SEAFDEC Training Department

Text/Reference Book

Southeast Asian Fisheries Development Center

November 1978

Manual on Treating Numerical Data

Otohiko Suzuki

Training Department

Southeast Asian Fisheries Development Center

Issued occasionally from the Training Department, Southeast Asian Fisheries Development Center.
P.O. Box 4, Phrapradaeng, Samutprakarn, Thailand.

FOREWARD

The present manual deals with basic knowledge concerning data analysis. Originally, this handbook was drafted as part of the Manual on Physics Experiments for undergraduates in the Department of Fisheries, Faculty of Agriculture, Kyoto University, Kyoto, Japan. On the occasion of this publication, the text has been revised to include further concrete examples. The author hopes that it will prove useful to those students and research scientists who engage in data processing in the course of their laboratory experiments or field observations.

Finally, the author wishes to thank Miss B. Mountfield for her devoted assistance in the compilation of the present manual.

Bangkok
November 1978

Otohiko Suzuki

Training Department
Southeast Asian Fisheries
Development Center

Manual on Treating Numerical Data

Contents

1. Random Error and Mean Value

When a quantity is measured, it is usually read by means of the scale of an apparatus. Generally, the measurement is taken by the eye to the nearest 1/10 of the smallest scale. In consequence, the last figure of the measured values contains a reading error, besides an instrumental error.

If there are n observed data $a_1, a_2, \ldots, a_n$ on a quantity, then the arithmetic mean M is given to be

$$M = \frac{1}{n} \Sigma a_i. \tag{1.1}$$

The difference α_i of each measured value a_i from M is called the residual or simply the error. The sum of these errors must be zero, i.e.,

$$M - a_1 = \alpha_1, \quad M - a_2 = \alpha_2, \ \ldots, \quad M - a_n = \alpha_n,$$

$$\Sigma \alpha_i = 0. \tag{1.2}$$

The noteworthy point is that the mean value is not a measured value but a value derived from measured data. When measuring a quantity, its true value is unknown and can never be determined exactly by observation. In cases containing random errors, the mean value is regarded as the best estimate that can approach closer to the true value as the number of observations increases[1/].

1/ In the present case, the true value is unknown but can be estimated by the arithmetic mean of observed data. The accuracy of this estimate is given as $\sigma/\sqrt{n}$, where σ and n are the standard deviation of the sample and the number of observations respectively (Refer Eq. 1.3). Therefore, if σ does not vary greatly with increasing n, we can roughly regard that the accuracy of the estimate increases in proportion to $\sqrt{n}$.

A peculiarity of random error is that the more observations are made the more frequently the measured values fall in the proximity of the mean value; the probability of making an error decreases very rapidly as the magnitude of the error increases. In other

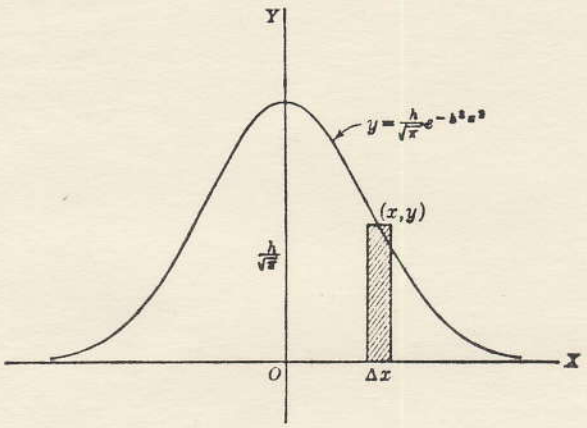

Fig. 1

words the frequency distribution of such data shows a symmetrical bell-shaped graph[2/] with the center at the mean (Fig. 1). This implies that the probability of a datum which falls below the mean is equal to that of the datum above this value.

The standard deviation σ of a set of observations is defined as the root-mean-square value of the deviation from the arithmetic mean, or the square root of the averaged squared residual. Thus if n is the number of observations, we have

$$\sigma = \sqrt{\frac{\Sigma\, \alpha_i^2}{n}} \,. \qquad (1.3)$$

2/ This type of curve is referred to as error curve or probability curve.

