Least Squares

FRAME 1

Introduction

If you and your friends each measure the length of a line, it is quite likely that you will not all get exactly the same result. For example, using a metre rule, the following figures for a particular length might be obtained by 10 different people: 5·127, 5·130, 5·125, 5·128, 5·126, 5·128, 5·129, 5·130, 5·126, 5·127 metres. In these circumstances, what is the best value to take for the length of the line? The answer you would probably give to this question is "The average of the 10 values which is 5·1276 m". However, you would not be dogmatic about it and say that the length of the line is exactly 5·1276 m.

FRAME 2

We are now going to look at this problem from a slightly different viewpoint. This may seem to complicate the situation somewhat, but the ideas involved are important where the problem is not quite so simple.

Suppose the various measurements obtained are denoted by $x_1, x_2, \ldots\ldots x_{10}$, and the best estimate of the length is x. The DEVIATIONS of the various readings from x are $x_1 - x$, $x_2 - x$, $x_3 - x$, etc. Let S be the sum of the squares of these deviations, then

$$S = (x_1 - x)^2 + (x_2 - x)^2 + \ldots\ldots (x_{10} - x)^2 \qquad (2.1)$$

As S is the sum of a number of squares, it can only be positive or zero, and it is extremely unlikely that it is zero. It certainly cannot be zero in the example that has been taken here. If x is chosen poorly, it is possible to make S quite large. For example, if x is taken as 5·1, then $S = 0{\cdot}007\,644$ and, if as 5, then $S = 0{\cdot}162\,844$. As a poor value of x makes S large, suppose x is chosen so that S becomes as small as possible.

What value of x will make S, as given in (2.1), as small as possible?

****

2A

For S to be least, we must have $\frac{dS}{dx} = 0$. This leads to the value of x as being

$$(x_1 + x_2 + \ldots\ldots + x_{10})/10$$

which, of course, is the average value, usually denoted by $\bar{x}$. $\frac{d^2S}{dx^2} = 20$ and this confirms that S is a minimum.

FRAME 3

For this x, $S = 0{\cdot}000\,026\,4$, and by comparison with this value, 0·007 644 and 0·162 844 are certainly large, however small they might have seemed to be when you were reading the last frame.

The process of making S as small as possible is called the METHOD OF LEAST SQUARES. As you have seen, when only one variable is involved, it leads to the mean value. If the number of x's is n instead of 10, then

FRAME 3 (continued)

$$S = (x_1 - x)^2 + (x_2 - x)^2 + \dots + (x_n - x)^2$$

and putting $\frac{dS}{dx} = 0$ leads to

$$\bar{x} = (x_1 + x_2 + x_3 + \dots\dots\dots\dots + x_n)/n$$

Alternatively, the Σ notation can be used and then

$$S = \sum_{i=1}^{n} (x_i - x)^2 \qquad \text{and} \qquad \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Can you recall another place in this book where we have taken the smallness of the sum of a number of squares as a criterion in deciding which of two solutions to a problem is better?

****

3A

In the solution of linear simultaneous equations. *See FRAME 18, page 91.*

FRAME 4

Fitting the 'Best' Straight Line to a Set of Points

There are many cases in practice where, if two variables are involved, they are connected by a relation of the form $y = a + bx$. One of the best known is the elongation y produced in a wire when it is subjected to a load x. This, of course, is only of the form $y = a + bx$ provided that the elastic limit of the wire has not been reached. Another example is the length of a rod which is heated to various temperatures.

If the constants a and b are known in a particular case, there is no problem. However, sometimes the only information that you may be given, or can obtain by experiment, is a number of corresponding values of x and y. If this happens, and there is justification for the assumption that the law connecting x and y is linear, the question arises as to what values should be given to a and b to get the straight line best fitting the points (x, y) when these are plotted on a graph.

You might be wondering how a knowledge of this line can help us. Firstly, taking the load-extension situation, it can be used to estimate the length of the wire for any other load, provided that load is within the range of values used in the original experiment. Secondly it can be used for calibration purposes.

One way in which this 'best' straight line can be found is to guess its position by means of a taut thread or transparent ruler placed over a plot of the points. But, just as measurements of the length of a line will vary when made by different people, so also will guesses as to the best straight line through a number of points. It is therefore desirable to place the definition of the best straight line on to a more mathematical basis.

