NOTE ON FITTING A STRAIGHT LINE WHEN BOTH VARIABLES ARE SUBJECT TO ERROR AND SOME APPLICATIONS

By Hirosi Hudimoto

(Received Jan. 23, 1956)

1. Introduction.

A simple method of fitting a straight line when both variables are subject to error was proposed by Wald [1] in 1940. The purpose of the present note is to give some practical device to that method and its application. If we have N observations (x_i, y_i) $(i=1, 2\cdots, N)$, the usually employed method of least squares determining the coefficients of the straight line, as is known, gives two different lines according to whether x or y is regarded as the independent variable. Concerning this much has been discussed in various papers, for example, in C. Eisenhart [3] and J. Berkson [4]. Here, we shall begin with considering the problem by Wald's method. Under some conditions he took up $b=\frac{b_2}{b_1}=\left(\sum_{i=m+1}^{N} y_i-\sum_{j=1}^{m} y_j\right)\Big/\left(\sum_{i=m+1}^{N} x_i-\sum_{j=1}^{m} x_j\right)$ as a consistent estimate of β in $Y=\alpha+\beta X$, and this was given a certain modification by Bartlett [2]. He classified the observations into three groups so that when the observations are arranged according to order, the first and last groups consist of the first and last k terms, respectively, where k denotes the integer nearest to $N/3$, while Wald classified the observations into two groups. To determine the slope the mean coordinates $\bar{x}_1$, $\bar{y}_1$ and $\bar{x}_3$, $\bar{y}_3$ for the two extreme groups were used, i.e., $b'=\frac{\bar{y}_3-\bar{y}_1}{\bar{x}_3-\bar{x}_1}$ was used. Bartlett's idea is that in the particular case where x has no errors and takes only positive integral values $1, 2, \cdots, 2l+1 (N=2l+1)$, the value of k equal to $(2l+1)/3$ maximizes the relative efficiency of b' between the variance for the method of least squares and the variance by this method.

In order to examine such a situation between their statistical constructions we shall take some actual examples. The following tables show coefficients $b_{x\cdot y}$, b and b' of the straight lines of the yield point Y (unit: kg/mm^2) on the ultimate tensile strength X (unit: kg/mm^2) of iron materials for ferro concrete given by the method of least squares, the Wald's method and the Bartlett's method. .

[25 mmϕ; diameters at a section]
$b_{y \cdot x}=0.600$, $b=0.583$, $b'=0.584$,
$N=440$, correlation coefficient $r=0.85$.

N' *	20	50	100	150	200
b	0.56	0.60	0.60	0.62	0.63
b'	0.63	0.64	0.60	0.66	0.64

* We selected $N'=20$, 50, 100, 150, 200 observations at random from a group of $N=440$ observations.

[16 mmϕ; diameters at a section]
$b_{y \cdot x}=0.79$, $b=0.76$, $b'=0.71$,
$N=560$, $r=0.89$.

N'	20	50	100	150	200
b	0.74	0.75	0.76	0.77	0.71
b'	0.85	0.80	0.77	0.74	0.67

[9 mmϕ; diameters at a section]
$b_{y \cdot x}=0.68$, $b=0.69$, $b'=0.70$,
$N=1724$, $r=0.86$

N'	20	50	100	150	200
b	0.76	0.78	0.76	0.79	0.78
b'	0.73	0.78	0.73	0.77	0.77

**

Those examples seem suitable for determination of the slopes b and b'. However, the problem to determine boundary points of this grouping precisely in the mathematical sence remains to us. A similar problem was discussed by Neyman and Scott [6].

2. A fitted straight line and some measure of trend.

Let us take a set of pairs of observed values (x_i, y_i), $i=1, 2, \cdots, N$. Denote the expected value $E(x_i)$ of x_i by X_i and the expected value

** These trials were carried through by Miss Kazuko Aihara and Miss.Eiko Ozaki.

$E(y_i)$ of y_i by Y_i, $i=1, 2, \cdots, N$. We shall call X_i the true value of x_i, Y_i the true value of y_i and accordingly consider $x_i - X_i = \varepsilon_i$ as the error in the i-th term of the X-set, and $y_i - Y_i = \varepsilon_i'$ as the error in the i-th term of the Y-set. Further assume that a single linear relation holds between the true values X and Y, i.e., $Y_i = \alpha + \beta X_i$, and all the random variable ε_i have the same distribution, all ε_i' also have the same one, and $E(\varepsilon_i \varepsilon_j)=0$, $E(\varepsilon_i' \varepsilon_j')=0$ for $i \neq j$, $E(\varepsilon_i \varepsilon_j')=0$, $(i=1, \cdots, N)$.

