THE FITTING OF STRAIGHT LINES IF BOTH VARIABLES ARE SUBJECT TO ERROR

By Abraham Wald

1. Introduction. The problem of fitting straight lines if both variables x and y are subject to error, has been treated by many authors. If we have $N > 2$ observed points (x_i, y_i) $(i = 1, \cdots, N)$, the usually employed method of least squares for determining the coefficients a, b, of the straight line $y = ax + b$ is that of choosing values of a and b which minimize the sum of the squares of the residuals of the y's, i.e. $\Sigma(ax_i + b - y_i)^2$ is a minimum. It is well known that treating y as an independent variable and minimizing the sum of the squares of the residuals of the x's, we get a different straight line as best fit. It has been pointed out[1] that if both variables are subject to error there is no reason to prefer one of the regression lines described above to the other. For obtaining the "best fit," which is not necessarily equal to one of the two lines mentioned, new criteria have to be found. This problem was treated by R. J. Adcock as early as 1877.[2]

He defines the line of best fit as the one for which the sum of the squares of the normal deviates of the N observed points from the line becomes a minimum. (Another early attempt to solve this problem by minimizing the sum of squares of the normal deviates was made by Karl Pearson.[3])

Many objections can be raised against this method. First, there is no justification for minimizing the sum of the squares of the *normal* deviates, and not the deviations in some other direction. Second, the straight line obtained by that method is not invariant under transformation of the coordinate system. It is clear that a satisfactory method should give results which do not depend on the choice of a particular coordinate system. This point has been emphasized by C. F. Roos. He gives[4] a good summary of the different methods and then proposes a general formula for fitting lines (and planes in case of more than two variables) which do not depend on the choice of the coordinate system.

---

[1] See for instance Henry Schultz' "The Statistical Law of Demand," *Jour. of Political Economy*, Vol. 33, Dec. (1925).

[2] *Analyst*, Vol. IV, p. 183 and Vol. V, p. 53.

[3] "On Lines and Planes of Closest Fit to Systems of Points in Space" *Phil. Mag.* 6th Ser. Vol. II (1901).

[4] "A General Invariant Criterion of Fit for Lines and Planes where all Variates are Subject to Error," *Metron*, February 1937. See also Oppenheim and Roos *Bulletin of the American Mathematical Society*, Vol. 34 (1928), pp. 140–141.

Roos' formula includes many previous solutions[5] as special cases. H. E. Jones[6] gives an interesting geometric interpretation of Roos' general formula.

It is a common feature of Roos' general formula and of all other methods proposed in recent years that the fitted straight line cannot be determined without *a priori* assumptions (independent of the observations) regarding the weights of the errors in the variables x and y. That is to say, either the standard deviations of the errors in x and in y are involved (or at least their ratio is included) in the formula of the fitted straight line and there is no method given by which those standard deviations can be estimated by means of the observed values of x and y.

R. Frisch[7] has developed a new general theory of linear regression analysis, when all variables are subject to error. His very interesting theory employs quite new methods and is not based on probability concepts. Also on the basis of Frisch's discussion it seems that there is no way of determining the "true" regression without *a priori* assumptions about the disturbing intensities.

T. Koopmans[8] combined Frisch's regression theory with the classical one in a new general theory based on probability concepts. Also, according to his theory, the regression line can be determined only if the ratio of the standard deviations of the errors is known.

In a recent paper R. G. D. Allen[9] gives a new interesting method for determining the fitted straight line in case of two variables x and y. Denoting by σ_ϵ the standard deviation of the errors in x, by σ_η the standard deviation of the errors in y and by ρ the correlation coefficient between the errors in the two variables, Allen emphasizes (p. 194)[9] that the fitted line can be determined only if the *values of two* of the three quantities σ_ϵ, σ_η, ρ are given *a priori*.

Finally I should like to mention a paper by C. Eisenhart,[10] which contains many interesting remarks related to the subject treated here.

In the present paper I shall deal with the case of two variables x and y in which the errors are uncorrelated. It will be shown that under certain conditions:

(1) The fitted straight line can be determined without making *a priori* assumptions (independent of the observed values x and y) regarding the standard deviations of the errors.

(2) The standard deviation of the errors can be well estimated by means of

---

[5] For instance also Corrado Gini's method described in his paper, "Sull' Interpolazione di una Retta Quando i Valori della Variable Independente sono Affecti da Errori Accidentalis," *Metron*, Vol. I, No. 3 (1921), pp. 63-82.

[6] "Some Geometrical Considerations in the General Theory of Fitting Lines and Planes," *Metron*, February 1937.

[7] *Statistical Confluence Analysis by Means of Complete Regression Systems*, Oslo, 1934.

[8] *Linear Regression Analysis of Economic Time Series*, Haarlem, 1937.

[9] "The Assumptions of Linear Regression," *Economica*, May 1939.

[10] "The interpretation of certain regression methods and their use in biological and industrial research," *Annals of Math. Stat.*, Vol. 10 (1939), pp. 162-186.

the observed values of x and y. The precision of the estimate increases with the number of the observations and would give the exact values if the number of observations were infinite. (See in this connection also condition V in section 3.)

2. Formulation of the Problem. Let us begin with a precise formulation of the problem. We consider two sets of random variables[11]

$$x_1, \cdots, x_N; \qquad y_1, \cdots, y_N.$$

Denote the expected value $E(x_i)$ of x_i by X_i and the expected value $E(y_i)$ of y_i by Y_i $(i = 1, \cdots, N)$. We shall call X_i the true value of x_i, Y_i the true value of y_i, $x_i - X_i = \epsilon_i$ the error in the i-th term of the x-set, and $y_i - Y_i = \eta_i$ the error in the i-th term of the y-set.

