Line with minimum weighted mean squared distance from a set of data points[1]

The sum of squared distances of n points $(x_{i\mathrm{con}}, y_{i\mathrm{con}})$, $i \in \{1, 2, 3, \ldots, n\}$, constrained to lie on a line of y-intercept a and slope b,

$$f(x_{i\mathrm{con}}, y_{i\mathrm{con}}; a, b) = a + b x_{i\mathrm{con}} - y_{i\mathrm{con}} = 0, \tag{1}$$

from n data points (x_i, y_i), with horizontal weights w_{x_i} and vertical weights w_{y_i},[2] is given by

$$S = \sum_{i=1}^{n} \left[w_{x_i} (x_{i\mathrm{con}} - x_i)^2 + w_{y_i} (y_{i\mathrm{con}} - y_i)^2 \right]. \tag{2}$$

Minimization of S subject to the constraints of Eq. (1), one for each point $(x_{i\mathrm{con}}, y_{i\mathrm{con}})$, is equivalent to minimization of

$$\mathcal{S} = S + \sum_{i=1}^{n} \lambda_i f(x_{i\mathrm{con}}, y_{i\mathrm{con}}; a, b) = \sum_{i=1}^{n} \left[w_{x_i} (x_{i\mathrm{con}} - x_i)^2 + w_{y_i} (y_{i\mathrm{con}} - y_i)^2 \right] + \sum_{i=1}^{n} \lambda_i (a + b x_{i\mathrm{con}} - y_{i\mathrm{con}}), \tag{3}$$

where the λ_i, $i \in \{1, 2, 3, \ldots, n\}$, are n Lagrange multipliers, one for each constraint. The combination of values of the $3n+2$ unknowns $x_{i\mathrm{con}}$, $y_{i\mathrm{con}}$, λ_i, a, and b that minimize $\mathcal{S}$ solve $3n+2$ equations obtained by setting the partial derivatives of $\mathcal{S}$ with respect to the unknowns equal to zero,

$$0 \stackrel{!}{=} \partial\mathcal{S}/\partial x_{i\mathrm{con}} = 2 w_{x_i} (x_{i\mathrm{con}} - x_i) + \lambda_i b, \tag{4}$$

$$0 \stackrel{!}{=} \partial\mathcal{S}/\partial y_{i\mathrm{con}} = 2 w_{y_i} (y_{i\mathrm{con}} - y_i) - \lambda_i, \qquad 0 \stackrel{!}{=} \partial\mathcal{S}/\partial \lambda_i = a + b x_{i\mathrm{con}} - y_{i\mathrm{con}}, \tag{5}$$

$$0 \stackrel{!}{=} \frac{\partial \mathcal{S}}{\partial a} = \sum_{i=1}^{n} \lambda_i, \qquad 0 \stackrel{!}{=} \frac{\partial \mathcal{S}}{\partial b} = \sum_{i=1}^{n} \lambda_i x_{i\mathrm{con}}. \tag{6}$$

Elimination of $y_{i\mathrm{con}}$ from Eq. (5) gives

$$\lambda_i = 2 w_{y_i} (a + b x_{i\mathrm{con}} - y_i). \tag{7}$$

Use of Eq. (7) to eliminate λ_i from Eqs. (4) and (6) gives

$$0 = \left(w_{x_i} + b^2 w_{y_i} \right) x_{i\mathrm{con}} - w_{x_i} x_i + a b w_{y_i} - b w_{y_i} y_i, \tag{8}$$

$$0 = \sum_{i=1}^{n} w_{y_i} (a + b x_{i\mathrm{con}} - y_i), \qquad 0 = \sum_{i=1}^{n} w_{y_i} (a + b x_{i\mathrm{con}} - y_i)\, x_{i\mathrm{calc}}. \tag{9}$$