Following Wald's method, if we take the statistic $b=\frac{b_2}{b_1}=\frac{\bar{y}_2-\bar{y}_1}{\bar{x}_2-\bar{x}_1}$ as an estimate of β, it can easily be seen that

$$b=\left(\beta+\frac{\bar{\varepsilon}_2'-\bar{\varepsilon}_1'}{\bar{X}_2-\bar{X}_1}\right)\left(1-\frac{\bar{\varepsilon}_1-\bar{\varepsilon}_2}{\bar{X}_2-\bar{X}_1}\right)^{-1}, \tag{2.1}$$

providing $\left(\frac{\bar{\varepsilon}_1-\bar{\varepsilon}_2}{\bar{x}_2-\bar{x}_1}\right)<1$, where, for the sake of simplicity, we assume that $N(>2)$ is even, and $\bar{x}_1, \bar{y}_1, \bar{\varepsilon}_1, \bar{\varepsilon}_1'$ and $\bar{x}_2, \bar{y}_2, \bar{\varepsilon}_2, \bar{\varepsilon}_2'$ denote the arithmetic means of the observed values x and y, and the error ε and ε' on the first half group and the second half group in the ordered arrangements $x_1 < x_2 < \cdots < x_N$, when $m=N/2$, i.e.,

$$\bar{x}_1=\frac{1}{m}\sum_{i=1}^{m} x_i\,, \qquad \bar{y}_1=\frac{1}{m}\sum_{i=1}^{m} y_i\,, \qquad \bar{\varepsilon}_1=\frac{1}{m}\sum_{i=1}^{m} \varepsilon_i\,,$$

$$\bar{\varepsilon}_1'=\frac{1}{m}\sum_{i=1}^{m} \varepsilon_i'\,;$$

and
$$\bar{x}_2=\frac{1}{m}\sum_{i=m+1}^{N} y_i\,, \qquad \bar{y}_2=\frac{1}{m}\sum_{i=m+1}^{N} y_i\,, \qquad \bar{\varepsilon}_2=\frac{1}{m}\sum_{i=m+1}^{N}\,,$$

$$\bar{\varepsilon}_2'=\frac{1}{m}\sum_{i=m+1}^{N} \varepsilon_i'\,.$$

Thus, if we take $\hat{Y}_i = a + bX_i$ as estimates of Y_i''s and write $y_i - \hat{Y}_i$ by w_i, we have

$$\begin{aligned} w_i &= y_i - (\alpha+\beta X_i) - \bar{\varepsilon}' + \beta\bar{\varepsilon} \\ &\quad - \frac{(\bar{\varepsilon}_2'-\bar{\varepsilon}_1')-\beta(\bar{\varepsilon}_2-\bar{\varepsilon}_1)}{\bar{X}_2-\bar{X}_1}\left(1-\frac{\bar{\varepsilon}_1-\bar{\varepsilon}_2}{\bar{X}_2-\bar{X}_1}\right)^{-1}\cdot(X_i-\bar{x})\,. \end{aligned} \tag{2.2}$$

Now, when we obtain a further observation (x_i', y_i') and put $w_i' = y_i' - \hat{Y}_i$, the expected value of w_i' for fixed y_i' is given as follows:

$$(2.3)\qquad E(w_i')=y_i'-(\alpha+\beta X_i')-\beta(X_i'-\overline{X})\left\{\underset{\varepsilon,\varepsilon'}{E}\left(\frac{\bar{\varepsilon}_1-\bar{\varepsilon}_2}{\overline{X}_2-\overline{X}_1}\right)^2 + \underset{\varepsilon,\varepsilon'}{E}\left(\frac{\bar{\varepsilon}_1-\bar{\varepsilon}_2}{\overline{X}_2-\overline{X}_1}\right)^3+\cdots\right\}$$

$$\doteqdot y_i'-(\alpha+\beta X_i')+\beta\Bigg[(\overline{X}_2-\overline{X}_i')\left\{\frac{\underset{\varepsilon}{E}(\bar{\varepsilon}_1^2)}{(\overline{X}_2-\overline{X}_1)^2}-\frac{\underset{\varepsilon}{E}(\bar{\varepsilon}_1^3)}{(\overline{X}_2-\overline{X}_1)^3}+\frac{\underset{\varepsilon}{E}(\bar{\varepsilon}_1^4)}{(\overline{X}_2-\overline{X}_1)^4}\right\}$$

$$-(\overline{X}_i'-\overline{X}_1)\left\{\frac{\underset{\varepsilon}{E}(\bar{\varepsilon}_2^2)}{(\overline{X}_2-\overline{X}_1)^2}-\frac{\underset{\varepsilon}{E}(\bar{\varepsilon}_2^3)}{(\overline{X}_2-\overline{X}_1)^3}+\frac{\underset{\varepsilon}{E}(\bar{\varepsilon}_2^4)}{(\overline{X}_2-\overline{X}_1)^4}\right\}\Bigg]$$

The result shows that the influence of this bias is small as far as X_i' fall between $\overline{X}_1$ and $\overline{X}_2$, and if X_i' has no error, the result is, of course, unbiased.

