PARUL UNIVERSITY
FACULTY OF ENGINEERING AND TECHNOLOGY
DEPARTMENT OF APPLIED SCIENCE AND HUMANITIES
4th SEMESTER B.TECH PROGRAMME
PROBABILITY, STATISTICS AND NUMERICAL METHODS
(203191251)
ACADEMIC YEAR 2019-2020

UNIT 1: CORRELATION AND REGRESSION

Correlation Analysis: we have studied problems relating to one variable only. In practice we come across a large number of problems involving the use of two or more variables. If two quantities vary in such a way that change in one variable are effects a change in the value of other. These quantities are correlated.

Types of correlation: There are three types of correlation.

(i) **Positive or Negative correlation:** If two variables are changing in the same direction, correlation is said to be **positive or direct correlation**. If two variables are changing in the opposite direction, correlation is said to be **negative or inverse correlation**.
For example: The correlation between heights and weights of group of people is positive and the correlation between pressure and volume of a gas is negative.

(i) **Simple, partial or multiple:** The difference between the simple, partial or multiple correlation is based on the number of variable studied. When only two variable are studied correlation is said to be **simple correlation**. When three or more variable are involved then the problem may be either **partial or multiple correlation.**

(ii) **Linear or Non-linear correlation:** If the amount of change in one variable tends to bear a constant ratio to the amount of change in the other variable then the correlation is said to be **linear correlation.**

For example: consider to variables X and Y

X	5	10	15	20	25	30
Y	5	10	15	20	25	30
	0	0	0	0	0	0

It is clear shows that the ratio of change in both the variables is same.

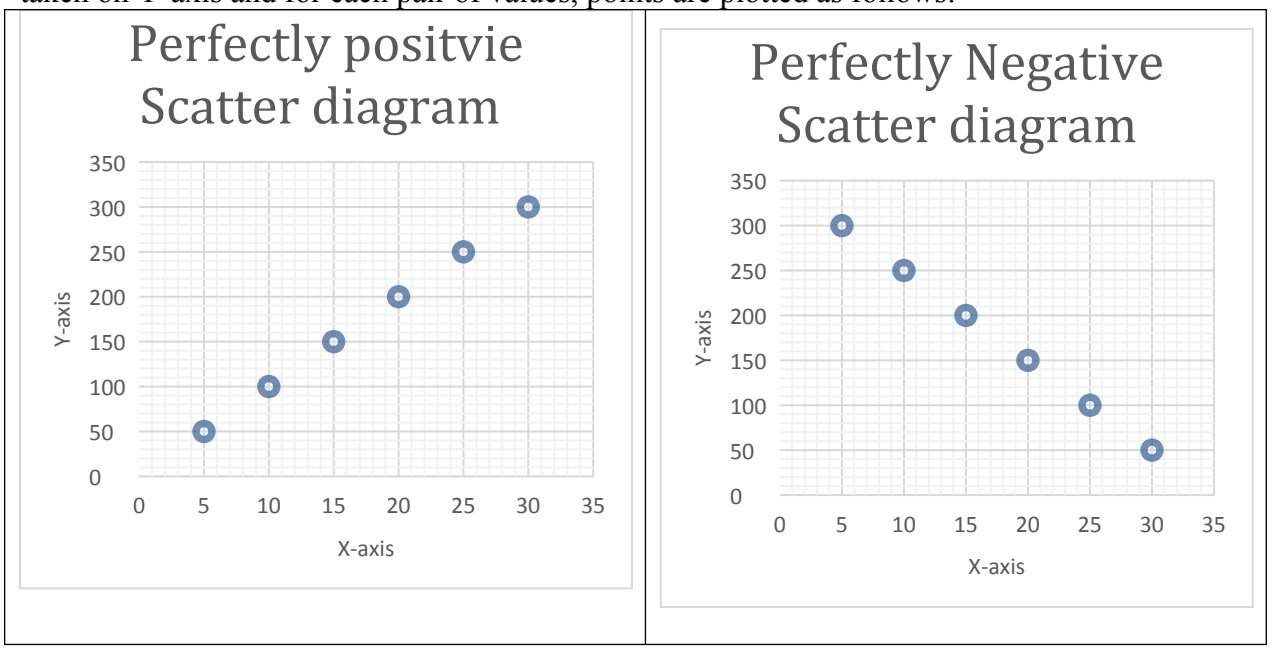
If the amount of change in one variable does not tend to bear a constant ratio to the amount of change in the other variable then the correlation is said to be **Non-linear correlation** or **curly linear correlation.**

Methods of studying correlation: There are mainly three types of methods.