Use of Eq. (8) to eliminate $x_{i\mathrm{con}}$ from Eq. (9) gives

$$a = \left[\sum_{i=1}^{n} \frac{w_{x_i} w_{y_i} y_i}{w_{x_i} + b^2 w_{y_i}} - b \sum_{i=1}^{n} \frac{w_{x_i} w_{y_i} x_i}{w_{x_i} + b^2 w_{y_i}} \right] \Bigg/ \sum_{j=1}^{n} \frac{w_{x_j} w_{y_j}}{w_{x_j} + b^2 w_{y_j}}, \tag{10}$$

$$0 = a^2 b \sum_{i=1}^{n} \frac{w_{x_i} w_{y_i}^2}{\left(w_{x_i} + b^2 w_{y_i} \right)^2} - a \sum_{i=1}^{n} \frac{w_{x_i} w_{y_i} \left(w_{x_i} x_i + 2 b w_{y_i} y_i - b^2 w_{y_i} x_i \right)}{\left(w_{x_i} + b^2 w_{y_i} \right)^2} + \sum_{i=1}^{n} \frac{w_{x_i} w_{y_i} \left(w_{x_i} x_i y_i - b w_{x_i} x_i^2 + b w_{y_i} y_i^2 - b^2 w_{y_i} x_i y_i \right)}{\left(w_{x_i} + b^2 w_{y_i} \right)^2}. \tag{11}$$

Multiplication of Eq. (11) by $\left(\sum_{k=1}^{n} \frac{w_{x_k} w_{y_k}}{w_{x_k} + b^2 w_{y_k}} \right)^2$ and use of Eq. (10) to eliminate a gives

$$\begin{aligned} 0 = {} & b^3 \sum_{k=1}^{n} \frac{w_{x_k} w_{y_k} x_k}{w_{x_k} + b^2 w_{y_k}} \left[\sum_{j=1}^{n} \frac{w_{x_j} w_{y_j} x_j}{w_{x_j} + b^2 w_{y_j}} \sum_{i=1}^{n} \frac{w_{x_i} w_{y_i}^2}{\left(w_{x_i} + b^2 w_{y_i} \right)^2} - \sum_{j=1}^{n} \frac{w_{x_j} w_{y_j}}{w_{x_j} + b^2 w_{y_j}} \sum_{i=1}^{n} \frac{w_{x_i} w_{y_i}^2 x_i}{\left(w_{x_i} + b^2 w_{y_i} \right)^2} \right] \\ & + b^2 \left\{ \sum_{k=1}^{n} \frac{w_{x_k} w_{y_k}}{w_{x_k} + b^2 w_{y_k}} \left[2 \sum_{j=1}^{n} \frac{w_{x_j} w_{y_j} x_j}{w_{x_j} + b^2 w_{y_j}} \sum_{i=1}^{n} \frac{w_{x_i} w_{y_i}^2 y_i}{\left(w_{x_i} + b^2 w_{y_i} \right)^2} + \sum_{j=1}^{n} \frac{w_{x_j} w_{y_j} y_j}{w_{x_j} + b^2 w_{y_j}} \sum_{i=1}^{n} \frac{w_{x_i} w_{y_i}^2 x_i}{\left(w_{x_i} + b^2 w_{y_i} \right)^2} \right] \right. \\ & \left. - 2 \sum_{k=1}^{n} \frac{w_{x_k} w_{y_k} y_k}{w_{x_k} + b^2 w_{y_k}} \sum_{j=1}^{n} \frac{w_{x_j} w_{y_j} x_j}{w_{x_j} + b^2 w_{y_j}} \sum_{i=1}^{n} \frac{w_{x_i} w_{y_i}^2}{\left(w_{x_i} + b^2 w_{y_i} \right)^2} - \left[\sum_{k=1}^{n} \frac{w_{x_k} w_{y_k}}{w_{x_k} + b^2 w_{y_k}} \right]^2 \sum_{i=1}^{n} \frac{w_{x_i} w_{y_i}^2 x_i y_i}{\left(w_{x_i} + b^2 w_{y_i} \right)^2} \right\} \end{aligned} \tag{12}$$

(continued)

---

[1] W. Edwards Deming, *Statistical Adjustment of Data*, Dover, New York 1964. (This is an unabridged and corrected republication of the Second Edition, Wiley, New York 1943; First Edition, 1938.)

