March 14, 2003

APPENDIX E: 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, measured (essentially) without error or with very small 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 $(X_1, Y_1), \ldots, (X_n, Y_n)$, with $n \geq 2$, let s_x be the sample standard deviation of the X_i, and s_y of the Y_i,

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

If $s_x = 0$ then the y-on-x line is not uniquely determined. Any line through $(\overline{X}, \overline{Y})$ 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, $(X_1, Y_1), \ldots, (X_n, Y_n)$, 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 $(\overline{X}, \overline{Y})$. The slope of the y-on-x line is $r \cdot s_y/s_x$ (Hogg and Tanis, 6th Ed., p. 241) 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 $r = -1$), as stated in Hogg and Tanis, p. 239, Q.E.D.

So, the two regression lines will in most cases be different. 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 $(x_1, y_1), \ldots, (x_n, y_n)$ 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 *bfd line (best-fitting by distance line)* if $E[d((X, Y), L)^2]$ is minimized at $L = L_o$. This will apply to data sets $(x_i, y_i),\ i = 1, \ldots, n$, by adding up probabilities $1/n$ at each point (x_i, y_i).

Let $\mathrm{Cov}(X, Y) = E(XY) - EXEY$, 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 $\rho = \rho_{X,Y} = \mathrm{Cov}(X, Y)/(\sigma_X \sigma_Y)$. Then $-1 \le \rho \le 1$.

Let $L_{a,b}$ be the line $y = ax + b$ 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 bfd 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 bfd 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 bfd line.

In all other cases the bfd line L is unique.

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

If $\sigma > 0,\ \tau > 0$ and $\rho \neq 0$ (the general case) then $L = L_{a_+,b_+}$ where

$$a_+ \ = \ [\tau^2 - \sigma^2 + \{(\sigma^2 - \tau^2)^2 + 4\rho^2\sigma^2\tau^2\}^{1/2}]/(2\rho\sigma\tau), \quad b_+ = EY - a_+ EX.$$

Proof. To find the distance from a point (X, Y) to a line L, if $L = L_{\infty;c}$ it's $|X - c|$. If $L = L_{0,b}$ it's $|Y - b|$. So suppose $L = L_{a,b}$ with $a \neq 0$. We first find the line through (X, Y) perpendicular to $L_{a,b}$, which has slope $-1/a$, so the line is $y - Y = -(x - X)/a$. The intersection of this with $L_{a,b}$ gives $ax + b = Y - (x - X)/a$,

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

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

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

$$(X - \xi)^2 + (Y - \eta)^2 \ = \ [(a^2 X - aY + ab)^2 + (Y - aX - b)^2]/(a^2 + 1)^2$$

$$= \ [a^2(Y - aX - b)^2 + (Y - aX - b)^2]/(a^2 + 1)^2 \ = \ (Y - aX - b)^2/(a^2 + 1).$$

So $E(d((X, Y), L_{a,b})^2) \ = \ E((Y - aX - b)^2)/(a^2 + 1)$. For fixed a, this is a quadratic function of b, and goes to $+\infty$ as $|b|$ does. So it will be minimized at the unique point where the partial derivative with respect to b is 0, which gives $-2E(Y - aX) + 2b = 0$, or $b = EY - aEX$. This says that the point $E(X, Y) = (EX, EY)$ is on the line $L_{a,b}$. Then we want to minimize

$$f(a) \ := \ E([Y - EY - a(X - EX)]^2)/(a^2 + 1) = (\tau^2 - 2a\mathrm{Cov}(X, Y) + a^2\sigma^2)/(a^2 + 1).$$

If $\sigma = \tau = 0$ then $f(a) \equiv 0$ and any line through (EX, EY) is bfd. Or if $\sigma = 0 < \tau$, then $f(a) = \tau^2/(a^2+1) > 0$ which is smallest as $a \to \pm\infty$. The bfd line is $L_{\infty,EX}$.

If $\sigma > 0 = \tau$ then $f(a) = a^2\sigma^2/(a^2+1)$ is clearly minimized when $a = 0$ and $L_{0,EY}$ is bfd.

Suppose then that $\sigma > 0$ and $\tau > 0$. Then

$$f(a) \;=\; [\tau^2 - 2a\rho\sigma\tau + a^2\sigma^2]/(a^2+1).$$

If $\rho = 0$ then $f(a) = \sigma^2 + (\tau^2 - \sigma^2)/(a^2+1)$, and:
(a) If $\sigma = \tau$ then $f(a) \equiv \sigma^2$ and all lines through (EX, EY) are bfd.
(b) If $\sigma^2 > \tau^2$ then f is minimized at $a = 0$ and $L_{0,EY}$ is the unique bfd line.
(c) If $\sigma^2 < \tau^2$ then f is smallest as $a \to \pm\infty$ and the unique bfd line is $L_{\infty,EX}$.

So suppose $\rho \neq 0$. Then setting $f'(a) = 0$ gives

$$0 = (a^2+1)(-2\rho\sigma\tau + 2a\sigma^2) - 2a(\tau^2 - 2a\rho\sigma\tau + a^2\sigma^2) = 2[\rho\sigma\tau a^2 + (\sigma^2 - \tau^2)a - \rho\sigma\tau],$$

and the factor of 2 on the right side can be cancelled since the expression equals 0. This quadratic in a has two distinct real roots,

$$a_{\pm} \;=\; [\tau^2 - \sigma^2 \pm \{(\sigma^2 - \tau^2)^2 + 4\rho^2\sigma^2\tau^2\}^{1/2}]/(2\rho\sigma\tau).$$

Next, $f'(0) = -2\rho\sigma\tau$. If $\rho > 0$ then $a_- < 0 < a_+$ and $f'(a) < 0$ for $a_- < a < a_+$ so a_+ gives a bfd line (minimum of $f(a)$). If $\rho < 0$ then $a_+ < 0 < a_-$ and $f'(a) > 0$ for $a_+ < a < a_-$ so again a_+ gives the bfd line, proving the Theorem. $\square$

When fitting a line to a finite sample $(x_1, y_1), \ldots, (x_n, y_n)$, EX is replaced by $\overline{x}$, EY by $\overline{y}$, σ^2 by $\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{n-1}{n}s_x^2$, τ^2 by $\frac{n-1}{n}s_Y^2$, and ρ by the sample correlation coefficient r.

If the distribution of (X, Y) is concentrated in a line $L_{a,b}$ with $\sigma > 0$ and $a \neq 0$, we have $\rho = +1$ if $a > 0$ and $\rho = -1$ if $a < 0$. Then $\tau = |a|\sigma$, $a_{\pm} = [\tau^2 - \sigma^2 \pm (\sigma^2 + \tau^2)]/(2\rho\sigma\tau)$, and $a_+ = \tau/(\rho\sigma) = |a|/\rho = a$ as it should.

REFERENCE

Hogg, R. V., and Tanis, E. A. (2001). *Probability and Statistical Inference*, 6th ed. Prentice Hall, Upper Saddle River, NJ.