A NEW APPROACH TO WALD'S ESTIMATOR IN THE ERROR IN VARIABLES MODEL

Graydon W. Bell

BU-812-M April 1983

BU-812-M in the Mimeo Series of the Biometrics Unit, Cornell University.

A NEW APPROACH TO WALD'S ESTIMATOR IN THE ERROR IN VARIABLES MODEL

Graydon W. Bell*

ABSTRACT

Estimating the true slope of a line in the presence of errors in both variables is an old problem. An estimator due to Wald involves dividing the data into two sets and estimating the slope of the underlying line as the slope between the centroids of these two parts. The natural appeal of this scheme is preserved in the present work where two centroids are obtained without dividing the data into two sets.

KEY WORDS: Errors in variables; Linear functional relationship; Wald's estimator; Least squares.

---

* Graydon W. Bell, Department of Mathematics, Northern Arizona University, Flagstaff, Arizona. This work was completed while the author was a visiting fellow with the Biometrics Unit, Department of Plant Breeding and Biometry, Cornell University.

1. INTRODUCTION

Many physical laws postulate a linear relationship between mathematical variables. They are particularly simple to work with and use for prediction. If measurement error is involved in both variables, estimation of the slope becomes a statistical problem. The likelihood principle is known to have unacceptable performance for estimation of the slope, unless assumptions can be made about the variances of the error distributions. If no assumptions are made Solari (1969) showed by a geometrical argument that the likelihood surface is saddle shaped in the neighborhood of its critical points. Kendall and Stuart (1967, Chapter 29) devote a good deal of discussion to the problems associated with estimation of the slope. The main difficulty is that each new sample point brings with it a new parameter; these are called incidental parameters, and must be estimated or eliminated. Wald (1940) gave an appealing method for estimating the slope. His approach is to divide the data into two sets, compute the centroid of each set, and estimate the slope of the line as the slope of a line segment between these centroids. A modification of this procedure, due to Bartlett (1949) is recommended by Draper and Smith (1981, p. 124). This entails dividing the data into three parts and discarding the middle group before computing the centroids.

In this article another alternative is explored, one that determines two centroids without splitting the data into subsets. This overcomes a major problem with Wald's procedure, namely that error-free partitioning of the data cannot be guaranteed on the basis of the observed points. It also alleviates concern about using less than the whole data set as in Bartlett's modification.

2. MODELS AND THEIR CONSEQUENCES

In this section the main models associated with the problem are reviewed. These include the usual regression model and the error in variables model.

In the case of linear regression, the data consist of n points, (x_1,y_1), $\cdots,(x_n,y_n)$, with $y_i = \alpha + \beta x_i + e_i$, where $e_i \sim N(0,\sigma^2)$ is assumed. The likelihood function is

$$\left(\frac{1}{2\pi\sigma^2} \right)^{\frac{n}{2}} \exp \left\{ -\frac{1}{2\sigma^2} \sum (y_i-\alpha-\beta x_i)^2 \right\} \tag{1}$$

and is maximized when β is estimated by

$$b = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{\Sigma(x_i-\bar{x})^2} \quad . \tag{2}$$

If errors of measurement occur in both variables this model is inadequate. It may be replaced by

$$x_i = \xi_i + e_{i1} \ ,$$

and

$$y_i = \alpha + \beta\xi_i + e_{i2} \ ,$$

where $e_{ij} \sim N(0,\sigma_j^2)$, $(i = 1,\cdots,n;\ j = 1,2)$, to express the relationship between, and errors in, true variables ξ and $\eta = \alpha + \beta\xi$. Implicit in the above is the assumption that e_{i1} and e_{i2} are uncorrelated. The likelihood function replacing (1) is

$$\left(\frac{1}{2\pi\sigma_1\sigma_2} \right)^n \exp \left[-\tfrac{1}{2} \left\{ \frac{\Sigma(x_i-\xi_i)^2}{\sigma_1^2} + \frac{\Sigma(y_i-\alpha-\beta\xi_i)^2}{\sigma_2^2} \right\} \right] \ . \tag{3}$$

An attempt to maximize (3) leads to saddle points on the surface; there are no overall maxima. Thus, no progress can be made without further assumptions, the usual one being that the error variances have a known ratio, λ. This restriction leads to an estimate of β by taking the positive square root in the solution of

$$b^2 \Sigma u_i v_i + b(\lambda \Sigma u_i^2 - \Sigma v_i^2) - \lambda \Sigma u_i v_i = 0 \quad , \tag{4}$$

where $u_i = x_i - \bar{x}$, $y_i - \bar{y}$.

