

Line-fitting by distance: errors-in-variables regression.

Regression of y on x is based on the idea that the points x_i are not random variables but some fixed points called “design points,” measured with little or no error, while the y_i are random variables. Thus y -on- x regression minimizes the sum of squared vertical deviations. One can also do x -on- y regression which assumes that the points y_i are some fixed points while x_i are random variables and/or are measured with errors. So x -on- y regression minimizes the sum of squares of horizontal deviations of the data points from a line.

For given $(X_1, Y_1), \dots, (X_n, Y_n)$, with $n \geq 2$, let s_x be the sample standard deviation of the X_i , and s_y of the Y_i ,

$$s_x = \left(\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \right)^{1/2}, \quad s_y = \left(\frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2 \right)^{1/2}.$$

If $s_x = 0$ then the y -on- x line is not uniquely determined. Any line through (\bar{X}, \bar{Y}) will minimize the sum of squares of vertical deviations of the points from the line. Likewise if $s_y = 0$ the x -on- y line is not unique. In all other cases these regression lines are defined and unique.

If all the points are on a line, then that line will clearly be the best-fitting line either for vertical deviations (y -on- x) or horizontal deviations (x -on- y) because these deviations will be 0 in that case. It may be surprising that these are the only times these two regressions agree:

Theorem 1. For given observations in the plane, $(X_1, Y_1), \dots, (X_n, Y_n)$, where $n \geq 2$, $s_x^2 > 0$ and $s_y^2 > 0$, the lines given by y -on- x and x -on- y regression only agree when all the points (X_i, Y_i) are on a line.

Proof. Both regression lines pass through the point (\bar{X}, \bar{Y}) . The slope of the y -on- x line is $r \cdot s_y/s_x$ (Rice, §14.2.3 p. 561) where r is the correlation coefficient of the observations.

The slope of the x -on- y line, if we take the y axis as horizontal and the x axis as vertical, is then $r \cdot s_x/s_y$. In the original orientation where the x axis is horizontal and the y axis is vertical, the slope is replaced by its reciprocal, which is $(1/r)s_y/s_x$. So, the two lines are only the same if $r = 1/r$ so $r^2 = 1$, $r = \pm 1$. Then the points (X_i, Y_i) are all on a line (with positive slope if $r = 1$ or negative slope if $r = -1$), Q.E.D.

About notation: Rice on p. 561 uses definitions of sample covariances and variances with a factor of $\frac{1}{n}$ rather than $\frac{1}{n-1}$. Note that in his next three displays after the definitions, such factors will appear both in the numerator and denominator, so they will divide out. One just needs to be consistent in using one factor or the other. Also note that what Rice calls s_{xx} and s_{yy} are estimators of sample variance (as opposed to standard deviation).

Theorem 1 implies that the two regression lines will in nearly all cases be different (if $n \geq 3$). If the y -on- x regression line has a positive slope, but the correlation $r < 1$, then the x -on- y line always has a larger slope, by a factor of $1/r^2$. In many situations, the assumptions for y -on- x and x -on- y regression may not hold. We need something better.

A third way of fitting a line to a set of points $(x_1, y_1), \dots, (x_n, y_n)$ is to minimize the sum of squared distances of the points to the line. This corresponds to what is sometimes called “errors-in-variables” regression. The idea is that both x_i and y_i are measured with error, so that both are random variables.

For any point p and line L in the plane, let $d(p, L)$ be the distance from p to L . Given a joint distribution of (X, Y) in the plane, where $E(X^2 + Y^2) < \infty$, a line L_o will be called a *bfsd line* (*best-fitting by squared distance line*) if $E[d((X, Y), L)^2]$ is minimized at $L = L_o$. This will apply to data sets (x_i, y_i) , $i = 1, \dots, n$, as will be explained at the end.