Now, let us introduce the random variable Z such that

$$(2.4)\qquad Z=\begin{cases}1 & \text{when } y_i'-(\alpha+\beta X_i')>0\\ 0 & \text{when } y'-(\alpha+\beta X_i')<0\end{cases}$$

and consider the problem to estimate the probability p that $Z=1$. For this problem we take further n observations $(x_1', y_1'), \cdots, (x_n', y_n')$, and compute w_i' to every $(x_i'; y_i')$. When we obtain n^* positive values of w_i, $\hat{p}=n^*/n$ will be appropriate for an estimate of p, provided that the terms of the higher order than the second moment of $\left(\frac{\bar{\varepsilon}_1-\bar{\varepsilon}_2}{\overline{X}_2-\overline{X}_1}\right)$ may be negligible.

Further, put $y_i-\alpha-\beta(\overline{X}_i-X)=\eta_i$. We then have

$$(2{\cdot}5)\qquad \underset{\varepsilon,\varepsilon'}{E}\left\{w_i-\eta_i\right\}^2=\underset{\varepsilon\varepsilon'}{E}\left\{(\bar{\varepsilon}'-\beta\bar{\varepsilon}) + \frac{(\bar{\varepsilon}_2'-\bar{\varepsilon}_1')-\beta(\bar{\varepsilon}_2-\bar{\varepsilon}_1)}{\overline{X}_2-\overline{X}_1}\left(1-\frac{\bar{\varepsilon}_1-\bar{\varepsilon}_2}{\overline{X}_2-\overline{X}_1}\right)^{-1}\cdot(X_i-\overline{X})\right\}^2$$

If the contribution of the terms of the higher order than the second moment of $(\bar{\varepsilon}_1-\bar{\varepsilon}_2)/(\overline{X}_2-\overline{X}_1)$ is negligible, we have

$$(2.6)\qquad \underset{\varepsilon,\varepsilon'}{E}[w_i-\eta_i]^2\doteqdot\underset{\varepsilon,\varepsilon'}{E}(\bar{\varepsilon}'-\beta\bar{\varepsilon})^2+\frac{\underset{\varepsilon,\varepsilon'}{E}[b_1^2(b-\beta)^2]}{(\overline{X}_2-\overline{X}_1)^2}\cdot(X_i-\overline{X})^2,$$

for in this case we have $E[\bar{\varepsilon}(\bar{\varepsilon}_1-\bar{\varepsilon}_2)=[E(\bar{\varepsilon}_1^2)-E(\bar{\varepsilon}_2^2)]/2\doteqdot 0$, and $E[\bar{\varepsilon}'(\bar{\varepsilon}_2'-\bar{\varepsilon}_1')]\doteqdot 0$. Further, if ε, ε' have the Gaussian distributions with zero mean and variance σ_ε^2 and $\sigma_{\varepsilon'}^2$, respectively, we obtain

$$(2.7)\qquad \underset{\varepsilon,\varepsilon'}{E}[w_i-\eta_i]^2 \doteqdot \frac{\sigma_{\varepsilon'}^2+\beta^2\sigma_\varepsilon^2}{N}\left\{1+\frac{(X_i-\overline{X})^2}{(\overline{X}_2-\overline{X}_1)^2}\right\}.$$

Let $F(\eta)$, $G(\xi)$ and $f(\eta)$, $g(\xi)$ denote the distribution functions (dist. f.) and the density functions (dens. f.) of η and ξ, respectively, and put $w=\eta+\xi$. We shall consider w for a sufficiently large value of N, and suppose $g(x)$ to have the Gaussian distribution with zero mean. Then $g(\xi)$ is symmetric.

