Chapter 5 - Least Square

- Best straight line: Linear regression:

1. Scatter diagram:

Let us suppose that n pairs of measurement \((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\) are mad on two variable \(x\) and \(y\). The first step in the investigation is to plot the data on a scatter diagram in order to get a rough idea of the relationship (if any) between \(x\) and \(y\).

Example (1)

An experiment was set up to investigate the variation of the specific heat of a certain chemical with temperature. Two measurement of the specific heat were taken at each of a series of temperature. The following result were obtained:

<table>
<thead>
<tr>
<th>Temperature (°C)</th>
<th>50</th>
<th>60</th>
<th>70</th>
<th>80</th>
<th>90</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Specific Heat</td>
<td>1.6</td>
<td>1.63</td>
<td>1.67</td>
<td>1.70</td>
<td>1.71</td>
<td>1.71</td>
</tr>
<tr>
<td></td>
<td>1.64</td>
<td>1.65</td>
<td>1.67</td>
<td>1.72</td>
<td>1.72</td>
<td>1.74</td>
</tr>
</tbody>
</table>

Plot the result on a scatter diagram.
Fig (1) - Scatter diagram of data.

The problem now is to fit a line or curve to the data in order to predict the mean value of the dependent variable for a given value of the controlled variable. If the dependent variable is denoted by $y$ and the controlled variable by $x$, this curve is called the regression curve, or line of $y$ on $x$.

A general method of estimating the parameter of a regression curve is by the method of least square.
2 - Least Square method: Best straight line.

we will begin by considering the problem of fitting a straight line to \( n \) pairs of measurement \((x_i, y_i), \ldots, (x_n, y_n)\) where the \( y_i \) are subject to scatter but the \( x_i \) are not. A straight line can be represented by the equation:

\[ y = a_0 + a_1 x \]

Our task is to find estimates of \( a_0 \) and \( a_1 \), such that line gives a good fit to the data. One way of doing this is by the method of least squares.

At any point \( x_i \), the corresponding point on the line is given by \( y = a_0 + a_1 x \), so the difference between the observed value of \( y \) and the predicted value is given by:

\[ \varepsilon_i = y_i - (a_0 + a_1 x_i) \]

The least squares estimates of \( a_0 \) and \( a_1 \) are obtained by choosing the values which minimize the sum of squares of these deviations.

The sum of the squared deviations is given by
\[ S = \sum_{i=1}^{n} e_i^2 \]

\[ = \sum_{i=1}^{n} \left[ y_i - (a_0 + a_1 x_i) \right]^2 \]

This quantity is a function of the unknown parameters \( a_0 \) and \( a_1 \).

It can be minimize by calculating \( \frac{dS}{da_0} \) and \( \frac{dS}{da_1} \), setting both these partial derivatives equal to zero and solving the two simultaneous equations to obtain the least square estimates \( a_0 \) and \( a_1 \),

\[
\frac{dS}{da_0} = \sum 2(y_i - a_0 - a_1 x_i)(-1) \\
\frac{dS}{da_1} = \sum 2(y_i - a_0 - a_1 x_i)(-x_i)
\]

When these are both zero we have

\[ \sum 2(y_i - a_0 - a_1 x_i)(-1) = 0 \]

\[ \sum 2(y_i - a_0 - a_1 x_i)(-x_i) = 0 \]

These can be solved for \( a_0 \) and \( a_1 \) given by

\[ a_1 = \frac{n \sum x_i y_i - \sum y_i \sum x_i}{n \sum x_i^2 - (\sum x_i)^2} \]
\[ a_0 = \left[ \frac{\sum x^2 \sum y - \sum x \sum xy}{n \sum x^2 - (\sum x)^2} \right] \]
Correlation:

We turn our attention to a different type of situation in which measurements are made simultaneously on two variables, neither of which can be controlled. In other words they are both random variables.

Example:

The following pairs of (cooled) measurements were taken of the temperature and thrust of a rocket engine while it was being run under the same operating conditions. Plot the results on a scatter diagram.

