LIII. *On Lines and Planes of Closest Fit to Systems of Points in Space.*

By KARL PEARSON, *F.R.S., University College, London* *.

(1) IN many physical, statistical, and biological investigations it is desirable to represent a system of points in plane, three, or higher dimensioned space by the "best-fitting" straight line or plane. Analytically this consists in taking

$$y = a_0 + a_1 x, \quad \text{or} \quad z = a_0 + a_1 x + b_1 y,$$
$$\text{or} \quad z = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_3 + \ldots + a_n x_n,$$

where y, x, z, x_1, x_2, $\ldots x_n$ are variables, and determining the "best" values for the constants a_0, a_1, b_1, a_0, a_1, a_2, a_3, $\ldots a_n$ in relation to the observed corresponding values of the variables. In nearly all the cases dealt with in the text-books of least squares, the variables on the right of our equations are treated as the independent, those on the left as the dependent variables. The result of this treatment is that we get one straight line or plane if we treat some one variable as independent, and a quite different one if we treat another variable as the independent variable. There is no paradox about this; it is, in fact, an easily understood and most important feature of the theory of a system of correlated variables. The most probable value of y for a given value of x, say, is not given by the same relation as the most probable value of x for a given value of y. Or, to take a concrete example, the most probable stature of a man with a given length of leg l being s, the most probable length of leg for a man of stature s will not be l. The "best-fitting" lines and planes for the cases of z up to n variables for a correlated system are given in my memoir on regression †. They depend upon a determination of the means, standard-deviations, and correlation-coefficients of the system. In such cases the values of the independent variables are supposed to be accurately known, and the probable value of the dependent variable is ascertained.

(2) In many cases of physics and biology, however, the "independent" variable is subject to just as much deviation or error as the "dependent" variable. We do not, for example, know x accurately and then proceed to find y, but both x and y are found by experiment or observation. We observe x and y and seek for a unique functional relation between them. Men of given stature may have a variety

* Communicated by the Author.

† Phil. Trans. vol. clxxxvii. A, pp. 301 *et seq.*

of leg-lengths; but a point at a given time will have one position only, although our observations of *both* time and position may be in error, and vary from experiment to experiment. In the case we are about to deal with, we suppose the observed variables—all subject to error—to be plotted in plane, three-dimensioned or higher space, and we endeavour to take a line (or plane) which will be the "best fit" to such a system of points.

Of course the term "best fit" is really arbitrary; but a good fit will clearly be obtained if we make the sum of the squares of the perpendiculars from the system of points upon the line or plane a minimum.

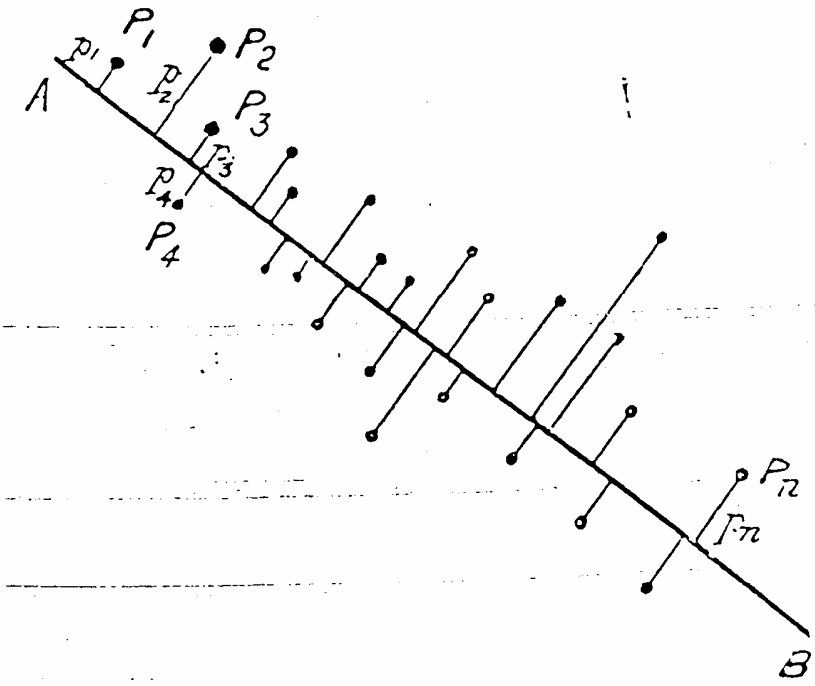
For example:—Let $P_1, P_2, \ldots P_n$ be the system of points with coordinates $x_1, y_1; x_2, y_2; \ldots x_n y_n$, and perpendicular distances $p_1, p_2, \ldots p_n$ from a line A B. Then we shall make

$$U = S(p^2) = \text{a minimum.}$$

If y were the dependent variable, we should have made

$$S(y'-y)^2 = \text{a minimum}$$

(y' being the ordinate of the theoretical line at the point x which corresponds to y), had we wanted to determine the best-fitting line in the usual manner.



