THE ESTIMATION OF LINEAR TRENDS

By G. W. Housner and J. F. Brennan

California Institute of Technology

1. Summary. This paper deals with the problem of bivariate regression where both variates are random variables having a finite number of means distributed along a straight line. A regression statistic is derived which is independent of change in scale so that a prior knowledge of the frequency distribution parameters is not required in order to obtain a unique estimate. The statistic is shown to be consistent. The efficiency of the estimate is discussed and its asymptotic distribution is derived for the case when the random variables are normally distributed. A numerical example is presented which compares the performance of the statistic of this paper with that of other commonly used statistics. In the example it is found that the method of estimation proposed in this paper is more efficient.

2. Introduction. A problem that often arises in statistical work is the estimation of linear trends. In the general problem it is known or presumed that a linear functional relation exists among a set of variables of the form,

$$a + b_1X + b_2Y + b_3Z \cdots = 0.$$

The observed values of the variables are of the form

$$x_{ik} = X_i + \epsilon_{ik}, \qquad y_{ik} = Y_i + \eta_{ik}, \quad \text{etc.}$$

That is, the x_{ik} are random variables with means X_i and $k = 1, 2, \cdots N_i$ observed values of x are associated with the mean X_i. The ordering of the X_i is according to magnitude. Similarly there are the observed values y_{ik}, z_{ik} and so forth. The ϵ_{ik} are random variables, with the same distribution for all i, with zero means. On the basis of a sample $O_n(x_{ik}, y_{ik}, z_{ik}, \cdots)$ it is desired to estimate the coefficients $a, b_1, b_2, b_3, \cdots$. One method used to estimate the coefficients is that of "weighted regression" which is essentially an application of the method of least squares. The problem has been studied by R. Allen, A. Wald and others.[1] The chief difficulty has been that the proposed methods of estimation require an a priori knowledge of the variances of the random variables. Wald has proposed a statistic which avoids this difficulty but which may have a relatively low efficiency in cases often encountered in practice. In this paper there is described a bivariate statistic which appears to have comparatively high precision and which does not require prior knowledge of the variances of the random variables. A numerical example is given at the end of the paper to illustrate the comparative performances of different methods of estimation.

[1] For a brief history of work done on this problem see the paper by A. Wald in the *Annals of Math. Stat.*, Vol. 11 (1940), p. 284.

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to *The Annals of Mathematical Statistics.* www.jstor.org

3. The Regression statistic. In the case of the bivariate problem, consider a sample

$$O_n(x_{ik}, y_{ik}), i = 1, 2, \cdots, n$$

and

$$k = 1, 2, \cdots, N_i,$$

where N_i sample values x_i, y_i are distributed about mean X_i, Y_i. Let the means be related by $Y_i = a + bX_i$ and let the random variables x_i be independent and have the same frequency distribution with variance σ_x^2 for all i and the random variables y_i have independent frequency distributions with variance σ_y^2 the same for all i. An appropriate statistic for estimating b is obtained by noting that a pair of sample points (x_{ik}, y_{ik}), (x_{jl}, y_{jl}) gives a sample value of the change in y corresponding to a change in x. It may thus be said that a sample value of b is

$$\hat{b}_{ik,jl} = \frac{y_{ik} - y_{jl}}{x_{ik} - x_{jl}}. \tag{1}$$

Making use of the fact that

$$y_{ik} = a + bx_{ik} + \eta_{ik} - b\epsilon_{ik} \tag{2}$$

equation (1) may be written

$$(x_{ik} - x_{jl})\,\hat{b}_{ik\,,jl} = (x_{ik} - x_{jl})\,b + (\eta_{ik} - \eta_{jl}) - b(\epsilon_{ik} - \epsilon_{jl}).$$

Summing this equation over all combinations of points there is obtained

$$b = \frac{\sum_i\sum_j\sum_k\sum_l (y_{ik} - y_{jl})}{\sum_i\sum_j\sum_k\sum_l (x_{ik} - x_{jl})} - \frac{\sum_i\sum_j\sum_k\sum_l ((\eta_{ik} - \eta_{jl}) - b(\epsilon_{ik} - \epsilon_{jl})}{\sum_i\sum_j\sum_k\sum_l (x_{ik} - x_{jl})}. \tag{3}$$

The summations in the above expression are to be carried out for

$l = 1, 2, \cdots, N_j$; $k = 1, 2, \cdots, N_i$; $j = 1, 2, \cdots, (i - 1)$; $i = 1, 2, \cdots, n$.

