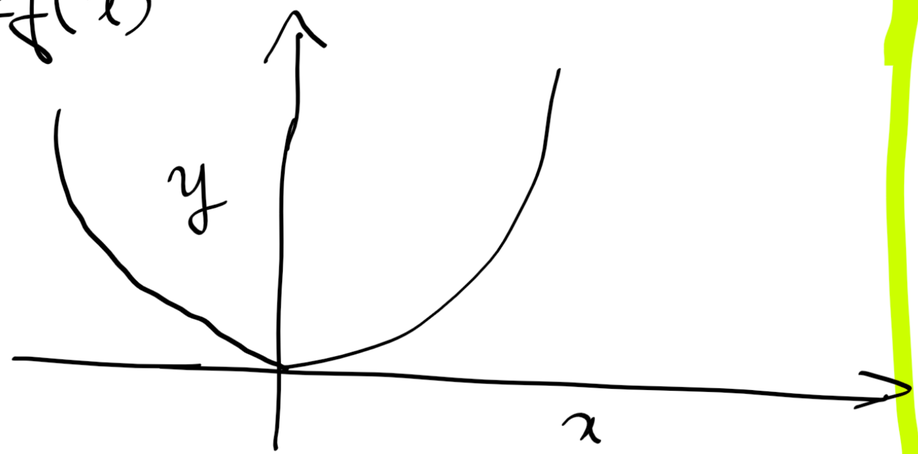
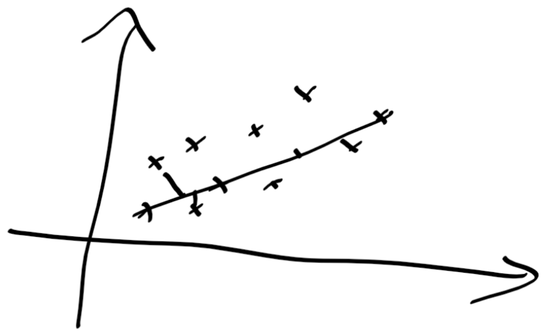
Curve Fitting

$y = x^2 \longrightarrow y = f(x)$

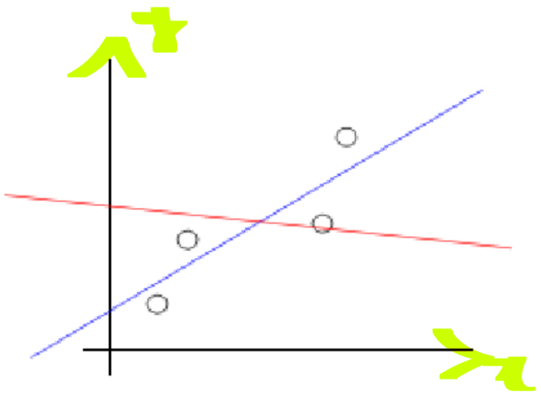
x	y
1	1
2	4
3	9
4	16

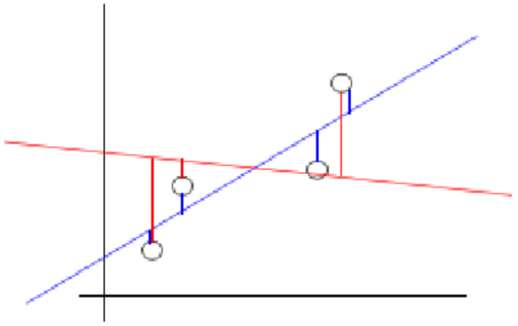


$$\left\{ \begin{array}{l} (x_i, y_i), \quad x_i = 1, 2, \ldots \\ y_i = f(x_i) \end{array} \right\} \quad y = f(x)$$