[2] Deming, Page 21, states that "By definition, the weight w_f of the function f is inversely proportional to the variance σ_f^2 of f. That is to say, $1/w_f$ is the variance coefficient of f. In symbols, $w_f = \sigma^2/\sigma_f^2$ or $\sigma_f^2 = \sigma^2/w_f$. σ^2 is simply a proportionality factor, and is evidently the variance of a function of unit weight."

$$+b\left\{\sum_{k=1}^{n}\frac{w_{x_k}w_{y_k}}{w_{x_k}+b^2w_{y_k}}\left[\sum_{j=1}^{n}\frac{w_{x_j}w_{y_j}x_j}{w_{x_j}+b^2w_{y_j}}\sum_{i=1}^{n}\frac{w_{x_i}^2w_{y_i}x_i}{\left(w_{x_i}+b^2w_{y_i}\right)^2}-2\sum_{j=1}^{n}\frac{w_{x_j}w_{y_j}y_j}{w_{x_j}+b^2w_{y_j}}\sum_{i=1}^{n}\frac{w_{x_i}w_{y_i}^2y_i}{\left(w_{x_i}+b^2w_{y_i}\right)^2}\right]\right.$$
$$\left.+\left[\sum_{j=1}^{n}\frac{w_{x_j}w_{y_j}}{w_{x_j}+b^2w_{y_j}}\right]^2\left[\sum_{i=1}^{n}\frac{w_{x_i}w_{y_i}^2y_i^2}{\left(w_{x_i}+b^2w_{y_i}\right)^2}-\sum_{i=1}^{n}\frac{w_{x_i}^2w_{y_i}x_i^2}{\left(w_{x_i}+b^2w_{y_i}\right)^2}\right]+\left[\sum_{j=1}^{n}\frac{w_{x_j}w_{y_j}y_j}{w_{x_j}+b^2w_{y_j}}\right]^2\sum_{i=1}^{n}\frac{w_{x_i}w_{y_i}^2}{\left(w_{x_i}+b^2w_{y_i}\right)^2}\right\}$$
$$+\sum_{k=1}^{n}\frac{w_{x_k}w_{y_k}}{w_{x_k}+b^2w_{y_k}}\left[\sum_{j=1}^{n}\frac{w_{x_j}w_{y_j}}{w_{x_j}+b^2w_{y_j}}\sum_{i=1}^{n}\frac{w_{x_i}^2w_{y_i}x_iy_i}{\left(w_{x_i}+b^2w_{y_i}\right)^2}-\sum_{j=1}^{n}\frac{w_{x_j}w_{y_j}y_j}{w_{x_j}+b^2w_{y_j}}\sum_{i=1}^{n}\frac{w_{x_i}^2w_{y_i}x_i}{\left(w_{x_i}+b^2w_{y_i}\right)^2}\right].$$

The occurences of b^2 in the denominators within the sums make Eq. (12) much more complicated than a cubic equation for the slope b. Except for special cases, no general algebraic solution of Eq. (12) exists. Also, a slope b that solves Eq. (12) does not necessarily minimize S.

If each vertical weight is the same multiple c of the corresponding horizontal weight,

$$w_{y_i}=cw_{x_i},\qquad\qquad i\in\{1,2,3,\ldots,n\}\,, \tag{13}$$

then $w_{x_i}/\left(w_{x_i}+b^2w_{y_i}\right)=1/\left(1+b^2c\right)$ factors out of sums and Eqs. (10) and (12) reduce to

$$a=\frac{1}{w}\sum_{i=1}^{n}w_{x_i}y_i-\frac{b}{w}\sum_{i=1}^{n}w_{x_i}x_i, \tag{14}$$

$$0=b^2cZ+b\left(X-cY\right)-Z, \tag{15}$$

where w is total horizontal weight,[3]

$$w\equiv\sum_{i=1}^{n}w_{x_i}=w_{x_1}+w_{x_2}+w_{x_3}+\cdots+w_{x_n}, \tag{16}$$

and X, Y, and Z are weighted x-variance, weighted y-variance, and weighted (x,y)-covariance,[4]

$$X\equiv\frac{1}{w}\sum_{i=1}^{n}w_{x_i}x_i^2-\left(\frac{1}{w}\sum_{i=1}^{n}w_{x_i}x_i\right)^2,\quad Y\equiv\frac{1}{w}\sum_{i=1}^{n}w_{x_i}y_i^2-\left(\frac{1}{w}\sum_{i=1}^{n}w_{x_i}y_i\right)^2,\quad Z\equiv\frac{1}{w}\sum_{i=1}^{n}w_{x_i}x_iy_i-\frac{1}{w}\sum_{i=1}^{n}w_{x_i}x_i\frac{1}{w}\sum_{j=1}^{n}w_{x_j}y_j. \tag{17}$$