If we denote the distribution function of $w=\eta+\xi$ by $\tilde{F}(x)$, we have

$$(2.8)\qquad \tilde{F}(x)=\int\!\!\int_{\eta+\xi\leqq x} f(\eta)g(\xi)d\xi d\eta=\int_{-\infty}^{\infty}F(x-\xi)g(\xi)d\xi\,.$$

In the following we shall restrict ourselves to the case where the dist. f. $\tilde{F}(x)$ satisfies the condition$\tilde{F}'(x)=\int_{-\infty}^{\infty}f(x-\xi)g(\xi)d\xi=\tilde{f}(x)$. Now, suppose, we have n values of w and among them n_1 values are less than x, one is between x and $x+dx$, and the remaining $n-n_1-1$ values are greater than $x+dx$. Then we have

$$(2.9)\qquad \omega(x)dx=\frac{n\,!}{n\,!(n-n_1-1)\,!}(\tilde{F}(x))^{n_1}(1-\tilde{F}(x))^{n-n_1-1}$$
$$\times\int_{-\infty}^{\infty}f(x-\xi)g(\xi)d\xi dx\,.$$

When we introduce a new variable τ by the substitution

$$\tau=n\tilde{F}(x)\,,$$

we have $0\leqq\tau\leqq n$, and the dens. f. $w^*(\tau)$ of the new variable is

$$(2.10)\qquad \omega^*(\tau)=\binom{n-1}{n_1}\left(\frac{\tau}{n}\right)^{n_1}\left(1-\frac{\tau}{n}\right)^{n-n_1-1}.$$

Now let us consider the relation between $\tilde{F}(x)$ and $F(\eta)$. However, the origin of η is unknown for us. Thus we shall select any y-axis parallel to ξ-axis and determine the origin on it, and if the origin of η agrees with the origin of y, we have

$$(2.11)\qquad \tilde{F}(x)=\int_{-\infty}^{\infty}F(x-\xi)g(\xi)d\xi=\int_{-\infty}^{\infty}F(x+\xi)g(\xi)d\xi=\tilde{\tilde{F}}(x)\,,$$

or

$$\tilde{F}(x)=\int_{-\infty}^{\infty}F(-x-\xi)g(\xi)d\xi=\int_{-\infty}^{\infty}F(-x+\xi)g(\xi)d\xi=\tilde{\tilde{F}}(-x)\,,$$

since $g(\xi)$ is symmetric. But when we can assume that the origin of ξ is at $-y_0$ or y_0, and η has its origin at y_0 or $-y_0$, where $y_0\geqq 0$, then $\tilde{\tilde{F}}(2y_0)$ changes to $\tilde{\tilde{F}}(0)$ and $\tilde{F}(2y_0)$ remains unchanged or $\tilde{\tilde{F}}(-2_0)$ changes to $\tilde{\tilde{F}}(0)$ and $\tilde{F}(-2y_0)$ remains unchanged, and we have

$$\tilde{F}(2y_0)\geqq\tilde{F}(0)=\tilde{\tilde{F}}(0)\,, \tag{2.12}$$
$$\tilde{F}(-2y_0)\leqq\tilde{F}(0)=\tilde{\tilde{F}}(0)\,.$$

According to this result, we have only to consider our problem for the interval $(-y_0, y_0)$ such that the probability that the origin of ξ falles into the interval has a sufficiently large value. Now if we assume $F(-(x-\xi))=1-F(x-\xi)$, i.e., $F(\eta)$ is symmetric with respect to its origin, we have $\tilde{F}(-x)=1-\tilde{F}(x)$ and $\tilde{\tilde{F}}(-x)=1-\tilde{\tilde{F}}(x)$. Thus if $\tilde{F}(x)$ or $\tilde{\tilde{F}}(x)$ is not symmetric, $F(\eta)$ is not symmetric, and if the origin of ξ is found and $\tilde{\tilde{F}}(0)$ is symmetric, we get $\tilde{\tilde{F}}(0)=F(0)$. Thus we can attain our purpose by considering $\tilde{F}(x)$ at $-2y_0$ and $2y_0$.

Our practical procedure was as follows. We inscribe two lines of the result computed according to the following procedure in the scatter diagram ploted to show the relation between the observations x and y. Since $\dfrac{b_1(b-\beta)\sqrt{N-2}}{\sqrt{(s_y')^2+\beta^2(s_x')^2-2s_{xy}'}}$ has the Student's t-distribution with $N-2$ degrees of freedom (see Wald [1]),

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

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

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

and, therefore, has assymptotically the Gaussian distribution for a sufficiently large N, we obtain the two lines which determine the values of β with the desired confidence coefficient. Then count the numbers n_1, n_2 of observations on y which will fall in the above sides of those two lines, respectively. We further put $p_1=\dfrac{n_1}{n}$ and $p_2=\dfrac{n_2}{n}$ and take

$\left(p_1-k_0\sqrt{\frac{1}{n}\left(\frac{1}{2}\right)^2}\,,\ p_2+k_0\sqrt{\frac{1}{n}\left(\frac{1}{2}\right)^2}\right)$ as a confidence interval for our purpose.