Wald's (1940) approach to the estimation of β utilized a division of the data into two sets, according to the values of ξ. Centroids are computed for each group, and the slope between centroids is used as an estimate of β. In practice the values of x must be used for partitioning and the ordering of the x_i may not agree with that of the ξ_i. Since an error-free division cannot be assured, Bartlett's (1949) procedure discards some points in the middle of the observed range and uses the outer groups in the manner of Wald.

3. A NEW CENTROID ESTIMATOR

In this section, a procedure is developed to avoid the problem of data splitting. The entire set is used to determine two points, C_1 and C_2, that may be thought of as centroids. The estimation of β will be the slope between these points as usual.

Let $(\bar{x}+h_1, \bar{y}+k_1)(\bar{x}+h_2, \bar{y}+k_2)$ be rectangular representations of C_1 and C_2 and (x_i, y_i) an observed point. Define d_{i1} and d_{i2} as the respective distances from (x_i, y_i) to C_1 and C_2. This yields a configuration as shown in Figure 1.

The points C_1 and C_2 are to be determined by variation in h_1, k_1, h_2 and k_2 to minimize

$$S = \Sigma(d_{i1}d_{i2})^2 \quad . \tag{5}$$

Their coordinates will be used in

$$b = \frac{k_2-k_1}{h_2-h_1} \quad , \tag{6}$$

as an estimator of β.

The choice of S at (5) may appear unusual, and deserves some comment. Strong motivation for this form comes from the simplicity of the resulting derivatives. A natural alternative would be to work with $\Sigma(d_{i1}+d_{i2})$, as if each point was on an ellipse with C_1 and C_2 as foci. On the other hand, an ellipse superimposed on the structure is artificial; there are, instead, n small ellipses, with centers at $(\xi_i, \alpha+\beta\xi_i)$. The use of the square of the product of distances does have a geometrical basis, however; such functions occur in the locus definition of a family of curves called the ovals of Cassini. These curves are defined, in the present notation, by $(d_{i1}d_{i2})^2$ = constant, and can have forms as shown in Figure 2. If the constant is small relative to the distance between C_1 and C_2, the case of two separated ovals is obtained (see Beyer, 1981, p. 271). Since S is to be minimized, this case may be visualized, though it is only the points C_1 and C_2 that are of interest.

A variety of constraints might be imposed on the quantity in (5). It would not be unreasonable to require that the segment from C_1 to C_2 also contain $(\bar{x},\bar{y})$. In addition, C_1 and C_2 could be constrained to lie equidistant from $(\bar{x},\bar{y})$. If both these conditions are met an interesting result obtains, namely, that minimizing (5) leads to (4) with $\lambda = 1$. This is also then equivalent to the use of normal deviations and to extracting principal components.

A more general result can be expected if no restrictions are placed on the coordinates. This gives, working with deviations from the observed means, $u_i = x_i - \bar{x}$ and $v_i = y_i - \bar{y}$,

$$S = \Sigma\{(u_i-h_1)^2 + (v_i-k_1)^2\}\{(u_i-h_2)^2 + (v_i-k_2)^2\} \tag{7}$$

as the objective function to be minimized. Equating the partial derivatives of S to zero and straightforward algebra yields, writing $\Sigma(u^2+v^2) = \Sigma r^2$ and dropping indices of summation,

$$\begin{aligned}
nh_1k_2^2 + nh_1h_2^2 + h_1\Sigma r^2 + 2k_2\Sigma uv + 2h_2\Sigma u^2 - \Sigma ur^2 &= 0 \\
nh_2k_1^2 + nh_2h_1^2 + h_2\Sigma r^2 + 2k_1\Sigma uv + 2h_1\Sigma u^2 - \Sigma ur^2 &= 0 \\
nk_1k_2^2 + nk_1h_2^2 + k_1\Sigma r^2 + 2k_2\Sigma v^2 + 2h_2\Sigma uv - \Sigma vr^2 &= 0 \\
nk_2k_1^2 + nk_2h_1^2 + k_2\Sigma r^2 + 2k_1\Sigma v^2 + 2h_1\Sigma uv - \Sigma vr^2 &= 0 \quad .
\end{aligned} \tag{8}$$