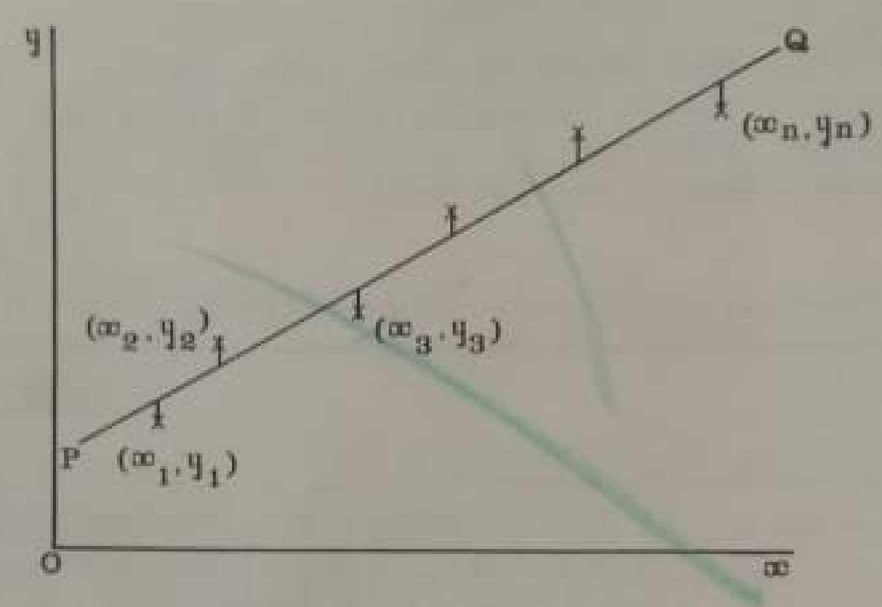
FRAME 5

A criterion that is often used for the best straight line through a number of points is based on the idea of minimising the sum of a number of squares. It will be assumed that the x measurements are correct and only the y

FRAME 5 (continued)

measurements are subject to experimental error. Thus it will be assumed that the load hung on the wire or the temperature of the rod are known exactly and only the measurements of the lengths are subject to error. In practice this will not be quite the case but it is hoped that any errors in the load and temperature are very small (so that they can be ignored) in comparison with those in the lengths.

FRAME 6



Suppose the points shown have been obtained, the x values being assumed correct, and that a line PQ has been drawn between them. The points may lie more nearly on the line than is apparent from the diagram but if they are shown very close to the line it is difficult to see what is happening. In any case, the appearance is only relative as it can be considerably altered simply by changing the scale on, say, the y axis.

Now, a small deviation d from the line can be associated with each point. These deviations are shown by the short vertical lines. Each d can be defined as

observed value of y at the point − value of y as given by the line for the same x.

The actual length shown in the figure for each point is $|d|$. It will be obvious from the figure that some d's are positive and others negative. If the equation of the line is $y = a + bx$ then, for the observed point (x_1, y_1), $d_1 = y_1 - (a + bx_1)$ and similarly for all the other points.

FRAME 7

Now, in FRAME 5, it was suggested that a criterion for the best straight line is based on the minimising of the sum of a number of squares and here it is Σd_i^2 that is used for this purpose. You might, however, think that something simpler than this could be used. The simplest sum, you would probably say, is Σd_i and wonder whether the use of this could lead anywhere. The trouble with this is that, as was previously remarked, some d's are positive, while others are negative. The tendency, then, would be for the various d's to cancel each other out in Σd_i. This might suggest that we should try to make Σd_i zero. If this is done, it is found that the only information obtained is that the line must pass through the centroid of the points, and this is insufficient to fix it completely.

Taking the squares of the deviations eliminates the effect of positive and negative d's. Another way in which this could be achieved would be to

FRAME 7 (continued)

take $\Sigma|d_i|$. The trouble with this, however, is that it is rather more difficult to deal with the mathematics of a process where moduli are involved.

As these two possibilities both have snags, it seems reasonable to take what you will probably agree is the next simplest sum, i.e. Σd_i^2, and see whether this will lead to any definite conclusion.

Denoting Σd_i^2 by S, we then have

$$S = (y_1 - a - bx_1)^2 + (y_2 - a - bx_2)^2 + \ldots\ldots + (y_n - a - bx_n)^2 \quad (7.1)$$

$$\text{or} \quad S = \sum_{i=1}^{n} (y_i - a - bx_i)^2 \quad (7.2)$$

In this expression all the x's and y's are fixed, being the coordinates of given points. a and b are regarded as the variables as it is by changing these that the position of the line is varied. Thus, if b is altered the slope of the line changes so that it effectively rotates, and if a is altered the line moves bodily upwards or downwards.

Now if u is a function of one variable t, it is necessary for $\frac{du}{dt}$ to be zero if u is to have a minimum value. The corresponding requirement for S, a function of the two variables a and b, to have a minimum value is that $\frac{\partial S}{\partial a}$ and $\frac{\partial S}{\partial b}$ should be zero simultaneously. Theoretically this will not distinguish between a maximum and a minimum just as $\frac{du}{dt} = 0$ doesn't either, but in practice this doesn't matter here as it is very easy to make Σd_i^2 as large as you like.

