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.