Linear regression

Alternative method

There is a second method method for doing the calculation of linear regression. It has certain advantages.

Recall that we have this data: three pairs of (x, y) values: (1,2), (2,2), (3,4). I will write these as:

$$X = (1, 2, 3), \qquad Y = (2, 2, 4)$$

So here we are using capital X as the notation for the list or *vector* of values x_i, and similarly Y contains the values y_i.

Then $\bar{X} = 2$ and $\bar{Y} = 8/3$.

Rather than subtract the mean from each value before multiplying, we do it afterward, as follows.

We use the dot product just as before, multiplying the corresponding elements from each of the two vectors together and then adding them up.

$$X \cdot Y = 1 \cdot 2 + 2 \cdot 2 + 3 \cdot 4 = 18$$

$$X \cdot X = 1 \cdot 1 + 2 \cdot 2 + 3 \cdot 3 = 14$$

Find the mean of $X \cdot Y$. Divide the result (18) by the number of elements (3) to get 6, and then subtract $\bar{X}\bar{Y}$. I get $6 - 16/3 = 2/3$.

Find the mean of $X \cdot X = 4 + 2/3$ and then subtract $\bar{X}\bar{X}$. I get 2/3.

Finally, divide the first result by the second to obtain the slope, which is $m = 1$, as we had before. The formula is:

$$m = \frac{\mu(X \cdot Y) - \bar{X}\bar{Y}}{\mu(X \cdot X) - \bar{X}\bar{X}}$$

Sigma notation

In statistics (and in all of mathematics) we try to simplify the notation, because it makes it easier to concentrate on what's important.

Compare these two lists of values:

$$x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}$$

$$x_1, x_2, \ldots x_{10}$$

The second is shorter, but the meaning (I think) is clear. The dots indicate the values that we didn't write explicitly.

In the general case there don't have to be exactly 10 values, so for n values we write

$$x_1, x_2, \ldots x_n$$

When we want to talk about a value from the list, without specifying exactly which one, call it x_i, with a subscript letter i written a little below the line.

Then we can write the sum of all the values using the capital Greek letter sigma (Σ) as

$$\sum x_i$$

where

$$\sum x_i = x_1 + x_2 \cdots + x_n$$

If we wanted to be really careful, we might write

$$\sum_{i=1}^{i=n} x_i = x_1 + x_2 \cdots + x_n$$

The i stands for *index.* Since there is only one index involved it doesn't really hurt to *suppress* or leave it out.

$$\sum x_i$$

And, in some cases we might even write

$$\sum x$$

and just remember that x ranges from x_1 to x_n.

Using this notation, the definition of the mean is

$$\bar{x} = \frac{1}{n} \sum x_i = \frac{1}{n}(x_1 + x_2 \cdots + x_n)$$

Sum up all the values and then divide by how many there are.

Rearranging

$$n \cdot \bar{x} = \sum x_i$$

Another important thing to remember about sum notation is that if we have two or more terms with the same index

$$\sum (x_i + y_i)$$

We can spread the sums out over the variables.

$$\sum x_i + \sum y_i$$

This is easy to show by writing out all the terms and just rearranging them.

Going back to the point about centering, once the data has been adjusted to have $\bar{x} = 0$ and $\bar{y} = 0$, the best fit line will go through the origin and the slope of the line will be simply

$$m = \frac{\sum x_i y_i}{\sum x_i x_i}$$

Variance

For a given set of values, we are often interested in how closely clustered they are. In the big picture overall, how different are the values from the mean?

We could subtract the mean from each value, as we did above

$$\sum (x_i - \bar{x}) = (x_1 - \bar{x}) + (x_2 - \bar{x}) \cdots + (x_n - \bar{x})$$

There is a problem, however. Generally, some values will be larger and some smaller than the mean, giving negative and positive values after subtracting the mean, and there will be cancellations when they are all added up.

In fact, the expression above always evaluates to zero. Can you rearrange terms to prove it? I can't resist. The formula contains n copies of the mean value of x:

$$\begin{aligned}\sum (x_i - \bar{x}) &= (x_1 - \bar{x}) + (x_2 - \bar{x}) \cdots + (x_n - \bar{x}) \\ &= x_1 + x_2 \cdots + x_n - n \cdot \bar{x} \\ &= \sum x_i - n \cdot \bar{x}\end{aligned}$$

But by definition $\bar{x} = (\sum x_i)/n$ so we have

$$= \sum x_i - n \cdot (\sum x_i)/n$$

$$= \sum x_i - \sum x_i = 0$$

To make all the differences positive or at least non-negative, a useful trick is to square them

$$\sum (x_i - \bar{x})^2 = (x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 \cdots + (x_n - \bar{x})^2$$

This is just what we called $SSxx$ before.