The first term on the right side of equation (3) is an estimate of b and the second term represents the deviation of the estimate from the true value. Accordingly, we take as an estimate of b the statistic

$$\hat{b} = \frac{\sum_i\sum_j\sum_k\sum_l (y_{ik} - y_{jl})}{\sum_i\sum_j\sum_k\sum_l (x_{ik} - x_{jl})}. \tag{4}$$

This requires, of course, that the denominator be not equal to zero. Summing out the subscripts k and l reduces (4) to

$$\hat{b} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{i-1} N_i N_j(\bar{y}_i - \bar{y}_j)}{\sum_{i=1}^{n}\sum_{j=1}^{i-1} N_i N_j(\bar{x}_i - \bar{x}_j)}$$

where $\bar{y}_i$ is the mean value of the y_{ik} and so forth. Summing out the subscript j gives

$$\hat{b} = \frac{\sum_{i} \left(N_i \bar{y}_i \sum_{1}^{i-1} N_j - N_i \sum_{1}^{i-1} N_j \bar{y}_j \right)}{\sum_{i^{`}} \left(N_i \bar{x}_i \sum_{1}^{i-1} N_j - N_i \sum_{1}^{i-1} N_j \bar{x}_j \right)}. \tag{5}$$

This expression may be put in a more convenient form by using the identity

$$\sum_{i=1}^{n} \left(N_i \sum_{1}^{i-1} N_j \bar{y}_j \right) = \sum_{i=1}^{n} \left(N_i \bar{y}_i \sum_{i+1}^{n} N_j \right) = \sum_{i=1}^{n} \left(N_i \bar{y}_i \left(\sum_{1}^{n} N_j - \sum_{1}^{i} N_j \right) \right).$$

With this substitution equation (5) becomes

$$\hat{b} = \frac{\sum_{i=1}^{n} \left[N_i \bar{y}_i \left(\sum_{1}^{n} N_j - 2 \sum_{1}^{i} N_j + N_i \right) \right]}{\sum_{i=1}^{n} \left[N_i \bar{x}_i \left(\sum_{1}^{n} N_j - 2 \sum_{1}^{i} N_j + N_i \right) \right]}. \tag{6}$$

This is the statistic for estimating the linear trend of bivariate data. It may be noted that its derivation is not based on the notion of fitting a line to the sample points. A line $y = \hat{a} + \hat{b}x$ may be fitted to the sample points by making it pass through the mean of the sample points, that is, by using the following estimate:

$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

where $\bar{y}$ and $\bar{x}$ are the means of all the y_{ik} and x_{ik} respectively.

4. Consistency of the estimate. Having established the statistics $\hat{b}$ and $\hat{a}$ it is desirable to examine the consistency and efficiency of the estimates, particularly for $\hat{b}$. To determine that $\hat{b}$ is a consistent estimate we investigate the behavior of (6) as the number of sample points increases, that is, as the $N_i \to \infty$. We wish first to establish the following identity. Consider the sum of the following array of terms:

$$\begin{array}{l} N_1(N_1 + N_2 + \cdots + N_n) \\ N_2(N_1 + N_2 + \cdots + N_n) \\ \vdots \\ N_n(N_1 + N_2 + \cdots + N_n) \end{array}$$

The sum may be written $\sum_{1}^{n} N_i \sum_{1}^{n} N_j$. Since the array is skew symmetrical the expression $2 \sum_{1}^{n} N_i \sum_{1}^{i} N_j$ also gives the sum of the array except for the fact that the terms along the principal diagonal are counted twice. We have, therefore

$$\sum_{1}^{n} N_i \sum_{1}^{n} N_j = 2 \sum_{1}^{n} N_i \sum_{1}^{i} N_j - \sum_{1}^{n} N_i^2.$$

Rearranging terms we obtain the identity

$$\sum_1^n \left[N_i\left(\sum_1^n N_j - 2\sum_1^i N_j + N_i\right)\right] = 0. \tag{7}$$

Now substituting (2) into (6) and making use of (7) there is obtained,

$$\hat{b} = b + \frac{\sum_1^n \left[N_i\left(\sum_1^n N_j - 2\sum_1^i N_j + N_i\right)(\bar{\eta}_i - b\bar{\epsilon}_i)\right]}{\sum_1^n \left[N_i\left(\sum_1^n N_j - 2\sum_1^i N_j + N_i\right)\bar{x}_i\right]}. \tag{8}$$

The $\bar{\eta}_i$ and $\bar{\epsilon}_i$ are random variables with zero means so that as $N_i \to \infty$ the sample means $\bar{\eta}_i$ and $\bar{\epsilon}_i$ converge in probability to zero. As $N_i \to \infty$, $\bar{x}_i$ converges in probability to its mean X_i. In view of (7) and that the denominator in (8) is not equal to zero the last term in (8) converges in probability to zero and $\hat{b} \to b$. The estimate is therefore consistant. A similar argument also shows the estimate $\hat{a}$ to be consistent.