Now clearly $U = S(p^2)$ is the moment of momentum, the second moment of the system of points, supposed equally loaded, about the line A B. But the second moment of a system about a series of parallel lines is always least for the

line going through the centroid. Hence: *The best-fitting straight line for a system of points in a space of any order goes through the centroid of the system.*

Now let there be n points each fixed by q variables $x_1, x_2 \ldots x_q$, and let

$$\bar{x}_1 = S(x_1)/n, \quad \bar{x}_2 = S(x_2)/n \ldots \bar{x}_q = S(x_q)/n. \quad . \quad \text{(i.)}$$

fix the centroid, or the *mean values* of the variables;

$$\sigma^2 x_1 = S(x_1^2)/n - \bar{x}_1^2, \quad \sigma^2 x_2 = S(x_2^2)/n - \bar{x}_2^2, \ldots$$
$$\sigma^2 x_q = S(x_q^2)/n - \bar{x}_q^2, \quad . \quad \text{(ii.)}$$

fix the *standard-deviations* (errors of mean square), or indirectly the moments of inertia or second-moments about the axes of coordinates, through the centroid parallel to the axes of the variables $x_1, x_2 \ldots x_q$. And, lastly, let

$$r_{x_u x_v} = \frac{S(x_u x_v) - n\bar{x}_u \bar{x}_v}{n\sigma_{x_u}\sigma_{x_v}} \quad . \quad . \quad . \quad . \quad . \quad \text{(iii.)}$$

for all pairs of values of u and v from 1, 2, 3, ... q, fix the *correlations* of the variables, or indirectly the products of inertia or product-moments about the axes.

Now let $l_1, l_2, l_3 \ldots l_q$ be the generalized direction-cosines of a plane at perpendicular distance p from the origin. We shall have

$$l_1^2 + l_2^2 + l_3^2 + \ldots + l_q^2 = 1. \quad . \quad . \quad . \quad . \quad \text{(iv.)}$$

Further, if U be the sum of the squares of the perpendicular distances of the system of n points from the plane

$$l_1 x_1 + l_2 x_2 + l_3 x_3 + \ldots + l_q x_q = p, \quad . \quad . \quad . \quad . \quad \text{(v.)}$$

we require to make a minimum of

$$U = S(l_1 x_1 + l_2 x_2 + l_3 x_3 + \ldots + l_q x_q - p)^2, \quad . \quad . \quad \text{(vi.)}$$

by variation of $l_1, l_2, \ldots l_q, p$ subject to (iv.). Differentiate first with regard to p and we have

$$l_1 S(x_1) + l_2 S(x_2) + l_3 S(x_3) + \ldots + l_q S(x_q) - np = 0;$$
$$\therefore \quad p = l_1 \bar{x}_1 + l_2 \bar{x}_2 + \ldots + l_q \bar{x}_q, \quad . \quad . \quad . \quad . \quad \text{(vii.)}$$

which shows us from (v.) that: *the best-fitting plane passes through the centroid of the system.*

Now vary (vi.) and add to it Q times the variation of (iv.), Q being an undetermined multiplier. We have, by equating to zero the coefficient of dl_u,

$$l_1 S(x_1 x_u) + l_2 S(x_2 x_u) + \ldots + l_u S(x_u^2) + \ldots + l_q S(x_q x_u)$$
$$- p S(x_u) + Q l_u = 0.$$

Pearson, K. 1901. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine* **2**:559-572.
http://pbil.univ-lyon1.fr/R/pearson1901.pdf

Or, substituting for p from (vii.) and using (ii.) and (iii.):

$$l_1\sigma_{x_1}\sigma_{x_u}r_{x_1x_u}+l_2\sigma_{x_2}\sigma_{x_u}r_{x_2x_u}+\ldots l_u\sigma_{x_u}{}^2+\ldots+l_q\sigma_{x_q}\sigma_{x_u}r_{x_ux_q}+\frac{Q}{n}l_u=0, \quad . \quad . \quad . \quad \text{(viii.)}$$

is the type equation.

Now (vi.) may be written :—

$$U=n\{l_1{}^2\sigma^2{}_{x_1}+l_2{}^2\sigma^2{}_{x_2}+\ldots+l_q{}^2\sigma^2{}_{x_q}+2l_1l_2\sigma_{x_1}\sigma_{x_2}r_{x_1x_2}+\ldots+2l_{q-1}l_q\sigma_{x_{q-1}}\sigma_{x_q}r_{x_{q-1}x_q}\} \quad . \quad \text{(viii.) } bis$$

Multiplying each type equation by its corresponding l_u, adding together and remembering (iv.), we find

$$\frac{U_m}{n}+\frac{Q}{n}=0, \quad \text{or} \quad Q=-U_m,$$

where U_m is the minimum value of U.

Now let Σ^2 be the mean square of the residuals, or

$$\Sigma^2=\frac{\Sigma(l_1x_1+l_2x_2+\ldots+l_qx_q-p)^2}{n}.$$

Then

$$\frac{Q}{n}=-\Sigma^2,$$

and a physical meaning has been given to Q, $\sqrt{-Q/n}$ is the "mean square residual,"—*i. e.*, the quantity, the square of which is the mean square of the residuals.

The type equation (viii.) may now be written:

$$l_1\sigma_{x_1}\sigma_{x_u}r_{x_1x_u}+l_2\sigma_{x_2}\sigma_{x_u}r_{x_2x_u}+\ldots+l_u(\sigma^2{}_{x_u}-\Sigma^2)+l_q\sigma_{x_q}\sigma_{x_u}r_{x_qx_u}=0. \quad . \quad \text{(ix.)}$$

We can eliminate the l's and dividing out row and column of resulting determinant by the corresponding σ, we have:

$$\begin{vmatrix} 1-\frac{\Sigma^2}{\sigma^2{}_{x_1}} & r_{x_1x_2}, & r_{x_1x_3}, \ldots r_{x_1x_q} \\ r_{x_2x_1}, & 1-\frac{\Sigma^2}{\sigma^2{}_{x_2}}, & r_{x_2x_3}, \ldots r_{x_2x_q} \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ r_{x_qx_1}, & r_{x_qx_2}, & r_{x_qx_3}, \quad 1-\frac{\Sigma^2}{\sigma^2{}_{x_q}} \end{vmatrix} =0, \quad . \quad . \quad \text{(x.)}$$

as a determinantal equation to find Σ^2. We must choose the least root of this equation, for the mean square residual must

be as small as possible. Substitute this value of Σ^2 in the type equations (viii.), and we find the required values of $l_1, l_2 \ldots : l_q$, using (iv.).