Let $\text{Cov}(X, Y) = E(XY) - EXEY$, the covariance of X and Y , for any random variables (X, Y) . If the standard deviations $\sigma_X > 0$ and $\sigma_Y > 0$ then the correlation of X and Y is defined by $\rho = \rho_{X,Y} = \text{Cov}(X, Y)/(\sigma_X\sigma_Y)$. We have $-1 \leq \rho \leq 1$.

Let $L_{a,b}$ be the line $y = a + bx$ for any real numbers a, b . Let $L_{\infty;c}$ be the vertical line $x \equiv c$, $-\infty < y < \infty$. So every line in the plane is either a line $L_{a,b}$ or a line $L_{\infty;c}$ for some a, b or c . Then bfsd lines are characterized as follows.

Theorem 2. For any random vector (X, Y) in the plane with $E(X^2 + Y^2) < \infty$ there is at least one bfsd line. All such lines go through the point (EX, EY) . Let $\sigma = \sigma_X$ and $\tau = \sigma_Y$. If $\sigma = \tau = 0$, or $\sigma = \tau > 0$ and $\rho = \rho_{X,Y} = 0$, then every line through (EX, EY) is a bfsd line.

In all other cases the bfsd line L is unique.

If $\sigma > 0 = \tau$ then $L = L_{EY,0}$, or if $\sigma = 0 < \tau$ then $L = L_{\infty;EX}$.

If $\sigma > 0$ and $\tau > 0$ then: if $\rho = 0$ and $\sigma^2 > \tau^2$ then $L = L_{EY,0}$, or if $\sigma^2 < \tau^2$ then $L = L_{\infty;EX}$.

If $\sigma > 0$, $\tau > 0$ and $\rho \neq 0$ (the general case) then $L_o = L_{a,b}$ has slope $b = \tan \theta$ (which, given θ , uniquely determines the line as $(y - EY) = b(x - EX)$) and θ is as follows:

If $\sigma > \tau$ then $\theta = \theta_I$ where

$$(1) \quad \theta_I = \frac{1}{2} \tan^{-1} \left[\frac{2\text{Cov}(X, Y)}{\sigma_X^2 - \sigma_Y^2} \right].$$

If $\sigma < \tau$ then $\theta = \theta_{II}$, defined as $\theta_I + \pi/2$.

If $\sigma = \tau$ then since $\rho \neq 0$, $\text{Cov}(X, Y) \neq 0$ and:

if $\text{Cov}(X, Y) > 0$, $\theta = \pi/4$, $b = 1$;

if $\text{Cov}(X, Y) < 0$, $\theta = -\pi/4$, $b = -1$.

Proof. In each of the following cases, the probability that (X, Y) is on the given line L is 1, so $E(d((X, Y), L)^2) = 0$ and L is a bfsd line: $\sigma > 0 = \tau$, so $Y \equiv EY$ is constant and $L = L_{EY,0}$ is horizontal; or $X \equiv EX$ is constant, $\sigma = 0 < \tau$ and $L = L_{\infty;EX}$, the vertical line $x \equiv EX$; or the distribution of (X, Y) is concentrated at one point (EX, EY) , i.e. $\sigma = \tau = 0$, and L is any line through (EX, EY) , either a line $y - EY = b(x - EX)$ for any finite slope b , or the vertical line $L_{\infty;EX}$.

To find the distance $d((X, Y), L)$ from a point (X, Y) to a line L , if $L = L_{\infty;c}$ it's $|X - c|$. If $L = L_{a,0}$ it's $|Y - a|$. So suppose $L = L_{a,b}$ with $b \neq 0$. Here are two ways of evaluating the distance. First, here's a geometric-trigonometric way. The vertical distance from (X, Y) to $L_{a,b}$ is clearly $|Y - a - bX|$. A line M through (X, Y) perpendicular to $L_{a,b}$ forms an angle θ at (X, Y) with a vertical line. Then the perpendicular distance from