Curve Fitting

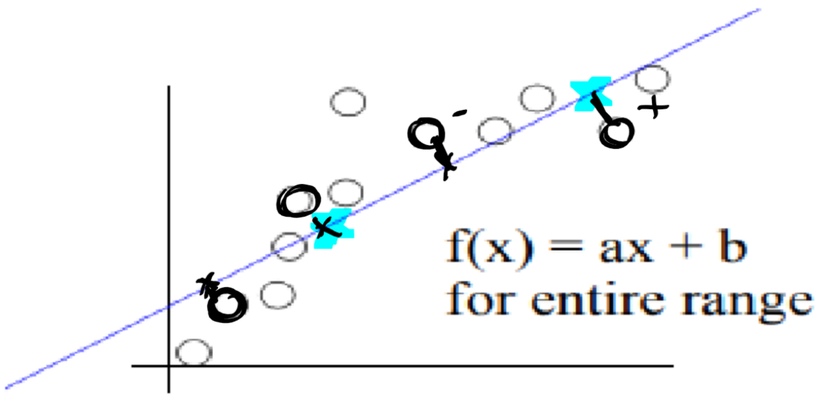




- Consider the distance between the data and points on the line
- Add up the length of all the red and blue verticle lines
- This is an expression of the 'error' between data and fitted line
- The one line that provides a <u>minimum</u> error is then the 'best' straight line

Method of curve fitting by the Principle of least Squares

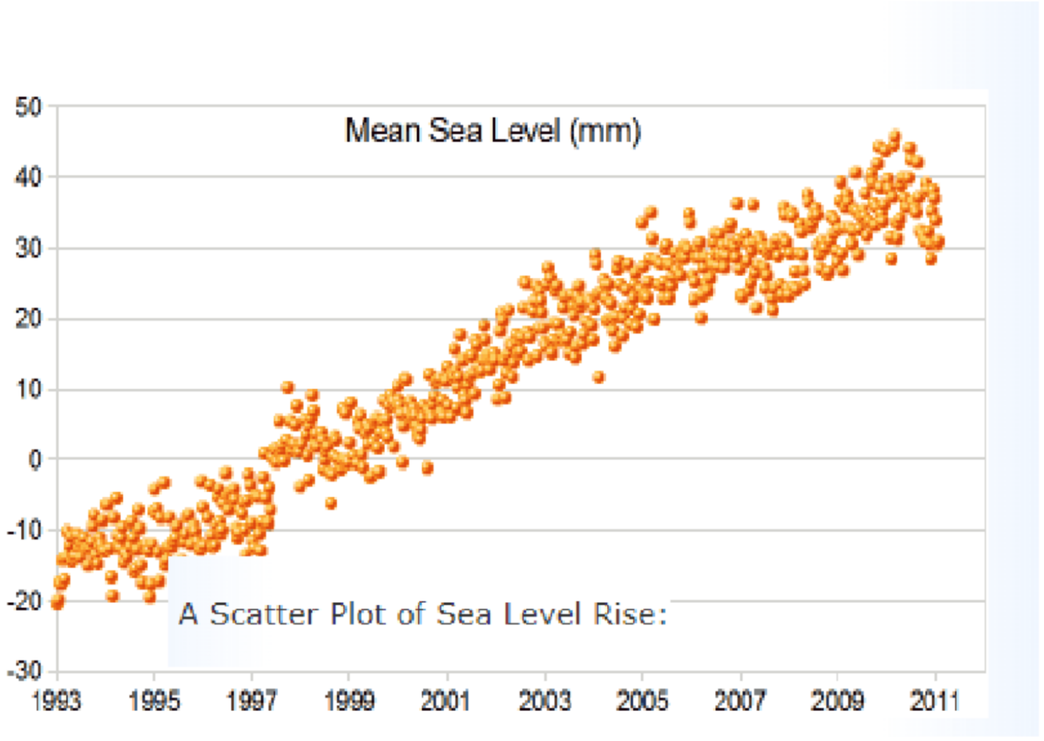
Curve fitting: Suppose that we are given ‘n” values x1, x2, x3, ... xn of an independent variable x and the corresponding values y1, y2, y3, ... yn of a variable y depending on x. Then the pairs of values (x1, y1), (x2, y2),..... (xn, yn) give n points in the xy plane. Generally it is not possible to find the actual curve f(x) which passes through these points. But we can find an approximate curve to this curve which passes through maximum no. of points. Such a curve is called “the curve of best fit”. The process of determining a curve of best fit is called curve fitting The method generally applied for curve fitting is known as the principle of least squares.

