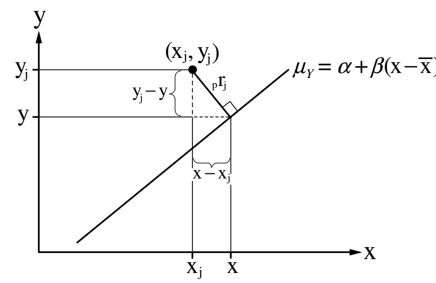
University of Waterloo W. H. Cherry

REGRESSION: Minimizing *Perpendicular* Distances

Regression models are widely used in the Analysis stage of the FDEAC cycle; most commonly, such models are fitted to bi- or multivariate data by the method of *least squares*, involving minimizing the sum of the squared residuals (*i.e.*, the *vertical* distances of the data points from the regression model). Routine use of such models, coupled with the idea that this is the case of 'best' fit to the data, makes it easy to forget that *this* use of 'least squares' is, in part, chosen for mathematical convenience and that there are *other* choices that could also reasonably be regarded as 'best' fit were they not less convenient. Figure 13.3 in the STAT 221 Course Materials (on page 13.21) compares, for a straight-line model, the regression of **Y** on **X**, the regression of **X** on **Y** and the so-called s.d. line – the latter is the line whose slope is the standard deviation of **Y** divided by that of **X**. At least for a data set (like the parent-child height data in Statistical Highlight #30) which can be modelled by a bivariate normal distribution, it can be argued that, of the three choices, the s.d. line is *visually* the line of 'best' fit – it seems to lie closest to the major axis of the elliptical point cloud, at least when the line has slope around 1.

The purpose of this Highlight #54 is to add to the previous comparison the line of 'best' fit from minimizing the squares of the *perpendicular* distances of the data points from the model line. Some aspects of the comparison are also pursued in Statistical Highlight #55.



1. Minimizing squared perpendicular distances

The diagram at the right shows part of the straight line model which, like like the two regression lines and the s.d. line, passes through the point of averages $(\bar{x}, \bar{y})$, which is *not* shown in the diagram. The j*th* data point (x_j, y_j) is shown above the model line and a line runs from this point perpendicular to the model line – it is labelled ${}_p r_j$ with the affix 'p' reminding us it is a *perpendicular* residual; the point of intersection of the two lines is (x, y). The diagram shows that ${}_p r_j^2 = (x - x_j)^2 + (y_j - y)^2$ [Pythagoras]. We set out the minimization process in thirteen steps.

Step 1: The equation of the model line is equation (HL54.1), which because this line passes through $(\bar{x}, \bar{y})$, becomes equation (HL54.2).

$$\mu_Y = \alpha + \beta(x_j - \bar{x}) \qquad \text{-----(HL54.1)}$$

$$\mu_Y = \bar{y} + \beta(x_j - \bar{x}) \qquad \text{-----(HL54.2)}$$

Step 2: The equation of the perpendicular line is equation (HL54.3). [The product of the slopes of perpendicular lines is −1.] Because this line passes through (x_j, y_j), it becomes equation (HL54.4).

$$y = \delta - \tfrac{1}{\beta}x \qquad \text{-----(HL54.3)}$$

$$y = y_j + \tfrac{1}{\beta}x_j - \tfrac{1}{\beta}x \qquad \text{-----(HL54.4)}$$

Step 3: At the point of intersection of the two lines, we can equate the right-hand sides of equations (HL54.2) and (HL54.4) to give equation (HL54.5).

$$\bar{y} + \beta(x_j - \bar{x}) = y_j + \tfrac{1}{\beta}x_j - \tfrac{1}{\beta}x \qquad \text{-----(HL54.5)}$$

or: $$x(\beta + \tfrac{1}{\beta}) = -\bar{y} + y_j + \tfrac{1}{\beta}x_j + \beta\bar{x} \qquad \text{-----(HL54.6)}$$

Step 4: Rearranging the terms of equation (HL54.6), and subtracting βx_j from each side, yields equation (HL54.7).

$$(x - x_j)(\beta + \tfrac{1}{\beta}) = (y_j - \bar{y}) - \beta(x_j - \bar{x}) \qquad \text{-----(HL54.7)}$$

Step 5: Isolating $(x - x_j)$ from equation (HL54.7) and squaring gives equation (HL54.8). [This is the squared length of the dashed horizontal side of the right-angled triangle containing ${}_p r_j$.]

$$(x - x_j)^2 = \frac{\beta^2}{(\beta^2+1)^2}\left[(y_j - \bar{y})^2 - 2\beta(y_j - \bar{y})(x_j - \bar{x}) + \beta^2(x_j - \bar{x})^2\right] \qquad \text{-----(HL54.8)}$$