The following assumptions will be made:

I. *The random variables* $\epsilon_1, \cdots, \epsilon_N$ *each have the same distribution and they are uncorrelated, i.e.* $E(\epsilon_i\epsilon_j) = 0$ *for* $i \neq j$. *The variance of* ϵ_i *is finite.*

II. *The random variables* $\eta_1, \cdots, \eta_N$ *each have the same distribution and are uncorrelated, i.e.* $E(\eta_i\eta_j) = 0$ *for* $i \neq j$. *The variance of* η_i *is finite.*

III. *The random variables* ϵ_i *and* η_j $(i = 1, \cdots, N; j = 1, \cdots, N)$ *are uncorrelated, i.e.* $E(\epsilon_i\eta_j) = 0$.

IV. *A single linear relation holds between the true values* X *and* Y, *that is to say* $Y_i = \alpha X_i + \beta$ $(i = 1, \cdots, N)$.

Denote by ϵ a random variable having the same probability distribution as possessed by each of the random variables $\epsilon_1, \cdots, \epsilon_N$, and by η a random variable having the same distribution as $\eta_1, \cdots, \eta_N$.

The problem to be solved can be formulated as follows:

We know only two sets of observations: $x_1', \cdots, x_N'; y_1', \cdots, y_N'$, where x_i' denotes the observed value of x_i and y_i' denotes the observed value of y_i. We know neither the true values $X_1, \cdots, X_N; Y_1, \cdots, Y_N$, nor the coefficients α and β of the linear relation between them. We have to estimate by means of the observations $x_1', \cdots, x_N'; y_1', \cdots, y_N'$, (1) the values of α and β, (2) the standard deviation σ_ϵ of ϵ, and (3) the standard deviation σ_η of η.

Problems of this kind occur often in Economics, where we are dealing with time series. For example, denote by x_i the price of a certain good G in the period t_i, and by y_i the quantity of G demanded in t_i. In each time period t_i there exists a normal price X_i and a normal demand Y_i which would obtain if the influence of some accidental disturbances could be eliminated. If we have reason to assume that there exists between the normal price and the normal demand a linear relationship we have to deal with a problem of the kind described above.

In the following discussions we shall use the notations x_i and y_i also for their

[11] A random or stochastic variable is a real variable associated with a probability distribution.

observed values x_i' and y_i' since it will be clear in which sense they are meant and no confusion can arise.

3. Consistent Estimates of the Parameters α, β, σ_ϵ, σ_η. For the sake of simplicity we assume that N is even. We consider the expression

$$(1)\qquad a_1 = \frac{(x_1 + \cdots + x_m) - (x_{m+1} + \cdots + x_N)}{N};$$
$$a_2 = \frac{(y_1 + \cdots + y_m) - (y_{m+1} + \cdots + y_N)}{N},$$

where $m = N/2$. As an estimate of α we shall use the expression

$$(2)\qquad a = \frac{a_2}{a_1} = \frac{(y_1 + \cdots + y_m) - (y_{m+1} + \cdots + y_N)}{(x_1 + \cdots + x_m) - (x_{m+1} + \cdots + x_N)}.$$

We make the assumption

V. *The limit inferior of*

$$\left|\frac{(X_1 + \cdots + X_m) - (X_{m+1} + \cdots + X_N)}{N}\right| \qquad (N = 2, 3, \cdots \text{ad. inf.}$$

is positive.

We shall prove that a is a consistent estimate of α, i.e. a converges stochastically to α with $N \to \infty$, if the assumptions I–V hold. Denote the expected value of a_1 by $\bar{a}_1$ and the expected value of a_2 by $\bar{a}_2$. It is obvious that

$$(3)\qquad \bar{a}_1 = \frac{(X_1 + \cdots + X_m) - (X_{m+1} + \cdots + X_N)}{N},$$
$$\bar{a}_2 = \frac{(Y_1 + \cdots + Y_m) - (Y_{m+1} + \cdots + Y_N)}{N}.$$

On account of the condition IV we have

$$(4)\qquad \bar{a}_2 = \alpha\bar{a}_1, \quad \text{or} \quad \frac{\bar{a}_2}{\bar{a}_1} = \alpha.$$

The variance of $a_1 - \bar{a}_1$ is equal to σ_ϵ^2/N and the variance of $a_2 - \bar{a}_2$ is equal to σ_η^2/N. Hence a_1 and a_2 converge stochastically towards $\bar{a}_1$ and $\bar{a}_2$ respectively. From that and assumption V it follows that also $\frac{a_2}{a_1}$ converges stochastically towards $\frac{\bar{a}_2}{\bar{a}_1} = \alpha$. The intercept β of the regression line will be estimated by

$$(5)\qquad b = \bar{y} - a\bar{x}, \qquad \text{where } \bar{x} = \frac{x_1 + \cdots + x_N}{N} \quad \text{and} \quad \bar{y} = \frac{y_1 + \cdots + y_N}{N}.$$

Denote by $\bar{X}$ the arithmetic mean of $X_1, \cdots, X_N$ and by $\bar{Y}$ the arithmetic mean of $Y_1, \cdots, Y_N$. Since $\bar{y}$ converges stochastically towards $\bar{Y}$, $\bar{x}$ towards

$\bar{X}$, and a towards α, b converges stochastically towards $\bar{Y} - \alpha\bar{X}$. From condition IV it follows that $\bar{Y} - \alpha\bar{X} = \beta$. Hence b converges stochastically towards β.

Let us introduce the following notations:

$$s_x = \sqrt{\Sigma \frac{(x_i - \bar{x})^2}{N}} = \text{sample standard deviation of the } x\text{-observations},$$

$$s_y = \sqrt{\Sigma \frac{(y_i - \bar{y})^2}{N}} = \text{sample standard deviation of the } y\text{-observations},$$

$$s_{xy} = \Sigma \frac{(x_i - \bar{x})(y_i - \bar{y})}{N} = \text{sample covariance between the } x\text{-set and } y\text{-set}.$$

s_X, s_Y and s_{XY} denote the same expressions of the true values $X_1, \cdots, X_N$; $Y_1, \cdots, Y_N$.

