NOTES

ON STANDARD ERROR FOR THE LINE OF MUTUAL REGRESSION

By Y. K. Wong

1. In Pearson's *On Lines and Planes of Closest Fit to System of Points in Space*, he establishes a formula for the mean square residual for the best fitting line in q-space:

$$(1) \qquad (\text{mean sq. residual})^2 = \sigma_{x_1}^2 + \cdots + \sigma_{x_q}^2 - \Delta R_{\max}^2$$

where $2R_{\max}$ is the length of the maximum axis of the correlation ellipse in q-space, and Δ is the correlation determinant.[1]

In the present paper, we consider a 2-dimensional case, and shall call the mean sq. residual as the standard error, denoted by S_N.

In 2-dimensional space, a correlation ellipse is

$$(2) \qquad ax^2 + 2hxy + by^2 + c = 0,$$

where

$$(2a) \quad a = \sigma_y^2, \quad b = \sigma_x^2, \quad h = -r_{xy}\sigma_x\sigma_y = -p_{xy} = -p_{yx}, \quad c = -\sigma_x^2\sigma_y^2.$$

Pearson gives in the 2-dimensional space the following formula for S_N:

$$(3) \qquad S_N = \sigma_x\sigma_y/\text{semi-major axis of equation (2)}.$$

Expression (3) can be readily deduced from (1). This paper aims to present some formulae for S_N, more convenient for practical computation, and also call attention to a misprint in Pearson's paper.

2. From analytic geometry, we see that the angle φ, between the major axis of the ellipse (2) and the x-axis is given by

$$(4) \qquad \tan 2\varphi = 2h/(a - b).$$

By rotation of the axes, equation (1) can be written in the form

$$(5) \qquad a'x^2 + b'y^2 + c = 0,$$

where

$$(5a) \qquad \begin{aligned} a' &= a\cdot\cos^2\varphi - 2h\cdot\sin\varphi\cdot\cos\varphi - b\cdot\sin^2\varphi > 0 \\ b' &= a\cdot\sin^2\varphi - 2h\cdot\sin\varphi\cdot\cos\varphi - b\cdot\cos^2\varphi > 0. \end{aligned}$$

[1] Philosophical Magazine, 6th Series, II (November, 1901), p. 559.

LEMMA 1. The value of a' given by (5a) is less than b'.

To prove this lemma, we find from (4) and (5)

$$a' - b' = a + b, \qquad a' - b' = 2h/\sin 2\varphi = -2p_{xy}/\sin 2\varphi,$$

and hence

$$(6) \qquad 2a' = a + b - 2p_{xy}/\sin 2\varphi, \qquad 2b' = a + b + 2p_{xy}/\sin 2\varphi.$$

Since both a and b are positive, the lemma will be proved if we can show that $p_{xy}/\sin 2\varphi$ is a positive quantity. By (2a), $p_{xy} = r_{xy}\sigma_x\sigma_y$, in which σ_x, σ_y are positive; hence the sign of p depends upon the sign of r. If $r_{xy} < 0$, then $\varphi > \frac{\pi}{2}$, and 2φ is of such a nature that $\frac{3\pi}{2} < 2\varphi < 2\pi$. It follows $\sin 2\varphi < 0$, and hence $p_{xy}/\sin 2\varphi$ is positive. On the other hand, if $r_{xy} > 0$, then φ is such that $0 < 2\varphi < \pi$, and hence $\sin 2\varphi > 0$. It follows that $p_{xy}/\sin 2\varphi$ is positive independent of the sign of r_{xy}.

LEMMA 2. The square of the mean square residual is equal to a', and hence

$$S_N^2 = \sigma_y^2 \cos^2\varphi - 2p_{xy}\sin\varphi\cos\varphi + \sigma_x^2\sin^2\varphi = \tfrac{1}{2}(\sigma_x^2 - \sigma_y^2) - p_{xy}/\sin 2\varphi.$$

For from (5), we obtain (semi-major axis)$^2 = -c/a' = +\dfrac{\sigma_x^2\sigma_y^2}{a'}$. Substituting this into (3), we obtain $S_N = a'$. The balance of the lemma follows from (5a), (6), and (2a).

LEMMA 3. For every r_{xy}, we have

$$(7) \qquad \sin 2\varphi = p_{xy}/\sqrt{K}, \qquad K = (\sigma_x^2 - \sigma_y^2)^2 + 4p_{xy}^2.$$

For, from (4), we find $\sin 2\varphi = -p_{xy}/\pm\sqrt{K} = r_{xy}\left(\dfrac{-\sigma_x\sigma_y}{\pm\sqrt{K}}\right)$. By the argument given in the demonstration of Lemma 1, we see that r_{xy} and $\sin 2\varphi$ should be of the same sign. Hence the negative sign is chosen before the radical.