What equations will result from either (7.1) or (7.2) when you put $\frac{\partial S}{\partial a} = 0$ and $\frac{\partial S}{\partial b} = 0$?

****

7A

i) Using (7.1) these give

$$-2(y_1 - a - bx_1) - 2(y_2 - a - bx_2) - \ldots -2(y_n - a - bx_n) = 0$$

$$\text{and } -2x_1(y_1 - a - bx_1) - 2x_2(y_2 - a - bx_2) - \ldots -2x_n(y_n - a - bx_n) = 0$$

$$\text{i.e., } (y_1 - a - bx_1) + (y_2 - a - bx_2) + \ldots + (y_n - a - bx_n) = 0 \quad (7A.1)$$

$$\text{and } x_1(y_1 - a - bx_1) + x_2(y_2 - a - bx_2) + \ldots + x_n(y_n - a - bx_n) = 0$$

ii) Using (7.2), these give

$$\sum_{i=1}^{n} \{(-2)(y_i - a - bx_i)\} = 0 \quad \text{and} \quad \sum_{i=1}^{n} \{(-2x_i)(y_i - a - bx_i)\} = 0$$

$$\text{i.e., } \sum_{i=1}^{n} (y_i - a - bx_i) = 0 \quad (7A.2) \quad \text{and} \quad \sum_{i=1}^{n} x_i(y_i - a - bx_i) = 0 \quad (7A.3)$$

FRAME 8

As you will realise, the pairs of results given in 7A are simply two slightly different ways of stating the same things. The second of them

FRAME 8 (continued)

is more compact and then:

From (7A.2), $\sum_{i=1}^{n} y_i - na - b\sum_{i=1}^{n} x_i = 0$ or $na + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$ (8.1)

From (7A.3), $\sum_{i=1}^{n} x_iy_i - a\sum_{i=1}^{n} x_i - b\sum_{i=1}^{n} x_i^2 = 0$

or $a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_iy_i$ (8.2)

If you had difficulty in seeing where na comes from in (8.1), have a look back at (7A.1) which is the expanded form of (7A.2). In it there are n brackets, each containing a.

(8.1) and (8.2) are two simultaneous equations for a and b. They are called the NORMAL EQUATIONS.

If (8.1.) is divided by n it gives information about one of the points through which the line $y = a + bx$ passes. Can you spot what this is?

****

8A

$$a + b\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$\therefore\ a + b\bar{x} = \bar{y}$ *as* $\frac{1}{n}\sum_{i=1}^{n} x_i$ *is the mean of the* x *values, i.e.,* $\bar{x}$ *and similarly* $\frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}$. *The straight line found in this way therefore passes through* $(\bar{x}, \bar{y})$, *i.e., through the centroid of the observed points.*

FRAME 9

As an example, let us take the case of a rod that is heated to various temperatures, its length being measured at intervals of 10°C from 10°C to 70°C. The following table shows the results obtained, T, assumed free from error, being the temperature in °C and ℓ the length in mm:

T	10	20	30	40	50	60	70
ℓ	962·3	962·5	962·6	962·9	963·0	963·2	963·4

It is known that the law connecting T and ℓ is linear. If it is of the form $\ell = a + bT$, what are the best values for a and b?

What will the normal equations be for this example? Give them in the forms corresponding to (8.1) and (8.2), i.e., without substituting the numerical values.

****

9A

$$na + b\sum_{i=1}^{n} T_i = \sum_{i=1}^{n} \ell_i, \qquad a\sum_{i=1}^{n} T_i + b\sum_{i=1}^{n} T_i^2 = \sum_{i=1}^{n} T_i\ell_i$$

FRAME 10

Here $n = 7$, $\Sigma T = 280$, $\Sigma T^2 = 14\,000$, $\Sigma \ell = 6739{\cdot}9$ and $\Sigma T\ell = 269\,647$.

(The abbreviated notation ΣT for $\sum_{i=1}^{n} T_i$ has now been used.)

The normal equations are thus

$$7a + 280b = 6739{\cdot}9 \qquad 280a + 14\,000b = 269\,647$$

from which $a = 962{\cdot}1$, $b = 0{\cdot}018\,214$, and so, to three significant figures,

$$\ell = 962 + 0{\cdot}0182T \qquad (10.1)$$

FRAME 11

There are two points of interest to notice about this example. The first concerns the arithmetic involved.