It is obvious that

$$E(s_x^2) = s_X^2 + \sigma_\epsilon^2 \frac{N-1}{N}, \tag{6}$$

$$E(s_y^2) = s_Y^2 + \sigma_\eta^2 \frac{N-1}{N}, \tag{7}$$

$$E(s_{xy}) = s_{XY}, \tag{8}$$

where $E(s_x^2)$, $E(s_y^2)$, and $E(s_{xy})$ denote the expected values of s_x^2, s_y^2, and s_{xy}.[12]

Since $Y_i = \alpha X_i + \beta$, we have

$$s_Y = \alpha s_X, \tag{9}$$

$$s_{XY} = \alpha s_X^2. \tag{10}$$

From (8), (9) and (10) we get

$$s_X^2 = \frac{E(s_{xy})}{\alpha}, \tag{11}$$

$$s_Y^2 = \alpha E(s_{xy}). \tag{12}$$

If we substitute in (6) and (7) for s_X^2 and s_Y^2 their values in (11) and (12), we get

$$\sigma_\epsilon^2 = \left[E(s_x^2) - \frac{E(s_{xy})}{\alpha}\right] N/(N-1), \tag{13}$$

$$\sigma_\eta^2 = [E(s_y^2) - \alpha E(s_{xy})] N/(N-1). \tag{14}$$

---

[12] I observe that the equations (6), (7) and (8) are essentially the same as those investigated by R. Frisch, *Statistical Confluence Analysis* pp. 51–52. See also Allen's equations (4) l.c. p. 194.

Since s_x^2, s_y^2, s_{xy} converge stochastically towards their expected values and a converges stochastically towards α, the expressions

$$\left[s_x^2 - \frac{s_{xy}}{a}\right] N/(N-1) \tag{15}$$

and

$$[s_y^2 - a s_{xy}] N/(N-1) \tag{16}$$

are consistent estimates of σ_ϵ^2 and σ_η^2 respectively.

4. Confidence Interval for α. In this section, as well as in sections 5 and 6, only the assumptions I–IV are assumed to hold. In other words, all statements made in these sections are valid independently of Assumption V, except where the contrary is explicitly stated.

Let us introduce the following notation:

$$\bar{x}_1 = \frac{x_1 + \cdots + x_m}{m}; \qquad \bar{y}_1 = \frac{y_1 + \cdots + y_m}{m}$$

$$\bar{x}_2 = \frac{x_{m+1} + \cdots + x_N}{m}; \qquad \bar{y}_2 = \frac{y_{m+1} + \cdots + y_N}{m}$$

$$(s_x')^2 = \frac{\sum_{i=1}^{m} (x_i - \bar{x}_1)^2 + \sum_{j=m+1}^{N} (x_j - \bar{x}_2)^2}{N}$$

$$(s_y')^2 = \frac{\sum_{i=1}^{m} (y_i - \bar{y}_1)^2 + \sum_{j=m+1}^{N} (y_j - \bar{y}_2)^2}{N}$$

$$s_{xy}' = \frac{\sum_{i=1}^{m} (x_i - \bar{x}_1)(y_i - \bar{y}_1) + \sum_{j=m+1}^{N} (x_j - \bar{x}_2)(y_j - \bar{y}_2)}{N}.$$

$\bar{X}_1$, $\bar{X}_2$, $\bar{Y}_1$, $\bar{Y}_2$, $(s_X')^2$, $(s_Y')^2$ and s_{XY}' denote the same functions of the true values $X_1, \cdots, X_N$, $Y_1, \cdots, Y_N$. The expressions s_x', s_y', and s_{xy}' are slightly different from the corresponding expressions s_x, s_y, and s_{xy}. The reason for introducing these new expressions is that the distributions of s_x, s_y, and s_{xy} are not independent of the slope $a = \frac{a_2}{a_1}$ of the sample regression line, but s_x', s_y' and s_{xy}' are distributed independently from a (assuming that ϵ and η are normally distributed). The latter statement follows easily from the fact that according to (1) and (2) $a = \frac{\bar{y}_1 - \bar{y}_2}{\bar{x}_1 - \bar{x}_2}$ and s_x', s_y', s_{xy}' are distributed independently of $\bar{x}_1$, $\bar{x}_2$, $\bar{y}_1$ and $\bar{y}_2$.

In the same way as we derived (13) and (14), we get

$$\sigma_\epsilon^2 = \left[E(s_x')^2 - \frac{E(s_{xy}')}{\alpha}\right]N/(N-2), \tag{13'}$$

$$\sigma_\eta^2 = [E(s_y')^2 - \alpha E(s_{xy}')]N/(N-2). \tag{14'}$$

These formulae differ from the corresponding formulae (13) and (14) only in the denominator of the second factor, having there $N-2$ instead of $N-1$. This is due to the fact that the estimates s_x, s_y, s_{xy} are based on $N-1$ degrees of freedom whereas s_x', s_y' and s_{xy}' are based only on $N-2$ degrees of freedom. From (13') and (14') we get the following estimates[13] for σ_ϵ^2 and σ_η^2:

$$\left[(s_x')^2 - \frac{s_{xy}'}{\alpha}\right]N/(N-2), \tag{17}$$

$$[(s_y')^2 - \alpha s_{xy}']N/(N-2). \tag{18}$$

Hence we get as an estimate of $\sigma_\eta^2 + \alpha^2\sigma_\epsilon^2$ the expression:

$$s^2 = [(s_y')^2 + \alpha^2(s_x')^2 - 2\alpha s_{xy}']N/(N-2)$$

$$= \frac{N}{N-2}\left\{\frac{\sum_{i=1}^{m}[(y_i - \alpha x_i) - (\bar{y}_1 - \alpha\bar{x}_1)]^2 + \sum_{j=m+1}^{N}[(y_j - \alpha x_j) - (\bar{y}_2 - \alpha\bar{x}_2)]^2}{N}\right\} \tag{19}$$