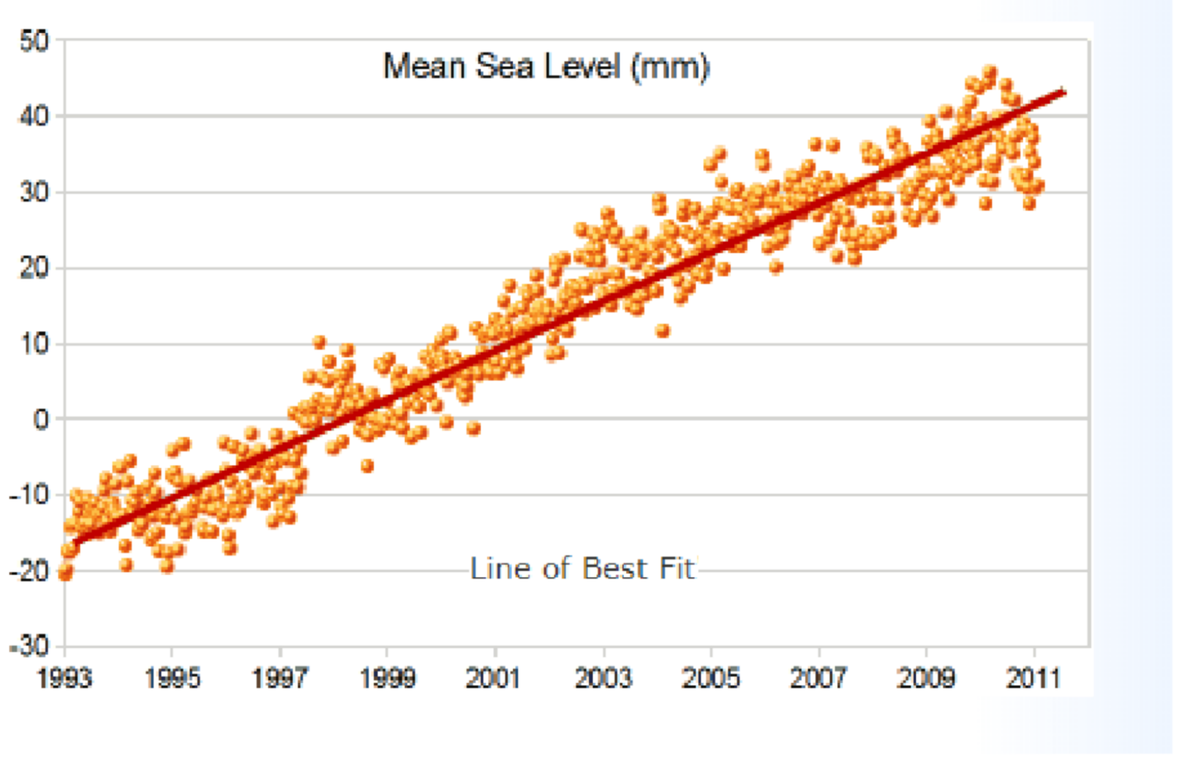


Curve Fitting

Scatter diagram:

Suppose we are given 'n' pairs of values (x_1, y_1), (x_2, y_2),…… (x_n, y_n) of two variable x and y then these n points may be plotted as dots in the xy-plane. The diagram of data so obtained is known as scatter diagram or the dot diagram.





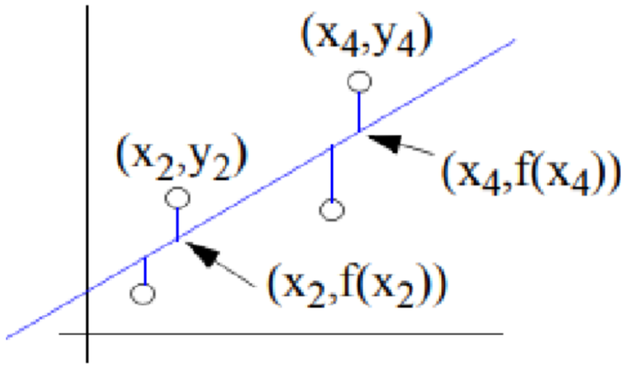
Principle of least squares:

Principle of least squares provides us an analytical or mathematical device to obtain an objective fit to the trend of the given time series. Most of the data relating to economic and business time series conform to definite laws of growth or decay and accordingly in such a situation analytical trend fitting will be more reliable for forecasting and predictions. This technique can be used to fit linear as well as non-linear trends.