From Lemma 2 and (7), we have the formula given by Pearson:

$$(8) \qquad 2S_N^2 = (\sigma_x^2 + \sigma_y^2)^2 - \sqrt{K}.$$

3. We are going to establish several more formulae for S_N. From (4), we have $2h\cdot\tan(\varphi) = -(a - b) \pm \sqrt{K}$. The sign before the radical is determined in such a way that $\tan(\varphi)$ has the same sign as r_{xy}. By the reasoning given in Lemma 1, the negative sign is chosen. Thus

$$-2p_{xy}\cdot\tan\varphi = -(\sigma_y^2 - \sigma_x^2) - \sqrt{K} = \sigma_x^2 + \sigma_y^2 - \sqrt{K} - 2\,\sigma_y^2$$

or

$$2(\sigma_y^2 - p_{xy}\tan\varphi) = \sigma_x^2 - \sigma_y^2 - \sqrt{K}.$$

This proves that

$$(9) \qquad S_N^2 = \sigma_y^2 - p_{xy}\tan\varphi.$$

Similarly, we have

$$(10) \qquad S_N^2 = \sigma_x^2 - p_{xy} \cot \varphi .$$

For computation, (9) and (10) are more convenient than (8). When the line of mutual regression is determined, it is known that $\tan \varphi$ (denoted by B) is equal to the slope of that line, and hence $\cot \varphi$ ($= 1/B$) is equal to the reciprocal of the slope. Then we can write (9) and (10) as follows:

$$(11) \qquad S_N^2 = \sigma_y^2 - p_{yx} \cdot B$$

$$(12) \qquad S_N^2 = \sigma_x^2 - p_{xy}/B .$$

The second formula given in Lemma 2 is simpler than (8), but not as simple as (11) and (12).

For computation, it is convenient to find φ from the equation

$$\tan 2\varphi = \frac{+2r_{xy}\sigma_x\sigma_y}{\sigma_x^2 - \sigma_y^2} = H,$$

i.e.,

$$2\varphi = \arctan H.$$

Since $\sin 2\varphi$ and r_{xy} are of the same sign, we can determine the value of φ from the preceding equation by inspection, though $\arctan H$ is a multiple-valued function. After the determination of φ, we can obtain

$$B = \tan \varphi .$$

Then we can compute S_N either from (9), (11), or (10), (12).

There is a very interesting fact furnished by (11) and (12). These two formulae are, in fact, generalizations of the following two well known ones:

$$(a) \qquad S_y^2 = \sigma_y^2(1 - r)$$

$$(b) \qquad S_x^2 = \sigma_x^2(1 - r),$$

where S_y is the standard error of the line of regression when y is used as dependent variable and x as independent variable, and similarly for S_x. It is clear that the line of mutual regression may be looked upon as a generalization of the other two lines of regression when we use y or x as dependent variable. So the slope B of the line of mutual regression is a generalization of $b_{yx} = r\frac{\sigma_y}{\sigma_x}$ and $b_{xy} = r\frac{\sigma_x}{\sigma_y}$. where the subscript yx means y on x and xy, x on y. If we use x as independent variable, then we must obtain b_{yx} instead of B. Hence substituting the formula of b_{yx} instead of B into (11), we obtain, after a simple reduction, the same result as given by (a). On the other hand, if we use y as independent variable, we must obtain b_{xy} instead of $1/B$. It will result (b) when b_{xy} is put in the place of $1/B$ in (12). The generalization perhaps can be seen more clearly if we write (a) and (b) into slightly different forms:

(a′) $$S_y^2 = \sigma_y^2 - p_{yx} \cdot b_{yx}$$

(b′) $$S_x^2 = \sigma_x^2 - p_{xy} \cdot b_{xy}.$$

4. The misprint in Pearson's paper is on the second formula of the following:

$$(MSR)^2 = \frac{\sigma_x^2 \sigma_y^2}{\cot^2 \varphi} = \frac{1}{2}\left(\sigma_x^2 - \sigma_y^2 - \sqrt{(\sigma_x^2 - \sigma_y^2)^2 - 4r^2 \sigma_x^2 \sigma_y^2}\right)$$

where $\tan 2\varphi = 2r_{xy}\sigma_x\sigma_y/(\sigma_x^2 - \sigma_y^2)$. $\cot^2 \varphi$ should read "square of semi-major axis of ellipse (2)." Professor Henry Schultz first noticed this misprint and suggested to the writer to investigate it.

In a recent letter to Schultz, Pearson pointed out that one of the simplest formula for S_N^2 or $(MSR)^2$ is given by

(α) $$S_N^2 = \sigma_x^2 \sin^2 \varphi + \sigma_y^2 \cos^2 \varphi,$$

where φ is defined by (4). However, Professor Schultz expressed doubt about its validity. From lemma 2, it is clear that (α) is also not true.

INSTITUTE OF SOCIAL SCIENCES,
ACADEMIA SINICA, PEIPING