There is one more manipulation. Divide the whole thing by the number of items, n, and give it a special name, the *variance* of x. Write V for variance:

$$V_x = \frac{1}{n} \sum (x_i - \bar{x})^2$$

When there are two variables, x and y, we can also compute the *co-variance*, C:

$$C_{xy} = \frac{1}{n} \sum (x_i - \bar{x}) \cdot (y_i - \bar{y})$$

We might reasonably call the variance of x the co-variance of x with itself, or C_{xx}.

$$V_x = C_{xx} = \frac{1}{n} \sum (x_i - \bar{x}) \cdot (x_i - \bar{x})$$

Divide one by the other, the factors of $1/n$ cancel, and we have

$$\frac{C_{xy}}{V_x} = \frac{SSxy}{SSxx} = m$$

where m is the slope of the linear regression equation. And as before, $b = \bar{y} - m\bar{x}$.

technical notes

It turns out to be better, when estimating the variance of a population from a small sample, to divide by $n-1$ rather than n. So there is a distinction between the *population* variance (divide by n) and the *sample* variance (divide by $n-1$). That doesn't matter for us since the divisor, whatever it is, always cancels here.

The variance as we just calculated it, is usually written as σ^2. The reason is that the square root of the variance is often used as a measure of spread in values. This is called the standard deviation, stdev or σ.

residuals or errors

The whole point of this topic is to take some bivariate data, plot it, and then also find a line that best fits the data. By "best fit" we mean something precise.

Draw a vertical from each data point to the line of best fit. Compute that distance as $y_i - (mx_i + b)$. y_i is the actual y-value from the data point, and $mx_i - b$ is the predicted value from the line of best fit.

The difference is called the *residual* or error for that data point. The smaller the total of all the errors, the more tightly clustered the data points are.

The formulas for the line of best fit are derived by asking for the line where if we take the square of each residual and add them all up, that total is smaller than for any other line.

Let's just look at the data again. Recall that for the toy example, we had data of $(1,2),(2,2),(3,4)$ and an equation of $y = x + 2/3$. Let us compute the residual for each point. Subtract the predicted value for each x using the equation, from the given data point for y.

$$2 - 5/3 = 1/3$$
$$2 - 8/3 = -2/3$$
$$4 - 11/3 = 1/3$$

The sum of the errors is zero. This is always true for linear regression.

correlation coefficient r and r^2

There are two more statistics listed by Desmos, r and r^2. These describe how well the line fits the data, or how well the data fits the line, depending on your point of view.

Pearson's correlation coefficient is defined to be:

$$r = \frac{C_{xy}}{\sigma_x \cdot \sigma_y} = \frac{C_{xy}}{\sqrt{V(x)} \cdot \sqrt{V(y)}}$$

We *could* first compute the covariance of x and y

$$C_{xy} = \frac{1}{n} \cdot SS_{xy}$$

Then compute the variance for both x and y

$$\sigma_x^2 = \frac{1}{n} \cdot SS_{xx}$$
$$\sigma_y^2 = \frac{1}{n} \cdot SS_{yy}$$

As we said, the square root of the variance is called the *standard deviation.*

However, we notice that each of the terms in the denominator has a factor of $1/\sqrt{n}$ which multiply to give a factor of $1/n$. So we see that the n's will cancel, when taking account of the numerator.

An equivalent formula is

$$\frac{SS_{xy}}{\sqrt{SS_{xx}} \cdot \sqrt{SS_{yy}}}$$

We already have $SS_{xy} = 2$ and $SS_{xx} = 2$. We just need SS_{yy}.

$$SS_{yy} = (-2/3)^2 + (-2/3)^2 + (4/3)^2$$

$$= 4/9 + 4/9 + 16/9 = 24/9 = 8/3$$

and then

$$\frac{2}{\sqrt{2} \cdot \sqrt{8/3}}$$

Pull $\sqrt{2}$ out of both square roots to cancel 2 on top and bottom. What is left is

$$\frac{1}{1 \cdot \sqrt{4/3}}$$

Invert to obtain

$$r = \sqrt{3/4} = 0.866$$

$$r^2 = 3/4 = 0.75$$

And that matches what Desmos gave.

If the data were perfectly correlated, then r and r^2 would be equal to one. For example, if the x-values were plotted against themselves, they would all lie exactly on a straight line with $r^2 = 1$.