5. Efficiency of the estimate. A general investigation of the efficiency of the estimate $\hat{b}$ is beyond the scope of this paper. We may note, however, that the efficiency of the estimate can be made to depend upon the grouping of the data, that is, the optimum efficiency of the estimate may depend upon the omission of some of the pairs $(y_{ik} - y_{jl})$ from the estimate. The maximum efficiency is obtained for $\hat{b}$ when the second term in (3) is minimized. This requires prior knowledge of the frequency distribution of the random variables x and y; however, in applications a recognition of (3) may often indicate a practical method of increasing the efficiency.

In what follows we make an investigation of the precision of the estimate $\hat{b}$ for a special case which is of some practical interest. Let x and y be random variables as defined in the first part of the paper and consider the new variables defined by $\hat{b} = \frac{v}{u}$ that is,

$$u = \sum_1^n \left[N_i\left(\sum_1^n N_j - 2\sum_1^i N_j + N_i\right)\bar{x}_i\right]$$

$$v = \sum_1^n \left[N_i\left(\sum_1^n N_j - 2\sum_1^i N_j + N_i\right)\bar{y}_i\right].$$

The random variables u and v are then independently distributed with joint probability element $f(u) f(v)\, du\, dv$. Making the change of variable $u = r \cos \theta$, $v = r \sin \theta$ the probability element becomes $f(r, \theta) dr\, d\theta$ where $\tan \theta = u$. Integrating out the variable r gives the probability element for θ. In what follows we investigate the distribution of θ for the case where x and y are normally distributed with the same variance. Since u and v are linear functions of x and y respectively they are also normally distributed with the same standard deviation.

We designate the means of u and v by m_1 and m_2 respectively and the standard deviation by σ. The probability element in u and v is then

$$\frac{1}{2\pi\sigma^2} \exp\left\{-\frac{1}{2\sigma^2}\left[(u - m_1)^2 + (v - m_2)^2\right]\right\} du\, dv. \tag{9}$$

Changing variables to r, θ and setting $m_1 = \bar{r} \cos \bar{\theta}$, $m_2 = \bar{r} \sin \bar{\theta}$ we obtain the following probability element:

$$\frac{r}{2\pi\sigma^2} \exp\left\{-\frac{1}{2\sigma^2}\left[(r \cos \theta - \bar{r} \cos \bar{\theta})^2 + (r \sin \theta - \bar{r} \sin \bar{\theta})\right]\right\} dr\, d\theta.$$

Completing the square in r and substituting $\phi = \theta - \bar{\theta}$ there is obtained

$$\frac{r}{2\pi\sigma^2} \exp\left\{-\frac{1}{2\sigma^2}(r - \bar{r} \cos \phi)^2\right\} \exp -\tfrac{1}{2}\left(\frac{\bar{r} \sin \phi}{\sigma}\right)^2\Bigg\} dr\, d\phi. \tag{10}$$

To integrate out r make further change of variable

$$t = \frac{r}{\sigma} - \frac{\bar{r}}{\sigma} \cos \phi.$$

Setting $\frac{\bar{r}}{\sigma} \cos \phi = w$ for convenience in notation there is obtained

$$\left(\frac{1}{2\pi} t \exp\left\{-\frac{t^2}{2}\right\} + \frac{w}{2\pi} \exp\left\{-\frac{t^2}{2}\right\}\right) \exp\left\{-\tfrac{1}{2}\left(\frac{\bar{r}^2}{\sigma^2} - w^2\right)\right\} dt\, d\phi.$$

The variable t is to be integrated out of this expression. The corresponding limits of integration are exhibited by

$$\begin{aligned} &\frac{w}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{\bar{r}^2}{\sigma^2} - w^2\right)\right\}\left(\frac{1}{\sqrt{2\pi}} \int_{-w}^{+\infty} \frac{t}{w} \exp\left\{-\frac{t^2}{2}\right\} dt\right. \\ &\qquad\qquad \left. + \frac{1}{\sqrt{2\pi}} \int_{-w}^{+\infty} \exp\left\{-\frac{t^2}{2}\right\} dt\right) d\phi. \end{aligned} \tag{12}$$

Now as the number of points in the original sample increases the value of $\bar{r}$ also increases and as $\frac{\sigma}{\bar{r}} \to 0$, with $|\phi| < \frac{\pi}{2}$, the value of $w \to \infty$. In this case then (12) approaches asymptotically to

$$\frac{\bar{r}}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}\left(\frac{\sin \phi^2}{\sigma/\bar{r}}\right)^2\right\} d(\sin \phi).$$