Suppose we are given 'n' pairs of values $(x_1, y_1), (x_2, y_2), \ldots\ldots (x_n, y_n)$ of two variable x and y. Let $y = f(x)$ be the curve to be fitted to this data. Then difference between the observed value y_i and the estimated value $f(x_i)$ is called the error t $x=x_i$ and is denoted by e_i i.e $e_i = y_i - f(x_i)$ $\quad e_i = y_i - f(x_i)$

The curve of best fit is that for which the e_i's are as small as possible and this is possible if the sum of their squares i.e

$e_1^2 + e_2^2 + - - - - - + e_n^2$ Is minimum. This principle is known as principle of least squares.



Fitting a Straight line

Fitting of straight line by the method of principle of least square

Suppose that the given pairs of values are $(x_1, y_1), (x_2, y_2), \ldots\ldots (x_n, y_n)$.

Let the equation of the line of best fit be $\boxed{y = a + b\,x}$ ---------- (1)

where a and b are arbitrary constants to be determined. The differences between observed values y_i and estimated values $a+bx_i$ are called errors.

The sum of the squares of these errors is

$$E = \sum_{i=1}^{n} [y_i - (a + bx_i)]^2 \qquad (2)$$

According to principle of least squares we have to determine the constants a and b such that E is minimum.

The necessary conditions for minimum of E are $\frac{\partial E}{\partial a} = 0$ and $\frac{\partial E}{\partial b} = 0$

$$\sum_{i=1}^{n} 2[y_i - (a + bx_i)](-1) = 0$$

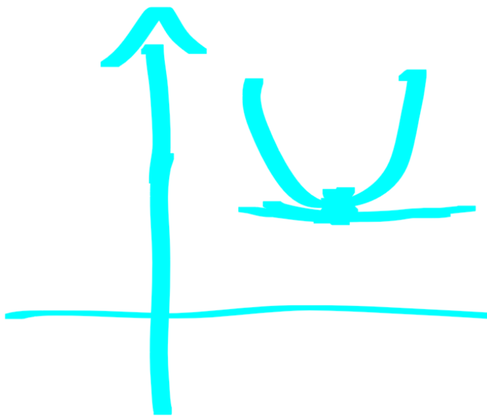
$$\sum [y_i - (a + bx_i)]^2 = \min \qquad (3)$$

minimum

$$\sum_{i=1}^{n} 2[y_i - (a + bx_i)](-x_i) = 0 \qquad (4)$$

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} (a + bx_i) = na + b\sum_{i=1}^{n} x_i$$

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



$$\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i$$

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

Hence the normal equations are

$$\sum_{i=1}^{n} y = na + b\sum_{i=1}^{n} x \tag{5}$$

$$\sum_{i=1}^{n} xy = a\sum_{i=1}^{n} x + b\sum_{i=1}^{n} x^2 \tag{6}$$

Solving these two equations we get the values of a and b and substituting these values in equation (1) give the line of best fit.

Problems

1. Fit a straight line from the following data and estimate the value of y when x = 3.5

x	1	2	3	4	6	8
y	2.4	3	3.6	4	5	6

Solution:

Let the line of best fit be y = a + b x

$$\sum y = na + b\sum x$$

$$\sum xy = a\sum x + b\sum x^2$$

24 = 6a + 24b

113.2 = 24a + 130b

a=1.9768 , b = 0.506

$$\boxed{y = 1.9768 + 0.506x}$$

y(3.5)=3.747

	x_i	y_i	x_i^2	$x_i y_i$
n				
	$\sum x_i =$	$\sum y_i =$	$\sum x_i^2 =$	$\sum x_i y_i =$

Normal equations

$$\sum y_i = na + b\sum x_i$$

$$\sum x_i y_i = a\sum x_i + b\sum x_i^2$$

2. Fit a straight line from the following data

x	0	0.5	1	1.5	2	2.5
y	0	1.5	3	4.5	6	7

Solution:

Let the line of best fit be y = a + b x

$$\sum y = na + b\sum x$$

$$\sum xy = a\sum x + b\sum x^2$$

22.5 = 6 a + 7.5b

41.25 = 7.5a + 13.75b

a = 0 , b = 3

y = 3 x

This fits the data exactly. That is, the error is zero.