The number ε , such that the probability that an error is between $-\varepsilon$ and $+\varepsilon$ is $\frac{1}{2}$, is called the probable error of a single observation. These σ and ε are related by the following equation,

$$0.6745\sqrt{\frac{\Sigma\,\alpha_i^2}{n}} = 0.6745\,\sigma = \varepsilon, \qquad (1.4)$$

The value of σ or of ε can be used as a measure of dispersion of a set of observations. For small σ or ε the curve of frequency distribution has a high peak and falls sharply, showing a small spread or dispersion in the observation, and, conversely, if σ and ε have a large value, the peak is low, the curve falls gradually and the spread is wide (Fig. 2).

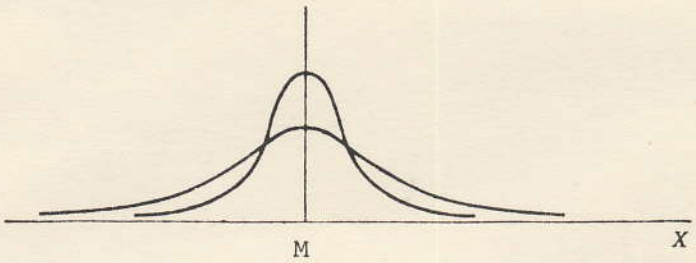

Fig. 2

2. Significant Figures

Since any measured value contains some errors, it is meaningless to write down simply all the figures derived from mathematical calculations. As an example, let us consider a volume of liquid in a vessel graduated to the order of 10 ml. and another volume of liquid in the other vessel graduated to the order of 1/10 ml. Suppose that the former value was measured to be 251 ml. and the latter 25.27 ml. and the liquids were mixed. The volume of the

mixed liquids should not be expressed as 276.27 ml. but should be 276 ml. The value should be rounded off to the nearest figure of the less precisely measured value (Fig. 3).

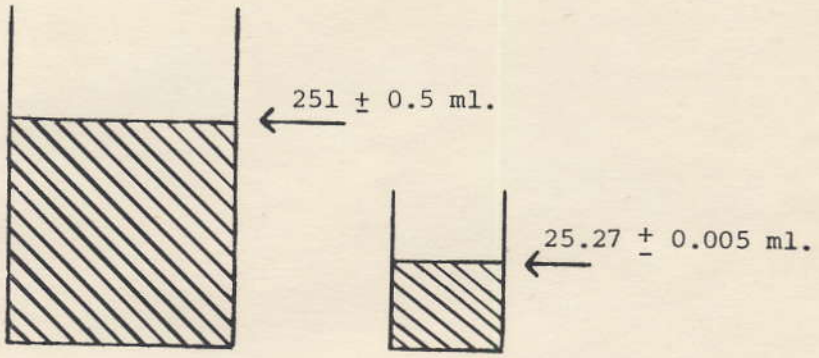

Fig. 3

When calculations are made using numerical values containing errors such as $(a \pm \delta a)$ and $(b \pm \delta b)$, the error $\delta\gamma$ exerting influence upon the calculated result γ can be obtained as follows:

1. $\delta\gamma = n\delta a$, for $\gamma = na$.

2. $\delta\gamma = n\delta a/a^2$, for $\gamma = n/a$.

3. $\delta\gamma = n(\delta a + \delta b)$, for $\gamma = n(a \pm b)$.

4. $\delta\gamma = n(b\delta a + a\delta b)$
   $= nab(\delta a/a + \delta b/b)$, for $\gamma = nab$.

5. $\delta\gamma = n(a/b)(\delta a/a - \delta b/b)$, for $\gamma = na/b$.