(i) Scatter Diagram
(ii) Karl Pearson's method

(iii) Spearman's method of rank correlation

(i) **Scatter diagram:** This is a very simple method studying the relationship between two variables. In this method one variable is taken on X-axis and the other variable is taken on Y-axis and for each pair of values, points are plotted as follows:



(ii) **Karl Pearson's coefficient of correlation:** The several mathematical methods of measuring correlation the Karl Pearson's popularly known as Pearson's coefficient of correlation is most widely used. It is denoted by r. The formula for computing the coefficient of correlation is as follows:

Where,

This formula also can be written as follow:

Correlation coefficient for the grouped data the formula can be written as follows:

OR

Properties of the coefficient of correlation:

(1) The coefficient of correlation always lies between -1 and 1 including -1 and 1.
i.e.
(2) The correlation coefficient is independent of change of origin and scale.
(3) The correlation coefficient is an absolute number and it is independent of units of measurements.

Example: Find the Pearson's Correlation Coefficient of the following data:

x	100	101	102	102	100	99	97	98	96	95
y	98	99	99	97	95	92	95	94	90	91

Solution:

x	y	$x-\bar{x}$	$y-\bar{y}$	$\left(x-\bar{x}\right)^2$	$\left(y-\bar{y}\right)^2$	$\left(x-\bar{x}\right)\left(y-\bar{y}\right)$
100	98	1	3	1	9	3
101	99	2	4	4	16	8
102	99	3	4	9	16	12
102	97	3	2	9	4	6
100	95	1	0	1	0	0
99	92	0	-3	0	9	0
97	95	-2	0	4	0	0
98	94	-1	-1	1	1	1
96	90	-3	-5	9	25	15
95	91	-4	-4	16	16	16
$\sum x$ =990	$\sum y$ =950	$\sum\left(x-\bar{x}\right)$ =0	$\sum\left(y-\bar{y}\right)$ =0	$\sum\left(x-\bar{x}\right)^2$ =54	$\sum\left(y-\bar{y}\right)^2$ =96	$\sum\left(x-\bar{x}\right)\left(y-\bar{y}\right)$ =61

$$\bar{x}=\frac{\sum x}{n}=\frac{990}{10}=99$$

$$\bar{y}=\frac{\sum y}{n}=\frac{950}{10}=95$$

$$\textit{Correlation Coefficient}\ ,\ r=\frac{\sum\left(x-\bar{x}\right)\left(y-\bar{y}\right)}{\sqrt{\left(x-\bar{x}\right)^2}\sqrt{\left(y-\bar{y}\right)^2}}=\frac{61}{\sqrt{54}\sqrt{96}}=0.85$$

Calculated by following formula:

Where, n=number of pairs

In case finding out **rank correlation coefficient** when the observations are paired the above formula can be written as:

In $\sum d^2$, $\frac{m}{12}\left(m^2-1\right)$ is added where m is the number of times an item is repeated.
The value of correlation coefficient by Spearman's method also lies between -1 and +1. If the ranks are same for each pair of two series then each value of d=0. Hence =0 and the value of r=+1, which shows that perfect positive correlation between the two variables. If the ranks are exactly in reverse order for each pair of two series, then the value of r = which shown perfect negative correlation between the variables.

Example: Two judges have given ranks to 10 students for their honesty. Find the rank correlation coefficient of the following data:

1st Judge	3	5	8	4	7	10	2	1	6	9
2nd judge	6	4	9	8	1	2	3	10	5	7

Solution:

Rank given by 1st judge	Rank given by 2nd judge	Difference in ranks d	d^2
3	6	-3	9
5	4	1	1
8	9	-1	1
4	8	-4	16
7	1	6	36
10	2	8	64
2	3	-1	1
1	10	-9	81
6	5	1	1
9	7	2	4
			$\sum d^2$ =214

$$Rank\ Correlation\ , r = 1-\frac{6\sum d^2}{n\left(n^2-1\right)} = 1-\frac{6*214}{10\left(100-1\right)} = 1-\frac{1284}{990} = 1-1.30 = -0.30$$