If a calculating machine is available, all or some of the sums ΣT^2, $\Sigma T\ell$, ΣT and $\Sigma \ell$ can be formed simultaneously on the machine without writing down any intermediate figures. If such a machine is not available, it is necessary to record the individual values of T^2 and $T\ell$. In this case, it is best to extend the table given in FRAME 9 as follows:

T	ℓ	T^2	$T\ell$
10	962·3	100	9623
20	962·5	400	19 250
30	962·6	900	28 878
40	962·9	1600	38 516
50	963·0	2500	48 150
60	963·2	3600	57 792
70	963·4	4900	67 438
280	6739·9	14 000	269 647

FRAME 12

The second point concerns the Physics of the problem. If you have studied this subject, you will know that the equation for the linear expansion of a rod is usually given in the form

$$\ell = \ell_0(1 + \alpha T)$$

ℓ_0 being the length of the rod at zero temperature and α the coefficient of expansion. Putting equation (10.1) in this form gives

$$\ell = 962(1 + 0{\cdot}000\,0189T)$$

Comparing this value of α with a table of values of α for different materials suggests that the rod could possibly have been made of brass.

FRAME 13

Now try the following example:

In a wind tunnel test, a set of helicoidal propellers of different pitch ratios gave, at constant thrust, torque ratios as in the following table:

FRAME 13 (continued)

Pitch ratio	0·3	0·5	0·7	0·9	1·2	1·5	1·8	2·2	2·6
Torque ratio	0·316	0·533	0·753	0·979	1·310	1·650	1·980	2·436	2·875

Find the best line of the form $y = a + bx$ to fit these values, where x denotes the pitch ratio (assumed free from error) and y the torque ratio. Compare the values given for y by the formula you obtain with those in the table.

****

13A

$\Sigma x = 11{\cdot}7, \quad \Sigma x^2 = 20{\cdot}17, \quad \Sigma y = 12{\cdot}832, \quad \Sigma xy = 22{\cdot}2147$

Normal equations are

$$9a + 11{\cdot}7b = 12{\cdot}832 \qquad 11{\cdot}7a + 20{\cdot}17b = 22{\cdot}2147$$

giving

$$a = -0{\cdot}024\,43, \qquad b = 1{\cdot}115\,64, \qquad y = -0{\cdot}0244 + 1{\cdot}116x$$

The values of y, *as given by this equation, are*

$0{\cdot}310, \quad 0{\cdot}534, \quad 0{\cdot}757, \quad 0{\cdot}980, \quad 1{\cdot}315, \quad 1{\cdot}650, \quad 1{\cdot}984, \quad 2{\cdot}431, \quad 2{\cdot}877.$

FRAME 14

When calculating the equation of the best straight line, it is advisable, within reason, not to round off individual calculations as these are performed. (An exception to this rule however is in division, where you can't usually help it.) Rather, it is better to wait right till the end and only round off the final values. Obviously, errors have occurred in the measurements of y and these are going to affect the result. As you saw in the first programme in Unit 1, rounding off can seriously affect the results of certain arithmetical operations and it is better not to introduce the possibility of this happening, thus making your result even more inaccurate.

FRAME 15

Extension of the Least Squares Process to Laws reducible to a Linear Form

Although some physical laws are of the straight line type, there are many that are not. In such cases the best straight line calculation can be performed but the result of doing so can be an extremely bad fit. As an example take the points (-2,4), (-1,1), (0,0), (1,1), (2,4), (3,9) and find the 'best' straight line passing through them. Then plot the points on a graph and also the line you have found.