Now we shall show that

$$\frac{(N-2)s^2}{\sigma_\eta^2 + \alpha^2\sigma_\epsilon^2} \tag{20}$$

has the χ^2-distribution with $N-2$ degrees of freedom, provided that ϵ and η are normally distributed. In fact,

$$(y_i - \alpha x_i) - (\bar{y}_1 - \alpha\bar{x}_1) = \eta_i - \alpha\epsilon_i - (\bar{\eta}_1 - \alpha\bar{\epsilon}_1) \qquad (i = 1, \cdots, m)$$

and

$$(y_j - \alpha x_j) - (\bar{y}_2 - \alpha\bar{x}_2) = \eta_j - \alpha\epsilon_j - (\bar{\eta}_2 - \alpha\bar{\epsilon}_2) \quad (j = m+1, \cdots, N),$$

where

$$\bar{\epsilon}_1 = \frac{\epsilon_1 + \cdots + \epsilon_m}{m}, \qquad \bar{\epsilon}_2 = \frac{\epsilon_{m+1} + \cdots + \epsilon_N}{m},$$

$$\bar{\eta}_1 = \frac{\eta_1 + \cdots + \eta_m}{m}, \qquad \bar{\eta}_2 = \frac{\eta_{m+1} + \cdots + \eta_N}{m}$$

Since the variance of $\eta_k - \alpha\epsilon_k$ is equal to $\sigma_\eta^2 + \alpha^2\sigma_\epsilon^2$ and since $\eta_k - \alpha\epsilon_k$ is uncorrelated with $\eta_l - \alpha\epsilon_l$ $(k \neq l)$ $(k, l = 1, \cdots, N)$, the expression (20) has the χ^2-distribution with $N-2$ degrees of freedom.

[13] An "estimate" is usually a function of the observations not involving any unknown parameters. We designate here as estimates also some functions involving the parameter α.

Now we shall show that

$$\frac{\sqrt{N}\, a_1(a-\alpha)}{\sqrt{\sigma_\eta^2 + \alpha^2\sigma_\epsilon^2}} \tag{21}$$

is normally distributed with zero mean and unit variance. In fact from the equations (1)–(4) it follows that

$$\begin{aligned} a_1(a-\alpha) &= \bar{a}_2 + \frac{\bar{\eta}_1 - \bar{\eta}_2}{2} - a_1\left(\frac{\bar{a}_2}{\bar{a}_1}\right) \\ &= \bar{a}_2 + \frac{\bar{\eta}_1 - \bar{\eta}_2}{2} - \left(\bar{a}_1 + \frac{\bar{\epsilon}_1 - \bar{\epsilon}_2}{2}\right)\left(\frac{\bar{a}_2}{\bar{a}_1}\right) \\ &= \frac{\bar{\eta}_1 - \bar{\eta}_2}{2} - \alpha\frac{\bar{\epsilon}_1 - \bar{\epsilon}_2}{2}. \end{aligned}$$

Since the latter expression is normally distributed (provided that ϵ and η are normally distributed) with zero mean and variance $\frac{\sigma_\eta^2 + \alpha^2\sigma_\epsilon^2}{N}$, our statement about (21) is proved.

Obviously (20) and (21) are independently distributed, hence $\sqrt{N-2}$ times the ratio of (21) to the square root of (20), namely,

$$t = \sqrt{N-2}\,\frac{\sqrt{N}\, a_1(a-\alpha)}{\sqrt{N-2}\, s} = \frac{a_1(a-\alpha)\sqrt{N-2}}{\sqrt{(s_y')^2 + \alpha^2(s_x')^2 - 2\alpha s_{xy}'}} \tag{22}$$

has the Student distribution with $N-2$ degrees of freedom. Denote by t_0 the critical value of t corresponding to a chosen probability level. The deviation of a from an assumed population value α is significant if

$$\left|\frac{a_1(a-\alpha)\sqrt{N-2}}{\sqrt{(s_y')^2 + \alpha^2(s_x')^2 - 2\alpha s_{xy}'}}\right| \geqq t_0.$$

The confidence interval for α can be obtained by solving the equation in α,

$$a_1^2(a-\alpha)^2 = [(s_y')^2 + \alpha^2(s_x')^2 - 2\alpha s_{xy}']\,\frac{t_0^2}{N-2}. \tag{23}$$

Now we shall show that if the relation

$$a_1^2 > \frac{(s_x')^2 t_0^2}{N-2}, \tag{24}$$

holds, the roots α_1 and α_2 are real and a is contained in the interior of the interval $[\alpha_1\alpha_2]$. From (19) it follows that

$$(s_y')^2 + \alpha^2(s_x')^2 - 2\alpha s_{xy}' > 0$$

for all values of α. Hence, for $\alpha = a$ the left hand side of (23) is smaller than the right hand side. On account of (24) there exists a value $a' > a$ and a

value $a'' < a$ such that the left hand side of (23) is greater than the right hand side for $\alpha = a'$ and $\alpha = a''$. Hence one root must lie between a and a' and the other root between a'' and a. This proves our statement. The relation (24) always holds for sufficiently large N if Assumption V is fulfilled. The confidence interval of α is the interval $[\alpha_1, \alpha_2]$. For very small N (24) may not hold.

Finally I should like to remark that no essentially better estimate of the variance of $\eta - \alpha\epsilon$ can be given than the expression s^2 in (19). In fact, we have $2N$ observations $x_1, \cdots, x_N; y_1, \cdots, y_N$. For the estimation of the variance of $\eta - \alpha\epsilon$ we must eliminate the unknowns $X_1, \cdots, X_N$ and β. (The unknowns $Y_1, \cdots, Y_N$ are determined by the relations $Y_i = \alpha X_i + \beta$ and α is involved in the expression whose variance is to be determined.) Hence we have at most $N - 1$ degrees of freedom and the estimate in (19) is based on $N - 2$ degrees of freedom.