As $\sigma/\bar{r} \to 0$ this distribution shows that ϕ converges in probability to zero and that the distribution approaches asymptotically to the normal form

$$\frac{1}{\sqrt{2\pi}\sigma/\bar{r}} \exp\left\{-\tfrac{1}{2}\left(\frac{\phi}{\sigma/\bar{r}}\right)^2\right\} d\phi. \tag{13}$$

It is required then to examine the conditions under which $\sigma/\bar{r}$ assumes small values. If the variance of the original variables x_i and y_i is designated by σ_1^2

then since u and v are linear functions of x_i and y_i respectively the variance of u and of v is

$$(14) \qquad \sigma^2 = \sigma_1^2 \sum_1^n \left\{ \left[N_i \left(\sum_1^n N_j - 2 \sum_1^i N_j + N_i \right) \right]^2 \left(\frac{1}{N_i} \right) \right\}.$$

Now $\bar{r}^2$ is the sum of the squares of the means of u and v so that

$$(15) \qquad \bar{r}^2 = (1 + b^2) \left\{ \sum_1^n \left[N_i \left(\sum_1^n N_j - 2 \sum_1^i N_j + N_i \right) X_i \right] \right\}^2.$$

Dividing (14) by (15) we obtain

$$(16) \qquad \left(\frac{\sigma}{\bar{r}} \right)^2 = \frac{\sigma_1^2}{1 + b^2} \frac{\sum_1^n \left\{ \left[N_i \left(\sum_1^n N_j - 2 \sum_1^i N_j + N_i \right) \right]^2 \left(\frac{1}{N_i} \right) \right\}}{\left\{ \sum_1^n \left[N_i \left(\sum_1^n N_j - 2 \sum_1^i N_j + N_i \right) X_i \right] \right\}^2}.$$

Inspection of (16) indicates that as the number of sample points N_i increases the value of $\left(\frac{\sigma}{\bar{r}} \right)^2$ decreases rapidly. To illustrate this we examine some particular cases. Consider first the case of four equally spaced means $X_i = 3i\sigma_1$, $(i = 1, 2, 3, 4)$ and let there be one sample point for each mean $(N_i = 1)$. With these values there is obtained,

$$\left(\frac{\sigma}{\bar{r}} \right)^2 = \frac{0.022}{1 + b^2}.$$

For $b = 1$ the rañge $-9° < \phi < +9°$ includes 95% of the population defined by (13). As the number of points N_i is increased or as the number of means X_i is increased the value of $\left(\frac{\sigma}{\bar{r}} \right)^2$ decreases rapidly. Consider now eight equally spaced means $X_i = 3_i\sigma_1$, $(i = 1, 2, \cdots, 8)$ with again one sample point for each mean $(N_i = 1)$. With these values there is obtained

$$\left(\frac{\sigma}{\bar{r}} \right)^2 = \frac{0.00045}{1 + b^2}.$$

For $b = 1$ the range $-1° < \phi < +1°$ includes 95% of the population defined by (13).

It is clear that a very high degree of precision is obtained with the estimate $\hat{b}$ when there is a considerable number of sample points. However, this will also be true in general of other statistics and it is really of interest to compare precisions in those cases where the statistics have a relatively low precision. A detailed comparison is beyond the scope of this paper. However, a direct comparison can be made very easily in the particular case when x_i is a fixed variate and only y_i is a random variable. For the sake of simplicity, let each $N_i = 1$ then the statistic for estimating b is

$$\hat{b} = \frac{\sum_1^n i(y_i - \bar{y})}{\sum_1^n i(x_i - \bar{x})} = \frac{\sum_1^n y_i(i - \bar{\imath})}{\sum_1^n x_i(i - \bar{\imath})}. \tag{17}$$

Since $\hat{b}$ is a linear function of the y_i by a well known theorem its variance is

$$\sigma_{\hat{b}}^2 = \sigma_y^2 \sum_1^n \left(\frac{i - \bar{\imath}}{\sum_1^n x_i(i - \bar{\imath})} \right)^2. \tag{18}$$

The customary least squares regression line of y on x gives for the estimate of b and its variance

$$\hat{b}_R = \frac{\sum_1^n y_i(x_i - \bar{x})}{\sum_1^n x_i(x_i - \bar{x})} \qquad \sigma_{\hat{b}_R}^2 = \sigma_y^2 \sum_1^n \left(\frac{x_i - \bar{x}}{\sum_1^n x_i(x_i - \bar{x})} \right)^2.$$