****

15A

$$\Sigma x = 3, \quad \Sigma x^2 = 19,$$
$$\Sigma y = 19, \quad \Sigma xy = 27$$

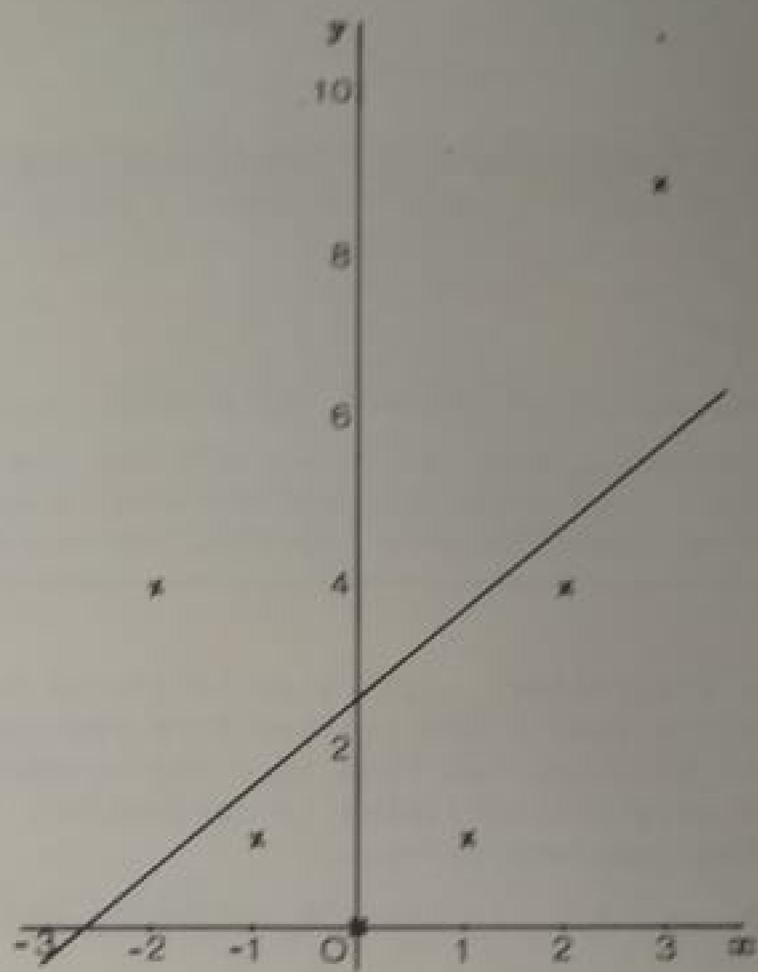
Normal equations are

$$6a + 3b = 19$$
$$3a + 19b = 27$$

giving $a = 8/3, \quad b = 1$

$$y = \frac{8}{3} + x$$

15A (continued)



FRAME 16

The points given in FRAME 15 lie exactly on the parabola $y = x^2$. It is obvious from your graph that the line you obtained is by no means a good representation of the plotted points. Fortunately, however, in practical cases, theory often provides us with a clue as to the type of curve on which a series of experimental points should lie. For example, if a gas is compressed isothermally, theory states that the pressure should be inversely proportional to the volume. On the other hand, if it is compressed adiabatically, the pressure and volume should be connected by a law of the form $pv^{\gamma} = c$.

FRAME 17

By making a suitable change of one or both variables it is sometimes possible to reduce the law to one which is of the straight line form. For example, if $pv = c$, then $p = \frac{c}{v}$ and if the substitution $\frac{1}{v} = V$ is made, $p = cV$ and the graph of p against V will be linear. In the case of $pv^{\gamma} = c$, $\log p + \gamma \log v = \log c$ and if you put $\log p = P$ and $\log v = V$, then $P + \gamma V = \log c$ and this is again linear.

Suggest alternative forms and/or suitable substitutions which would make the following laws linear:

i) Variables i and p: law $p = ki^n$
ii) Variables x and y: law $y = ke^{cx}$
iii) Variables w and I: law $w = \frac{a}{I} + b$
iv) Variables A and d: law $A = ad^2 + bd$

****

17A

i) $\log p = \log k + n \log i$: $\log p = P$ and $\log i = I$
ii) $\log y = \log k + cx \log e$: $\log y = Y$ and $c \log e = C$
iii) $\frac{1}{I} = i$
iv) $A/d = ad + b$: $\frac{A}{d} = Y$

The letters used in any substitutions are, of course, a matter of choice.

Note: When logs are involved, base 10 is usually the most convenient choice. In (ii) natural logs might be used instead, then the equation becomes $\ln y = \ln k + cx$ and the only substitution necessary is $\ln y = Y$.

FRAME 18

Having carried out the preliminary algebra as indicated in FRAME 17, the best straight line for the law in the revised form can easily be found. Finally this law has to be converted back into the original form, so that a connection between the original variables is obtained. The following example illustrates the complete process.

Find a law of the form $pv^{\gamma} = c$ to fit the following data. Assume v is measured accurately.

v	10	12	14	16	18	20	22	24
p	39·7	30·9	24·9	20·5	17·5	15·1	13·1	11·6

As was seen in FRAME 17, the linear law corresponding to $pv^{\gamma} = c$ is of the form $P + \gamma V = \log c$ or $P = C - \gamma V$ where $P = \log p$, $V = \log v$, $C = \log c$. The base of the logs is immaterial here but 10 is probably the most convenient.

Form the table showing corresponding values of P and V and find the best straight line law $P = C - \gamma V$ or $P = C + \beta V$ where $\beta = -\gamma$.

****

18A

V	1·0000	1·0792	1·1461	1·2041	1·2553	1·3010	1·3424	1·3802
P	1·5988	1·4900	1·3962	1·3118	1·2430	1·1790	1·1173	1·0645

Here 4 figure logs have been used.