3. Numerical Examples.

The following examples show some results tested on the compressive strength and the tensile strength by test-pieces made by the mortar. The amount of cement used for the testing was weighed out of the total with the prescribed proportion when it was bought from the maker, and several numbers of test-pieces were produced from mortar (water cement ratio 65%, cement/sand=1/2) that contained this cement, and the test-pieces were kept in water, and examined at the 3rd-day, 7th-day and 28th-day, respectively.*** (We were furnished with these data from the material test room in Nippon Telephone and Telegram Public Coorporation.)

⟨The compressive strength⟩ $N=285$

X; strength at the 3rd-day

Y; strength at the 28th-day (unit: kg/cm^2)

arithmetic mean $\bar{x}=118.59$, $\bar{y}=382.14$,

standard deviation $s_x=22.61$ $s_y=38.11$.

$b_{y,x}=0.520$, $b=0.599$, $r=0.308$.

⟨The tensile strength⟩

X; strength at the 3rd-day,

Y; strength at the 28th-day (unit: kg/cm^2)

arithmetic mean $\bar{x}=29.42$, $\bar{y}=73.61$,

standard deviation $s_x=5.65$, $s_y=8.78$,

$b_{y,x}=0.834$, $b=0.842$, $r=0.537$.

When we designate by t_0 the critical value of t corresponding to 95% probability level, the following results are obtained for β_1^* and β_2^*, that is, for the compressive strength between the 3rd-day and 28th-day, (0.348, 0.854), and for the tensile strength, (0.654, 1.032).

*** All numerical works were done by Miss Kazuko Aihara and Miss Eiko Ozaki.

Compressive strength. $p_1=N_1/N=0.52$, $p_2=N_2/N=0.56$			Tensile strength. $p_1=N_1/N=0.44$, $p_2=N_2/N=0.47$		
Maker's name	n_1/n	n_2/n		n_1/n	n_2/n
O	0.58	0.63		0.56	0.63
I	0.63	0.65		0.26	0.32
T	0.28	0.40		0.33	0.49
N	0.50	0.56		0.52	0.56
U	0.54	0.58		0.70	0.71

The similar work is performed for the results between the 7-th day and the 28-th day. (β_1^*, β_2^*) is (0.0338, 0.639) for the compressive strength between the 7th day and the 28th day and (0.431, 0.800) for the tensile strength.

Compressive strength. $p_1=0.47$ $p_2=0.50$			Tensile strength $p_1=0.47$ $p_2=0.48$		
Maker's name	n_1/n	n_2/n		n_1/n	n_2/n
O	0.52	0.61		0.55	0.58
I	0.54	0.58		0.18	0.28
T	0.23	0.30		0.40	0.60
N	0.48	0.52		0.52	0.60
U	0.53	0.58		0.75	0.80

The above tables seem to show the certain tendency of behaviour on the consolidation of the manufactures of the various makers. For example, if we use at the same time the following table of their mean strengths about each maker's goods, we can possibly guess from the tables of compressive strength that *T* maker's goods become solid quickly, and from the tables of tensile strength that *U* maker's goods have the

Sample mean strength of each maker's manufactured goods,
unit: kg/mm^2

	Compressive strength			Tensile strength		
Maker's name	3rd-day	7th-day	28th-day	3rd-day	7th-day	28th-day
O	123.57	239.99	392.57	31.11	53.90	78.40
I	103.75	215.49	375.30	24.60	45.49	67.20
T	134.74	253.59	381.24	32.37	53.58	75.37
N	118.60	228.82	382.16	29.46	50.18	73.67
U	122.13	229.83	391.52	30.98	51.82	78.82

highest firmness against the brittleness among these mak ers. We can regard these properties of O and U as excellent among them.

THE INSTITUTE OF STATISTICAL MATHEMATICS

REFERENCES

[1] Wald, A., The fitting of straight lines if both variables are subject to error, *Ann. Math. Stat.* Vol. 11 (1940).

[2] Bartlett, M. S., Fitting a straight line when both variables are subject to error, *Biometrics*, Vol. 5 (1949).

[3] Eisenhart, C., The interpretation of certain regression methods and their use in biological and industrial research, *Ann. Math. Stat.* Vol. 10 (1939).

[4] Berkson, J., Are there two regressions? *Jour. Amer. Stat. Assoc.* Vol. 45 (1950).

[5] *The report of material test room* in Nippon Telephone and Telegram Public Corporation, No. 7 (1956).

[6] Neyman, J. and Elizabeth, L. Scott, On certain method of estimating the linear structural relation, *Ann. Math. Stat.* Vol. 22 (1951).