3. Fit a straight line from the following data

x	0	0.5	1	1.5	2	2.5
y	-0.4326	-0.1656	3.1253	4.7877	4.8535	8.6909

Solution:

Let the line of best fit be y = a + b x

$$\sum y = na + b\sum x$$

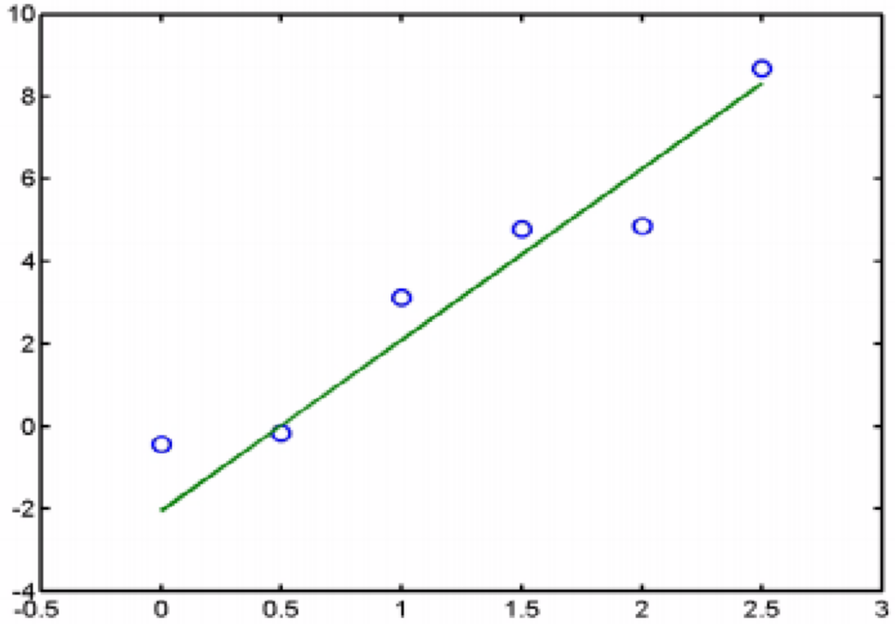
$$\sum xy = a\sum x + b\sum x^2$$

20.8593 = 6 a + 7.5b

41.6584 = 7.5a + 13.75b

a = -0.975 , b = 3.561

y = -0.975+3.561x



Exercise

4. Fit a straight line using the principle of least squares using the following data

x	2	4	6	8	10	12
y	7.32	8.24	9.2	10.19	11.01	12.05

5. Fit a straight line using the principle of least squares using the following data

x	6	7	4	3
y	8	10	4	2

HW

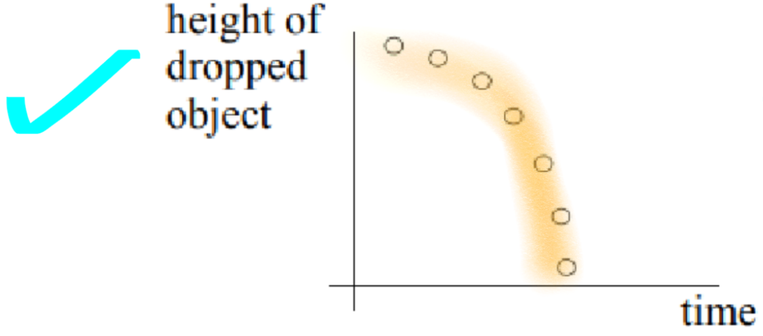
Fitting a Parabola

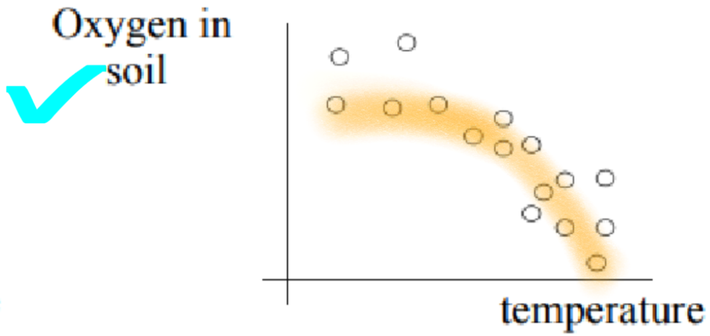
$y = a + bx$

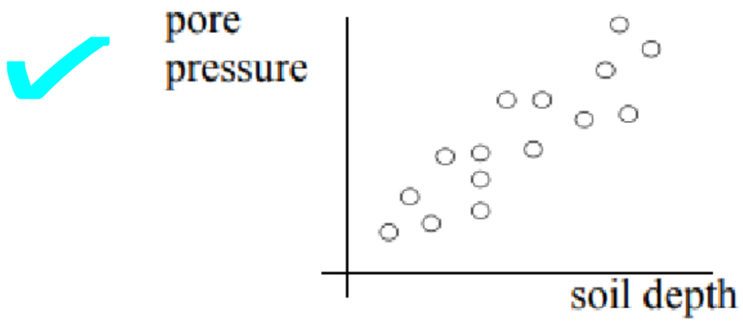
So...what do we do when a straight line is not suitable for the data set?

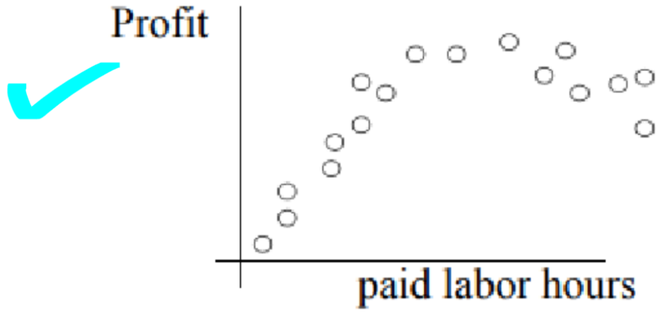
Straight line will not predict diminishing returns that data shows

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

Fitting of a parabola :

Suppose that the given pairs of values are $(x_1, y_1), (x_2, y_2), \ldots\ldots (x_n, y_n)$. Let the equation of the line of best fit be

$$\boxed{y = a + b\,x + c\,x^2} \qquad (1)$$

where a, b and c are arbitrary constants to be determined. The differences between observed values y_i and estimated values $a+bx_i + cx_i^2$ called errors.

The sum of the squares of these errors is