$\Sigma V = 9{\cdot}7083$, $\Sigma V^2 = 11{\cdot}9034$, $\Sigma P = 10{\cdot}4006$, $\Sigma VP = 12{\cdot}4498$

Normal equations are

$$8C + 9{\cdot}7083\beta = 10{\cdot}4006$$
$$9{\cdot}7083C + 11{\cdot}9034\beta = 12{\cdot}4498$$

from which $C = 3{\cdot}008$ *and* $\beta = -1{\cdot}407$. *Thus* $\gamma = 1{\cdot}407$ *and* $c = 1019$ *and the law is* $pv^{1{\cdot}407} = 1019$.

FRAME 19

The actual working in this example has not been carried through to the full number of decimal places, as each V^2 and VP would be to 8 decimal places. This is a case where the "within reason" qualification of FRAME 14 would be used.

FRAME 20

Extension of Least Squares Process to Polynomial Laws

There are some non-linear laws, such as $y = a + bx + cx^2$, for example, where the method of least squares can be applied directly. Indeed, in the case of a law such as this, it would be impossible to find an associated law of a straight line form. The reason for this is that here there are three unknown constants whereas, in the equation of a straight line, only two constants occur. Now as the normal equations are solved simultaneously to find the constants in the equation of the straight line or curve being found, it follows that in the case of the parabola $y = a + bx + cx^2$ there must be three normal equations. It is necessary then to see how these are formed.

FRAME 21

As before, suppose n points $(x_1, y_1), (x_2, y_2), \ldots\ldots, (x_n, y_n)$ have been found by experiment, all the x's being assumed free from error. To each point (x_i, y_i) there will be an associated deviation $y_i - (a + bx_i + cx_i^2)$, this being the difference between the observed value of y_i and the value of y_i as calculated from $y = a + bx + cx^2$ when $x = x_i$. The sum of the squares of the deviations, i.e.,

$$S = \sum_{i=1}^{n} (y_i - a - bx_i - cx_i^2)^2$$

is now minimised as before. However, this time S is regarded as a function of three independent variables a, b and c. The values of these will serve to fix the parabola.

For S to be a minimum when it is a function of a, b and c, it is necessary that the conditions $\frac{\partial S}{\partial a} = 0$, $\frac{\partial S}{\partial b} = 0$, $\frac{\partial S}{\partial c} = 0$ are satisfied simultaneously. Three normal equations will thus be obtained for a, b and c. Find these equations, expressing them in forms similar to (8.1) and (8.2).

****

21A

$$\frac{\partial S}{\partial a} = \sum_{i=1}^{n} \left\{-2(y_i - a - bx_i - cx_i^2)\right\}$$

$$\frac{\partial S}{\partial b} = \sum_{i=1}^{n} \left\{-2x_i(y_i - a - bx_i - cx_i^2)\right\}$$

$$\frac{\partial S}{\partial c} = \sum_{i=1}^{n} \left\{-2x_i^2(y_i - a - bx_i - cx_i^2)\right\}$$

Putting each of these equal to zero leads to

$$na + b\Sigma x_i + c\Sigma x_i^2 = \Sigma y_i \qquad (21A.1)$$
$$a\Sigma x_i + b\Sigma x_i^2 + c\Sigma x_i^3 = \Sigma x_i y_i \qquad (21A.2)$$
$$a\Sigma x_i^2 + b\Sigma x_i^3 + c\Sigma x_i^4 = \Sigma x_i^2 y_i \qquad (21A.3)$$

The summation limits are here understood to be from $i = 1$ to n.

FRAME 26 (continued)

are $\begin{pmatrix} 8 & 13 & 33 \\ 13 & 33 & 225 \\ 33 & 225 & 978 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} 381 \\ 1088 \\ 5656 \end{pmatrix}$

In the coefficient matrix, you will notice that the elements on the trailing diagonal (from bottom left to top right) are equal to each other (33). Also the elements on any line parallel to this are equal to each other (13 on one and 225 on the other). If you examine the equations (21A.1), (21A.2) and (21A.3) you will see that this must always occur.

FRAME 27

The method of extension of this technique to higher degree polynomials follows immediately.

Can you say what the normal equations would be if it is desired to fit the curve

$$y = a_0 + a_1x + a_2x^2 + a_3x^3$$

to a set of points (x_i, y_i)? If not, see if you can work them out.