5. Confidence Interval for β if α is Given. In this case the best estimate of β is given by the expression:

$$b_\alpha = \bar{y} - \alpha\bar{x} \text{ where } \bar{x} = \frac{x_1 + \cdots + x_N}{N} \text{ and } \bar{y} = \frac{y_1 + \cdots + y_N}{N}.$$

We have

$$b_\alpha - \beta = (\bar{y} - \bar{Y}) - \alpha(\bar{x} - \bar{X}) = \bar{\eta} - \alpha\bar{\epsilon}$$

where

$$\bar{\epsilon} = \frac{\epsilon_1 + \cdots + \epsilon_N}{N}, \text{ and } \bar{\eta} = \frac{\eta_1 + \cdots + \eta_N}{N}.$$

Hence,

$$\frac{\sqrt{N}\,(b_\alpha - \beta)}{\sqrt{\sigma_\eta^2 + \alpha^2\sigma_\epsilon^2}} \tag{25}$$

is normally distributed with zero mean and unit variance. It is obvious that the expressions (20) and (25) are independently distributed. Hence $\sqrt{N-2}$ times the ratio of (25) to the square root of (20), i.e.

$$t = \sqrt{N-2}\,\frac{\sqrt{N}\,(b_\alpha - \beta)}{\sqrt{N-2}\,s} = \frac{\sqrt{N-2}\,(b_\alpha - \beta)}{\sqrt{(s_y')^2 + \alpha^2(s_x')^2 - 2\alpha s_{xy}'}}$$

has the Student distribution with $N - 2$ degrees of freedom. Denoting by t_0 the critical value of t according to the chosen probability level, the confidence interval for β is given by the interval:

$$\left[b_\alpha + \frac{\sqrt{(s_y')^2 + \alpha^2(s_x')^2 - 2\alpha s_{xy}'}}{\sqrt{N-2}}\,t_0,\quad b_\alpha - \frac{\sqrt{(s_y')^2 + \alpha^2(s_x')^2 - 2\alpha s_{xy}'}}{\sqrt{N-2}}\,t_0\right].$$

6. Confidence Region for α and β Jointly. In most practical cases we want to know confidence limits for α and β jointly. A pair of values α, β can be represented in the plane by the point with the coordinates α, β. A region R of this plane is called confidence region of the true point (α, β) corresponding to the probability level P if the following two conditions are fulfilled.

(1) The region R is a function of the observations $x_1, \cdots, x_N; y_1, \cdots, y_N$, i.e. it is uniquely determined by the observations.

(2) Before performing the experiment the probability that we shall obtain observed values such that (α, β) will be contained in R, is exactly equal to P. P is usually chosen to be equal to .95 or .99.

We have shown that the expressions (21) and (25), i.e.

$$\frac{\sqrt{N}\, a_1(a-\alpha)}{\sqrt{\sigma_\eta^2 + \alpha^2\sigma_\epsilon^2}}, \qquad \frac{\sqrt{N}\,(b_\alpha - \beta)}{\sqrt{\sigma_\eta^2 + \alpha^2\sigma_\epsilon^2}}$$

are normally distributed with zero mean and unit variance. Now we shall show that these two quantities are independently distributed. For this purpose we have only to show that $\bar{x}$, $\bar{y}$, a_1 and a_2 are independently distributed (a_1 and a_2 are defined in (1)), but since

$$\begin{aligned} a_1 - E(a_1) &= (\bar{\epsilon}_1 - \bar{\epsilon}_2)/2 \\ a_2 - E(a_2) &= (\eta_1 - \eta_2)/2 \\ \bar{x} - E(\bar{x}) &= \bar{\epsilon} \\ \bar{y} - E(\bar{y}) &= \bar{\eta}, \end{aligned}$$

we have only to show that $\bar{\epsilon}$, $\bar{\eta}$, $\bar{\epsilon}_1 - \bar{\epsilon}_2$, $\eta_1 - \eta_2$ are independently distributed. We obviously have

$$\bar{\epsilon} = \frac{\bar{\epsilon}_1 + \bar{\epsilon}_2}{2}, \qquad \bar{\eta} = \frac{\eta_1 + \eta_2}{2}.$$

It is evident that $\bar{\epsilon}_1$, $\bar{\epsilon}_2$, η_1 and η_2 are independently distributed. Hence, $E[\bar{\epsilon}(\bar{\epsilon}_1 - \bar{\epsilon}_2)] = (E\bar{\epsilon}_1^2 - E\bar{\epsilon}_2^2)/2 = 0$ and also $E[\bar{\eta}(\eta_1 - \eta_2)] = (E\eta_1^2 - E\eta_2^2)/2 = 0$. Since $\bar{\epsilon}_1 - \bar{\epsilon}_2$, $\eta_1 - \eta_2$, and $\bar{\epsilon}$ and $\bar{\eta}$ are normally distributed, the independence of this set of variables is proved, and therefore also (21) and (25) are independently distributed. It is obvious that the expression (20) is distributed independently of (21) and (25). From this it follows that

$$\frac{N-2}{2} \cdot \frac{N[a_1^2(a-\alpha)^2 + (\bar{y} - \alpha x - \beta)^2]}{(N-2)s^2} = \frac{(N-2)[a_1^2(a-\alpha)^2 + (\bar{y} - \alpha\bar{x} - \beta)^2]}{2[(s_y')^2 + \alpha^2(s_x')^2 - 2\alpha s_{xy}']} \tag{26}$$

has the F-distribution (analysis of variance distribution) with 2 and $N-2$ degrees of freedom. The F-distribution is tabulated in Snedecor's book: *Calcu-*

lation and Interpretation of Analysis of Variance, Collegiate Press, Ames, Iowa, 1934. The distribution of $\frac{1}{2} \log F = z$ is tabulated in R. A. Fisher's book: *Statistical Methods for Research Workers*, London, 1936. Denote by F_0 the critical value of F corresponding to the chosen probability level P. Then the confidence region R is the set of points (α, β) which satisfy the inequality

$$\frac{N-2}{2} \cdot \frac{a_1^2(a-\alpha)^2 + (\bar{y} - \alpha\bar{x} - \beta)^2}{(s_y')^2 + \alpha^2(s_x')^2 - 2\alpha s_{xy}'} < F_0. \tag{27}$$