This is the complete analytical solution of the problem of drawing the best-fitting plane through n non-coplanar points. We see that it depends only on a knowledge of the means, standard-deviations, and correlations of the q variables.

Whenever we may suppose that variation is due to errors of observation or measurement,—*i. e.*, is not organic, but there exists a unique functional relation between the true values of the variables,—then, assuming it of the first degree, we may determine the best values of the constants in the manner given above.

(3) A geometrical interpretation is of course to be found from (viii.) *bis*. Consider the quadric

$$\sigma^2_{x_1} x_1^2 + \sigma^2_{x_2} x_2^2 + \ldots + \sigma^2_{x_q} x_q^2 + 2\sigma_{x_1}\sigma_{x_2} r_{x_1 x_2} x_1 x_2 + \ldots + 2\sigma_{x_{q-1}}\sigma_{x_q} r_{x_{q-1} x_q} x_{q-1} x_q = \epsilon^4, \quad \ldots \quad \text{(xi.)}$$

where ϵ is any line. Then this quadric will be "ellipsoidal" since the coefficients of $x_1^2 \ldots x_q^2$ are all positive quantities. Let R be its radius-vector measured in the direction $l_1, l_2, \ldots l_q$, or perpendicular to the plane from which we are measuring the residuals; then clearly:

$$U = n\epsilon^4/R^2,$$

$$\text{or} \qquad \Sigma^2 = \epsilon^4/R^2. \quad \ldots \ldots \quad \text{(xii.)}$$

Thus the inverse square of the radius of this "ellipsoid" measures the square of the mean square residual. We shall speak of the ellipsoid as the *ellipsoid of residuals*. Since Σ is to be a minimum, R must be a maximum; or we conclude: *that the best-fitting plane is perpendicular to the greatest axis of the ellipsoid of residuals and the minimum mean square residual varies inversely as the length of this axis.*

A case of failure can only arise if the ellipsoid of residuals degenerates into an "oblate spheroid," *i. e.*, when every plane through its shorter axis is one of "best fit," or into a sphere, when every plane through the centroid of the system of points is an equally good fit. This sphericity of distribution of points in space involves the vanishing of all the correlations between the variables and the equality of all their standard-deviations. It corresponds to isotropic inertia in the theory of moments in dynamics.

(4) The theory of the best-fitting straight line need not

Pearson, K. 1901. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine* **2**:559-572.
http://pbil.univ-lyon1.fr/R/pearson1901.pdf

detain us long. Let its equation be

$$\frac{x_1-x_1'}{l_1}=\frac{x_2-x_2'}{l_2}=\frac{x_3-x_3'}{l_3}=\ldots\frac{x_q-x_q'}{l_q}=\rho. \quad \text{(xiii.)}$$

Draw the plane perpendicular to this line through x_1', x_2' $x_3'\ldots x_q'$; i. e.,

$$l_1x_1+l_2x_2+l_3x_3+\ldots+l_qx_q=\mathrm{H},$$

where $\mathrm{H}=l_1x_1'+l_2x_2'+l_3x_3'+\ldots+l_qx_q'$.

Then if p be the perpendicular from any point in space or the line (xiii.):

$$p^2=(x_1-x_1')^2+(x_2-x_2')^2+\ldots+(x_q-x_q')^2$$
$$-\{l_1(x_1-x_1')+l_2(x_2-x_2')+l_3(x_3-x_3')+\ldots+l_q(x_q-x_q')\}^2.$$

Now x_1', $x_2'\ldots x_q'$ and l_1, $l_2\ldots l_q$, subject to the relation $l_1^2+l_2^2+l_3^2+\ldots+l_q^2=1$, are the constants at our disposal. Sum p^2 and differentiate to find when $\mathrm{U}=\mathrm{S}(p^2)$ is a minimum. We have for type equation

$$\mathrm{S}(x_u-x_u')-l_u[\mathrm{S}\{l_1(x_1-x_1')+l_2(x_2-x_2')+\ldots+l_q(x_q-x_q')\}$$

whence we see:

$$\frac{\mathrm{S}(x_u)-nx_u'}{l_u}=\text{symmetrical function of } x\text{'s}.$$

Or, we must have

$$\frac{\bar{x}_1-x_1'}{l_1}=\frac{\bar{x}_2-x_2'}{l_2}=\ldots=\frac{\bar{x}_q-x_q'}{l_q},$$

which show us that the straight line passes (as we have already noted) through the centroid of the system. We can accordingly take x_1', $x_2'\ldots x_q'$ to be that centroid, and we find:

$$\Sigma'^2=\frac{\mathrm{U}}{n}=\frac{\mathrm{S}(p^2)}{n}=\sigma^2_{x_1}+\sigma^2_{x_2}+\ldots+\sigma^2_{x_q}$$
$$-[l_1^2\sigma^2_{x_1}+l_2^2\sigma^2_{x_2}+\ldots+l^2_q\sigma^2_{x_q}$$
$$+2l_1l_2\sigma_{x_1}\sigma_{x_2}r_{x_1x_2}+\ldots+2l_{q-1}l_q\sigma_{x_{q-1}}\sigma_{x_q}r_{x_{q-1}r_{x_q}}].$$