****

27A

$$\begin{aligned} na_0 + a_1\Sigma x_i + a_2\Sigma x_i^2 + a_3\Sigma x_i^3 &= \Sigma y_i \\ a_0\Sigma x_i + a_1\Sigma x_i^2 + a_2\Sigma x_i^3 + a_3\Sigma x_i^4 &= \Sigma x_i y_i \\ a_0\Sigma x_i^2 + a_1\Sigma x_i^3 + a_2\Sigma x_i^4 + a_3\Sigma x_i^5 &= \Sigma x_i^2 y_i \\ a_0\Sigma x_i^3 + a_1\Sigma x_i^4 + a_2\Sigma x_i^5 + a_3\Sigma x_i^6 &= \Sigma x_i^3 y_i \end{aligned}$$

FRAME 28

The pattern in these equations should now be obvious, as also should be the property mentioned in FRAME 26 for the example there. It is now very easy to see what the normal equations will be for even higher degree polynomials, and a flow diagram for these is given on page 174, taking the case of a polynomial of degree m and n(> m) data points. Unfortunately, however, there can be some difficulty with the arithmetic, including a tendency towards ill-conditioning, when the degree of the polynomial is increased to more than about four.

FRAME 29

Use of a False Origin or Coding

Sometimes, when doing an example manually, it is helpful to subtract a number from all the x readings, from all the y readings, or numbers from both sets. This is simply to reduce the amount of arithmetic involved.

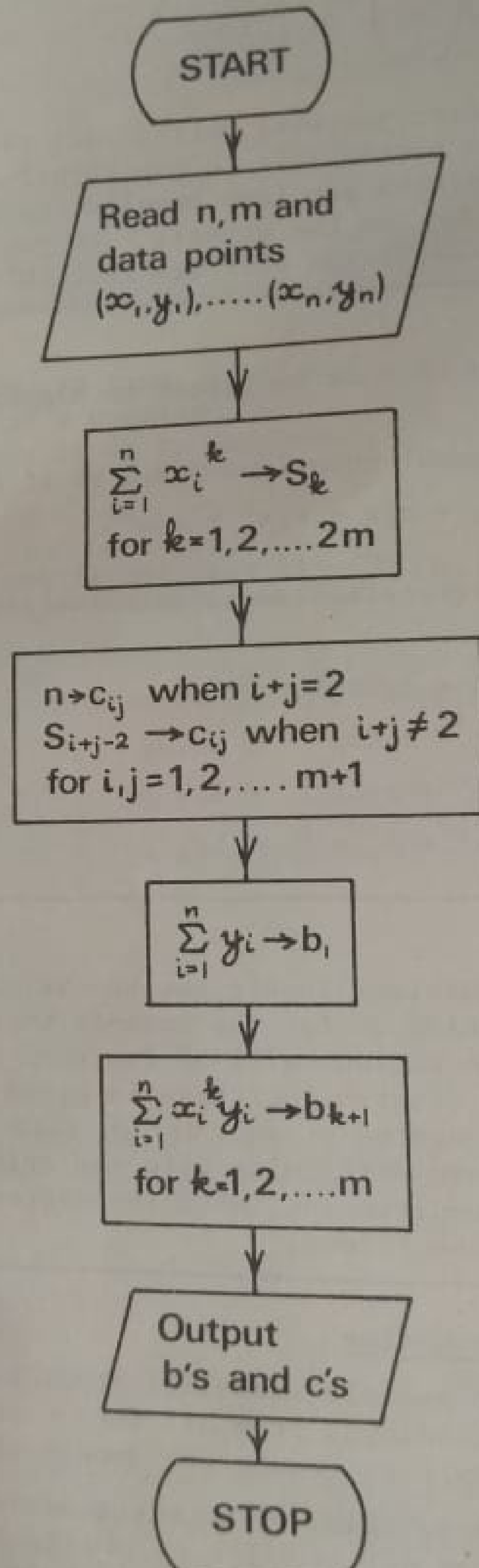
Returning to the example in FRAME 9, we might decide to subtract 40 from all the T's and 963 from all the ℓ's. 40 is the mean of the T's and, when the mean is a simple figure, this is the best number to use. If X = T − 40 and Y = ℓ − 963, then a table showing values of X and Y is

X	−30	−20	−10	0	10	20	30
Y	−0·7	−0·5	−0·4	−0·1	0·0	0·2	0·4

Find the best straight line through these points.

****

Flow diagram for FRAME 28.



(Programs using the method of least squares can be found in references (3), (7) and (8).)

FRAME 32 (continued)

3. The following table gives index numbers of food retail prices of two countries for a number of years.

Year	1960	1962	1964	1966	1968	1970	1972
Country A	182	176	169	186	200	227	232
Country B	157	161	163	172	197	201	212

Denoting readings for Country A by x and those for B by y, find the two regression lines.

4. The conductive heating of steel pipes at a given temperature was tested and gave the following readings for the power p and the current i:

i	25	20	15	10	5
p	0·432	0·305	0·191	0·103	0·035

Find the best law of the form $p = ci^k$ to fit this data, assuming the readings of i to be correct.