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

According to principle of least squares we have to determine the constants a, b and c such that E is minimum.

The necessary conditions for minimum of E are $\boxed{\frac{\partial E}{\partial a} = 0}$, $\boxed{\frac{\partial E}{\partial b} = 0}$ and $\boxed{\frac{\partial E}{\partial c} = 0}$

$$\sum_{i=1}^{n} 2\left[y_i - (a + bx_i + cx_i^2) \right](-1) = 0 \qquad (3)$$

$$y_i = a + bx_i + cx_i^2$$

$$\sum_{i=1}^{n} 2\left[y_i - (a + bx_i + cx_i^2) \right](-x_i) = 0 \qquad (4)$$

$$E = \sum \left[y_i - (a + bx_i + cx_i^2) \right]^2$$

$$\sum_{i=1}^{n} 2\left[y_i - (a + bx_i + cx_i^2) \right](-x_i^2) = 0 \qquad (5)$$

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} (a + bx_i + cx_i^2) = na + b\sum_{i=1}^{n} x_i + c\sum_{i=1}^{n} x_i^2$$

$$\sum_{i=1}^{n} x_i y_i = \sum_{i=1}^{n} x_i (a + bx_i + cx_i^2) = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 + c\sum_{i=1}^{n} x_i^3$$

$$\sum_{i=1}^{n} x_i^2 y_i = \sum_{i=1}^{n} x_i^2 (a + bx_i + cx_i^2) = a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i^3 + c\sum_{i=1}^{n} x_i^4$$

Hence the normal equations are

$$\sum y = na + b\sum x + c\sum x^2 \tag{6}$$

$$\sum xy = a\sum x + b\sum x^2 + c\sum x^3 \tag{7}$$

$$\sum x^2 y = a\sum x^2 + b\sum x^3 + c\sum x^4 \tag{8}$$

Solving these two equations we get the values of a , b and c substituting these values in equation (1) give the ~~line~~ of best fit