But the expression in square brackets is precisely the square of the mean square residual with regard to the plane,

$$l_1(x_1-\bar{x}_1)+l_2(x_2-\bar{x}_2)+\ldots+l_q(x_q-\bar{x}_q)=0,$$

or Σ^2. Thus we have:

$$\Sigma'^2=\sigma^2_{x_1}+\sigma^2_{x_2}+\ldots+\sigma^2_{x_q}-\Sigma^2.$$

Now clearly $\sigma^2_{x_1}+\sigma^2_{x_2}+\ldots+\sigma^2_{x_q}$ is a constant. Hence Σ'^2 will be a minimum when Σ^2 is a maximum, or when the

Pearson, K. 1901. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine* **2**:559-572.
http://pbil.univ-lyon1.fr/R/pearson1901.pdf

plane perpendicular to the best-fitting line is perpendicular to the least axis of the ellipsoid of residuals. Thus we find: *That the line which fits best a system of n points in q-fold space passes through the centroid of the system and coincides in direction with the least axis of the ellipsoid of residuals.*

The mean square residual (which measures of course the closeness of the fit) is given by

$$\Sigma' = \sqrt{\sigma^2 x_1 + \sigma^2 x_2 + \ldots + \sigma^2 x_q - \frac{\epsilon^4}{R^2}}, \quad \ldots \quad \text{(xiv.)}$$

where R is the least radius of the ellipsoid of residuals. The direction-cosines of the line can be found from (ix.) by giving Σ^2 the least value among the roots of (x.).

Clearly the plane of best fit passes through the line of best fit, and is further perpendicular to the greatest radius, the maximum axis of the ellipsoid of residuals.

(5) While the geometry of lines and planes of best fit is thus seen to be very simple from the standpoint of inertia ellipsoids,—particularly from the consideration of the surface which, for the theory of errors, I have termed the ellipsoid of residuals,—they most frequently occur, perhaps, in the case of correlated variations or errors, and it is thus of interest to consider them in relation to the ellipses and ellipsoids which arise as "contours" in correlation surfaces.

Now take the case of two variables x and y only, the type-ellipse of the contours of the correlation surface is, when referred to its centroid as origin:

$$\frac{x^2}{\sigma^2 x} + \frac{y^2}{\sigma^2 y} - \frac{2 r_{xy}\, xy}{\sigma_x\, \sigma_y} = 1.$$

Compare this with the ellipse of residuals

$$\sigma^2 x\, x'^2 + \sigma^2 y\, y'^2 + 2\sigma_x \sigma_y r_{xy}\, x'y' = \epsilon^4.$$

Clearly if we take $x' = y$, $y' = -x$, and $\epsilon^4 = \sigma^2 x \sigma^2 y$ the ellipse of residuals becomes the correlation type-ellipse. Further, $x'^2 + y'^2 = x^2 + y^2$, or the two ellipses have equal rays, but they are at right-angles to each other. Thus the best-fitting straight line for the system of points coincides in direction with the major axis of the correlation ellipse, and the mean square residual for this line

$$= \frac{\text{product of standard deviations}}{\text{semi-major axis of correlation ellipse}}.$$

The geometry of these results is indicated in the accompanying diagram :—

EE′ is found by making $S(y'-y)^2$ a minimum,
FF′ ,, ,, $S(x'-x)^2$,,
AA′ ,, ,, $S(p^2)$,,