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>19</td>
<td>1.2</td>
</tr>
<tr>
<td>15</td>
<td>1.5</td>
</tr>
<tr>
<td>35</td>
<td>1.5</td>
</tr>
<tr>
<td>58</td>
<td>3.3</td>
</tr>
<tr>
<td>35</td>
<td>2.5</td>
</tr>
<tr>
<td>X</td>
<td>Y</td>
</tr>
<tr>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>33</td>
<td>2.1</td>
</tr>
<tr>
<td>30</td>
<td>2.5</td>
</tr>
<tr>
<td>57</td>
<td>3.2</td>
</tr>
<tr>
<td>49</td>
<td>2.8</td>
</tr>
<tr>
<td>26</td>
<td>1.5</td>
</tr>
<tr>
<td>X</td>
<td>Y</td>
</tr>
<tr>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>45</td>
<td>2.2</td>
</tr>
<tr>
<td>39</td>
<td>2.2</td>
</tr>
<tr>
<td>25</td>
<td>1.9</td>
</tr>
<tr>
<td>40</td>
<td>1.8</td>
</tr>
<tr>
<td>40</td>
<td>2.8</td>
</tr>
</tbody>
</table>

The problem in this sort of situation is to see if the two variables are inter-related, and, if so, to find a measure of the degree of association or correlation between them.
To simplify the equations, we shall use,

\[ x_1' = x_1 - x_n, \quad y_1' = y_1 - y_n \]

Then

\[ \Sigma x_1' = \Sigma y_1' = 0 \]

Then the straight line will be

\[ y_1' = a x_1' \]

\( y_1' \) are subject to error.
\( x_1' \) are assumed precise.

Then the best estimate of \( a \)

\[ a_n = \frac{\Sigma x_1 y_1'}{\Sigma x_1'^2} \]

The standard error is

\[ S_n(a) = \sqrt{\frac{\Sigma x_1'^2 \Sigma y_1'^2 - (\Sigma x_1 y_1')^2}{(n-2) \Sigma x_1'^2}} \]

\[ S_n(a) = \sqrt{\left[ \frac{(1-r^2) \Sigma y_1'^2}{(n-2) \Sigma x_1'^2} \right]} \]

where \( r = \frac{\Sigma x_1 y_1'}{(\Sigma x_1'^2 \Sigma y_1'^2)}^{\frac{1}{2}} \)

\( r \) = Coefficient of Correlation.
Interchanging \( x' \) and \( y' \) enables us to deal with the line of regression of \( x' \) on \( y' \):

\[
x' = a^* y'
\]

The best estimate of \( a^* \) is

\[
a^*_n = \frac{\sum x' y'}{\sum y'^2}
\]

\[
S_n(a^*) = \left[ \frac{(1-r^2) \sum x'^2}{(n-2) \sum y'^2} \right]^{1/2}
\]

- **Perfect Correlation:**

Suppose the points in Figure (1) all lay exactly on a straight line. In this case, the two lines of regression would be the same.

The slope of the two lines will be reciprocal:

\[
\therefore a_n = \frac{1}{a^*_n}
\]

\[
S_n(a) = S_n(a^*) = 0
\]

\[
r = \frac{\sum x' y'}{(\sum x'^2 \sum y'^2)^{1/2}} = \pm 1
\]

For these values of the coefficient of correlation we would say it denotes perfect correlation.
For \( r = 1 \), all the observed points lie on a straight line which has positive slope. For \( r = -1 \), all the observed points lie on a straight line which has a negative slope.

The correlation is said to be positive if large values of both variables tend to occur together, and said to be negative if large values of one variable tend to occur with small values of the other variable.