The boundary of the region is given by the equation

$$a_1^2(a-\alpha)^2 + (\bar{y} - \alpha\bar{x} - \beta)^2 = \frac{2F_0}{N-2}[(s_y')^2 + \alpha^2(s_x')^2 - 2\alpha s_{xy}']. \tag{28}$$

This is the equation of an ellipse. Hence the region R is the interior of the ellipse defined by the equation (28). If Assumption V holds, the length of the axes of the ellipse are of the order $1/\sqrt{N}$, hence with increasing N the ellipse reduces to a point.

7. The Grouping of the Observations. We have divided the observations in two equal groups G_1 and G_2, G_1 containing the first half $(x_1, y_1), \cdots, (x_m, y_m)$ and G_2 the second half $(x_{m+1}, y_{m+1}), \cdots, (x_N, y_N)$ of the observations. All the formulas and statements of the previous sections remain exactly valid for any arbitrary subdivision of the observations in two equal groups, provided that the subdivision is defined independently of the errors $\epsilon_1, \cdots, \epsilon_N$; $\eta_1, \cdots, \eta_N$. The question of which is the most advantageous grouping arises, i.e. for which grouping will a be the most efficient estimate of α (will lead to the shortest confidence interval for α). It is easy to see that the greater $|a_1|$ the more efficient is the estimate a of α. The expression $|a_1|$ becomes a maximum if we order the observations such that $x_1 \leq x_2 \leq \cdots \leq x_N$. That is to say $|a_1|$ becomes a maximum if we group the observations according to the following:

RULE I. *The point (x_i, y_i) belongs to the group G_1 if the number of elements x_j $(j \neq i)$ of the series $x_1, \cdots, x_N$ for which $x_j \leq x_i$ is less than $m = N/2$. The point (x_i, y_i) belongs to G_2 if the number of elements x_j $(j \neq i)$ for which $x_j \leq x_i$ is greater than or equal to m.*

This grouping, however, depends on the observed values $x_1, \cdots, x_N$ and is therefore in general not entirely independent of the errors $\epsilon_1, \cdots, \epsilon_N$. Let us now consider the grouping according to the following:

RULE II. *The point (x_i, y_i) belongs to the group G_1 if the number of elements X_j of the series $X_1, \cdots, X_N$ for which $X_j \leq X_i$ $(j \neq i)$ is less than m. The point (x_i, y_i) belongs to G_2 if the number of elements X_j for which $X_j \leq X_i$ $(j \neq i)$ is equal to or greater than m.*

The grouping according to Rule II is entirely independent of the errors $\epsilon_1, \cdots, \epsilon_N$; $\eta_1, \cdots, \eta_N$. It is identical with the grouping according to Rule I in the following case: Denote by x the median of $x_1, \cdots, x_N$; assume that ϵ can take values only within the finite interval $[-c, +c]$ and that all the values $x_1, \cdots, x_N$ fall outside the interval $[x - c, x + c]$. It is easy to see that in this case $x_i \leq x$ $(i = 1, \cdots, N)$ holds if and only if $X_i \leq X$, where X denotes the median of $X_1, \cdots, X_N$. Hence the grouping according to Rule II is identical to that according to Rule I and therefore the grouping according to Rule I is independent of the errors $\epsilon_1, \cdots, \epsilon_N$. In such cases we get the best estimate of α by grouping the observations according to Rule I. Practically, we can use the grouping according to Rule I and regard it as independent of the errors $\epsilon_1, \cdots, \epsilon_N$; $\eta_1, \cdots, \eta_N$ if there exists a positive value c for which the probability that $|\epsilon| \geq c$ is negligibly small and the number of observations contained in $[x - c, x + c]$ is also very small.

Denote by a' the value of a which we obtain by grouping the observations according to Rule I and by a'' the value of a if we group the observations according to Rule II. The value a'' is in general unknown, since the values $X_1, \cdots, X_N$ are unknown, except in the special case considered above, when we have $a'' = a'$. We will now show that an upper and a lower limit for a'' can always be given. First, we have to determine a positive value c such that the probability that $|\epsilon| \geq c$ is negligibly small. The value of c may often be determined before we make the observations having some *a priori* knowledge about the possible range of the errors. If this is not the case, we can estimate the value of c from the data. It is well known that if we have errors in both variables and fit a straight line by the method of least squares minimizing in the x-direction, the sum of the squared deviations divided by the number of degrees of freedom will overestimate σ_ϵ^2. Hence, if ϵ is normally distributed, we can consider the interval $[-3v, 3v]$ as the possible range of ϵ, i.e. $c = 3v$, where v^2 denotes the sum of the squared residuals divided by the number of degrees of freedom. If the distribution of ϵ is unknown, we shall have to take for c a somewhat larger value, for instance $c = 5v$. After having determined c, upper and lower limits for a'' can be given as follows: we consider the system S of all possible groupings satisfying the conditions:

(1) If $x_i \leq x - c$ the point (x_i, y_i) belongs to the group G_1.

(2) If $x_i \geq x + c$ the point (x_i, y_i) belongs to the group G_2.

We calculate the value of a according to each grouping of the system S and denote the minimum of these values by a^*, and the maximum by a^{**}. Since the grouping according to Rule II is contained in the system S, a^* is a lower and a^{**} an upper limit of a''.