Both solutions of Eq. (15) for the slope b are given by the quadratic formula,

$$b=\frac{1}{2cZ}\left[cY-X\pm\sqrt{\left(X-cY\right)^2+4cZ^2}\,\right], \tag{18}$$

and the sum of Eq. (2) may be expressed as

$$\begin{aligned}S&=\frac{cw}{1+b^2c}\left[b^2\frac{1}{w}\sum_{i=1}^{n}w_{x_i}x_i^2-2b\frac{1}{w}\sum_{i=1}^{n}w_{x_i}x_iy_i+\frac{1}{w}\sum_{i=1}^{n}w_{x_i}y_i^2+2ab\frac{1}{w}\sum_{i=1}^{n}w_{x_i}x_i+a^2-2a\frac{1}{w}\sum_{i=1}^{n}w_{x_i}y_i\right]\\&=\frac{cw}{1+b^2c}\left(b^2X-2bZ+Y\right)=\frac{w}{2}\left[X+cY\mp\sqrt{\left(X-cY\right)^2+4cZ^2}\,\right],\end{aligned} \tag{19}$$

where the first equality uses Eqs. (5), (8), (13), and (16), the second equality uses Eqs. (14) and (17), and the third equality uses Eq. (18). Upper signs in front of the square roots are for one line and lower signs are for another line.

The line with minimum sum of weighted squared distances from the data points has sum of weighted squared distances

$$S=\frac{w}{2}\left[X+cY-\sqrt{\left(X-cY\right)^2+4cZ^2}\,\right], \tag{20}$$

slope[5]

$$b=\frac{1}{2cZ}\left[cY-X+\sqrt{\left(X-cY\right)^2+4cZ^2}\,\right], \tag{21}$$

and y-intercept given by use of b from Eq. (21) in Eq. (14).

---

[3] Use of Eqs. (13) and (16) gives total vertical weight $\sum_{i=1}^{n}w_{y_i}=\sum_{i=1}^{n}cw_{x_i}=c\sum_{i=1}^{n}w_{x_i}=cw$.

[4] Use of Eq. (13) to replace horizontal weights in Eq. (17) by vertical weights gives $X=\frac{1}{cw}\sum_{i=1}^{n}w_{y_i}x_i^2-\left(\frac{1}{cw}\sum_{i=1}^{n}w_{y_i}x_i\right)^2$, $Y=\frac{1}{cw}\sum_{i=1}^{n}w_{y_i}y_i^2-\left(\frac{1}{cw}\sum_{i=1}^{n}w_{y_i}y_i\right)^2$, and $Z=\frac{1}{cw}\sum_{i=1}^{n}w_{y_i}x_iy_i-\frac{1}{cw}\sum_{i=1}^{n}w_{y_i}x_i\frac{1}{cw}\sum_{j=1}^{n}w_{y_j}y_j$.

[5] Deming, Page 184, states that "This is equivalent to a result obtained by Kummell in 1876, Karl Pearson in 1901, and Gini in 1921." Deming, as on pages 64, 140, 145, might have intended to refer to "Reduction of observation equations which contain more than one observed quantity", Charles H. Kummell, *The Analyst* (Des Moines) **6**, 97-105 (1879).

The slope of the residual line segment from point $(x_{i\mathrm{con}}, y_{i\mathrm{con}})$ to data point (x_i, y_i) is given by

$$\frac{y_i - y_{i\mathrm{con}}}{x_i - x_{i\mathrm{con}}} = -\frac{w_{x_i}}{b w_{y_i}} = -\frac{1}{bc}, \tag{22}$$

where the first equality uses Eqs. (4) and (5) and the second equality uses Eq. (13). Residual line segments that have the same ratio w_{y_i}/w_{x_i} have the same slope.