(X, Y) to $L_{a,b}$ is $|Y - a - bX| \cos \theta$. On the other hand, one can see by drawing a diagram or otherwise that the line $L_{a,b}$ forms the same angle θ with a horizontal line. Thus the slope of $L_{a,b}$, namely b , equals $\tan \theta$. From the trigonometric identity $1 + \tan^2 \theta \equiv \sec^2 \theta$ we get $\cos \theta = 1/\sqrt{1 + b^2}$, and the squared distance

$$(2) \quad d((X, Y), L_{a,b})^2 = \frac{(Y - a - bX)^2}{1 + b^2}.$$

Here two different although symmetric diagrams could be needed depending on whether $b > 0$ or $b < 0$. (By the way the French group of mathematical authors with the pseudonym Bourbaki decided that diagrams couldn't be part of proofs and decided to have no diagrams in their books.) Anyhow, here's another way to evaluate $d((X, Y), L_{a,b})^2$. We first find the line M through (X, Y) perpendicular to $L_{a,b}$, which has slope $-1/b$, so M is $y - Y = -(x - X)/b$. The intersection of M with $L_{a,b}$ gives $a + bx = Y - (x - X)/b$,

$$x = \xi = (Y - a + X/b)/(b + b^{-1}) = (bY - ab + X)/(b^2 + 1),$$

$$y = \eta = \xi + b\xi = (b^2Y + a + bX)/(b^2 + 1).$$

So the square of the distance from (X, Y) to $L_{a,b}$ is

$$\begin{aligned} (X - \xi)^2 + (Y - \eta)^2 &= [(b^2X - bY + ab)^2 + (Y - bX - a)^2]/(b^2 + 1)^2 \\ &= [b^2(Y - bX - a)^2 + (Y - bX - a)^2]/(b^2 + 1)^2 = (Y - bX - a)^2/(b^2 + 1), \end{aligned}$$

agreeing with the squared distance found in (2).

So $E(d((X, Y), L_{a,b})^2) = E((Y - bX - a)^2)/(b^2 + 1)$. For fixed b , the minimization to find a in terms of the other quantities is exactly as in y -on- x regression and gives the same result. Namely, we have a quadratic function of a , which goes to $+\infty$ as $|a|$ does. So it will be minimized at the unique point where the partial derivative with respect to a is 0, which gives $-2E(Y - bX) + 2a = 0$, or $a = EY - bEX$. This says that the point $E(X, Y) = (EX, EY)$ is on the line $L_{a,b}$, again, just as for y -on- x regression.

The line $L_{a,b}$ through (EX, EY) and the horizontal line $L_{EY,0}$ form some angles θ . As already indicated, we will take a θ such that the slope b equals $\tan \theta$. Recall that for any real number x , $\tan^{-1} x$ is an angle ϕ such that $\tan \phi = x$ and $-\pi/2 < \phi < \pi/2$. Then $\tan^{-1} x$ is uniquely defined since the tangent function is strictly increasing for $-\pi/2 < \theta < \pi/2$ and takes all real values there. The tangent function is periodic of period π . Thus, all angles ϕ such that $\tan \phi = x$ are of the form $\tan^{-1} x + m\pi$ where m is an integer, positive, negative or 0. On any interval of length π , containing just one of its endpoints, the tangent function takes all real values once each, and also goes to $\pm\infty$ at one point. It's convenient for present purposes to choose θ such that $-\pi/4 \leq \theta < 3\pi/4$, which is an interval of length π containing only its lower endpoint. For any real number (slope) b this gives a unique θ such that $\tan \theta = b$.