5. The following table gives observations of the stopping distance of a vehicle travelling at speed v

v(km/h)	30	40	50	60	70
s(m)	90	138	206	292	396

Write down the normal equations for a least-squares fit of form

(a) $s = a + bv + cv^2$, (b) $s = bv + cv^2$

Select one of these formulae, giving a reason for your choice and hence determine the relevant parameters of the regression. (C.E.I.)

(Note: The parameters of the regression are a, b and c. Readings of v are assumed to be free of error.)

6. In certain cases, y might be a function of two independent variables, x and t, say. If the relationship between the variables is $y = a + bt + cx$, find the normal equations by minimising $S = \Sigma(y - a - bt - cx)^2$. Assume that x and t are without error.

7. The yield of a chemical process was measured at three temperatures, each with two concentrations of a particular reactant, as recorded below:

Temperature, $t^{o}C$	40	40	50	50	60	60
Concentration, x	0·2	0·4	0·2	0·4	0·2	0·4
Yield, y	38	42	41	46	46	49

Use the method of least squares to find the best values of the coefficients a, b, c in the equation $y = a + bt + cx$, and from your equation estimate the yield at $70^{o}C$ with concentration 0·5. (L.U.)

FRAME 33

Answers to Miscellaneous Examples

1. Writing the law in the form $p_t = p_o + \beta t$ leads to $p_t = 10 + 0{\cdot}036\,37t$ which gives $p_t = 10(1 + 0{\cdot}003\,637t)$.

If coding is used in this example, it would be reasonable to subtract

20, say, from all the values of t and 10 or 11 from all the values of p_t. The mean of the t's, which is $22\frac{1}{4}$, is not such a simple figure to use as 20.

2. $S = \sum_{i=1}^{n} (x_i - a' - b'y_i)^2$

Normal equations are

$$na' + b'\Sigma y_i = \Sigma x_i \quad (33.1) \qquad a'\Sigma y_i + b'\Sigma y_i^2 = \Sigma y_i x_i$$

You will notice that these equations are similar to (8.1) and (8.2) but with the x's and y's interchanged. As $\Sigma y_i = n\bar{y}$ and $\Sigma x_i = n\bar{x}$, (33.1) is equivalent to $a' + b'\bar{y} = \bar{x}$ and so this regression line also passes through the centroid of the points.

3. $n = 7$, $\Sigma x = 1372$, $\Sigma x^2 = 272\,610$
$\Sigma y = 1263$, $\Sigma y^2 = 230\,877$, $\Sigma xy = 250\,660$

Line of regression of y on x is $y = 15{\cdot}488 + 0{\cdot}8415x$
Line of regression of x on y is $x = 8{\cdot}568 + 1{\cdot}0388y$

If coding is used one could subtract, say, 200 from x and 180 from y. Then, if $X = x - 200$, $Y = y - 180$,

X	-18	-24	-31	-14	0	27	32
Y	-23	-19	-17	-8	17	21	32

$n = 7$, $\Sigma X = -28$, $\Sigma X^2 = 3810$
$\Sigma Y = 3$, $\Sigma Y^2 = 2997$, $\Sigma XY = 3100$

These lead to the same results. For manual working these numbers are much easier to deal with.

4. $\log p = \log c + k \log i$ or $P = C + kI$

I	1·3979	1·3010	1·1761	1·0000	0·6990
P	-0·3645	-0·5157	-0·7190	-0·9872	-1·4559

$P = -2{\cdot}5481 + 1{\cdot}560\,53I$
$p = 0{\cdot}002\,83i^{1{\cdot}56}$, quoting to 3 significant figures.

5. The normal equations for (a) are

$$\begin{aligned} na + b\Sigma v + c\Sigma v^2 &= \Sigma s \\ a\Sigma v + b\Sigma v^2 + c\Sigma v^3 &= \Sigma vs \\ a\Sigma v^2 + b\Sigma v^3 + c\Sigma v^4 &= \Sigma v^2 s \end{aligned}$$

and for (b)

$$\begin{aligned} b\Sigma v^2 + c\Sigma v^3 &= \Sigma vs \\ b\Sigma v^3 + c\Sigma v^4 &= \Sigma v^2 s \end{aligned}$$

In each case $\sum_{i=1}^{n} v_i$, etc., have been abbreviated to Σv, etc.

As $s = 0$ obviously corresponds to $v = 0$, the relationship between them should not contain a and so it would appear possible to use the simpler (b) equation. However the point (0,0) is outside the range of the figures and the best parabola through the given points may not pass through the origin. The calculation given here is for (a).