Each equation in (8) defines a circle in either (h_1,k_1) or (h_2,k_2) space, with center and radius determined by one of the other variables. Common chords and the like are easy to find, but an explicit solution for h_1, h_2, k_1, and k_2 does not seem feasible. The system can easily be solved numerically, due to the structure of the equations (each involves only three variables) and since good starting values can be based on the solution of (4) with $\lambda = 1$ or even on (2). Remarkably, it is not necessary to solve (8) in order to evaluate the slope in (6). It is possible to derive, from (8) and (6), the following quadratic equation in b:

$$b^2 + b\left(\frac{n(\Sigma ur^2)^2 - n(\Sigma vr^2)^2 + 4(\Sigma r^2)^2(\Sigma u^2-\Sigma v^2)}{n\Sigma vr^2\Sigma ur^2 + 4(\Sigma r^2)^2\Sigma uv}\right) - 1 = 0 \quad . \tag{9}$$

For estimating a slope then, it is only necessary to compute the roots of (9) rather than solving (8). Notice that the roots of (9), like (4) with $\lambda = 1$, are negative reciprocals. This is appropriate in any procedure that treats x and y symmetrically. To obtain a single solution the positive square root is used in solving (9), just as it is in (4).

Consider now the coefficient of b in (9); it can be expressed as

$$\frac{\frac{n(\Sigma ur^2)^2 - n(\Sigma vr^2)^2}{4(\Sigma r^2)^2} + \Sigma u^2 - \Sigma v^2}{\frac{n\Sigma ur^2 \Sigma vr^2}{4(\Sigma r^2)^2} + \Sigma uv} \tag{10}$$

In this form the relationship to (4) is apparent, with the usual least squares quantities of Σu^2, Σv^2, and Σuv being augmented or adjusted by higher moments of the data.

4. THE CASE OF NO ERRORS

If no errors are present the points (x_i, y_i) are actually (ξ_i, η_i). The quantities estimated by (8) and (9) may then be determined. In this case $\eta_i - \bar{\eta} = \beta(\xi_i - \bar{\xi})$ and u_i and v_i are replaced accordingly. The solution to (9) is easily found to be β, so the quadratic equation in (9) will identify the slope of a line from points along that line.

More can be learned, however, for in this case (8) may be solved explicitly. The solution is

$$h_{1,2} = \frac{1}{2m_2}\{m_3 \mp \sqrt{(m_3^2 + 4m_2^3)}\} ,$$

$$k_1 = \beta h_1 \qquad k_2 = \beta h_2 \quad ,$$

where $m_2 = \Sigma(\xi_i - \bar{\xi})^2/n$ and $m_3 = \Sigma(\xi_i - \bar{\xi})^3/n$. If the points are symmetrically placed along the line $m_3 = 0$, and so $h_1 = \sqrt{m_2}$. The centroids then are $(-\sqrt{m_2}, -\beta\sqrt{m_2})$ and $(\sqrt{m_2}, \beta\sqrt{m_2})$, a particularly appealing form. If symmetry is not present, simple manipulation shows h_1 can be expressed as

$$\frac{\sqrt{m_2}}{2}\left\{\frac{m_3}{m_2^{3/2}} - \sqrt{\left(\frac{m_3^2}{m_2^3} + 4\right)}\right\} .$$

The lack of symmetry, as measured by the moment ratio, adjusts the centroids to either the right or left.

5. NUMERICAL CONSIDERATIONS

Brown (1957) gives a set of simulated data consisting of 9 points. These 'observed' points are

x_i:	1.8	4.1	5.8	7.5	9.3	10.6	13.4	14.7	18.9
y_i:	6.9	12.5	20.0	15.7	24.9	23.4	30.2	35.6	39.1 .

These were generated from a linear equation with $\beta = 2$, $\sigma_1^2 = \sigma_2^2 = 1$. Computing the necessary functions of the deviations gives

$$\Sigma u^2 = 237.96 \qquad \Sigma v^2 = 905.14 \qquad \Sigma uv = 451.28$$

$$\Sigma ur^2 = 853.47 \qquad \Sigma vr^2 = 862.45 \qquad \Sigma r^2 = 1143.10 ,$$

and substituting in (9), $b = 1.98$. Working with the form at (10) the coefficient of b becomes

$$\frac{-.026 - 667.18}{1.267 + 451.28} ,$$

showing that quite a minor adjustment is being made to the solution to (4).

Other estimates that might be advanced for this data are listed below:

a. Regression slope, y on x : 1.90 ,

b. Regression slope, x on y : 2.00 ,

c. Solution of (4) with $\lambda = 1$: 1.98 ,

d. Wald's estimates

(i) 4-5 grouping : 1.96 ,

(ii) 5-4 grouping : 1.85 ,

e. Discarding middle sets

(i) 3-3-3 : 1.86 ,

(ii) 4-1-4 : 1.91 .