The equation to EE′ referred to C is $y=\frac{r_{xy}\sigma_y}{\sigma_x}x$,

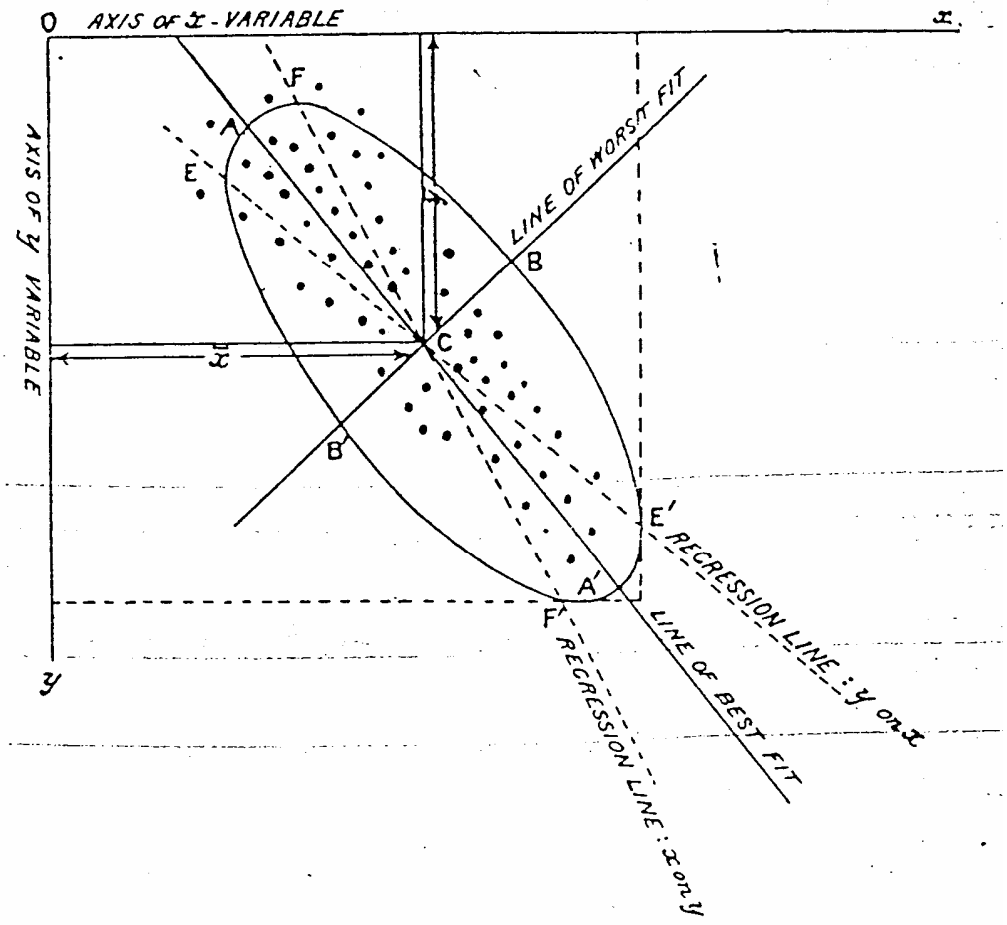
,, ,, FF′ ,, ,, $x=\frac{r_{xy}\sigma_x}{\sigma_y}y$.

The angle θ which AA′ makes with Ox is determined by

$$\tan 2\theta=\frac{2r_{xy}\sigma_x\sigma_y}{\sigma^2_x-\sigma^2_y}.$$

Further:

$$(\text{Mean sq. residual})^2=\sigma^2_x\sigma^2_y/\cot^2\theta$$
$$=\tfrac{1}{2}(\sigma^2_x+\sigma^2_y)-\tfrac{1}{2}\sqrt{(\sigma^2_x-\sigma^2_y)^2+4r^2_{xy}\sigma^2_x\sigma^2_y}.$$



Physically the axes of the correlation type-ellipse are the directions of independent or uncorrelated variation. Hence the line of best fit is a direction of uncorrelated variation.

(6) We turn to the correlation type-"ellipsoid" for q variables. It is*:

$$\Delta_{11}\frac{x_1^2}{\sigma^2_{x_1}}+\Delta_{22}\frac{x_2^2}{\sigma^2_{x_2}}+\ldots\Delta_{qq}\frac{x_q^2}{\sigma^2_{x_q}}+2\Delta_{12}\frac{x_1x_2}{\sigma_{x_1}\sigma_{x_2}}$$
$$+\ldots\ldots+2\Delta_{q-1q}\frac{x_{q-1}x_q}{\sigma_{x_{q-1}}\sigma_{x_q}}=1, \quad \ldots \text{(xv.)}$$

where Δ_{11}, Δ_{22}, $\Delta_{12}\ldots\Delta_{q-1\,q}$, Δ_{qq} are the minors corresponding to the constituents marked by the same subscripts of the determinant:

$$\Delta=\begin{vmatrix} 1 & r_{12} & r_{13} & \ldots\ldots & r_{1q} \\ r_{21} & 1 & r_{23} & \ldots\ldots & r_{2q} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ r_{q1} & r_{q2} & r_{q3} & \ldots\ldots & 1 \end{vmatrix} \quad \ldots\ldots \text{(xvi.)},$$

Now let us find the directions and magnitudes of the principal axes of this ellipsoid. We must make

$$u^2=x_1^2+x_2^2+\ldots+x_q^2$$

a maximum. Or if Q be an indeterminate multiplier, we have:

$$(\Delta_{11}+Q\sigma^2_{x_1})\frac{x_1}{\sigma_{x_1}}+\Delta_{12}\frac{x_2}{\sigma_{x_2}}+\Delta_{13}\frac{x_3}{\sigma_{x_3}}+\ldots+\Delta_{1q}\frac{x_q}{\sigma_{x_q}}=0,$$

$$\Delta_{1q}\frac{x_1}{\sigma_{x_1}}+(\Delta_{22}+Q\sigma^2_{x_2})\frac{x_2}{\sigma_{x_2}}+\Delta_{23}\frac{x_3}{\sigma_{x_3}}+\ldots+\Delta_{2q}\frac{x_q}{\sigma_{x_q}}=0,$$

$$\cdots\cdots\cdots\cdots\cdots\cdots$$

$$\cdots\cdots\cdots\cdots\cdots\cdots$$

$$\Delta_{1q}\frac{x_1}{\sigma_{x_1}}+\Delta_{2q}\frac{x_2}{\sigma_{x_2}}+\Delta_{3q}\frac{x_3}{\sigma_{x_3}}+\ldots+(\Delta_{qq}+Q\sigma^2_{x_q})\frac{x_q}{\sigma_{x_q}}=0 \text{(xvii.)}$$

Multiply the last $q-1$ of these equations by $r_{12}, r_{13}, \ldots r_{1q}$ respectively and add them to the first, then we know that:

$$\Delta_{11}+r_{12}\Delta_{12}+r_{13}\Delta_{13}+\ldots+r_{1q}\Delta_{1q}=\Delta,$$

and if u be not 1:

$$\Delta_{1u}+r_{12}\Delta_{2u}+r_{13}\Delta_{3u}+\ldots+r_{1q}\Delta_{qu}=0.$$

* Phil. Trans. vol. clxxxvii. A, p. 302.