Example: Find the Coefficient of rank correlation of the following data:

x	35	40	42	43	40	53	54	49	41	55
y	102	101	97	98	38	101	97	92	95	95

Solution:

x	y	Ranks in x	Ranks in y	Difference d	d^2
35	102	10	1	9	81
40	101	8.5	2.5	6	36
42	97	6	5.5	0.5	0.25
43	98	5	4	1	1
40	38	8.5	10	-1.5	2.25
53	101	3	2.5	0.5	0.25
54	97	2	5.5	-3.5	10.25
49	92	4	9	-5	25
41	95	7	7.5	-0.5	0.25
55	95	1	7.5	-6.5	42.25
					$\sum d^2$ =200.25

$$Rank\ Correlation\ , r = 1 - \frac{6\left\{\sum d^2 + \frac{m}{12}\left(m^2-1\right) + \frac{m}{12}\left(m^2-1\right) + \frac{m}{12}\left(m^2-1\right) + \frac{m}{12}\left(m^2-1\right)\right\}}{n\left(n^2-1\right)}$$

$$= 1 - \frac{6\{200.50 + 0.5 + 0.5 + 0.5 + 0.5\}}{990}$$

$$= -0.227$$

Regression Analysis: By studying the correlation we can know the existence degree and direction of relationship between two variables but we can not the answer the question of the type if there is a certain amount of change in one variable, what will be the corresponding change in the other variable. The above type of question can be answered if we can establish a quantitative relationship between two related variables.

The statistical tool by which it is possible to predict or estimate the unknown values of one variable from known values of another variable is called regression. A line of regression is straight line.

This equation is called regression line on and is called regression coefficient. The formula can be computed as:

Where,

This formula can be used to compute the value of y for given value of x.

Similarly, the regression line on and is called regression coefficient. The formula can be computed as;

=

Where,

This formula can be used to compute the value of x for the given value of y.

NOTE:

(1) and are also computed using the following formula

and

(2) Angle between the two regression lines are as follows:

When and in this case both the regression lines are perpendicular to each other. If and in this case both the regression lines are same line because point is common point.

Properties of regression coefficient:

(1) , the sign of r should be taken before the square root is that of the regression coefficient.

(2) Since both the regression coefficient cannot be greater than unity (1).

(3) Arithmetic mean of regression coefficients is greater than or equal to the coefficient of correlation.

i.e.

(4) Regression coefficient are independent of origin but not of scale.

Example: The following data regarding the heights (y) and weights (x) of 100 college students are given: $\sum x = 15000$, $\sum x^2 = 2272500$, $\sum xy = 1022250$, $\sum y = 6800$, $\sum y^2 = 463025$ Find the coefficient of correlation between height and weight and also the equation of regression of height and weight.

Solution:

Here, n=100

$$b_{yx} = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - \left(\sum x\right)^2} = 0.1$$

$$b_{xy} = \frac{n\sum xy - \sum x \sum y}{n\sum y^2 - \left(\sum y\right)^2} = 3.6$$

$$r = \sqrt{b_{xy} * b_{yx}} = \sqrt{3.6 * 0.1} = 0.6$$

$$\bar{x} = \frac{\sum x}{n} = \frac{15000}{100} = 150$$

$$\bar{y} = \frac{\sum y}{n} = \frac{6800}{100} = 68$$

The equation of the line of regression of y on x is:

$$y - \bar{y} = b_{yx}\left(x - \bar{x} \right)$$

$$y - 68 = 0.1(x - 150)$$

$$y = 0.1\ x + 53$$

The equation of the line of regression of x on y is:

$$x - \bar{x} = b_{xy}\left(y - \bar{y} \right)$$

$$x - 150 = 3.6(y - 68)$$

$$x = 3.6\ y - 94.8$$

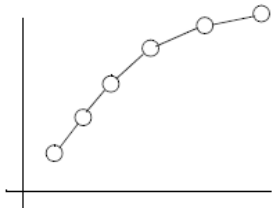
Example: Find the equation of regression line from the following data and also estimate y for $x = 1$ and x for $y = 4$.

Curve Fitting

Q: Where does this given function come from in the first place?

