A COEFFICIENT OF LINEAR CORRELATION BASED ON THE METHOD OF LEAST SQUARES AND THE LINE OF BEST FIT.

By J. B. Coleman

Given N points in a plane, corresponding to N pairs of values for two variables, X and Y, we find the line of best fit and the line of *worst* fit, by the method of least squares*. Then we derive a cofficient of correlation based on the sum of the squares of the distances of the points from these two lines.

The line of best fit is in the line such that the sum of the squares of the distances of the points from it is a minimum. The line of *worst* fit is the one from which the sum of the squares of the distances of the points is a maximum. We shall refer to them, respectively, as the *minimum* and *maximum* lines.

For convenience we take the origin at the centre of gravity of the points, letting x and y denote deviations of X and Y, respectively, from their arithmetic means.

1. The two lines pass thru the arithmetic means of the X's, and of the Y's.

$y = mx + b$ may represent any line of the plane. The distance, d_i, of a point, (x_i, y_i), from the line is

$\frac{y_i - mx_i - b}{\sqrt{1+m^2}}$. The sum of the squares of the distances of the N points from the line will be

*For a general discussion of this method of fitting when q variables are involved, see, Pearson, Karl, "On Lines and Planes of Closest Fit to Systems of Points in Space", Phil. Mag., 6th series, vol. ii, 1901, P. 559.

$$(1) \qquad \Sigma \underline{d}^2 = \frac{\Sigma y^2 + m^2\Sigma x^2 + Nb^2 - 2m\Sigma xy - 2b\Sigma y + 2mb\Sigma x}{1+m^2}$$

Using $\Sigma x = \Sigma y = O$, in the condition for maximum or minimum values in (1), the condition reduces to $\underline{b} = O$, and the theorem follows.

2. To find the slopes of the two lines.
   Equation (1) now becomes

$$(2) \qquad \Sigma \underline{d}^2 = \frac{\Sigma y^2 - 2m\Sigma xy + m^2\Sigma x^2}{1+m^2}.$$

The condition under which (2) will have a maximum or minimum values, is that

$$m^2\Sigma xy + m(\Sigma x^2 - \Sigma y^2) - \Sigma xy = O.$$

This condition is satisfied by two values of $\underline{m}$, namely;

$$(3) \qquad m_1 = \frac{\Sigma y^2 - \Sigma x^2 + \sqrt{(\Sigma x^2 - \Sigma y^2)^2 + 4(\Sigma xy)^2}}{2\Sigma xy},$$

$$(4) \qquad m_2 = \frac{\Sigma y^2 - \Sigma x^2 - \sqrt{(\Sigma x^2 - \Sigma y^2)^2 + 4(\Sigma xy)^2}}{2\Sigma xy}.$$

It is found by considering the second derivative that (4) is the condition under which $\Sigma \underline{d}^2$ will have a maximum value, and (3) is the condition for a minimum value.

The equation of the *minimum* line is $y = m_1 x$, and that of the *maximum* line is $y = m_2 x$. The value of m_1 and m_2 are those given in (3) and (4).

3. The *maximum* and *minimum* lines are perpendicular to each other.

That $m_1 = -1/m_2$ is easily shown from (3) and (4).

Further m_1 has the same sign as Σxy, and m_2, the opposite sign, since their numerators are, respectively, positive and negative.

4. The *minimum* line lies between the two lines of regression, or coincides with them.

If $\Sigma xy > 0$,

$$m_1 \leqq \frac{\Sigma y^2 - \Sigma x^2 + \sqrt{(\Sigma x^2 - \Sigma y^2)^2 + 4\Sigma x^2 \Sigma y^2}}{2\Sigma xy}$$

since $\Sigma x^2 \Sigma y^2 \geqq (\Sigma xy)^2$.

Hence $m_1 \leqq \Sigma y^2 / \Sigma xy = m_{xy}$, the slope of the line of regression of x on y.

Rationalizing the numerator of (3), and noting that $\Sigma x^2 \Sigma y^2 \geqq (\Sigma xy)^2$, we obtain

$$m_1 \geqq \frac{2\Sigma xy}{\Sigma x^2 - \Sigma y^2 + \sqrt{(\Sigma x^2 - \Sigma y^2)^2 + 4\Sigma x^2 \Sigma y^2}} = \frac{\Sigma xy}{\Sigma x^2} = m_{yx},$$

the slope of the line of regression of y on x.

In the same way it may be shown that if $\Sigma xy < 0$, then $m_{xy} \leqq m_1 \leqq m_{yx}$.

The condition that m_1 be equal to the slope of one line of regression is the same that it be equal to the other, so that the *minimum* line coincides with both lines of regression, or else lies between the two.