Hence:

$$(\Delta+Q\sigma^2{}_x)\frac{x_1}{\sigma_{x_1}}+r_{12}Q\sigma^2{}_{x_2}\frac{x_2}{\sigma_{x_2}}+r_{13}Q\sigma^2{}_{x_3}\frac{x_3}{\sigma_{x_3}}+\ldots+r_{1q}Q\sigma^2{}_{x_q}\frac{x_q}{\sigma_{x_q}}=($$

which may be written:

$$\left(\frac{\Delta}{Q\sigma^2{}_{x_1}}+1\right)x_1\sigma_{x_1}+r_{12}x_2\sigma_{x_2}+r_{13}x_3\sigma_{x_3}+\ldots+r_{1q}x_q\sigma_{x_q}=0.$$

Or system (xvii.) may be replaced by:

$$x_1\sigma^2{}_{x_1}\left(1-\frac{\Delta R^2}{\sigma^2{}_{x_1}}\right)+r_{12}x_2\sigma_{x_2}\sigma_{x_1}+r_{13}x_3\sigma_{x_3}\sigma_{x_1}+\ldots+r_{1q}x_q\sigma_{x_q}\sigma_{x_1}=($$

$$r_{12}x_1\sigma_{x_2}\sigma_{x_1}+x_2{}^2\sigma^2{}_{x_2}\left(1-\frac{\Delta R^2}{\sigma^2{}_{x_2}}\right)+r_{23}x_3\sigma_{x_3}\sigma_{x_2}+\ldots+r_{2q}x_q\sigma_{x_q}\sigma_{x_2}=($$

. . . . . .

$$r_{1q}x_1\sigma_{x_q}\sigma_{x_1}+r_{2q}x_2\sigma_{x_2}\sigma_{x_q}+r_{3q}x_3\sigma_{x_3}\sigma_x+\ldots+x_q{}^2\sigma^2{}_x\left(1-\frac{\Delta R}{\sigma^2{}_{x_q}}\right)^2=$$

. . . (xvi

For multiplying (xvii.) by $x_1, x_2, \ldots x_q$ respectively and adding, we find

$$1+Qu^2=0;$$

or if R be a maximum or minimum value of u, $Q=-1/R^2$.

Now compare these equations with those we obtain for the directions and magnitudes of axes of the ellipsoid of residuals:

$$\sigma^2{}_{x_1}x_1{}'^2+\sigma^2{}_{x_2}x_2{}'^2+\ldots+\sigma^2{}_{x_q}x_q{}'^2+2r_{12}\sigma_{x_1}\sigma_{x_2}x_1{}'x_2{}'+\ldots$$
$$+2r_{q-1q}\sigma_{x_{q-1}}\sigma_{x_q}x'_{q-1}x_q{}'=\epsilon^4. \quad \text{(xix.)}$$

These are:

$$x_1{}'\sigma^2{}_{x_1}\left(1-\frac{\epsilon^4}{R'^2\sigma^2{}_{x_1}}\right)+r_{12}x_2{}'\sigma_{x_2}\sigma_{x_1}+\ldots+r_{1q}x_q{}'\sigma_{x_q}\sigma_{x_1}=0,$$

$$r_{12}x_1{}'\sigma_{x_1}\sigma_{x_2}+x_2{}'\sigma^2{}_{x_2}\left(1-\frac{\epsilon^4}{R'^2\sigma^2{}_{x_2}}\right)+\ldots+r_{2q}x_q{}'\sigma_{x_q}\sigma_{x_2}=0,$$

$$r_{1q}x_1{}'\sigma_{x_1}\sigma_{x_q}+r_{2q}x_2{}'\sigma_{x_2}\sigma_{x_q}+\ldots+x_q{}'\sigma^2{}_{x_q}\left(1-\frac{\epsilon^4}{R_1{}^2\sigma^2{}_{x_q}}\right)=0$$

. . . (xx.)

Now eliminate the x's from (xviii.) and the x''s from (xx.) and we have precisely the same determinant to find

ΔR^2 and ϵ^4/R'^2. Hence for the semi-axes or max.-min. values:

$$\Delta R^2 = \epsilon^4/R'^2 \quad . \quad . \quad . \quad . \quad . \quad . \quad \text{(xxi.)}$$

Equations (xviii.) and (xx.) will now give the same values for the ratios of the x's and of the x''s, or for any axis:

$$x_1/x_1' = x_2/x_2' = x_3/x_3' = \ldots = x_q/x_q' \quad . \quad . \quad \text{(xxii.)}$$

But (xxii.) combined with (xxi.) gives us:

$$x_1 = \frac{\epsilon^2}{\sqrt{\Delta}} \frac{x_1'}{R'^2}, \quad x_2 = \frac{\epsilon^2}{\sqrt{\Delta}} \frac{x_2'}{R'^2}, \ldots, x_q = \frac{\epsilon^2}{\sqrt{\Delta}} \frac{x_q'}{R'^2} \quad \text{(xxiii.)}$$

In other words, if we define points given by (xxiii.) to be corresponding points,—*i. e.*, if corresponding points lie on the same line at distances inversely as each other from the origin,—then the ends of the principal axes of the two ellipsoids are corresponding points. Thus the principal axes of the correlation ellipsoid coincide with those of the ellipsoid of residuals in direction, and a minimum axis of the one is a maximum axis of the other and *vice versa*. We therefore conclude:

(i.) That the best fitting plane to a system of points is perpendicular to the least axis of the correlation ellipsoid, and that if $2R_{min.}$ be the length of this axis the mean square residual $= \sqrt{\Delta} \times R_{min.}$ where Δ is the well-known determinant of the correlation coefficients.