![Graphs](image)

**a - positive correlation**

**b - negative correlation**

Many pocket calculators which have a facility to calculate \( r \) use the algebraically equivalent formula:

\[
 r = \frac{\sum x_i y_i - (\sum x_i \sum y_i)}{\sqrt{\left( \sum x_i^2 - \frac{(\sum x_i)^2}{n} \right) \left( \sum y_i^2 - \frac{(\sum y_i)^2}{n} \right)}}^{1/2}
\]

Example: \( r = 0.897 \)
No Correlation

Suppose the points scattered about so as to give for every measurement point at \((X_i, Y_i)\) another at \((X_i', Y_i')\)
or one at \((-X_i, Y_i)\) then

\[
\sum x_i^2 \neq 0 \quad \text{and} \quad \sum y_i^2 \neq 0
\]

\[
\therefore \ r = 0
\]

Then \(a_n = 0\), \(a_n^* = 0\)

\[
S_n (a) = \left[ \frac{\sum y_i^2}{(n-2) \sum x_i^2} \right]^{1/2}
\]

\[
S_n (a^*) = \left[ \frac{\sum x_i^2}{(n-2) \sum y_i^2} \right]^{1/2}
\]

Thus the two lines of regression, in this case are \(x\) and \(y\) axes, are contradictory as possible.

This contradictory is reinforced when, because of \((n-2)^2\) dependence of the errors, their slopes are apparently determine with greater accuracy as \(n\) of measurements.
Estimating the regression lines:

If the size of the sample correlation coefficient indicates that the random variables are interdependent, then we may want to predict the value of one of the variable from a given value of the other variable. In this situation, it is important to realize that there are two regression lines, one to predict y from x and one to predict x from y.

If the variable are linearly related, then the regression line of y on x can be denoted:

\[ y = a_0 + b_1 x \]

and the regression line of x on y by:

\[ x = b_0 + b_2 y \]

to estimated regression line of y on x is given by:

\[ y - \bar{y} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}} (x - \bar{x}) \]

and the estimated regression line of x on y is given by:

\[ x - \bar{x} = \frac{\sum xy - \frac{\sum x \sum y}{15}}{\sum y^2 - \frac{(\sum y)^2}{15}} (y - \bar{y}) \]
Chapter 4 Causes of error

- Systematic errors

It is clear how some errors may arise. For example an instrument may have zero error, reading 0.02 mm suppose five measurement of length of metal rod, \( l \), were (8.76, 8.76, 8.75, 8.74, 8.74) mm.

The mean value is then

\[
\frac{8.76 + 8.76 + 8.75 + 8.74 + 8.74}{5} = 8.75 \text{ mm}
\]

This mean value is not the best estimate of \( l \) we might to subtract the zero error (either from measurement or from the average) to give mean value equal to 8.73 mm.

Moreover, it might have happened that the temperature was falling as the experiment proceeded so that the true value was not a fixed quantity, \( l \), but one that varied with temperature, so we must know the coefficient of expansion of the metal, correct the measurements accordingly!

Such error are called systematic errors. There is no strict definition of systematic errors. Since as systematic errors for one experiment may not be for another.

They are usually constant
Random errors:

When all systematic errors have either been eliminated or corrected, we still do not obtain identical "true" measurements or repeated set-up readings.

These errors are called random errors, they may arise from ambiguities or uncertainties in the process, or from fluctuations which are too irregular or fast to be observed in detail.

For example random error that might arise in measuring the position of a galvanometer spot on a scale graduated in millimetre divisions. An error comes from the person reading the scale and one from the galvanometer itself. Suppose the observer’s eyesight is such that he can really only see that the spot is between two millimetre graduations, so that he is equally likely to say that the spot is at 11 or 12 mm. When in fact it is anywhere between the two. This is not an unreasonable assumption although for most of us this uncertainty sets in at some what less than 1mm division. Suppose in addition that the galvanometer has a faulty movement that either stop the spot 0.5 mm below its true value or allows it to overswing and stop 0.5 mm above. This effect is irregular or
random and is un biased, by which we mean that an error of either sign equally likely to occur when a measurement is made.

**Matching error:**

Suppose that we decide to improve the precision of the apparatus by changing the movement, and that this brings the galvanometer error down to \( \pm 0.01 \text{ mm} \).

The error diagram will now be as in Fig. Below the final row is unchanged. There is a \( \frac{1}{4} \), \( \frac{1}{2} \), \( \frac{1}{4} \) probability of reading the value 10, 11, 12 mm. and \( \sigma = \left( \frac{1}{2} \right)^{\frac{1}{2}} = 0.707 \).

We therefore improve the precision of apparatus (remember that the observer is counted as part of the apparatus).