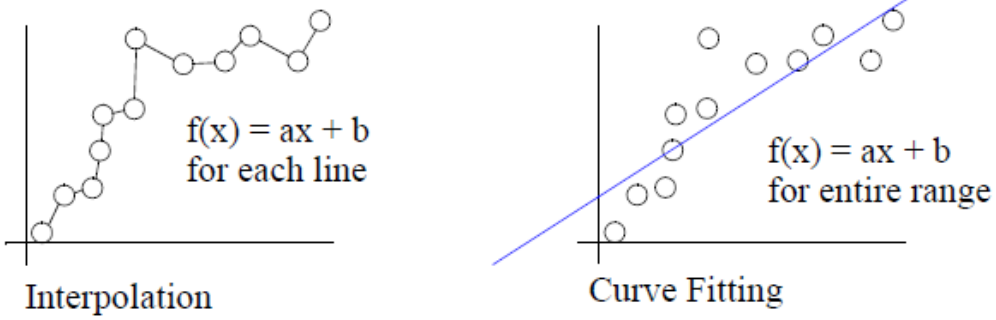
- Analytical models of phenomena (e.g. equations from physics)
- Create an equation from observed data 1)

Interpolation (connect the data-dots) If data is reliable, we can plot it and connect the dots This is piece-wise, linear interpolation.



This has limited use as a general function Sinceits really a group of small functions, connecting one point to the next it doesn't work very well for data that has built in random error (scatter) 2)

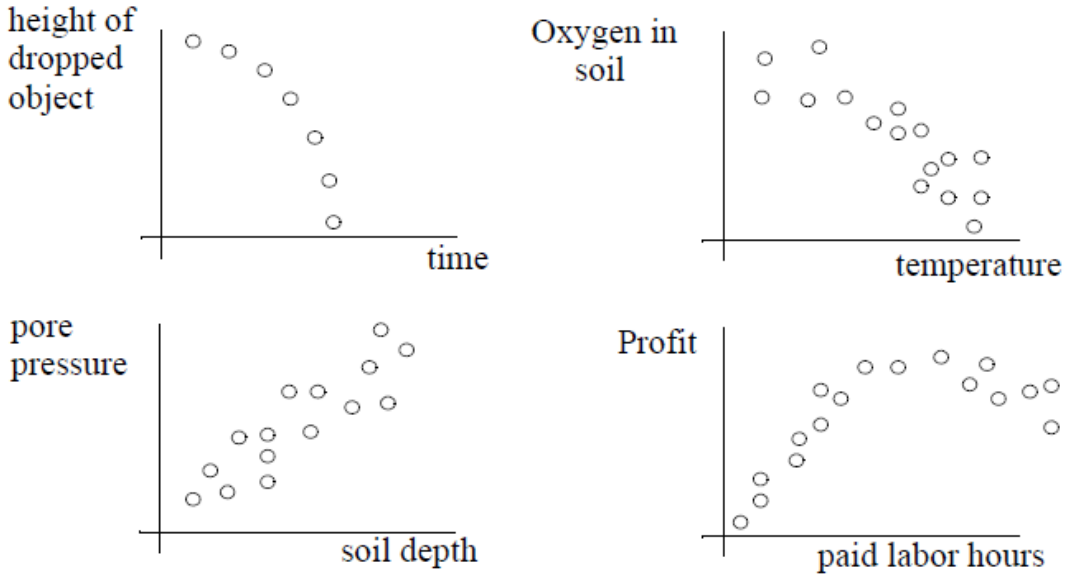
Curve fitting - capturing the trend in the data by assigning a single function across the entire range. The example below uses a straight line function



A straight line is described by $f(x) = ax + b$

The goal is to identify the coefficients 'a' and 'b' such that '$f(x)$' fits the data well

Other examples of data sets that we can fit a function to

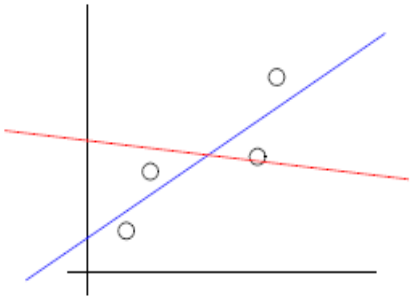


Is a straight line suitable for each of these cases ?