Let g be a grouping contained in S and denote by I_g the confidence interval for α which we obtain from formula (23) using the grouping g. Denote further by I the smallest interval which contains the intervals I_g for all elements g of S. Then I contains also the confidence interval corresponding to the grouping according to Rule II. If we denote by P the chosen probability level (say

$P = .95$), then we can say: If we were to draw a sample consisting of N pairs of observations $(x_1, y_1), \cdots, (x_N, y_N)$, the probability is greater than or equal to P that we shall obtain a system of observations such that the interval I will include the true slope α.

The computing work for the determination of I may be considerable if the number of observations within the interval $[x - c, x + c]$ is not small. We can get a good approximation to I by less computation work as follows: First we calculate the slope a' using the grouping according to Rule I and determine the confidence interval $[a' - \delta, a' + \Delta]$ according to formula (23). Denote by $a(g)$ the value of the slope, i.e. the value of $\frac{\bar{y}_1 - \bar{y}_2}{\bar{x}_1 - \bar{x}_2}$, corresponding to a grouping g of the system S, and by $[a(g) - \delta_g, a(g) + \Delta_g]$ the corresponding confidence interval calculated from (23). Neglecting the differences $(\delta_g - \delta)$ and $(\Delta_g - \Delta)$, we obtain for I the interval $[a^* - \delta, a^{**} + \Delta]$.

If the difference $a^{**} - a^*$ is small, we can consider $I = [a^* - \delta, a^{**} + \Delta]$ as the correct confidence interval of α corresponding to the chosen probability level P. If, however, $a^{**} - a^*$ is large, the interval I is unnecessarily large. In such cases we may get a much shorter confidence interval by using some other grouping defined independently of the errors $\epsilon_1, \cdots, \epsilon_N$; $\eta_1, \cdots, \eta_N$. For instance if we see that the values $x_1, \cdots, x_N$ considered in the order as they have been observed, show a monotonically increasing (or decreasing) tendency, we shall define the group G_1 as the first half, and the group G_2 as the second half of the observations. Though we decide to make this grouping after having observed that the values $x_1, \cdots, x_N$ show a clear trend, the grouping can be considered as independent of the errors $\epsilon_1, \cdots, \epsilon_N$. In fact, if the range of the error ϵ is small in comparison to the true part X, the trend tendency of the value $x_1, \cdots, x_N$ will not be affected by the size of the errors $\epsilon_1, \cdots, \epsilon_N$. We may use for the grouping also any other property of the data which is independent of the errors.

The results of the preceding considerations can be summarized as follows: We use first the grouping according to Rule I, calculate the slope $a' = \frac{\bar{y}_1 - \bar{y}_2}{\bar{x}_1 - \bar{x}_2}$ and the corresponding confidence interval $[a' - \delta, a' + \Delta]$ (formula (23)). This confidence interval cannot be considered as exact since the grouping according to Rule I is not completely independent of the errors. In order to take account of this fact, we calculate a^* and a^{**}. If $a^{**} - a^*$ is small, we consider $I = [a^* - \delta, a^{**} + \Delta]$ with practical approximation as the correct confidence interval. If, however, $a^{**} - a^*$ is large, the interval I is unnecessarily large. We can only say that I is a confidence interval corresponding to a probability level greater than or equal to the chosen one. In such cases we should try to use some other grouping defined independently of the errors, which eventually will lead to a considerably shorter confidence interval.

Analogous considerations hold regarding the joint confidence region for α and β. We use the grouping according to Rule I and calculate from (27) the

corresponding confidence region R. If $| a^{**} - a^{*} |$ and $| b^{**} - b^{*} |$ are small ($b^{*} = \bar{y} - a^{*}\bar{x}$ and $b^{**} = \bar{y} - a^{**}\bar{x}$) we enlarge R to a region $\bar{R}$ corresponding to the fact that a and b may take any values within the intervals $[a^{**}, a^{*}]$ and $[b^{**}, b^{*}]$ respectively. The region $\bar{R}$ can be considered with practical approximation as the correct confidence region. If $| a^{**} - a^{*} |$ or $| b^{**} - b^{*} |$ is large, we may try some other grouping defined independently of the errors, which may lead to a smaller confidence region. In any case $\bar{R}$ represents a confidence region corresponding to a probability level greater than or equal to the chosen one.

8. Some Remarks on the Consistency of the Estimates of α, β, σ_ϵ, σ_η. We have shown in section 3 that the given estimates of α, β, σ_ϵ and σ_η are consistent if condition V is satisfied.

If the values $x_1, \cdots, x_N$ are not obtained by random sampling, it will in general be possible to define a grouping which is independent of the errors and for which condition V is satisfied. We can sometimes arrange the experiments such that no values of the series $x_1, \cdots, x_N$ should be within the interval $[x - c, x + c]$ where x denotes the median of $x_1, \cdots, x_N$ and c the range of the error ϵ. In such cases, as we saw, the grouping according to Rule I is independent of the errors. Condition V is certainly satisfied if we group the data according to Rule I.

Let us now consider the case that $X_1, \cdots, X_N$ are random variables independently distributed, each having the same distribution. Denote by X a random variable having the same probability distribution as possessed by each of the random variables $X_1, \cdots, X_N$. Assuming that X has a finite second moment, the expression in condition V will approach zero stochastically with $N \to \infty$ for any grouping defined independently of the values $X_1, \cdots, X_N$. It is possible, however, to define a grouping independent of the errors (but not independent of $X_1, \cdots, X_N$) for which the expression in V does not approach zero, provided that X has the following property: There exists a real value λ such that the probability that X will lie within the interval $[\lambda - c, \lambda + c]$ (c denotes the range of the error ϵ) is zero, the probability that $X > \lambda + c$ is positive, and the probability that $X < \lambda - c$ is positive. The grouping can be defined, for instance, as follows:

The i-th observation (x_i, y_i) belongs to the group G_1 if $x_i \leq \lambda$ and to G_2 if $x_i > \lambda$. We continue the grouping according to this rule up to a value i for which one of the groups G_1, G_2 contains already $N/2$ elements. All further observations belong to the other group.