(ii.) The best-fitting straight line to a system of points coincides in direction with the maximum axis of the correlation ellipsoid, and the mean square residual

$$= \sqrt{\sigma^2_{x_1} + \sigma^2_{x_2} + \sigma^2_{x_3} + \ldots + \sigma^2_{x_q} - \Delta . R^2_{max.}}, \quad \text{(xxiv.)}$$

where $2R_{max.}$ is the length of the maximum axis.

We have thus the properties of the best-fitting plane and line in terms of the correlation ellipsoid, which is the one generally adopted for variation problems. At the same time our investigation shows us that the q directions of independent variation and the standard-deviations of the independent variables may be found from the ellipsoid of residuals, which will usually be a process involving much simpler arithmetic.

(7) *Numerical Illustrations.*

Case (i.). Find the best fitting straight line to the following system of points supposed of equal weight:

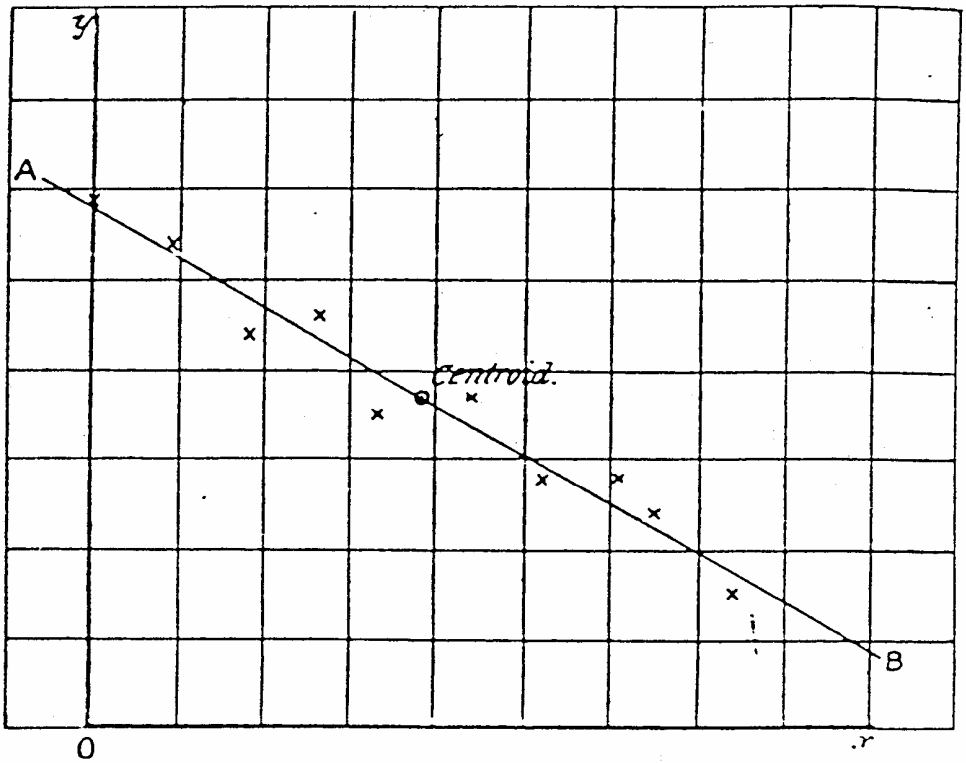
$x=0$	$y=5\cdot9$	$x=4\cdot4$	$y=3\cdot7$
$x=\cdot9$	$y=5\cdot4$	$x=5\cdot2$	$y=2\cdot8$
$x=1\cdot8$	$y=4\cdot4$	$x=6\cdot1$	$y=2\cdot8$
$x=2\cdot6$	$y=4\cdot6$	$x=6\cdot5$	$y=2\cdot4$
$x=3\cdot3$	$y=3\cdot5$	$x=7\cdot4$	$y=1\cdot5$

Pearson, K. 1901. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine* **2**:559-572.
http://pbil.univ-lyon1.fr/R/pearson1901.pdf

We have at once:

$$\bar{x}=3{\cdot}82 \qquad \bar{y}=3{\cdot}70$$
$$\sigma_x=2{\cdot}3748 \qquad \sigma_y=1{\cdot}31225$$
$$r_{xy}=-{\cdot}9765$$
$$\tan 2\theta=2r_{xy}\sigma_x\sigma y/(\sigma_x^2-\sigma_y^2)=-1{\cdot}5535.$$

Hence: $\tan\theta=-{\cdot}54556$, or the best-fitting line passes through the point 3·82, 3·70 at a slope of $-{\cdot}546$. This line is shown in the accompanying diagram by AB. The mean



square residual is ·2484. Had we made $S(y-y')^2$ a minimum, the slope of the "best-fitting" line would have been $-{\cdot}5396$ and the mean square residual ·2828; had we made $S(x-x')^2$ a minimum, the slope of the "best-fitting" line would have been $-{\cdot}5659$, and the mean square residual ·5118. These lines are of course the regression lines of slopes $r_{xy}\sigma_y/\sigma_x$ and $\sigma_y/(r_{xy}\sigma_x)$ to the horizontal and mean square vertical and horizontal residuals of $\sigma_y\sqrt{1-r^2}$ and $\sigma_x\sqrt{1-r^2}$ respectively.