In the limit $c \to 0$, all residual line segments are vertical and S divided by total vertical weight cw becomes the minimum weighted mean squared vertical distance of a line from the set of data points,

$$d^2_{\mathrm{vrmsmin}} = \lim_{c\to 0} \frac{S}{cw} = Y - \frac{Z^2}{X}. \tag{23}$$

In the limit $c \to \infty$, all residual line segments are horizontal and S divided by total horizontal weight w becomes the minimum weighted mean squared horizontal distance of a line from the set of data points,

$$d^2_{\mathrm{hrmsmin}} = \lim_{c\to \infty} \frac{S}{w} = X - \frac{Z^2}{Y}. \tag{24}$$

If horizontal dimension and vertical dimension are equal, then b and c are dimensionless and the squared length of the residual line segment is

$$d_i^2 = (x_i - x_{i\mathrm{con}})^2 + (y_i - y_{i\mathrm{con}})^2 = (x_i - x_{i\mathrm{con}})^2 + (x_i - x_{i\mathrm{con}})^2/\left(b^2c^2\right) = (a + bx_i - y_i)^2\left(1 + b^2c^2\right)/\left(1 + b^2c\right)^2, \tag{25}$$

where the second equality uses Eq. (22), the third equality uses Eqs. (8) and (13).

The minimum weighted mean squared length of the residual line segments, or minimum weighted mean squared distance of the line from the set of data points, is

$$\begin{aligned} d^2_{\mathrm{crmsmin}} &= \frac{1}{w}\sum_{i=1}^{n} w_{x_i} d_i^2 = \frac{1}{w}\sum_{i=1}^{n} w_{x_i} \frac{1+b^2c^2}{(1+b^2c)^2}(a + bx_i - y_i)^2 = \frac{1+b^2c^2}{(1+b^2c)^2}\frac{1}{w}\sum_{i=1}^{n} w_{x_i}(a + bx_i - y_i)^2 \\ &= \frac{1+b^2c^2}{(1+b^2c)^2}\left(b^2X - 2bZ + Y\right) = \frac{1}{2}(X+Y) - \frac{(X-Y)(X-cY) + 2(1+c)Z^2}{2\sqrt{(X-cY)^2 + 4cZ^2}}, \end{aligned} \tag{26}$$

where the second equality uses Eq. (25), the fourth equality uses Eqs. (14) and (17), and the fifth equality uses Eq. (21).

Comparison of the factor $b^2X - 2bZ + Y$ in Eqs. (19) and (26) shows that d^2_{crmsmin} may be expressed as

$$d^2_{\mathrm{crmsmin}} = \frac{1+b^2c^2}{(1+b^2c)\,c}\frac{S}{w}, \tag{27}$$

where S is given in Eq. (20) and b is given in Eq. (21); that is, d^2_{crmsmin} equals S divided by a total weight

$$\frac{\left(1+b^2c\right)c}{1+b^2c^2}\,w = \frac{(X-cY)^2 + 4cZ^2 - (X-cY)\sqrt{(X-cY)^2 + 4cZ^2}}{(X-cY)^2 + 2(1+c)Z^2 - (X-cY)\sqrt{(X-cY)^2 + 4cZ^2}}\,w. \tag{28}$$

The ratio of this total weight divided by total vertical weight cw equals unity in the limit $c \to 0$ and the ratio of this total weight divided by total horizontal weight w equals unity in the limit $c \to \infty$.

In the limit $c \to 0$, all residual line segments are vertical and, as in Eq. (23), Eq. (26) becomes the minimum weighted mean squared vertical distance of a line from the set of data points,

$$d^2_{\mathrm{vrmsmin}} = \lim_{c\to 0} d^2_{\mathrm{crmsmin}} = \lim_{c\to 0} \frac{1+b^2c^2}{(1+b^2c)\,c}\frac{S}{w} = Y - Z^2/X. \tag{29}$$

In the limit $c \to \infty$, all residual line segments are horizontal and, as in Eq. (24), Eq. (26) becomes the minimum weighted mean squared horizontal distance of a line from the set of data points,

$$d^2_{\mathrm{hrmsmin}} = \lim_{c\to \infty} d^2_{\mathrm{crmsmin}} = \lim_{c\to \infty} \frac{1+b^2c^2}{(1+b^2c)\,c}\frac{S}{w} = X - Z^2/Y. \tag{30}$$

For $c = 1$, all residual line segments have slope $-1/b$ and are perpendicular to the line of slope b and Eq. (26) reduces to the minimum weighted mean squared perpendicular distance of a line from the set of data points,

$$d^2_{\mathrm{rmsmin}} = \lim_{c\to 1} d^2_{\mathrm{crmsmin}} = \lim_{c\to 1} \frac{1+b^2c^2}{(1+b^2c)\,c}\frac{S}{w} = \frac{1}{2}\left[X + Y - \sqrt{(X-Y)^2 + 4Z^2}\,\right]. \tag{31}$$