Step 6: Using the notation: $SS_x = \sum_{j=1}^{n}(x_j - \bar{x})^2$, $SS_{xy} = \sum_{j=1}^{n}(x_j - \bar{x})(y_j - \bar{y})$ and $SS_y = \sum_{j=1}^{n}(y_j - \bar{y})^2$ and summing over the (sample of) n data points, equation (HL54.8) becomes equation (HL54.9).

$$\sum_{j=1}^{n}(x - x_j)^2 = \frac{1}{(\beta^2+1)^2}\left[\beta^2 SS_y^2 - 2\beta^3 SS_{xy} + \beta^4 SS_x^2\right] \qquad \text{-----(HL54.9)}$$

Step 7: Using the expression for x from equation (HL54.6) in equation (HL54.4) gives equation (HL54.10).

$$y = y_j + \tfrac{1}{\beta}x_j - \tfrac{1}{\beta}x - \frac{1}{(\beta^2+1)^2}\left[(-\bar{y} + y_j + \tfrac{1}{\beta}x_j + \beta\bar{x}\right] \qquad \text{-----(HL54.10)}$$

or: $$(y_j - y)(\beta^2+1) = (y_j - \bar{y}) + \beta(x_j - \bar{x}) \qquad \text{-----(HL54.11)}$$

Step 8: Isolating $(y_j - y)$ from equation (HL54.11) and squaring gives equation (HL54.12). [This is the squared length of the dashed verticalal side of the right-angled triangle containing ${}_p r_j$.]

$$(y_j - y)^2 = \frac{1}{(\beta^2+1)^2}\left[(y_j - \bar{y})^2 - 2\beta(y_j - \bar{y})(x_j - \bar{x}) + \beta^2(x_j - \bar{x})^2\right] \qquad \text{-----(HL54.12)}$$

Step 9: Using the notation of Step 6 and summing over the (sample of) n data points, equation (HL54.12) becomes equation (HL54.13).

$$\sum_{j=1}^{n}(y_j - y)^2 = \frac{1}{(\beta^2+1)^2}\left[SS_y - 2\beta SS_{xy} + \beta^2 SS_x\right] \qquad \text{-----(HL54.13)}$$

Step 10: The sum of the squares of the (perpendicular) residuals is the *combination* of equations (HL54.9) and (HL54.13), which simplifies to equation (HL54.14).

$$\sum_{j=1}^{n}{}_p r_j^2 = \frac{1}{\beta^2+1}\left[SS_y - 2\beta SS_{xy} + \beta^2 SS_x\right] \qquad \text{-----(HL54.14)}$$

(*continued overleaf*)

Step 11: Differentiating $\Sigma_p r_j^2$ from Step 10 with respect to β, we obtain (after simplifying the right-hand side) equation (HL54.15).

$$\frac{d\Sigma_p r_j^2}{d\beta} = \frac{2}{(\beta^2+1)^2}[\beta^2 SS_{xy} + \beta(SS_x - SS_y) - SS_{xy}] \qquad \text{-----(HL54.15)}$$

Step 12: The derivative in equation (HL54.15) is a quadratic equation in β with roots (which make it *zero*) as in equation [HL54.16) [These roots multiply to –1 – they are the slopes of *perpendicular* lines.]

$$\beta = \frac{1}{2SS_{xy}}[SS_y - SS_x \pm \sqrt{(SS_x - SS_y)^2 + 4SS_{xy}^2}] \qquad \text{-----(HL54.16)}$$

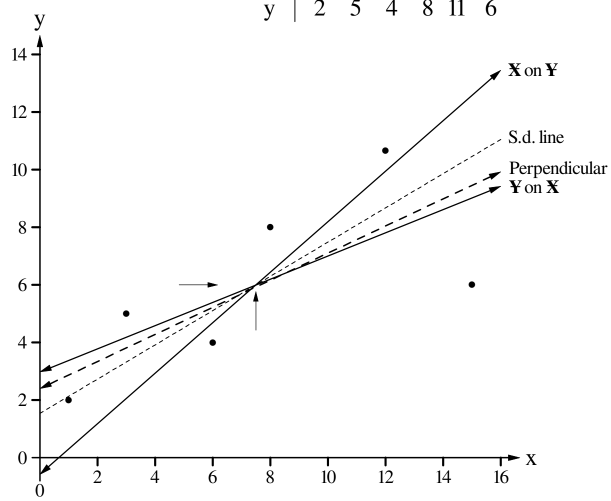
Step 13: Differentiating in equation (HL54.15) to obtain the *second* derivative yields (after simplification) equation (HL54.17). [It can be used to confirm (when *positive*) that a root in equation (HL54.16) *is* at a minimum.]

