

The Geometry of Linear Regression

The *least-squares* method of fitting a straight line $y = mx + b$ to a collection of data points $(x_1, y_1), (x_2, y_2), (x_3, y_3), \dots, (x_n, y_n)$ is routinely available on graphing calculators. The simple geometric algorithm that is used to calculate m and b seems not to be well-known, however. As we shall see, the data-analysis problem to be solved is essentially the following:

Given points A, B , and C , find the point D on \overline{AB} that is closest to C .

It is evident that we must make \overline{CD} perpendicular to \overline{AB} . Thus \overline{AD} is a perpendicular *projection* of \overline{AC} on \overline{AB} . Its unknown length may be expressed trigonometrically in terms of known quantities:

$$AD = AC \cdot \cos \angle CAB$$

At this time, let us recall the familiar *dot product* of vectors (which is also known as the *scalar product* of vectors). For instance, given vectors $\mathbf{u} = [a, b, c]$ and $\mathbf{v} = [p, q, r]$, the dot product is defined by

$$\mathbf{u} \cdot \mathbf{v} = ap + bq + cr.$$

For our data-analysis purposes, the crucial property of the dot product of vectors is that

$$\frac{\mathbf{u} \cdot \mathbf{v}}{|\mathbf{u}| \cdot |\mathbf{v}|} = \cos \theta,$$

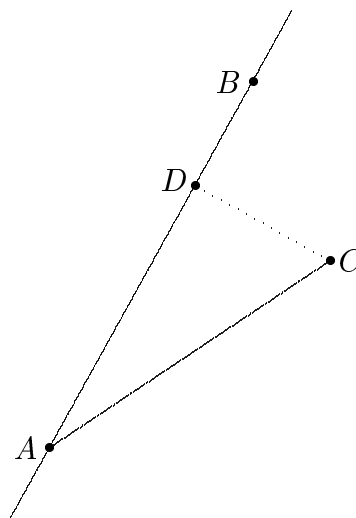
where θ is the angle made by placing \mathbf{u} and \mathbf{v} tail to tail, and $|\mathbf{u}|$ and $|\mathbf{v}|$ are the lengths of \mathbf{u} and \mathbf{v} , respectively. In particular, the projection problem above is solved by calculating

$$AD = AC \cdot \frac{\overrightarrow{AB} \cdot \overrightarrow{AC}}{AB \cdot AC} = \frac{\overrightarrow{AB} \cdot \overrightarrow{AC}}{AB}$$

This in principle locates D exactly. If coordinate information is desired, however, it is necessary to make further calculations with the lengths AD and AB . Namely, to calculate the coordinates of D from the coordinates of A , we must first calculate

$$\overrightarrow{AD} = \frac{AD}{AB} \cdot \overrightarrow{AB},$$

and then the equation $D = A + \overrightarrow{AD}$ finishes the job.



The Geometry of Linear Regression

One of the interesting aspects of the preceding problem and solution is that it is unaffected by the *dimension* of the space containing the three given points. As suggested by the picture, it could be two, or — as suggested by the dot-product example — it could be three. In fact, it could be any positive integer n . It is necessary only to adjust the companion formulas for dot product and length. Namely, the dot product of vectors $\mathbf{u} = [u_1, u_2, u_3, \dots, u_n]$ and $\mathbf{v} = [v_1, v_2, v_3, \dots, v_n]$ is defined to be

$$\mathbf{u} \bullet \mathbf{v} = u_1v_1 + u_2v_2 + u_3v_3 + \dots + u_nv_n,$$

and the length of vector $\mathbf{u} = [u_1, u_2, u_3, \dots, u_n]$ is

$$|\mathbf{u}| = \sqrt{u_1^2 + u_2^2 + u_3^2 + \dots + u_n^2} = \sqrt{\mathbf{u} \bullet \mathbf{u}}.$$

The important fact is that $\mathbf{u} \bullet \mathbf{v}$ lies between $-|\mathbf{u}| \cdot |\mathbf{v}|$ and $|\mathbf{u}| \cdot |\mathbf{v}|$ (which is proved below), showing that the equation

$$\frac{\mathbf{u} \bullet \mathbf{v}}{|\mathbf{u}| \cdot |\mathbf{v}|} = \cos \theta,$$

makes sense in any dimension.

Is there a fourth dimension? The skeptical student may wonder whether there is any purpose to pushing geometry into higher dimensions in this formal way. The data analysis examples below provide a convincing and understandable application.

The five data points $(-8, -6)$, $(-3, -3)$, $(-1, 2)$, $(5, 3)$, and $(7, 4)$ are not collinear. Because there is no obvious answer, it is therefore an interesting question to ask for the straight line that best fits these points. (This requires that we also decide what *best fits* means.)

For reasons that will soon become clear, let us first ask a narrower question: Which line $y = mx$ through the origin is best? To understand the least-squares answer to this question, it is necessary for us to look at the data as *two vectors*, not as *five pairs*. In other words, let $\mathbf{u} = [-8, -3, -1, 5, 7]$ be the five-component vector of x -values and let $\mathbf{v} = [-6, -3, 2, 3, 4]$ be the five-component vector of corresponding y -values.