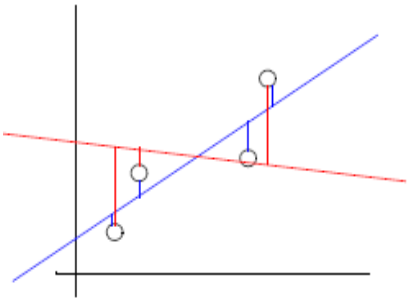
No. But we're not stuck with just straight line fits. We'll start with straight lines, then expand the concept.

Linear curve fitting (linear regression)
Given the general form of a straight line

$$f(x) = ax + b$$



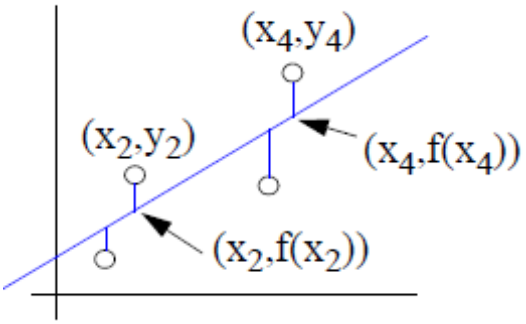
How can we pick the coefficients that best fits the line to the data?
First question: What makes a particular straight line a 'good' fit?
Why does the blue line appear to us to fit the trend better?
• Consider the distance between the data and points on the line
• Add up the length of all the red and blue verticle lines
• This is an expression of the 'error' between data and fitted line
• The one line that provides a minimum error is then the 'best' straight line



Quantifying errors in a curve fit

Assumption:
(1) positive or negative error have the same value(data point is above or below the line)

(2) Weight greater errors more heavily
we can do both of these things by squaring the distancedenote data values as (x, y) ======>>
denote points on the fitted line as (x, f(x))
sum the error at the four data points



$$err = \sum_{i=1}^{n} d_i^{\,2} = (y_1 - f(x_1))^2 + (y_2 - f(x_2))^2 + \ldots\ldots(y_n - f(x_n))^2$$
$$= (y_1 - (ax_1 + b))^2 + (y_2 - (ax_2 + b))^2 + \ldots\ldots + (y_n - (ax_n + b))^2$$
$$= \sum_{i=1}^{n} (y_i - (ax_i + b))^2$$

Error is minimum if first ordered partial derivatives=0

$$\frac{\partial(err)}{\partial a} = \sum_{i=1}^{n} -2x_i\,(y_i - (ax_i + b)) = 0 \qquad \frac{\partial(err)}{\partial b} = \sum_{i=1}^{n} -2(y_i - (ax_i + b)) = 0$$

$$\therefore \sum_{i=1}^{n} x_i y_i - a\sum_{i=1}^{n} x_i^{\,2} - b\sum_{i=1}^{n} x_i = 0 \qquad \therefore \sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i - b\sum_{i=1}^{n} 1 = 0$$

$$\therefore \sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i^{\,2} + b\sum_{i=1}^{n} x_i \qquad \textbf{and} \qquad \therefore \sum_{i=1}^{n} y_i = a\sum_{i=1}^{n} x_i + nb$$

Solve the equations

$$\sum_{i=1}^{n} y_i = a\sum_{i=1}^{n} x_i + nb \qquad (1)$$

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i^{\,2} + b\sum_{i=1}^{n} x_i \qquad (2)$$

Example: Fit a straight line using least square method

x_i	0	0.5	1	1.5	2	2.5
y_i	0	1.5	3	4.5	6	7.5

Solution:

x_i	y_i	$x_i^{\,2}$	$x_i y_i$
0	**0**	**0**	**0**
0.5	**1.5**	**0.25**	**0.75**
1	**3**	**1**	**3**
1.5	**4.5**	**2.25**	**6.75**
2	**6**	**4**	**12**
$\sum_{i=1}^{n} x_i$ **=2.5**	$\sum_{i=1}^{n} y_i$ **=7.5**	$\sum_{i=1}^{n} x_i^{\,2}$ **=6.25**	$\sum_{i=1}^{n} x_i y_i$ **=18.75**