The squared distance from $L_{a,b}$ to (X, Y) is

$$[Y - EY - (\tan \theta)(X - EX)]^2/(1 + \tan^2 \theta)$$

which since $1 + \tan^2 \theta = 1/\cos^2 \theta$ equals

$$[(Y - EY) \cos \theta - (X - EX) \sin \theta]^2.$$

We want to find θ to minimize the expectation of this, which is

$$f(\theta) \equiv \sigma_Y^2 \cos^2 \theta - 2\text{Cov}(X, Y) \sin \theta \cos \theta + \sigma_X^2 \sin^2 \theta.$$

Since f is smooth and periodic of period 2π (actually, of period π because of the product and squaring), setting $f'(\theta) = 0$ we can expect to find at least one minimum and at least one maximum. They will turn out to be in perpendicular directions. We get

$$f'(\theta) = 2 \sin \theta \cos \theta (\sigma_X^2 - \sigma_Y^2) - 2 \cos(2\theta) \text{Cov}(X, Y).$$

If $\sigma_X^2 \neq \sigma_Y^2$ this gives $\tan(2\theta) = 2\text{Cov}(X, Y)/(\sigma_X^2 - \sigma_Y^2)$. There are two solutions for θ , namely θ_I given by (1) and $\theta_{II} = \theta_I + (\pi/2)$, since $\tan(\phi + \pi) = \tan \phi$ for any ϕ . Then $-\pi/4 < \theta_I < \pi/4 < \theta_{II} < 3\pi/4$, so both θ_I and θ_{II} are in the chosen interval for θ . A point where $f'(\theta) = 0$ will be a relative minimum if $f''(\theta) > 0$. We have

$$f''(\theta) = 2 \cos(2\theta)(\sigma_X^2 - \sigma_Y^2) + 2 \sin(2\theta) \cdot 2\text{Cov}(X, Y).$$

At a point where $f'(\theta) = 0$ this becomes $f''(\theta) = 2(\sigma_X^2 - \sigma_Y^2)/\cos(2\theta)$. If $\sigma_X^2 > \sigma_Y^2$ we want $\theta = \theta_I$ since $\cos(2\theta_I) > 0$ for a minimum, and θ_{II} will give a maximum. If $\sigma_X^2 < \sigma_Y^2$ we want $\theta = \theta_{II}$ so that $\pi/2 < 2\theta < 3\pi/2$ and $\cos(2\theta_{II}) < 0$ for a minimum, while θ_I then gives a maximum.

Now, what if $\sigma_X^2 = \sigma_Y^2$? In that case $f'(\theta) = -2 \cos(2\theta) \text{Cov}(X, Y)$. If $\text{Cov}(X, Y) = 0$, f is a constant and all θ , in other words all lines through (EX, EY) , are equally good. (Consider for example a bivariate normal distribution with mean 0 whose covariance matrix is σ^2 times the identity matrix. This distribution is rotationally invariant, so clearly all lines through the origin fit it equally well.)

If $\text{Cov}(X, Y) \neq 0$ then we need $\cos(2\theta) = 0$, so we can take $\theta = \pm\pi/4$. Again we need to consider the second derivative, which is $f''(\theta) = 4 \sin(2\theta) \text{Cov}(X, Y)$. To have $f''(\theta) > 0$ for a minimum of f , if $\text{Cov}(X, Y) > 0$ we want $\theta = \pi/4$, giving a line with slope 1. If $\text{Cov}(X, Y) < 0$ we want $\theta = -\pi/4$, giving a line with slope -1 . This completes the proof of Theorem 2. \square

When fitting a line to a finite sample $(x_1, y_1), \dots, (x_n, y_n)$, EX is replaced by \bar{x} , EY by \bar{y} , $\sigma^2 = \sigma_X^2$ by the sample variance s_X^2 , $\tau^2 = \sigma_Y^2$ by the sample variance s_Y^2 , and the covariance $\text{Cov}(X, Y)$ by the sample covariance defined as

$$\text{cov}(X, Y) = \frac{1}{n-1} \sum_{j=1}^n (X_j - \bar{X})(Y_j - \bar{Y})$$

in this case, using $1/(n-1)$ to fit with the usual definition of sample variances s_X^2 and s_Y^2 . (If $1/(n-1)$ is replaced by $1/n$ in both sample variances and the sample covariance, the result is the same since these factors cancel out, appearing both in the numerator and denominator of (1), as long as it's done consistently.)

Acknowledgment. Daniel Kane suggested the trigonometric formulation and result in February, 2005.