Curve of best fit

Problems

1. Fit a Parabola from the following data and estimate the value of y when x = 3.5

x	0	0.5	1	1.5	2	2.5
y	0	0.25	1	2.25	4.0	6.25

Solution:

$$\sum y = na + b\sum x + c\sum x^2$$

$$\sum xy = a\sum x + b\sum x^2 + c\sum x^3$$

$$\sum x^2 y = a\sum x^2 + b\sum x^3 + c\sum x^4$$

$$13.75 = 6a + 7.5\,b + c\,13.75$$

$$28.125 = 7.5\,a + 13.75\,b + c\,28.125$$

$$61.1875 = 13.75\,a + 28.125\,b + c\,61.1875$$

a= 0, b= 0, c= 1

$$y = 0 + 0\,x + 1\,x^2$$

Y(3.5)= ?

The normal equations

$$\sum y_i = na + b\sum x_i + c\sum x_i^2$$

$$\sum x_i y_i = a\sum x_i + b\sum x_i^2 + c\sum x_i^3$$

$$\sum x_i^2 y_i = a\sum x_i^2 + b\sum x_i^3 + c\sum x_i^4$$

x_i	y_i	x_i^2	x_i^3	x_i^4	$x_i y_i$	$x_i^2 y_i$
$\sum x_i =$	$\sum y_i =$	$\sum x_i^2$	$\sum x_i^3$	$\sum x_i^4$	$\sum x_i y_i$	$\sum x_i^2 y_i$

Problems

2. Fit a Parabola from the following data and estimate the value of y when x = 3.5

x	0	1	2	3	4
y	1	5	10	22	38

Solution:

$$\sum y = na + b\sum x + c\sum x^2$$

$$\sum xy = a\sum x + b\sum x^2 + c\sum x^3$$

$$\sum x^2 y = a\sum x^2 + b\sum x^3 + c\sum x^4$$

a= 1.41, b= 9.1, c= 2.21

Answer : $\boxed{y = 1.41 + 0.26\,x + 2.21\,x^2}$

Problems

3. Fit a Parabola from the following data and estimate the value of y when x = 3.5

x	0	0.5	1	1.5	2	2.5
y	0.0674	-0.9156	1.6253	3.0377	3.3535	7.9409

Solution:

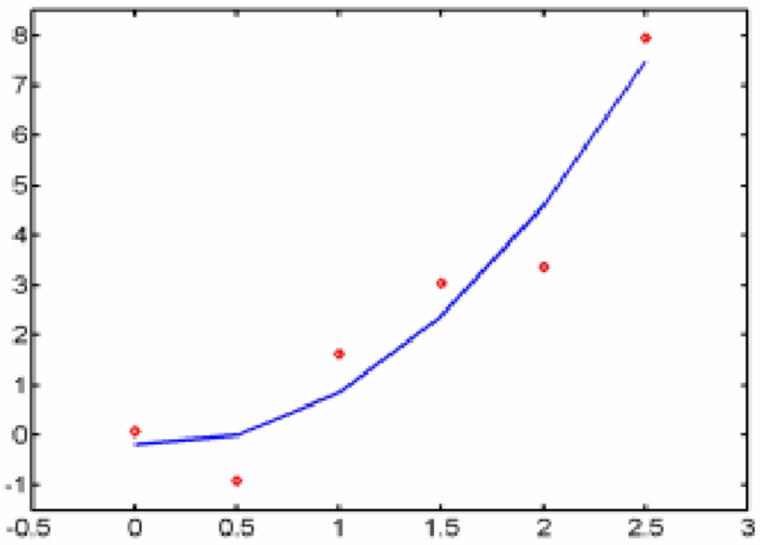
$$\sum y = na + b\sum x + c\sum x^2 \qquad 15.1093 = 6a + 7.5\,b + 13.75\,c$$

$$\sum xy = a\sum x + b\sum x^2 + c\sum x^3 \qquad 32.2834 = 7.5\,a + 13.75\,b + 28.125c$$

$$\sum x^2 y = a\sum x^2 + b\sum x^3 + c\sum x^4 \qquad 71.276 = 13.75\,a + 28.125\,b + 71.276\,c$$