Illustration (2).—The following system gives four values of a certain function z:

	$x=2.$	$x=4.$
$y=16$	219	127
$y=26$	261	231

Let us find the best-fitting plane, treating these as four points in three-dimensioned space. We have at once

$$\bar{x} = 3, \qquad \bar{y} = 21, \qquad \bar{z} = 209{\cdot}5$$
$$\sigma_x = 1, \qquad \sigma_y = 5, \qquad \sigma_z = 50{\cdot}0275$$
$$r_{zy}\sigma_y\sigma_z = 182{\cdot}5, \quad r_{xz}\sigma_x\sigma_z = -30{\cdot}5, \quad r_{xy}\sigma_x\sigma_y = 0.$$

Thus the ellipsoid of residuals is:

$$x^2 + 25y^2 + 2502{\cdot}75z^2 + 365yz - 61xz = \epsilon^4.$$

The equations to find the direction-cosines are:

$$\left(2 + 2\frac{Q}{n}\right)l_1 + 0\,.\,l_2 - 61\,.\,l_3 = 0,$$

$$0\,.\,l_1 + \left(50 + 2\frac{Q}{n}\right)l_2 + 365\,l_3 = 0$$

$$-61\,l_1 + 365\,l_2 + \left(5005{\cdot}5 + 2\frac{Q}{n}\right)l_3 = 0.$$

Whence writing $2\frac{Q}{n} = \chi$ (n=number of points=4) the cubic for χ is:

$$C = \chi^3 + 5057{\cdot}5\chi^2 + 123{,}440\chi + 48{,}050 = 0.$$

We want the *least* root:

$\chi = 0$, C=+; $\chi = -{\cdot}5$, C=−; $\chi = -100$, C=+; $\chi = -\infty$, C=−. Thus the required root lies between 0 and $-{\cdot}5$. It is easily found to be

$$\chi_1 = -{\cdot}395{,}660.$$

Thus $-\frac{Q}{n} = {\cdot}197830$, and the mean square residual

$$= \sqrt{-\frac{Q}{n}} = {\cdot}4448.$$

We easily deduce:

$$\frac{l_1}{38{\cdot}02187} = -\frac{l_2}{7{\cdot}35823} = \frac{l_3}{1}.$$

Thus the best-fitting plane is:

$$38{\cdot}02187\ (x-3) - 7{\cdot}35823\ (y-21) + z - 209{\cdot}5 = 0,$$

or:

$$z + 38{\cdot}02187x - 7{\cdot}35823y - 169{\cdot}03778 = 0. \quad . \quad \text{(xxv.)}$$

If we find the values for z for given x and y, say those of the four points, which are

$$211{\cdot}7, \quad 135{\cdot}7, \quad 283{\cdot}3, \quad 207{\cdot}3,$$

we should not be impressed by the goodness of the fit. But

the small value of the mean square residual shows how close to each of the points the plane really goes when we measure its distance from a point not by the vertical intercept, but by the perpendicular from the point on the plane. Thus the *vertical* distance from $x=4$, $y=16$, $z=127$, to the plane is 8·7, but the *perpendicular* distance is only ·1988.

If χ_2 and χ_3 be the other two roots of the cubic C we easily find:

$$\chi_2\chi_3=121,442{\cdot}65, \quad \chi_2+\chi_3=-5057{\cdot}10434.$$

Thus we have the quadratic to find χ_2 and χ_3

$$\chi^2+5057{\cdot}10434\chi+121,442{\cdot}65=0,$$

or:

$$\chi_2=-24{\cdot}12895, \quad \chi_3=-5032{\cdot}97445.$$

χ_3 gives the least axis of the ellipsoid of residuals; hence the direction-cosines of this axis are given by

$$\frac{l_1}{-{\cdot}012,125}=\frac{l_2}{{\cdot}073,249}=\frac{l_3}{1}.$$

We have accordingly for the equation of the best-fitting straight line to the four points:

$$\frac{x-3}{-12{\cdot}125}=\frac{y-21}{73{\cdot}249}=\frac{z-209{\cdot}5}{1000} \quad . \quad . \quad . \quad \text{(xxvi.)}$$

The mean square residual for this line Σ' is given by (xiv.)

$$\begin{aligned}\Sigma' &= \sqrt{\sigma^2{}_x+\sigma^2{}_y+\sigma^2{}_z+\tfrac{1}{2}\chi_3} \\ &= \sqrt{2528{\cdot}75-2516{\cdot}487225} \\ &= 3{\cdot}5018.\end{aligned}$$

This again is remarkably small, considering how far our four points are from being co-linear.

The reader will easily prove directly that the best line (xxvi.) really lies in the best plane (xxv.).

These two illustrations may suffice to show that the methods of this paper can be easily applied to numerical problems; the labour is not largely increased if we have a considerable number of points. It becomes more cumbersome if we have four, five, or more variables or characters which involve the determination of the least (or greatest) root (as the case may be) or an equation of the fourth, fifth, or higher order. Still, the coefficients being numerical and all the roots real and negative, it is not very difficult to localize them.

Pearson, K. 1901. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine* **2**:559-572.
http://pbil.univ-lyon1.fr/R/pearson1901.pdf