It is easy to see that the probability is equal to 1 that the relation $x_i \leq \lambda$ is equivalent to the relation $X_i < \lambda - c$ and the relation $x_i > \lambda$ is equivalent to the relation $X_i > \lambda + c$. Hence this grouping is independent of the errors. Since for this grouping condition V is satisfied, our statement is proved.

If X has not the property described above, it may happen that for every grouping defined independently of the errors, the expression in condition V con-

verges always to zero stochastically. Such a case arises for instance if X, ϵ and η are normally distributed.[14] It can be shown that in this case no consistent estimates of the parameters α and β can be given, unless we have some additional information not contained in the data (for instance we know *a priori* the ratio $\sigma_\epsilon/\sigma_\eta$).

9. Structural Relationship and Prediction.[15] The problem discussed in this paper was the question as to how to estimate the relationship between the true parts X and Y. We shall call the relationship between the true parts the structural relationship. The problem of finding the structural relationship must not be confused with the problem of prediction of one variable by means of the other. The problem of prediction can be formulated as follows: We have observed N pairs of values $(x_1, y_1), \cdots, (x_N, y_N)$. A new observation on x is given and we have to estimate the corresponding value of y by means of our previous observations $(x_1, y_1), \cdots, (x_N, y_N)$. One might think that if we have estimated the structural relationship between X and Y, we may estimate y by the same relationship. That is to say, if the estimated structural relationship is given by $Y = aX + b$, we may estimate y from x by the same formula: $y = ax + b$. This procedure may lead, however, to a biased estimate of y. This is, for instance, the case if X, ϵ and η are normally distributed. It can easily be shown in this case that for any given x the conditional expectation of y is a linear function of x, that the slope of this function is different from the slope of the structural relationship, and that among all unbiased estimates of y which are linear functions of x, the estimate obtained by the method of least squares has the smallest variance. Hence in this case we have to use the least square estimate for purposes of prediction. Even if we would know exactly the structural relationship $Y = \alpha X + \beta$, we would get a biased estimate of y by putting $y = \alpha x + \beta$.

Let us consider now the following example: X is a random variable having a rectangular distribution with the range [0, 1]. The random variable ϵ has a rectangular distribution with the range $[-0.1, +0.1]$. For any given x let us denote the conditional expectation of y by $E(y \mid x)$ and the conditional expectation of X by $E(X \mid x)$. Then we obviously have

$$E(y \mid x) = \alpha E(X \mid x) + \beta.$$

Now let us calculate $E(X \mid x)$. It is obvious that the joint distribution of X and ϵ is given by the density function:

$$5\, dX\, d\epsilon,$$

---

[14] I wish to thank Professor Hotelling for drawing my attention to this case.

[15] I should like to express my thanks to Professor Hotelling for many interesting suggestions and remarks on this subject.

where X can take any value within the interval [0, 1] and ϵ can take any value within $[-0.1, +0.1]$. From this we obtain easily that the joint distribution of x and X is given by the density function

$$5\,dx\,dX,$$

where x can take any value within the interval $[-0.1, 1.1]$ and X can take any value lying in both intervals [0, 1] and $[x - 0.1, x + 0.1]$ simultaneously. Denote by I_x the common part of these two intervals. Then for any fixed x the relative distribution of X is given by the probability density

$$\frac{dX}{\int_{I_x} dX}.$$

Hence, we have

$$E(X \mid x) = \frac{\int_{I_x} X\,dX}{\int_{I_x} dX}.$$

We have to consider 3 cases:

(1) $$0.1 \leq x \leq 0.9.$$

In this case $I_x = [x - 0.1, x + 0.1]$ and

$$E(X \mid x) = \frac{\int_{x-0.1}^{x+0.1} X\,dX}{\int_{x-0.1}^{x+0.1} dX} = x.$$

(2) $-0.1 < x \leq 0.1$. Then $I_x = [0, x + 0.1]$ and

$$E(X \mid x) = \frac{\int_{0}^{x+0.1} X\,dX}{\int_{0}^{x+0.1} dX} = .5x + .05.$$

(3) $0.9 \leq x < 1.1$. Then $I_x = [x - 0.1, 1]$ and

$$E(X \mid x) = \frac{\int_{x-0.1}^{1} X\,dX}{\int_{x-0.1}^{1} dX} = .5x + .45.$$

Since

$$E(y \mid x) = \alpha E(X \mid x) + \beta,$$

we see that the structural relationship gives an unbiased prediction of y from x if $0.1 \leq x \leq 0.9$, but not in the other cases.

The problem of cases for which the structural relationship is appropriate also for purposes of prediction, needs further investigation. I should like to mention a class of cases where the structural relationship has to be used also for prediction. Assume that we have observed N values $(x_1, y_1), \cdots, (x_N, y_N)$ of the variables x and y for which the conditions I–IV of section 2 hold. Then we make a new observation on x obtaining the value x'. We assume that the last observation on x has been made under changed conditions such that we are sure that x' does not contain error, i.e. x' is equal to the true part X'. Such a situation may arise for instance if the error ϵ is due to errors of measurement and the last observation has been made with an instrument of great precision for which the error of measurement can be neglected. In such cases the prediction of the corresponding y' has to be made by means of the estimated structural relationship, i.e. we have to put $y' = ax' + b$.

The knowledge of the structural relationship is essential for constructing any theory in the empirical sciences. The laws of the empirical sciences mostly express relationships among a limited number of variables which would prevail exactly if the disturbing influence of a great number of other variables could be eliminated. In our experiments we never succeed in eliminating completely these disturbances. Hence in deducing laws from observations, we have the task of estimating structural relationships.

COLUMBIA UNIVERSITY,
NEW YORK, N. Y.