5. To find the sum of the squares of the distances of the points from the *minimum* line; also from the *maximum* line.

Let d be the distance of a point from the line $y = m_1 x$

$$\Sigma d^2 = \Sigma\left(\frac{y-m_1x}{\sqrt{1+m^2}}\right)^2 = \frac{\Sigma y^2 - 2m_1\Sigma xy + m_1^2\Sigma x^2}{1+m_1^2}$$

Substituting for m from (3) and reducing, we obtain

(5)

$$\Sigma d^2 = \left[\Sigma x^2+\Sigma y^2-\sqrt{(\Sigma x^2-\Sigma y^2)^2+4(\Sigma xy)^2}\right]/2.$$

Similarly, if D represents the distance of a point from the line $y = m_2 x$, we obtain

(6) $$\Sigma D^2 = \left[\Sigma x^2+\Sigma y^2+\sqrt{(\Sigma x^2-\Sigma y^2)^2+4(\Sigma xy)^2}\right]/2.$$

6. To find a coefficient of linear correlation.

Let $q = \sqrt{\Sigma d^2/\Sigma D^2}$ Substituting from (5) and (6), and reducing, we obtain

(7a) $$q = \frac{\Sigma x^2+\Sigma y^2-\sqrt{(\Sigma x^2-\Sigma y^2)^2+4(\Sigma xy)^2}}{2\sqrt{\Sigma x^2\Sigma y^2-(\Sigma xy)^2}}$$, or

(7b) $$q = \frac{2\sqrt{\Sigma x^2\Sigma y^2-(\Sigma xy)^2}}{\Sigma x^2+\Sigma y^2+\sqrt{(\Sigma x^2-\Sigma y^2)^2+4(\Sigma xy)^2}}$$

q represents the ratio of the root-mean-squares of the distances of the point from the *minimum* and *maximum* lines. This ratio is a measure of the closeness of fit of the points to a line, and should furnish a measure of linear correlation. The value of q may vary from 0 to 1. $q = 0$ indicates that the points are all on a straight line, hence that the correlation is perfect. It is of interest to note that when $q = 0$, (7b) gives $\Sigma xy/N\sigma_x\sigma_y$ $[= r] = \pm 1$. When q is 1 the mean squares of the distances of the points from two lines at right angles is the same, and linear correlation is lacking. Hence $1-q$ would furnish a coefficient conforming to the customs that it have the value 1 for perfect correlation, and 0 for lack of correlation.

Values of q found from (7a) or (7b) would involve the units in which X and Y are given. Hence these forms would be objectionable, in that q could be made to assume different values for the same data, by changing the units in which x and y are expressed. However, this objection may be removed by taking σ_x and σ_y as units in which to express X and Y (7b) then reduces to

$$q = \frac{\sqrt{N^2 - (\Sigma xy)^2/\sigma_x^2\sigma_y^2}}{N + |\Sigma xy/\sigma_x\sigma_y|}$$

The coefficient $1-q$ may now be expressed as

$$(8) \qquad r_c = 1 - q = 1 - \sqrt{\frac{N - |\Sigma xy/\sigma_x\sigma_y|}{N + |\Sigma xy/\sigma_x\sigma_y|}} \,.$$

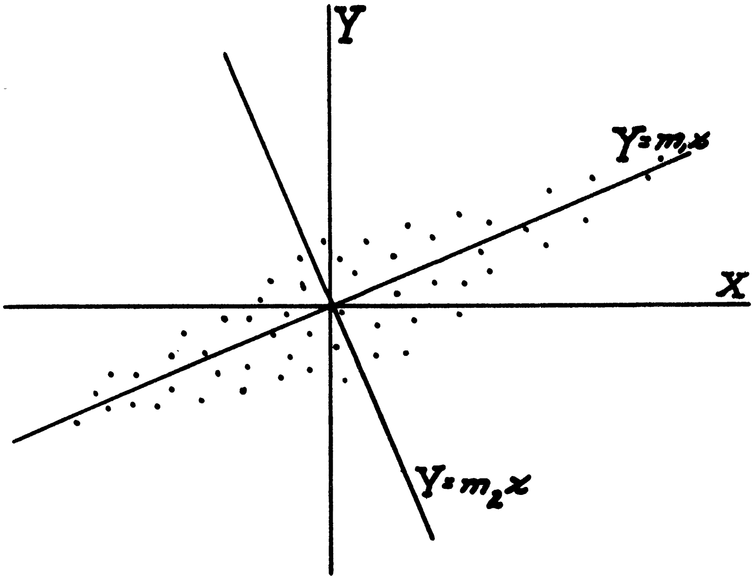
The sign of the coefficient should agree with that of the slope of the line to which the points are fitted. Hence, when the value for $1-q$ has been found it should be given the same sign as the slope of the ***minimum*** line. But the slope of the ***minimum*** line is determined by that of Σxy, so that the sign given to $1-q$ should be that of Σxy.

The coefficient $1-q$ may be expressed immediately in terms of the Pearson coefficient, r, which is equal to $\Sigma xy/N\sigma_x\sigma_y$. Making this substitution in (8) we have

$$r_c = 1-q = 1-\sqrt{\frac{1-|r|}{1+|r|}}\ .$$

In the table are shown values of $1-q$ corresponding to some given values of r . The values for $1-q^2$ have also been listed corresponding to the same set of values for r . The maximum difference occurs between $1-q$ and r when $r=.839$ and $1-q=$ 704, a difference of .135 by which $1-q$ is smaller.

r	$r_c=1-q$	$1-q^2$
1	1	1
.99	.929	.995
.95	.840	.974
.9	.771	.947
.839	.704	.912
.8	.667	.889
.7	.580	.824
.6	.500	.750
.5	.423	.667
.4	.345	.571
.3	.266	.462
.2	.183	.333
.1	.095	.182
0	0	0



J. B Coleman