Generally, if the equation is given to be

$$\gamma = f(a, b, c, \ldots), \tag{2.1}$$

the error $\delta\gamma$ is expressed as

$$\delta\gamma = \frac{\partial f}{\partial a}\delta a + \frac{\partial f}{\partial b}\delta b + \frac{\partial f}{\partial c}\delta c + \ldots \tag{2.2}$$

Since the calculated γ contains the error $\delta\gamma$, those figures below the place affected by the error are insignificant. It must also be noted that the precision of observations should be taken into account with the relative error $\delta\gamma/\gamma$, not with the absolute error $\delta\gamma$.

When writing numbers containing many zeros before or after the decimal point, it is convenient to employ powers of 10. For example, 135,000 or 0.000135 have three significant figures in common and can be expressed as 1.35×10^{5} or 1.35×10^{-4} respectively. Numbers associated with enumerations or countings, as opposed to measurements, are of course exact and so have an unlimited number of significant figures. In some of these cases, however, it may be difficult to decide which figures are significant without further information. For example, the number 186,000,000 may have 3, 4, ..., 9 significant figures. If it is known to have five significant figures, it would be better to record the number as 1.8600×10^{8}.

Bearing the exact meaning of significant figures in mind, we can often obtain a sufficient accuracy even by approximate calculation. In cases where the error is small, it may be sufficient to take merely the infinitesimal terms of the first degree given by expansion. The following are expansions usually employed under such circumstances. Here δ , γ , ... denote the infinitesimal values.

1. $(1 \pm \delta)^n = 1 \pm n\delta.$

2. $\sqrt{1 \pm \delta} = 1 \pm \tfrac{1}{2}\delta.$

3. $\dfrac{1}{(1 \pm \delta)^m} = 1 \mp m\delta.$

4. $\dfrac{1}{\sqrt{1 \pm \delta}} = 1 \mp \tfrac{1}{2}\delta.$

5. $\dfrac{(1 + \delta)^m}{(1 + \varepsilon)^n} = 1 + m\delta - n\varepsilon.$

6. $\dfrac{(1 \pm \delta)(1 \pm \zeta)\ldots}{(1 \pm \varepsilon)(1 \pm \eta)\ldots} = 1 \pm \delta \pm \zeta \pm \ldots \mp \varepsilon \mp \eta \mp \ldots$

7. $\sqrt{\rho_1 \rho_2} = \tfrac{1}{2}(\rho_1 + \rho_2),$

where $\rho_1 \doteqdot \rho_2.$

8. $\sin(x \pm \delta) = \sin x \pm \delta \cos x,$

$\cos(x \pm \delta) = \cos x \mp \delta \sin x,$

$\tan(x \pm \delta) = \dfrac{\tan x \pm \delta}{1 \pm \delta \tan x},$

$\sin \delta = \tan \delta = \delta,$

$\cos \delta = 1,$

where δ is given in radian.

3. Graphical Representation of Data (Linearization of Plots)

When studying the variation or the distribution of numerical values, graphical representation is mostly used to show the quantitative relationship between two variables. By this representation the general trend in a group of measured values can be understood at a glance. Some noteworthy points in using graph papers are described below.

There are different types of graph papers. The most commonly used are those in which the rectangular coordinate axes are graduated in equi-intervals. Logarithmic graph papers are also used and they can be in the form of semi-log or log-log papers.

The method of using the axes of the graph papers is as follows:

1. Take the independent variable on the abscissa.

2. Place marks on the axes in such a manner that the plots of data are distributed over a wide enough space.

3. Select the variables in such a way that the plots are distributed almost linearly. This includes the transformation of original variables.

4. Mark the units of the variables on the respective axis.

5. When the logarithms of the measured values are plotted, the ordinary linear section paper can of course be used but if logarithmic section papers are used the measured values can be plotted without any modification. In the latter case the decimal places of the figures such as 0.1, 1.0, 10, 100, etc., must be taken in equi-intervals

on the standard stubs of the axis (Fig. 4). Therefore, there is no zero on the logarithmic scale.