The system in (8) may also be solved numerically to give

$$h_1 = -4.67 \qquad k_1 = -9.60$$

$$h_2 = 5.41 \qquad k_2 = 10.35 .$$

Adding these to the observed means gives centroids of (4.90,13.54) and (14.98,33.50). These are shown, along with the data, in Figure 3. One might compare these to the centroids in the last case above; they are (4.80,13.78), (14.40,32.08).

Some simulation has been done to evaluate the performance of the new estimator. In a typical run n = 5 points were drawn with values of ξ equally spaced along a range of 15 units. Error variances were set at $\sigma_1^2 = \sigma_2^2 = .25$. Repeating the sampling 10,000 times with $\beta = 2.5$ gave $\bar{b} = 2.503$ $S_b^2 = .0202$. Approximate normality was suggested by moment ratios of .0018 and 3.094 in comparison to 0 and 3 for a normal.

6. CONCLUSION

The technique of determining two centroids without splitting the data seems to work nicely in cases where both variables involve error. Estimating the slope of a line by the slope between these centroids will almost certainly give biased results. In simulation work though, the bias of the new estimator has been found consistently to be smaller than if the regression slope for y on x is used.

While simulation has suggested that the centroid estimator is approximately normally distributed, it seems unlikely that an exact distribution can be found. It is also not known whether the centroid estimator is consistent; chances are that conditions on the range of the independent variable will have to be included to establish consistency.

REFERENCES

BARTLETT, M. S. (1949), "Fitting a Straight Line When Both Variables are Subject to Error," Biometrics, 5, 207-212.

BEYER, W. H., editor (1981), CRC Standard Mathematical Tables, 26th edition, CRC Press, Inc., Boca Raton, Florida.

BROWN, R. L. (1957), "Bivariate Structural Relation," Biometrika, 44, 84-96.

DRAPER, N. R. and SMITH, H. (1981), Applied Regression Analysis, 2nd edition, John Wiley : New York.

KENDALL, M. G. and STUART, A. (1967), The Advanced Theory of Statistics, vol. II, 2nd edition, Hafner : New York.

Solari, M. E. (1969), "The 'Maximum Likelihood Solution' of the Problem of Estimating a Linear Functional Relationship," Journal of the Royal Statistical Society, Series B, 31, 372-375.

Wald, M. (1940), "The Fitting of Straight Lines if Both Variables are Subject to Error," Annals of Mathematical Statistics, 11, 284-300.

LEGENDS

Figure 1. Centroids and Distances.

The relationship of an observed data point to the centroids is shown.

Figure 2. Ovals of Cassini.

Different shapes are possible depending on the distance between fixed points and the constant in $(d_1 d_2)^2 = \text{constant}$.

Figure 3. The Data of Brown (1957) with Centroids Determined.

