight
0	12	50
1	13	55
2	16	48
3	19	58
4	13	59
5	14	55
6	17	49
7	18	52
8	14	42
9	16	56

Dataframe (d) created manually depicting the age and Weight of students in a school/ college.

We will not look at the distribution of the data here as the purpose it to understand the calculation, so let's quickly jump into the working. For this purpose, we need the following information:

n — Number of records

ΣX — sum of X

Σy — sum of y

Σxy — sum of X*y

Σx2 — sum of X squared

Σy2 — sum of y squared

```
X = d['Age']
y = d['Weight']
```

```
n = len(d)
ΣX = sum(X)
Σy = sum(y)
Σx2 = sum(X**2)
Σy2 = sum(y**2)
Σxy = sum(d['Age']*d['Weight'])
```

```
print(
'Number of records ', n, '\n',
'Sum of X ', ΣX, '\n',
'Sum of y ', Σy, '\n',
'Sum of X squared ', Σx2, '\n',
'Sum of y squared ', Σy2, '\n',
'Sum of X*y ', Σxy)
```

```
Number of records  10
 Sum of X  152
 Sum of y  524
 Sum of X squared  2360
 Sum of y squared  27704
 Sum of X*y  7975
```

Required information

We will now proceed to find *m* which is the slope for the line also known as the coefficient. This simply means that for a single unit change in x, y will change by *m*. This shows a correlation between X and y.

Once we find *m,* we will calculate the value of *c* which is the constant value at y-intercept. This means that even when there is no X present at for the equation, a minimum of *c* on the *y* axis can be attained. For example if we are trying to find a linear relationship between Years of Experience and Salary, the minimum Salary that the company offers despite the years of experience will be a constant value *c*.

Please Note: These statements are not always practically true in all cases but the logic stands true. The value of *c* in some cases can also be negative and it should not be confused as the minimum value of *y* with no independent variables in the picture.

Similarly, the value of *m* can also be a negative value which simply means a negative correlation between *X* and *y*. With every unit of increase in *X, y* decreases by *m*.

To calculate the slope/ coefficient m :

```
#To calculate m
m = ((n*Σxy) - (ΣX*Σy)) / ((n*Σx2) - (ΣX**2))
print(round(m,2))
```

```
0.21
```

Value of coefficient (m)

Thus our value for $m = 0.21$ after rounding up. We will now calculate the value of c using the mean of X *as* $\bar{X}$ and y *as* $\bar{y}$ and compute in the formula –

$c = \bar{y} - m*\bar{X}$

```
#To calculate c
ȳ = Σy/n
X̄ = ΣX/n


c = ȳ - m*X̄
print(round(c,2))
```

```
49.27
```

Value for constant (c)

We now have our equation for this line:

$y = 0.21*X + 49.27$

Say for a given age of 15 years we need to calculate the weight we simply compute in the above equation:

```
y_15 = (m*15) + c
print(round(y_15, 2))
```

```
52.36
```

Calculation for y when Age= 15 years.

Lets quickly confirm this using the inbuilt Linear Regression function from the sklearn library

```
from sklearn.linear_model import LinearRegression
LR = LinearRegression()
LR = LR.fit(X,y)

print(round(LR.intercept_,2))

print(LR.coef_)

print(LR.predict([[15]]))
```

```
49.27
[0.20564516]
[52.35887097]
```

sklearns' Linear Regression

We see that the results are exactly the same as calculated by hand.

Summary:

In this article we looked at the calculated behind the simple linear regression equation with only 1 dependent variable. m being the slope of the line and c is the overall constant.

Next article we will look at the calculation for multiple linear equation.

This is the first article from a series I'm trying to do called: 'Decoding'. My idea with this series is to understand the formulae of the most basic concepts of Machine Learning. Next time when you apply them, you will definitely have a better idea of what is happening in the backend.

Any feedback is most welcome. Give me a clap if you like this article.

Linear Regression | Simple Linear Regression | Data Science | Data Journalism | Machine Learning