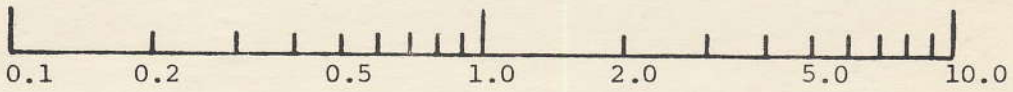

Fig. 4

When a theoretical equation is to be compared with measured data on a graph, it is preferable to plot them as straight lines rather than curves. Many types of curves can be converted into straight lines by suitable transformation of variables on the coordinate axes.

3.1. Straight Line

When plots are distributed almost linearly, the relation of the two quantities x and y is expressed by

$$y = a + bx, \tag{3.1}$$

where a and b are constants whose values can be determined from the observed data. The constants can be obtained by solving two simultaneous equations which are given by the readings of two pairs x and y from the line fitted visually. However, the more convenient method is a graphical solution.

In the equation (3.1), $y = a$, when $x = 0$; therefore a shows the intercept on the y-axis (Fig. 5). Since, moreover,

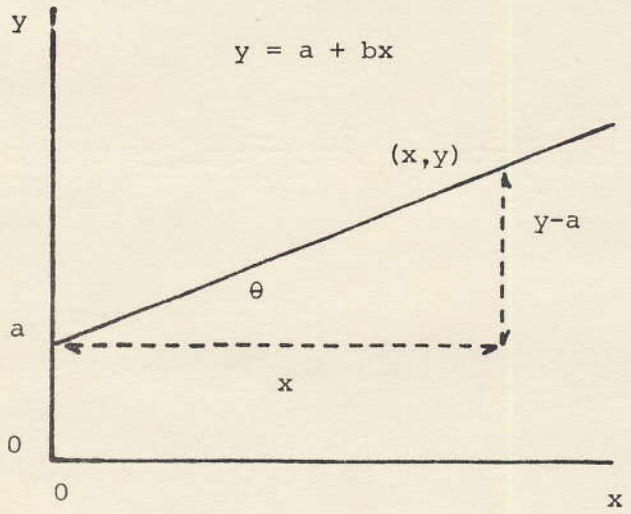

Fig. 5

$$b = \frac{y-a}{x} ,$$

or

$$b = \tan \theta ,$$

where θ is the angle between the x-axis and the straight line visually fitted, we can easily obtain b.

3.2. Parabola or Hyperbola

Consider the case where the two quantities x and y are related by

$$y = ax^n, \tag{3.3}$$

where a and n are constants. Taking logarithms, the original equation becomes

$$\log y = \log a + n \log x. \tag{3.4}$$

Therefore, when we plot log x and log y instead of x and y, we can get the following linear relation

$$Y = c + nX, \tag{3.4'}$$

This relationship is of course obtainable by plotting log x and log y on a linear section paper. When using log-log papers, however, we can easily obtain the relationship.

For the determination of the constant a, taking the fact that log x = 0, for x = 1 into consideration, we have the relation

$$\log y \Big]_{x=1} = \log a. \tag{3.5}$$

Therefore, we can get log a by reading the value on the Y-axis for x = 1. When there is no point x = 1 on the abscissa, we can also obtain the value by using the Y-intercepts corresponding to x = 10, 100, ... For example, denoting by $\log a_1$ the reading of the Y-intercept for x = 1000 (Fig. 6), we have