![Error Diagram](image)
Suppose, alternatively, we reduce the observer errors in the ratio 1:5 by putting subdivision at 0.2 mm intervals and supplying a magnifying glass so he can see when the galvo spot lies between two of these. Then we have the diagram below which indicates a \( \frac{1}{8} \), \( \frac{1}{4} \), \( \frac{1}{4} \) probability for the values 10.4, 10.6, 11.4, 11.6 mm.

The mean will be 11 mm and the standard deviation \( \sigma = 0.509 \)

Thus neither single improvement given the apparatus any marked greater precision.

However, if we now reduce both errors simultaneously by the same ratio 1:5 we shall have the figure below.
Combination of two Random errors

Suppose the two errors described above affect the reading of the galvanometer spot where \(11\) is its true value. Because of the random error, the spot position will be either 10.5 mm or 11.5 mm, with equal chance for the former 10 or 11 and the latter as (11 mm) or 12 mm. Reading 10 will come from an negative galvanometer error combined with an negative observer error, and negative The reading 11 either from positive galv. error or negative observer error or negative galv. error and positive observer error, and 12 mm from a combination of two positive errors see the Figure below.

![Diagram showing the combination of two random errors affecting a galvanometer reading. The diagram includes vertices labeled 10, 11, and 12 with error directions indicated by arrows and a diamond shape connecting the points.](image-url)
Binomial Distribution:

Suppose there are $n$ sources of error, each of which can contribute very small errors $+e$ or $-e$ to the true value $x$. We can construct an error diagram for this as shown in the figure.

The final measurements, shown in the bottom line, can clearly range from $x-nc$ to $x+nc$ in steps of $2e$. Any particular combination of errors, or path in the diagram from $x$ to one of the possible measurements, can be represented by a product such as $a_1a_2...a_n b_1 b_2...b_n$ where $a_1$ means an error to the left ($-e$), $b_1$ means an error to the right ($+e$). The product $a_1a_2...a_nb_1 b_2...b_n$ would mean an error $-e$, followed by $-e$, followed by $+e$,..., followed by $re$, followed by $-e$. There are $n$ factors altogether. So if $r$ of them are $b$'s, $n-r$ will be $a$'s and the product can be written as $a^{n-r} b^r$.

This will give a measurement error $(n-r)(-e) + re = (2r-n)e$. 
The number of different combinations of \((n-r)\) factors \(a_i\) and \(r\) factors \(b\) is given by:

\[
n C_{n-r} = n C_r = \frac{n!}{r! (n-r)!}
\]

and this is just the coefficient of \(a^{n-r} b^r\) in the binomial expansion

\[(a+b)^n = a^n + n a^{n-1} b + \ldots + n C_r a^{n-r} b^r + \ldots + n b^n .\]

The absolute probability of obtaining an measurement error \((2r-n)\)

\[
\frac{n!}{2^r r! (n-r)!}
\]

Putting \((2r-n) = m\) or \(r = \frac{1}{2} (n + m)\)

we have for the probability of obtaining an error \(E = me\)

or measurement \(x = X + E\)

The expression

\[
Q(m,n) = \frac{n!}{2^r (n+m)! (n-m)!}
\]

Examples of binomial distribution for various values of \(n\) are given in the figure below.
The scale has been varied so that $e^n$ decreases as $\frac{e^n}{n^k}$ increases to give each of the distribution a comparable 'width'. We can see that as $n$ increases we approach a smooth, symmetrical single peaked distribution of just the kind discussed before. It may be remarked that $m$ can take only even or odd values according to whether $n$ is even or odd, and that $\Phi(n)$ can therefore vary only in steps of $2\pi$. However in what follows we shall allow $e$ to decrease towards zero while $n$ increases to wards infinity, in which case these restrictions become unimportant.

Some properties of the Binomial distribution

1. Mean: $\mu = np$

2. Variance: $\sigma^2 = npq$, $q = 1 - p$

3. Standard deviation: $\sigma = \sqrt{npq}$

$$P(r) = \binom{n}{r} p^r q^{n-r} = n C_r p^r q^{n-r}$$