Linear regression works step-by-step.

Sample Dataset

|  |  |
| --- | --- |
| X | Y |
| 1 | 1 |
| 2 | 3 |
| 4 | 3 |
| 3 | 2 |
| 5 | 5 |

The attribute x is the input variable and y is the output variable that we are trying to

predict. If we got more data, we would only have x values and we would be interested in

predicting y values.

With simple linear regression we want to model our data as follows:

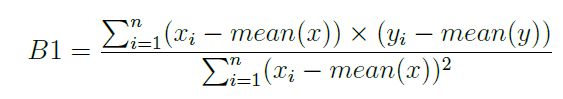
y = B0 + B1 \* x

This is a line where y is the output variable we want to predict, x is the input variable

we know and B0 and B1 are coefficients that we need to estimate that move the line around.

Technically, **B0 is called the intercept because it determines where the line intercepts the y-axis. In machine learning we can call this the bias, because it is added to offset all predictions that we make. The B1 term is called the slope because it defines the slope of the line or how x**

**translates into a y value before we add our bias.**

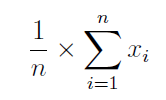




**Estimating the slope:**

Let's start with the top part of the equation, the numerator. First we need to calculate the

mean value of x and y. The mean is calculated as:



Where n is the number of values (5 in this case). You can use the AVERAGE() function in

your spreadsheet. Let's calculate the mean value of our x and y variables:

mean(X) = 3

mean(Y) = 2.8

Now we need to calculate the error of each variable from the mean. Let's do this with x first:

X – 1,2,4,3,5

Mean(X) = 3

X -mean(X)

1. 1-3 = -2
2. 2-3 = -1
3. 4-3 = 1
4. 3-3 = 0
5. 5-3 = 2

**Square of X-mean(X) = X-mean(X) \* X-mean(X)**

1. 4
2. 1
3. 1
4. 0
5. 4

Denominator Formula - 

**Denominator = 4+1+1+0+4 = 10**

Now let's do that for the y variable.

Y – 1,3,3,2,5

Mean(Y) = 2.8

Y-Mean(Y)

1. 1-2.8 = -1.8
2. 3-2.8 = 0.2
3. 3-2.8 = 0.2
4. 2-2.8 = -0.2
5. 5-2.8 = 2.2

Now Numerator = 

We have X -mean(X) and Y -mean(Y), now X -mean(X) \* Y -mean(Y)

1. -2 \* -1.8 = 3.6
2. -1 \* 0.2 = -0.2
3. 1 \* 0.2 = 0.2
4. 0 \* 0.8 = 0.8
5. 2 \* 2.2 = 4.4

Sum of X -mean(X) \* Y -mean(Y) = 3.6-0.2+0.2+0.8+4.4 = 8

So B1 = 8/10 = 0.8

B0 = mean(Y) – B1 \* mean(X) = 2.8 – 0.8 \* 3

B0 = 0.4

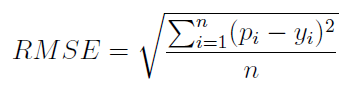
Now we know B0 and B1, so we can predict the Y values

|  |  |
| --- | --- |
| X | Predicted Y |
| 1 | 1.2 |
| 2 | 2 |
| 4 | 3.6 |
| 3 | 2.8 |
| 5 | 4.4 |

**Estimating Error**

We can calculate an error score for our predictions called the Root Mean Squared Error or

RMSE.



|  |  |  |  |
| --- | --- | --- | --- |
| **Predicted** | **Y** | **Predict - Y** | **Square of (Predict -Y)** |
| 1.2 | 1 | 0.2 | 0.04 |
| 2 | 3 | -1 | 1 |
| 3.6 | 3 | 0.6 | 0.36 |
| 2.8 | 2 | 0.8 | 0.64 |
| 4.4 | 5 | -0.6 | 0.36 |

The sum of Square of (Predict – Y) = 2.4

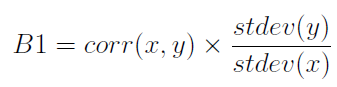
No of values

RMSE = Square root of 2.4 / 5

RMSE = 0.692

So each prediction is on average wrong by 0.692

Shortcut to calculate the B1



|  |  |  |  |
| --- | --- | --- | --- |
| X | Y | Pearson | 0.852802865 |
| 1 | 1 | STDEV Y | 1.483239697 |
| 2 | 3 | STDEV X | 1.58113883 |
| 4 | 3 | B1 | 0.8 |
| 3 | 2 |  |  |
| 5 | 5 |  |  |

Pearson =PEARSON(A2:A6,B2:B6)