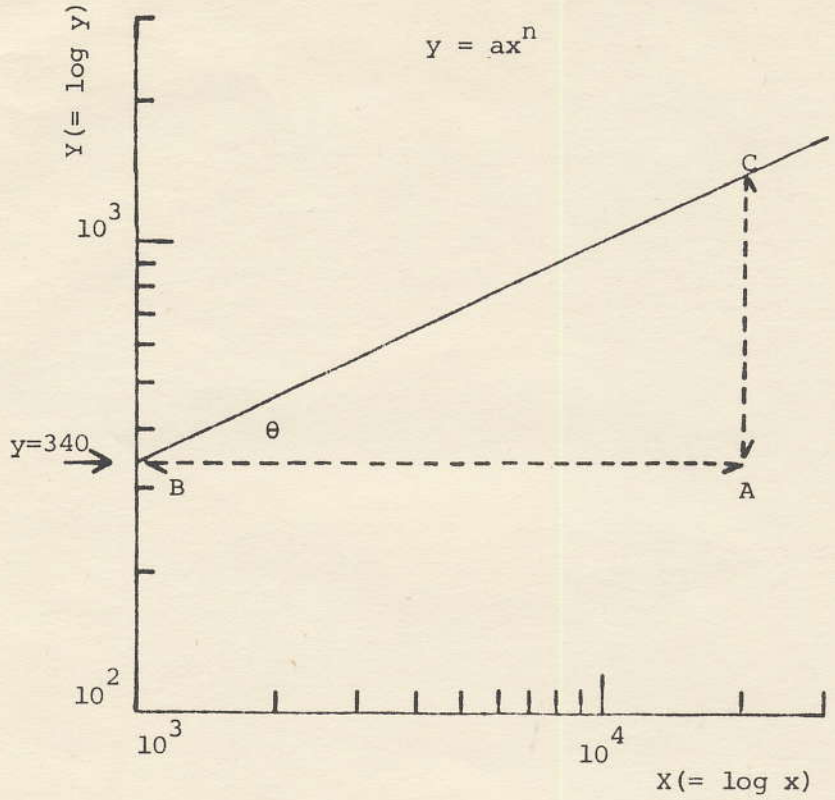

Fig. 6

$$\log a_1 = \log a + 3n,$$

or

$$\log a = \log a_1 - 3n. \qquad (3.5')$$

The constant a is thus determined, and n can be obtained from the gradient of the given straight line[3/].

3.3. Exponential Curve

When the relation between the two observed quantities x and y is given by $y = a.10^{nx}$ or $y = a.e^{nx}$ and accordingly in turn

$$\log y = \log a + nx, \qquad (3.6)$$

or

$$\log y = \log a + M\,nx, \qquad (3.6')$$

it is convenient to plot log y against x, where M = 0.4343 $(= \frac{\log_{10} N}{\log_e N})$. In this case the original exponential curve is converted into a straight line which forms

$$Y = b + cx, \qquad (3.6'')$$

where b and c are constants.

In the equations (3.6 and 3.6''), denoting the increments of Y and x by ΔY and Δx respectively, we have

$$n = \Delta Y/\Delta x. \qquad (3.7)$$

3/ In Fig. 6, the slope n is given numerically to be 0.467 (using an appropriate unit of length for BA and AC), and a_1 is read to be 340 for x = 1000. Therefore, from the equation (3.5'), we have

$$\log a = \log 340 - 3 \times 0.467$$
$$= 2.532 - 1.401 = 1.131$$

Thus, the constant a can be determined to be 13.52.