Now, Solve the equations

$$\sum_{i=1}^{n} y_i = a\sum_{i=1}^{n} x_i + nb \qquad (1)$$

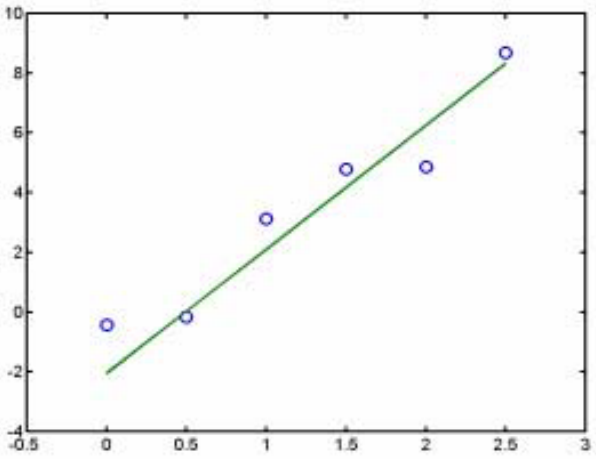
$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i^{\,2} + b\sum_{i=1}^{n} x_i \qquad (2)$$

Substitute the values from the table, here n=6.

$$7.5 = 2.5a + 6b$$
$$18.75 = 6.25a + 2.5b$$

$$a = 3.561 \text{ and } b = -0.975$$

Hence, the best fit line is .

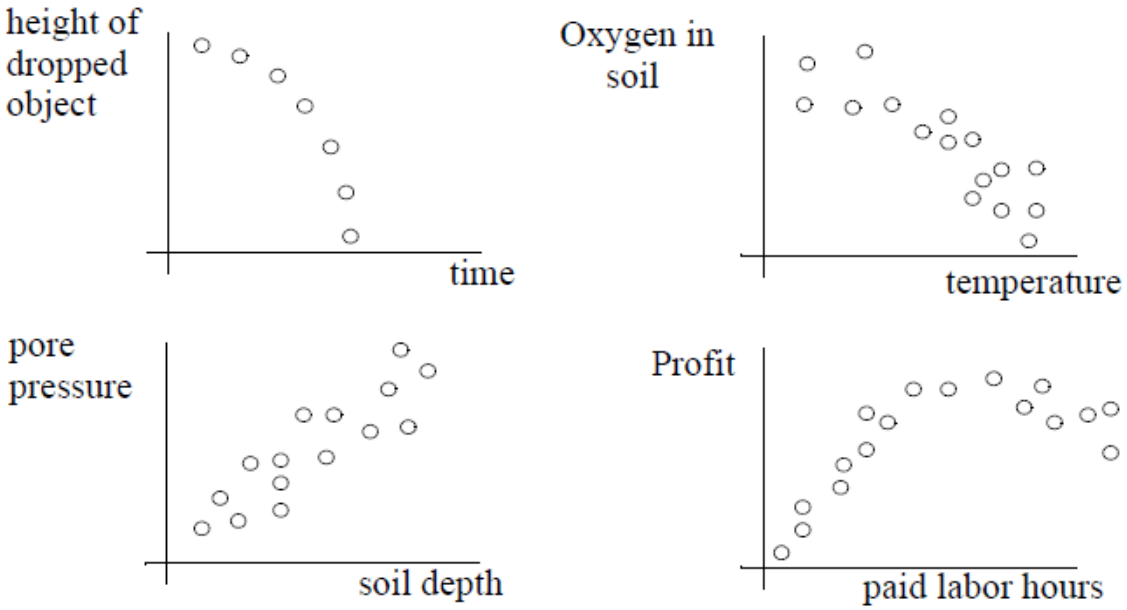


So, what we do if the straight line is not suitable for the data?

Straight line will not predict diminishing returns that data shows

Curve fitting - higher order polynomials

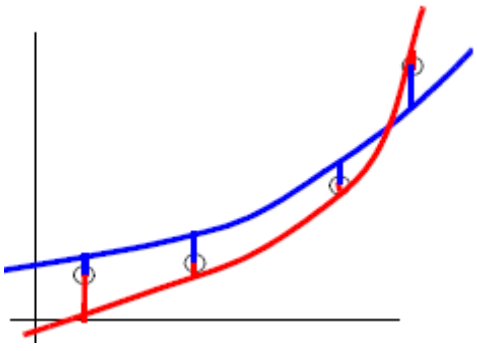
We started the linear curve fit by choosing a generic form of the straight line f(x) = ax + b
This is just one kind of function. There are an infinite number of generic forms we could choose from for almost any shape we want. Let's start with a simple extension to the linear regression concept recall the examples of sampled data