In the particular case when the x_i are equally spaced, $x_i = ci + d$, the estimates $\hat{b}$ and $\hat{b}_R$ are identical:

$$\hat{b} = \hat{b}_R = \frac{12}{cn(n^2 - 1)} \sum_1^n y_i(i - \bar{\imath}). \tag{19}$$

6. Numerical example. From a practical point of view the case where x and y are random variables is of greater interest than where x is a fixed variate. We give a numerical example of this case comparing the statistic $\hat{b}$ with several other statistics. Consider the case where there is one sample point for each mean X_i. We shall evaluate the following:

1). The statistic of this paper which for this case is

$$\hat{b}_1 = \frac{\sum_1^n y_i(i - \bar{\imath})}{\sum_1^n x_i(i - \bar{\imath})}.$$

2). The statistic obtained by minimizing the sum of the squares of the y deviations only

$$\hat{b}_2 = \frac{\sum_1^n y_i(x_i - \bar{x})}{\sum_1^n x_i(x_i - \bar{x})}.$$

3). The statistic obtained by minimizing the sum of the squares of the orthogonal deviations

$$\hat{b}_3 = \frac{\sum_1^n (y_i - \bar{y})^2 - \sum_1^n (x_i - \bar{x})^2 + \left[n \sum_1^n (y_i - \bar{y})^2 - n \sum_1^n (x_i - \bar{x})^2 + 4\left(\sum_1^n (y_i - \bar{y})(x_i - \bar{x}) \right)^2 \right]^{\frac{1}{2}}}{\sum_1^n (y - \bar{y})(x - \bar{x})}$$

TABLE I

Set	x_1	y_1	x_2	y_2	x_3	y_3	x_4	y_4
1	1.1	1.4	2.4	2.0	3.0	2.7	3.6	4.3
2	1.2	1.4	2.2	2.0	3.4	3.1	3.8	4.2
3	1.0	1.4	1.6	2.1	2.8	3.2	4.4	4.3
4	0.6	0.7	1.8	2.0	3.3	2.6	3.8	4.0
5	0.7	1.4	1.7	1.7	2.7	3.4	4.1	4.1
6	1.0	1.2	1.6	2.1	2.9	2.6	3.6	4.0
7	1.3	0.7	1.7	2.1	2.7	2.9	4.0	3.6

TABLE II

Set	b_1	b_2	b_3	b_4
1	1.160	1.068	1.220	1.162
2	1.056	1.009	1.059	1.027
3	0.860	0.843	0.803	0.870
4	0.946	0.896	0.924	0.830
5	0.875	0.867	0.913	1.000
6	0.978	0.939	0.981	0.846
7	1.044	0.959	1.045	1.000
Mean............................	0.990	0.940.	0.996	0.962
7 × Sample Var..................	0.0686	0.0373	0.1058	0.0834

4). The statistic proposed by Wald[2]

$$\hat{b}_4 = \frac{\sum_1^{n/2} y_i - \sum_{n/2}^{n} y_i}{\sum_1^{n/2} x_i - \sum_{n/2}^{n} x_i} .$$

We apply these statistics to sample data having four means $X_i = i$ and $Y_i = i$, $(i = 1, 2, 3, 4)$. By means of a table of random numbers seven sets of data were

[2] Loc. cit.

obtained, each set having one sample point corresponding to each mean. These sample points are described by Table I where it will be noted that the sample points were drawn from a discrete distribution. The estimates obtained from the four statistics are exhibited in Table II.

If the 28 sample points are treated as a single set of data and the four statistics in their appropriate forms are applied, there is obtained the following set of estimates:

$\hat{b}_1$	$\hat{b}_2$	$\hat{b}_3$	$\hat{b}_4$
0.9768	0.9183	0.9786	0.9496.

The preceding computations show that the estimate $\hat{b}_2$ is inferior to the other estimates, as would be expected. The estimate $\hat{b}_3$ is most accurate when the 28 sample points are treated as a single set of data with the estimate $\hat{b}_1$ being only very slightly less accurate, $\hat{b}_1 = 0.9768$ as compared to $\hat{b}_3 = 0.9786$. When the individual sets of sample points 1 to 7 are considered it is seen that the estimate $\hat{b}_1$ is most accurate with the estimate $\hat{b}_3$ rather less accurate; the estimate $\hat{b}_1$ is more precise than $\hat{b}_3$, the sample variances being in the ratio $0.0686 \div 0.1058 = 0.65$. From a practical viewpoint we may also point out that the computation of $\hat{b}_1$ requires very much less labor than the computation of $\hat{b}_3$.