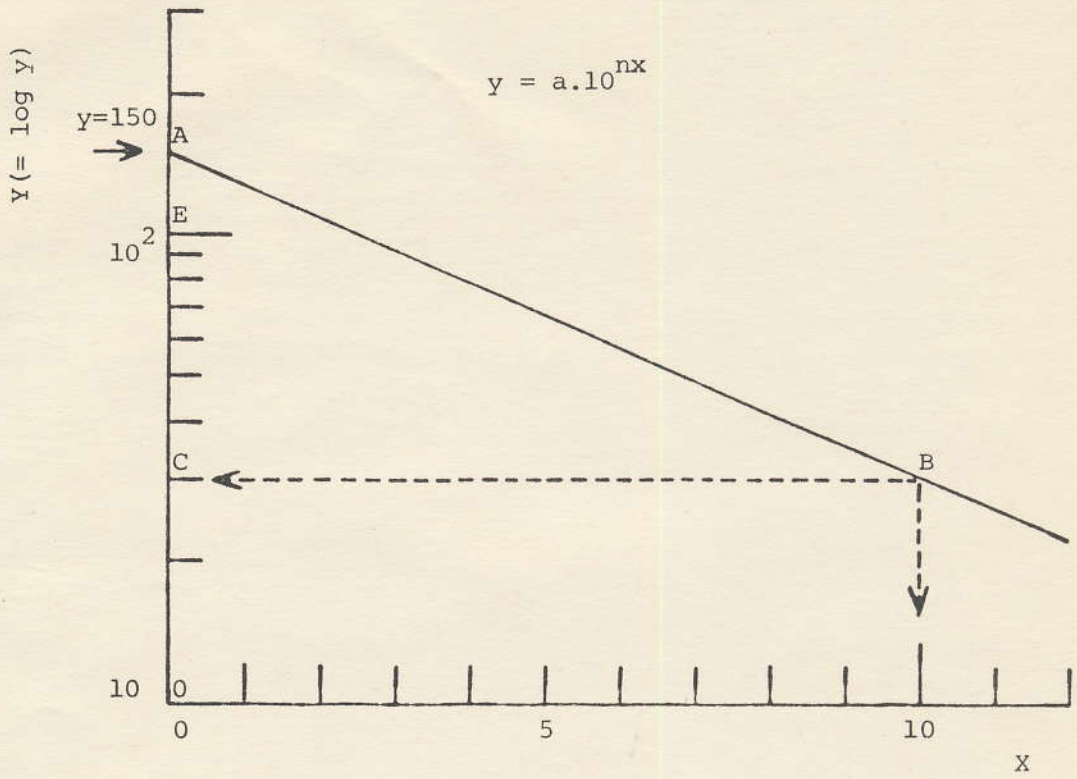

Fig. 7

For example, referring to Fig. 7, consider the $\Delta x = 10$, then the corresponding ΔY is given to be the linear segment $\overline{AC}$ on the Y-axis. Since $\overline{OE}$ = log 100 - log 10 = 1 in this figure, by measuring the linear segments $\overline{AC}$ and $\overline{OE}$, we obtain

$$\Delta Y = - \overline{AC} / \overline{OE} ,$$

and therefore,

$$n = \frac{- \overline{AC}/\overline{OE}}{10} . \qquad (3.8)$$

Since $y = a$ when $x = 0$ in the original form $y = a.10^{nx}$ the value of a is obtained from the intercept on the Y(=log y)-axis[4/].

When we have no information as to the theoretical equation for purposes of comparison, it is usually very difficult to ascertain the algebraical expression for a series of observed data, because there is no definite method by which to arrive at the final expression. However, the following table may be helpful in suggesting the type of mathematical equations which express the shape of the curve on section paper.

Equation	Characteristics
1. $y = a + bx$	Straight line on linear section paper.
2. $y = a + bx^2$	Curved line whose gradient is proportional to x on linear section paper.
3. $y = a + bx + cx^2$	Curved line whose gradient is proportional to x on linear section paper.

4/ In Fig. 7, the constant a can immediately be read to be 150. The slope n is given numerically to be -0.069_8 from the equation (3.8).

Equation	Characteristics
4. $y = \frac{a + bx}{x}$	Straight line when y is plotted against 1/x on linear section paper.
5. $y = \frac{ax}{1 + bx}$	Straight line when 1/y is plotted against 1/x on linear section paper.
6. $y = \frac{ax^2}{1 + bx^2}$	Straight line when 1/y is plotted against $1/x^2$ on linear section paper.
7. $y = a.10^{nx}$, $y = a.e^{nx}$	Straight line when y is plotted against x on semi-log section paper.
8. $y = ax^n$	Straight line when y is plotted against x on log-log section paper.

4. Least Square Method

When a series of observed data (x_i, y_i) follows the relation $y = mx + b$, as a natural consequence the following equations must be valid:

$$mx_i + b - y_i = 0, \tag{4.1}$$

where $i = 1, 2, 3, \ldots, n$. Since, however, the errors v_i always accompany the measured data, the above relation should be rewritten as

$$mx_i + b - y_i = v_i. \tag{4.2}$$

For the determination of constants m and b which are the nearest to their ideal values, the least square method can be employed. In the equation (4.2), y_i take the most likely values when $\Sigma \nu_i^2$ takes the least value. Putting $\Sigma \nu_i^2 = S$ for the sake of simplicity, we obtain

$$\begin{aligned} S &= \Sigma(x_i m + b - y_i)^2 \\ &= m^2 \Sigma x_i^2 + 2bm \Sigma x_i - 2m \Sigma x_i y_i + nb^2 \\ &\quad - 2b \Sigma y_i + \Sigma y_i^2 . \end{aligned} \tag{4.3}$$