Is a straight line suitable for each of these cases ? Top left and bottom right don't look linear in trend, so why fit a straight line? No reason to, let's consider other options. There are lots of functions with lots of different shapes that depend on coefficients. We can choose a form based on experience and trial/error. Let's develop a few options for non-linear curve fitting. We'll start with a simple extension to linear regression...higher order polynomials

Curve fitting – Quadratic polynomial

Let the general form of second order polynomial $f(x) = a + bx + cx^2$.

Just as was the case for linear regression, we ask: How can we pick the coefficients that best fits the curve to the data? We can use the same idea: The curve that gives minimum error between y data and the fit $f(x)$ is 'best' Quantify the error for these two second order curves... • Add up the length of all the red and blue verticle lines • pick curve with minimum total error	

Error - Least squares approach

$$err = \sum_{i=1}^{n} d_i^{\,2} = (y_1 - f(x_1))^2 + (y_2 - f(x_2))^2 + \ldots\ldots(y_n - f(x_n))^2$$

$$= \left(y_1 - \left(a + bx_1 + cx_1^{\,2}\right)\right)^2 + \left(y_2 - \left(a + bx_2 + +cx_2^{\,2}\right)\right)^2 + \ldots\ldots + \left(y_n - \left(a + bx_n + cx_n\right)\right)^2$$

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

To minimize the error, derivatives with respect to a, b and c equal to 0.

$$\frac{\partial(err)}{\partial a} = \sum_{i=1}^{n} -2\left(y_i - \left(a + bx_i + cx_i^{\,2}\right)\right) = 0$$

$$\frac{\partial(err)}{\partial b} = \sum_{i=1}^{n} -2x_i\left(y_i - \left(a + bx_i + cx_i^{\,2}\right)\right) = 0$$

$$\frac{\partial(err)}{\partial b} = \sum_{i=1}^{n} -2x_i^{\,2}\left(y_i - \left(a + bx_i + cx_i^{\,2}\right)\right) = 0$$

Simplify these equations,We get

$$\sum_{i=1}^{n} y_i = a\,n + b\sum_{i=1}^{n} x_i + c\sum_{i=1}^{n} x_i^{\,2}$$

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^{\,2} + c\sum_{i=1}^{n} x_i^{\,3}$$

$$\sum_{i=1}^{n} x_i^{\,2} y_i = a\sum_{i=1}^{n} x_i^{\,2} + b\sum_{i=1}^{n} x_i^{\,3} + c\sum_{i=1}^{n} x_i^{\,4}$$

Example: Fit a second order polynomial equation to following data

x_i	0	0.5	1.0	1.5	2.0	2.5
y_i	0	0.25	1.0	2.25	4.0	6.25

Solution:

x_i	y_i	$x_i^{\,2}$	$x_i^{\,3}$	$x_i^{\,4}$	$x_i y_i$	$x_i^{\,2} y_i$
0	0	0	0	0	0	0
0.5	0.25	0.25	0.125	0.0625	0.125	0.0625
1	1	1	1	1	1	1
1.5	2.25	2.25	3.375	5.0625	3.375	5.0625
2	4	4	8	16	8	16
2.5	6.25	6.25	15.625	39.0625	15.625	39.0625
$\sum x_i$ =7.5	$\sum y_i$ =13.75	$\sum x_i^{\,2}$ =13.75	$\sum x_i^{\,3}$ =28.125	$\sum x_i^{\,4}$ =61.1875	$\sum x_i y_i$ =28.125	$\sum x_i y_i$ =61.1875

Substitute these values in equations

$$\sum_{i=1}^{n} y_i = a\,n + b\sum_{i=1}^{n} x_i + c\sum_{i=1}^{n} x_i^{\,2}$$

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^{\,2} + c\sum_{i=1}^{n} x_i^{\,3}$$

$$\sum_{i=1}^{n} x_i^{\,2} y_i = a\sum_{i=1}^{n} x_i^{\,2} + b\sum_{i=1}^{n} x_i^{\,3} + c\sum_{i=1}^{n} x_i^{\,4}$$

Hence, $y = x^2$ is required equation which fits the data.