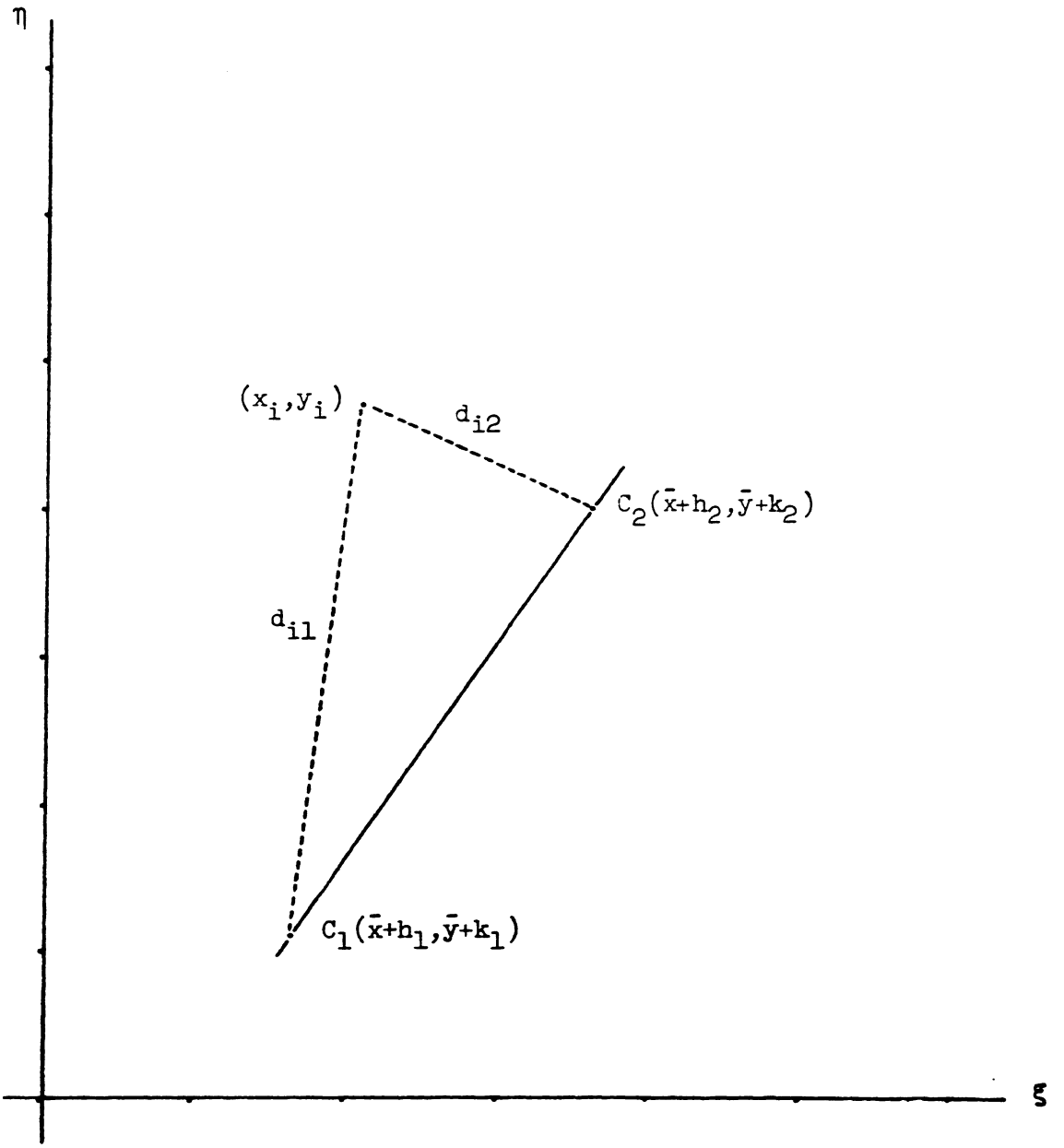

Figure 1. Centroids and Distances.

The relationship of an observed data point to the centroids is shown.

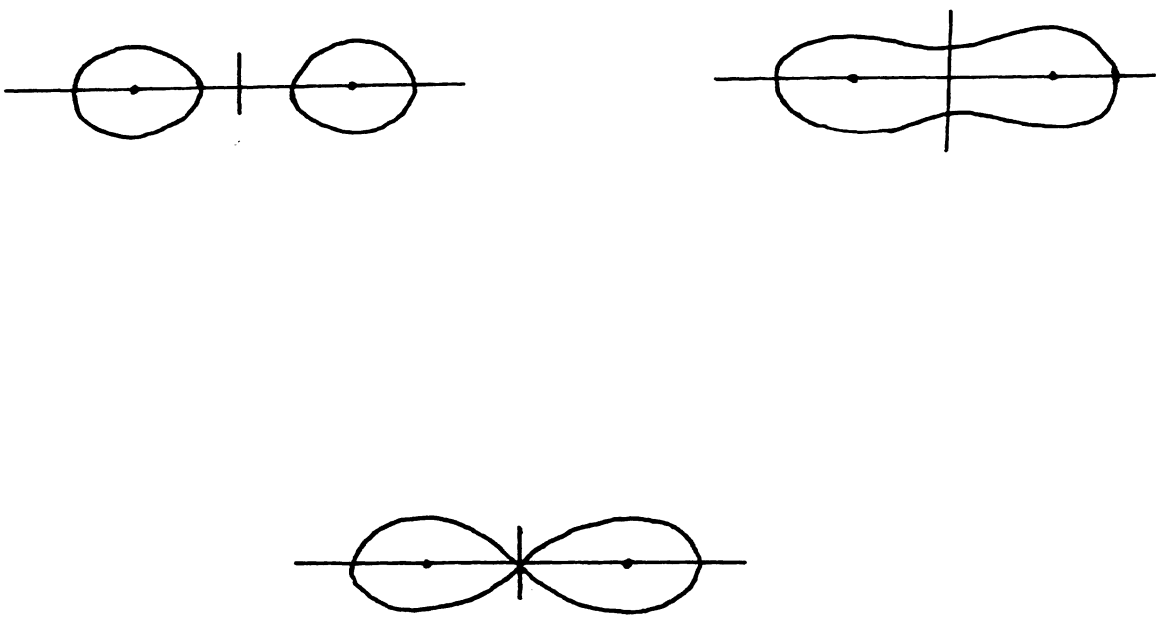

Figure 2. Ovals of Cassini.