In order that a group of points all fit an equation $y = mx$, it is necessary and sufficient that the vector \mathbf{v} of y -values be a multiple of the vector \mathbf{u} of x -values.

This is the key observation. For example, $(-8, -6)$, $(-3, -2.25)$, $(-1, -0.75)$, $(5, 3.75)$, and $(7, 5.25)$ all lie on the line $y = 0.75x$, and the vector $\mathbf{v} = [-6, -2.25, -0.75, 3.75, 5.25]$ is 0.75 times the vector $\mathbf{u} = [-8, -3, -1, 5, 7]$. Notice in particular that the multiplier is just the *slope* of the line.

The Geometry of Linear Regression

The technique of linear data analysis alters the y -values to achieve the best fit, leaving the x -values alone. Each data point can be adjusted upward or downward, so long as the new points are all aligned (with the origin, in this example). In effect, the vector \mathbf{v} of *observed* y -values is replaced by a vector \mathbf{w} of *hypothetical* y -values. This new vector \mathbf{w} is a multiple of \mathbf{u} , the multiplier being the best-fit slope: $\mathbf{w} = m\mathbf{u}$.

Because there are many possible multipliers to choose from, the original question returns, but in a different form: Which multiple of \mathbf{u} is the *best* replacement for the given \mathbf{v} ? The simple answer is to obtain \mathbf{w} by *projecting* \mathbf{v} onto \mathbf{u} .

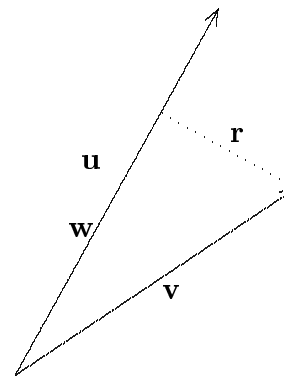
Here are the projection calculations for $\mathbf{u} = [-8, -3, -1, 5, 7]$ and $\mathbf{v} = [-6, -3, 2, 3, 4]$, which is the current example:

$$\mathbf{u} \cdot \mathbf{v} = 98, \text{ and } |\mathbf{u}| \cdot |\mathbf{v}| = \sqrt{148} \cdot \sqrt{74}, \text{ thus } \cos \theta = \frac{98}{\sqrt{148}\sqrt{74}}.$$

To obtain the length of the projection of \mathbf{v} on \mathbf{u} , multiply this by $|\mathbf{v}| = \sqrt{74}$. This is $|\mathbf{w}|$. To obtain the multiplier, divide $|\mathbf{w}|$ by $|\mathbf{u}| = \sqrt{148}$. The result is $m = 98/148$. At this point, it might be a good idea to ask your calculator for the slope of the least-squares line that fits the five data points $(-8, -6)$, $(-3, -3)$, $(-1, 2)$, $(5, 3)$, and $(7, 4)$. It should respond $m = 0.66216$, approximately.

The above method converted into a formula: $m = \frac{\mathbf{u} \cdot \mathbf{v}}{|\mathbf{u}|^2}$

Why is this best? It makes the difference between the two y -value vectors (\mathbf{v} and \mathbf{w}) as small as possible. This is a meaningful criterion, because the difference between these vectors is $\mathbf{r} = \mathbf{v} - \mathbf{w}$, the vector of *residuals*. In other words, each component of \mathbf{r} is the difference between an observed y -value and the corresponding hypothetical (adjusted) y -value. Because the Pythagorean length formula for $|\mathbf{r}|$ involves summing the squares of all the residuals, the projection method is usually called the method of *least squares*.



In summary: Any choice of slope other than the least-squares value $m = \frac{98}{148}$ produces a greater sum of squared residuals, because it makes the projection \mathbf{w} longer or shorter (though still aligned with \mathbf{u}), thus *lengthening* \mathbf{r} (see the figure above).

The next thing to do is to discuss the effect of varying the y -intercept b . Notice first a significant feature of the five data points chosen for our first example: The sum of the x -values is 0, and so is the sum of the y -values.

The Geometry of Linear Regression

Now that we have decided what *best fit* means — the sum of the squared residuals should be as small as possible — the inevitable question arises: Must the line $y = mx + b$ of best fit have $b = 0$? In cases where the sum of the observed x -values and the sum of the observed y -values *are both zero*, the answer is “yes”. Here is the reason:

Let the data points be $(x_1, y_1), (x_2, y_2), (x_3, y_3), \dots, (x_n, y_n)$, and assume that the sums $x_1 + x_2 + x_3 + \dots + x_n$ and $y_1 + y_2 + y_3 + \dots + y_n$ are both zero. Because a residual is the difference between an observed value y_i and its companion hypothetical value $mx_i + b$, the goal is to choose m and b to make the sum

$$(y_1 - mx_1 - b)^2 + (y_2 - mx_2 - b)^2 + (y_3 - mx_3 - b)^2 + \dots + (y_n - mx_n - b)^2$$

as small as possible. Each of the n terms can be expanded

$$(y_i - mx_i - b)^2 = (y_i - mx_i)^2 - 2b(y_i - mx_i) + b^2,$$

and the $3n$ pieces rearranged to make the sum look like

$$\begin{aligned} & (y_1 - mx_1)^2 + (y_2 - mx_2)^2 + (y_3 - mx_3)^2 + \dots + (y_n - mx_n)^2 \\ & - \{ 2b(y_1 - mx_1) + 2b(y_2 - mx_2) + 2b(y_3 - mx_3) + \dots + 2b(y_n - mx_n) \} \\ & + b^2 + b^2 + b^2 + \dots + b^2. \end{aligned}$$