S takes its least value when the following conditions are satisfied,

$$\frac{\partial s}{\partial m} = 0 = 2m \Sigma x_i^2 + 2b \Sigma x_i - 2 \Sigma x_i y_i , \tag{4.4}$$

$$\frac{\partial s}{\partial b} = 0 = 2m \Sigma x_i + 2bn - 2 \Sigma y_i . \tag{4.5}$$

On solving the simultaneous equations (4.4) and (4.5), we find

$$m = \frac{n \Sigma x_i y_i - \Sigma x_i \Sigma y_i}{n \Sigma x_i^2 - (\Sigma x_i)^2} , \tag{4.6}$$

$$b = \frac{\Sigma x_i^2 \Sigma y_i - \Sigma x_i \Sigma x_i y_i}{n \Sigma x_i^2 - (\Sigma x_i)^2} . \tag{4.7}$$

Generally, when an empirical expression is given in the form

$$L = F(x, y, z, \ldots),$$

the residual ν_i are expressed as follows

$$\nu_i = F(x_i, y_i, z_i, \ldots) - L_i ,$$

where x_i, y_i, z_i ... are observed values. The constants A, B, C, ... contained in the function F are selected in such a manner as to minimize Σv_i^2 ;

$$\frac{\partial(\Sigma v_i^2)}{\partial A} = 0,$$

$$\frac{\partial(\Sigma v_i^2)}{\partial B} = 0,$$

.....

Here, the number of these simultaneous equations must be equal to that of the constants contained in the original equation. When these equations are solved we can determine the constants.

To facilitate the understanding of the least square method, let us consider the following exampe: Table 1 gives experimental values of the pressure P of a given mass of gas corresponding to various values of the volume V. According to thermodynamic principles, a relationship having the form $PV^{\gamma} = C$, where γ and C are constants, should exist between the variables. (a) Find the values of γ and C. (b) Write the equation connecting P and V.

Table 1

V	54.3	61.8	72.4	88.7	118.6	194.0
P	61.2	49.5	37.6	28.4	19.2	10.1

Since $PV^{\gamma} = C$, we have

$$\log P + \gamma \log V = \log C,$$

or

$$\log P = \log C - \gamma \log V.$$

Taking Log V = X and log P = Y, the last equation can be written

$$Y = b + mX, \tag{4.8}$$

where $b = \log C$ and $m = -\gamma$.

Table 2 below gives X = log V and Y = log P corresponding to the values of V and P in Table 1 and also indicates the calculations involved in computing the least square line (4.8).

Table 2

X (= log V)	Y (= log P)	X^2	XY
1.7348	1.7868	3.0095	3.0997
1.7910	1.6946	3.2077	3.0350
1.8597	1.5752	3.4585	2.9294
1.9479	1.4533	3.7943	2.8309
2.0741	1.2833	4.3019	2.6617
2.2878	1.0043	5.2340	2.2976
$\Sigma X = 11.6953$	$\Sigma Y = 8.7975$	$\Sigma X^2 = 23.0059$	$\Sigma XY = 16.8543$

From the equations (4.6) and (4.7), we have

$$m = \frac{n \Sigma XY - \Sigma X \Sigma Y}{n \Sigma X^2 - (\Sigma X)^2} = -1.40,$$

and

$$b = \frac{\Sigma X^2 \Sigma Y - \Sigma X \Sigma XY}{n \Sigma X^2 - (\Sigma X)^2} = 4.20.$$

Therefore, the equation (4.8) becomes $Y = 4.20 - 1.40\,X$. Since $b = 4.20 = \log C$ and $m = -1.40 = -\gamma$, $C = 1.60 \times 10^4$ and $\gamma = 1.40$. The required equation in terms of P and V can be written $PV^{1.40} = 16{,}000$.

