Linear regression - derivation

derivation of the formulas

Suppose we have three points (x_1, y_1), (x_2, y_2) and (x_3, y_3). They nearly lie on one line, but not exactly. (If we had only two points, they must lie on the same line).

We want to find a linear regression equation $y = Ax + B$, where A and B are the slope and intercept that give the "best fit" to the data, defined in the following way.

For each x_i, y_i is what's in the data. The error of the predicted value is different by

$$e_i = y_i - (Ax_i + B)$$

The error squared is

$$e_i^2 = [\, y_i - (Ax_i + B)\,]^2$$

We want the total of all the squared errors to be a minimum. Note that we can either subtract y_i from $Ax_i + B$, or the reverse, and obtain the same result for the square. That's because $(c - d)^2 = (d - c)^2$.

The total error e is

$$e = [\, y_1 - (Ax_1 + B)\,]^2 + [\, y_2 - (Ax_2 + B)\,]^2 + [\, y_3 - (Ax_3 + B)\,]^2$$

For n points

$$e = \sum_{i=1}^{i=n} [\, y_i - (Ax_i + B)\,]^2$$

1

The x and y values are *constants* for a given problem, but A and B are variables. We are going to take the partial derivatives of e with respect to A and B.

Do not be nervous because we have two variables. Partial derivatives are just regular derivatives in disguise. When taking a partial derivative, say $\partial e/\partial A$, the other variable B is treated as a constant.

Nor does the sum cause any difficulty, since the derivative of a sum is the sum of the derivatives.

Thus (suppressing the index on the summation sign)

$$\frac{\partial}{\partial A} \sum [\, y_i - (Ax_i + B)\,]^2$$

$$= \sum 2\,[\, y_i - (Ax_i + B)\,]\,(-x_i)$$

This is elementary calculus. The derivative of a^n is na^{n-1}, times the derivative of a itself

The factor of $-x_i$ comes from the chain rule applied to the term $-(Ax_i + B)$. x_i is the constant cofactor of the variable A, with a minus sign out front.

The factor of 2 can go before the summation, and multiplying through

$$\frac{\partial e}{\partial A} = 2 \sum (Ax_i^2 - x_i y_i + Bx_i)$$

The second equation is the same except the factor of x_i is missing, since B just stands by itself without any cofactor.

$$\frac{\partial e}{\partial B} = 2 \sum [\, y_i - (Ax_i + B)\,](-1)$$

$$= 2 \sum [\, Ax_i - y_i + B\,]$$

We set both of these partial derivatives equal to zero to find a minimum, and write the individual sums.

The factors of 2 now go away and we have

$$A\sum x_i^2 - \sum x_i y_i + B \sum x_i = 0$$

$$A\sum x_i - \sum y_i + \sum B = 0$$

Note that $\sum B$ is B taken n times, or just nB, and that the second derivatives are both positive, which indicates that these extreme values are minima.

We also reverse the order in the last term of the first equation, for clarity below.

$$A\sum x_i^2 - \sum x_i y_i + \sum x_i B = 0$$

$$A\sum x_i - \sum y_i + nB = 0$$

I just solve the second equation for B.

$$B = \frac{1}{n} \cdot (-A\sum x_i + \sum y_i)$$

Substitute back into the first equation:

$$A\sum x_i^2 - \sum x_i y_i - \sum x_i \cdot (\frac{1}{n}) \cdot (A\sum x_i + \sum y_i) = 0$$

Gather terms, rearrange, and divide by the cofactors of A

$$A = \frac{\sum x_i y_i - (1/n)\sum x_i \sum y_i}{\sum x_i^2 - (1/n)\sum x_i \sum x_i}$$

If we multiply through on the top and bottom by $1/n$ we obtain just the right number of n's to turn all those sums into means.

$$A = \frac{\mu(x_i y_i) - \mu(x)\mu(y)}{\mu(x_i x_i) - \mu(x)\mu(x)}$$

3

We have derived one of the formulas to calculate the slope of a linear regression equation. We used a second one earlier in this series.

That formula is just SS_{xy}/SS_{xx} where SS_{xy} and SS_{xx} are defined to be

$$SS_{xy} = (x_i - \bar{x}) \cdot (y_i - \bar{y}), \qquad SS_{xx} = (x_i - \bar{x})^2$$

The two formulas have advantages in different respects. We will show that they are equivalent in a bit.

https://stats.stackexchange.com/questions/4446/sum-of-squares-two-ways-

intercept

Recall that

$$nB = -A \sum x_i + \sum y_i$$
$$B = -A\bar{x} + \bar{y}$$
$$A\bar{x} + B = \bar{y}$$

This not only gives a way to calculate B but shows that $(\bar{x}, \bar{y})$ is on the line.

another method

We will use the first formula to compute the slope and intercept of the linear regression equation.

$$\frac{\mu(x_i y_i) - \bar{x}\bar{y}}{\mu(x_i x_i) - \bar{x}\bar{x}}$$

Compute $x \cdot y$ for each position and then find the mean of those values.

$$\frac{1 \cdot 2 + 2 \cdot 2 + 3 \cdot 4}{3} = 6$$

4

Subtract the product of the means $\bar{x} \cdot \bar{y}$

$$6 - 2 \cdot 8/3 = 2/3$$

Similarly, compute $x \cdot x$ for each position and find the mean of those values.

$$\frac{1 \cdot 1 + 2 \cdot 2 + 3 \cdot 3}{3} = \frac{14}{3} = 4 + 2/3$$

Subtract $\bar{x} \cdot \bar{x}$

$$14/3 - 4 = 2/3$$

Divide the first result by the second to obtain $m = 1$, once again.

We derive the formulas used for these computations in the next part. It is a little more sophisticated and uses a tiny bit of calculus.

connecting the two formulas

Start with the denominator of this formula

$$A = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

using m for $\bar{x}$ and suppress the index i. We have

$$\sum (x - m)^2$$

Expand the square

$$= \sum x^2 - \sum 2mx + \sum m^2$$

Since for any given data set the x_i and y_i are constants, so the mean m is also a constant

$$= \sum x^2 - 2m \sum x + nm^2$$

But $\sum x = nm$ and then

$$= \sum x^2 - nm^2$$

$$= \sum x^2 - m \sum x$$

We will multiply both the numerator and denominator by $1/n$ and use μ for the mean:

$$1/n \cdot \sum (x - m)^2 = \frac{1}{n}\left(\sum x^2 - m \sum x\right)$$

$$= \mu(x^2) - \mu(x)\mu(x)$$

This matches the denominator of the previous formula.

For the numerator:

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = \sum (x_i y_i - x_i \bar{y} - y_i \bar{x} + \bar{x}\bar{y})$$

$$= \sum x_i y_i - \sum x_i \bar{y} - \sum y_i \bar{x} + \sum \bar{x}\bar{y}$$

$$= \sum x_i y_i - \bar{y} \sum x_i - \bar{x} \sum y_i + n\bar{x}\bar{y}$$

Multiplying by $1/n$ we have then

$$= \mu(x_i y_i) - \bar{y}\bar{x} - \bar{x}\bar{y} + \bar{x}\bar{y}$$

$$= \mu(x_i y_i) - \bar{x}\bar{y}$$

So the whole thing is a match.