Notice that the contribution of the middle line is zero, because of our assumption that $x_1 + x_2 + x_3 + \dots + x_n = 0 = y_1 + y_2 + y_3 + \dots + y_n$. This leaves us with only the sum of squares

$$(y_1 - mx_1)^2 + (y_2 - mx_2)^2 + (y_3 - mx_3)^2 + \dots + (y_n - mx_n)^2 + nb^2$$

to make as small as possible. It is now clear that b must be zero, and that m can be found by the projection method already described.

What if the x -values and the y -values do not sum to zero? For example, what if the data points were $(4, 1), (7, 4), (8, 6), (11, 8),$ and $(15, 11)$? Applied directly to this data, the projection method does not give the proper line. There is a simple modification of the method that always works, however. Calculate the *centroid* of the data, which is the point whose x -coordinate is the average of the given x -coordinates, and whose y -coordinate is the average of the given y -coordinates. In this example, the centroid is $(9, 6)$. Subtract the centroid coordinates from each of the data points, thereby *translating* the data to $(-5, -5), (-2, -2), (-1, 0), (2, 2),$ and $(6, 5)$. Notice that the centroid of this data is $(0, 0)$, so the projection method will produce the correct least-squares slope, which you can verify is $\frac{9}{10}$. *This is also the slope of the least-squares line for the given data points (why?).* Because this line must go through the centroid of the data (why?), we see that the line of best fit is $y = 0.9(x - 9) + 6$.

The Geometry of Linear Regression

The formulas for m and b , as they appear in many mathematics books:

$$m = \frac{n(\sum x_i y_i) - (\sum x_i)(\sum y_i)}{n(\sum x_i^2) - (\sum x_i)^2} \quad \text{and} \quad b = \frac{(\sum x_i^2)(\sum y_i) - (\sum x_i)(\sum x_i y_i)}{n(\sum x_i^2) - (\sum x_i)^2}$$

The sums all run from $i = 1$ to $i = n$. It is unfortunate that multivariable calculus is so often used to derive these formulas, leaving the intrinsic geometry barely visible.

The so-called *correlation coefficient* is a standard statistical measure of how well aligned the data points are. It varies between -1 and 1 , inclusive. It may come as a surprise that this statistic is just another name for $\cos \theta$, where θ is the angle formed by the vectors \mathbf{u} and \mathbf{v} of (translated) x -coordinates and y -coordinates. Given that $\cos 18^\circ = 0.951$, one must be careful not to attach too much significance to readings that are close to 1 .

Appendix

Proof that $-|\mathbf{u}| \cdot |\mathbf{v}| \leq \mathbf{u} \cdot \mathbf{v} \leq |\mathbf{u}| \cdot |\mathbf{v}|$:

Divide the inequality by $|\mathbf{u}| \cdot |\mathbf{v}|$ to obtain the equivalent version $-1 \leq \frac{\mathbf{u}}{|\mathbf{u}|} \cdot \frac{\mathbf{v}}{|\mathbf{v}|} \leq 1$, and notice that $\frac{\mathbf{u}}{|\mathbf{u}|}$ and $\frac{\mathbf{v}}{|\mathbf{v}|}$ are vectors of unit length. It is therefore sufficient to prove that $-1 \leq \mathbf{u} \cdot \mathbf{v} \leq 1$ for all unit vectors \mathbf{u} and \mathbf{v} . This can be accomplished by applying routine algebraic manipulations to the inequalities

$$0 \leq (\mathbf{u} - \mathbf{v}) \cdot (\mathbf{u} - \mathbf{v}) \quad \text{and} \quad 0 \leq (\mathbf{u} + \mathbf{v}) \cdot (\mathbf{u} + \mathbf{v}).$$

The first implies that $0 \leq \mathbf{u} \cdot \mathbf{u} - 2 \cdot \mathbf{u} \cdot \mathbf{v} + \mathbf{v} \cdot \mathbf{v}$, hence that $2 \cdot \mathbf{u} \cdot \mathbf{v} \leq 2$, or $\mathbf{u} \cdot \mathbf{v} \leq 1$. The second implies that $0 \leq \mathbf{u} \cdot \mathbf{u} + 2 \cdot \mathbf{u} \cdot \mathbf{v} + \mathbf{v} \cdot \mathbf{v}$, hence that $-2 \leq 2 \cdot \mathbf{u} \cdot \mathbf{v}$, or $-1 \leq \mathbf{u} \cdot \mathbf{v}$.

Richard Parris
Phillips Exeter Academy
Exeter, NH 03833-2460