Curve fitting - Other nonlinear fits (exponential)

Q: Will a polynomial of any order necessarily fit any set of data?

A: Nope, lots of phenomena don't follow a polynomial form. They may be, for example, exponential

(1) General exponential equation $f(x) = C\,e^{Ax}$

Now, take log on both side, we get

$$\ln y = \ln C + Ax$$

$$Y = b + aX; \qquad \text{where } Y = \ln y,\ X = x, \ln C = b \text{ and } a = \ln A$$

Which is equation of line, the original data in xy- plane mapped into XY-plane. This is called *linearization.*

The data (x, y) transformed as $(x, \ln y)$.

To find the value of a and b we will use the equations

$$\sum_{i=1}^{n} Y_i = a\sum_{i=1}^{n} X_i + n\,b \qquad (1)$$

$$\sum_{i=1}^{n} X_i Y_i = a\sum_{i=1}^{n} X_i^{\,2} + b\sum_{i=1}^{n} X_i \qquad (2)$$

After getting values of a and b, $A = \text{antilog}\, a,\ C = \text{antilog}\, b$.

Example: An experiment gave the following values:

X	1	5	7	9
Y	10	15	12	21

Fit an exponential curve $y = Ce^{Ax}$

Solution:

$X_i = y_i$	y_i	$Y_i = \ln y_i$	$X_i^{\,2}$	$X_i Y_i$
1	10	2.302585	1	2.302585
5	15	2.70805	25	13.54025
7	12	2.484906	49	17.39435
9	15	2.70805	81	24.37245
12	21	3.044522	144	36.53427
$\sum_{i=1}^{5} X_i$ =34		$\sum_{i=1}^{5} Y_i$ =13.24811	$\sum_{i=1}^{5} X_i^{\,2}$ =300	$\sum_{i=1}^{5} X_i Y_I$ =94.1439

$$13.24811 = 34A + 5B$$
$$94.1439 = 300A + 34B$$

A=2.00479, B=2.248664
a=antilog2.00479=7.424536

b=antilog(2.248664)=9.475068

Hence, best fit curve is $y = 9.475068\,e^{2.248664}$

(2) $y = bx^{a}$

Taking $\log_{10}$ on both the side

$$\log_{10} y = \log_{10} b + a \log_{10} x$$
$$Y = B + AX; \qquad \text{where } Y = \log_{10} y,\ X = \log_{10} x \text{ and } a = A, B = \log_{10} b$$

$$\sum_{i=1}^{n} Y_i = nB + A\sum_{i=1}^{n} X_i \qquad (1)$$
$$\sum_{i=1}^{n} X_i Y_i = B\sum_{i=1}^{n} X_i + A\sum_{i=1}^{n} X_i^{2} \qquad (2)$$

Example: An experiment gave the following values:

v (ft/min)	350	400	500	600
t (min)	61	26	7	2.6

It is known that *v* and *t* are connected by the relation $v = bt^{a}$, find the best possible values of a and b.

v	*t*	*Y=logv*	*X=logt*	*X²*	*XY*
350	61	2.544068	1.78533	3.18740262	4.542001
400	26	2.60206	1.414973	2.002149575	3.681846
500	7	2.69897	0.845098	0.714190697	2.280894
600	2.6	2.778151	0.414973	0.17220288	1.152859
		$\sum_{i=1}^{4} Y_i$ 10.62325	$\sum_{i=1}^{4} X_i$ =4.460375	$\sum_{i=1}^{4} X_i^{2}$ =6.075945772	$\sum_{i=1}^{4} X_i^{3}$ =11.6576

Substitute in given equation,

$$\sum_{i=1}^{n} Y_i = nB + A\sum_{i=1}^{n} X_i \qquad (1)$$
$$\sum_{i=1}^{n} X_i Y_i = B\sum_{i=1}^{n} X_i + A\sum_{i=1}^{n} X_i^{2} \qquad (2)$$

$$10.62325 = 4B + 4.460375A$$
$$11.6575 = 4.460375B + 6.075945772A$$

On solving these equations B=2.845 A=a= - 0.17.

$$b = anti\log(2.845) = 699.842$$