This problem can be solved as well by using the method in Section 3.2. For each pair of values of P and V in Table 1, we obtain a point which is plotted on a log-log graph paper as shown in Fig. 8. A line (drawn freehand) approximating these points is also indicated. The resulting graph shows that there is a linear relationship between log P and log V which can be represented by the equation

$$\log P = b + m \log V \quad \text{or} \quad Y = b + m X.$$

The slope m, which is negative in this case, is given numerically by the ratio of the lengths of AB to AC (using an appropriate unit of length). Measurement in this case yields $m = -1.4$. To obtain b, one point on the line is needed.

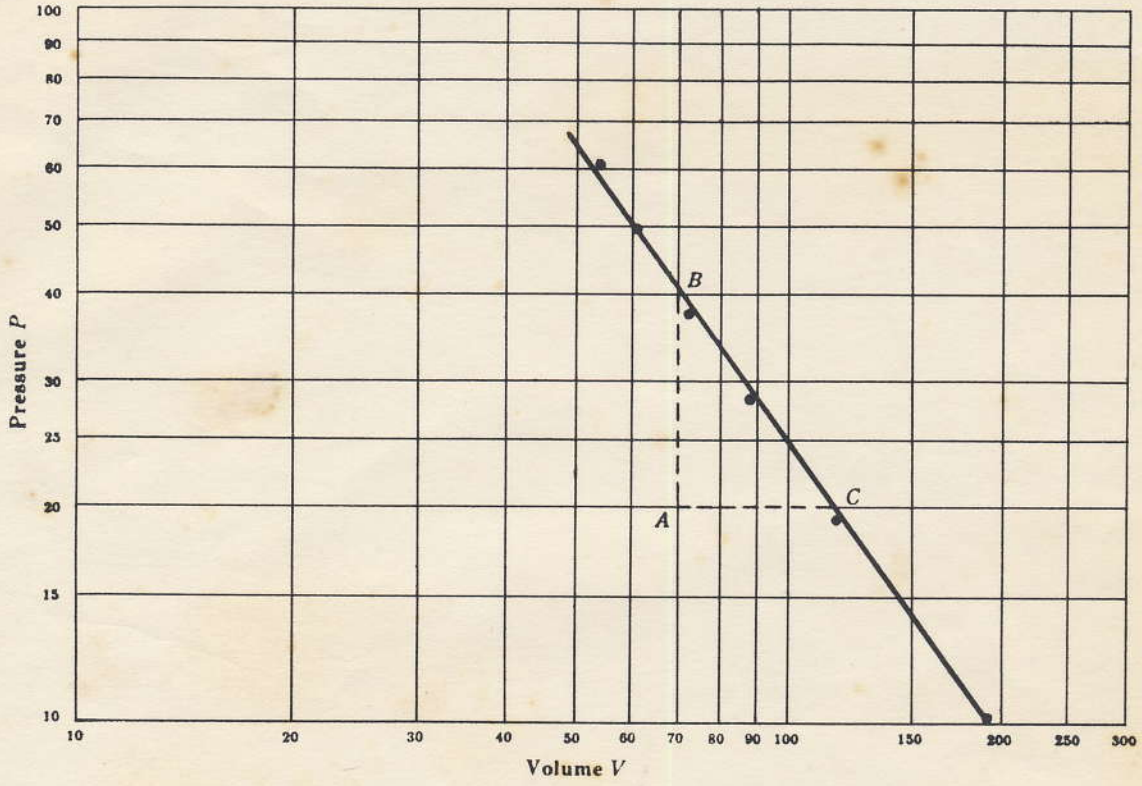

Fig. 8

For example when V = 100, P = 25 from the graph. Then

$$b = \log P - m \log V = \log 25 + 1.4 \log 100$$
$$= 1.4 + (1.4)(2) = 4.2$$

so that

$$\log P + 1.4 \log V = 4.2, \text{ and } PV^{1.4} = 16{,}000.$$