Different shapes are possible depending on the distance between fixed points and the constant in $(d_1 d_2)^2$ = constant.

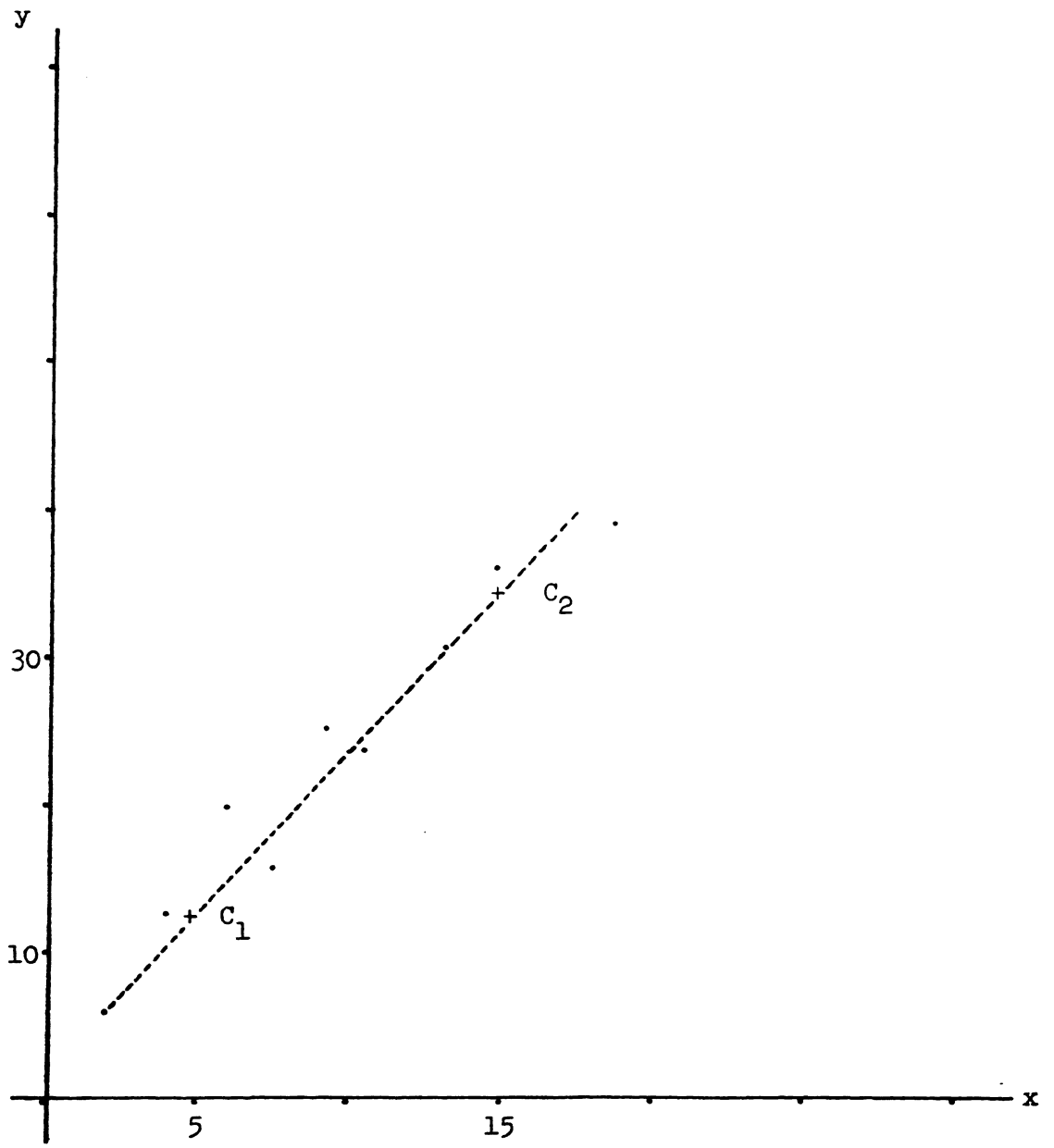

Figure 3. The Data of Brown (1957) with Centroids Determined.