$$\frac{d^2\Sigma_p r_j^2}{d\beta^2} = \frac{-2}{(\beta^2+1)^3}[2\beta(\beta^2+1)SS_{xy} + (3\beta^2-1)(SS_x - SS_y)] \qquad \text{-----(HL54.17)}$$

2. A comparison of four lines fitted to the same data

A small (hypothetical) data set is tabulated and plotted as a scatter diagram at the right, together with four lines discussed below; its numerical summaries are:

x	1	3	6	8	12	15
y	2	5	4	8	11	6



$\Sigma x_j = 45$, $\Sigma x_j^2 = 479$; $\Sigma x_j y_j = 327$; $SS_{xy} = 327 - 45{\times}36/6 = 57$,
$\Sigma y_j = 36$, $\Sigma y_j^2 = 266$; $(n = 6)$ $SS_x = 479 - 45^2/6 = 141½$,
$\overline{x} = 7½$, $\overline{y} = 6$, $SS_y = 266 - 36^2/6 = 50$;

As shown on page 13.19 in STAT 221, the estimates of β_1 (the slope) and β_0 (the intercept) of the regression of **Y** on **X** are:

$b_1 = \frac{57}{141½} = 0.402\,826\,855$, $b_0 = 6 - b_1{\times}7½ = 2.978\,798\,587$;

so the equation of the fitted straight line is:

${}_{\text{reg}}\overline{y} = 2.9788 + 0.4028x = 6 + 0.4028(x - 7½)$.

Similarly, from page 13.21 in STAT 221, the estimates of γ_1 (the slope) and γ_0 (the intercept) of the regression of **X** on **Y** are:

$g_1 = \frac{57}{50} = 1.14$, $g_0 = 7½ - g_1{\times}6 = 0.66$;

so the equation of the fitted straight line is:

${}_{\text{reg}}\overline{x} = 0.66 + 1.14y = 7½ + 1.14(y - 6)$, or: ${}_{\text{reg}}\overline{y} = -0.5789 + 0.8772x$.

For perpendicular residuals, the positive root from equation (HL54.16) gives slope $\beta = 0.479\,638\,849$ and intercept $6 - \beta{\times}7½ = 2.402\,708\,630$;
so the equation of the fitted straight line is: ${}_{\text{reg}}\overline{y} = 2.4027 + 0.4796x = 6 + 0.4796(x - 7½)$.

For the s.d. line, the slope is $\sqrt{50/141½} = 0.594\,438\,298$ and the intercept is $6 - 0.594\,438\,298{\times}7½ = 1.541\,712\,760$;
so its equation is: $y = 1.5417 + 0.5944x = 6 + 0.5944(x - 7½)$.

The diagram at the right above illustrates several noteworthy matters – the six data points were chosen with sufficient scatter to separate the lines enough for clarity in the diagram (the correlation coefficient is $r = 0.677\,659\,660$).

- The intersection of the four lines is (of course) at the *point of averages*, indicated by the two horizontal and vertical arrows.
- Even for this simple daa set, the s.d. line seems visually to be the 'best' fit to the six data points.
- The two regressions lines, **Y** on **X** and **X** on **Y**, fall symmetrically on either side of the s.d. line – their slopes are, respectively, r and 1/r times that of the s.d. line (as discussed on page 13.21 in the STAT 221 Course Materials).
- The regression based on perpendicular residuals is similar to that of **Y** on **X** but closer to the s.d. line and so is visually a somewhat better fit.
  - More quantitatively, a measure of the 'success' of a regression model is to express, as a percentage (between 0 and 100) of SS_y, the total initial variation in the ys (as measured by SS_y) minus the sum of the squared residuals – this is the *coefficient of determination.* The respective values here are about 54.7% and 45.9%, confirming our more favourable *visual* assessment of the perpendicular residual model. However, both percentages are low enough to show that neither model is particularly 'successful' (reflecting the wide scatter of the data about a straight-line model).
  - It seems likely that the differing 'success' of the regression of **Y** on **X** and that using *perpendicular* residuals depends on *slope* – the greater its magnitude (*i.e.*, the steeper the line), the more the coefficient of determination will favour the 'success' of the line based on perpendicular (rather than vertical) residuals.
  - If the model is a *curve* rather than a straight line, the greater sensitivity to slope of vertical (as opposed to perpendicular) residual length may affect fitted curve *shape* under the two criteria; the greater effect of steeper slopes would be on the left (smaller x values) for curves that are concave down, on the right (larger x values) for concave up curves.
    - More complex curve shapes (like a sine wave) could be considered as a succession of the two simpler shapes.

The foregoing discussion in this Highlight #54 is a reminder of the maxim of the late Dr. George E.P. Box, a respected U.S. statistician: *All models are wrong, some are useful*, to which we might now add: *but not necessarily equally so.*