a = - 0.812, b= - 0.3221, c = 1.3537

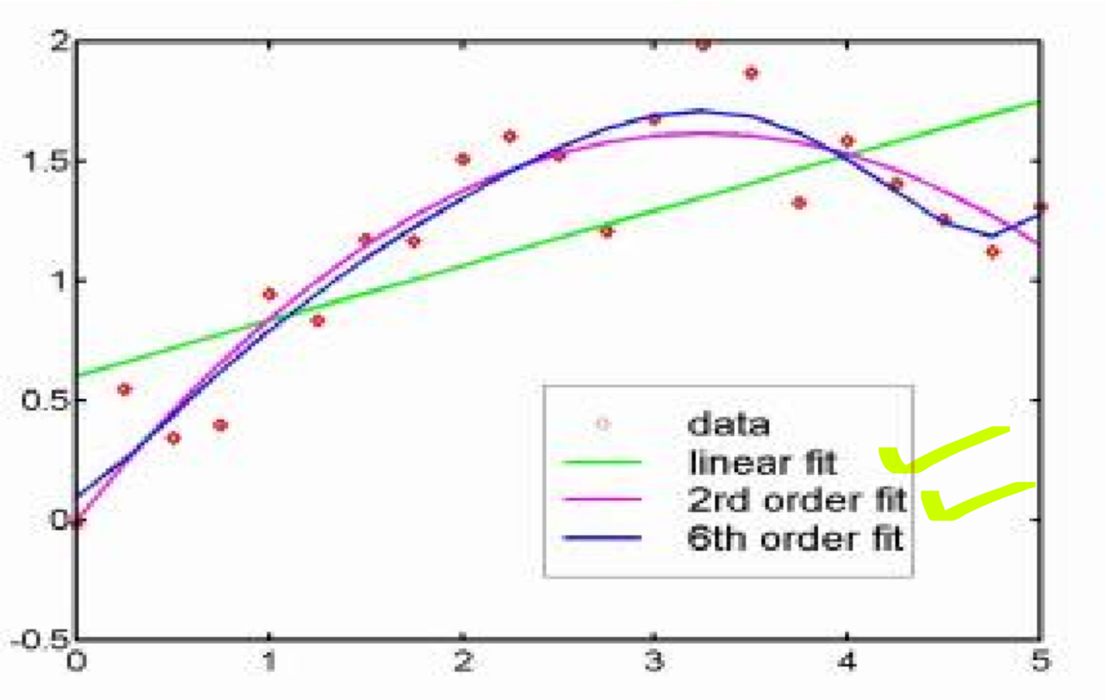
$$\boxed{y = -0.1812 - 0.3221\,x + 1.3537\,x^2}$$



Example :

For a certain data points here we have drawn three different fits In this example, we're not sure which order will fit well, so we try three different polynomial orders

* Linear regression, or first order curve fitting is just the general polynomial form we just saw,
* 2nd and 6th order look similar, but 6th has a 'squiggle to it. We may not want that...



$$y = a + bx + cx^2 + dx^3 + ex^4 + fx^5 + gx^6$$

Fitting of $y = C\,e^{Ax}$ and $y = ab^{x}$

$y = C\,e^{Ax}$ $\qquad$ $y = ab^{x}$

Suppose that the given pairs of values are $(x_1, y_1), (x_2, y_2), \ldots\ldots (x_n, y_n)$. Let the equation of the line of best fit be $\boxed{y = Ce^{Ax}}$

1) Take logarithm of both sides to get rid of the exponential $\log y = \log C + Ax$

2) Introduce the following change of variables $\boxed{\log y = Y}$, $\boxed{B = \log C}$, $\boxed{x = X}$

Now we have: Y=AX+B

which is a LINEAR equation. The original data points in the x- y plane get mapped into the X-Y plane. This is called data linearization.

The data is transformed as : (x,y) ⇒ (X,Y)=(x, logy)

Now we use the method for solving a first order linear curve fit for A and B. Where $\log y = Y$, $x = X$

Finally, we operate on $B = \log C$ to solve $C = e^{B}$ and we now have the coefficients for $y = C\,e^{Ax}$

$A =$

$B = \log C \Rightarrow C = e^{B}$

$y = Ce^{Ax}$

Problems

1. Fit a $y = ab^x$ from the following data by the method of least square

x	1	2	3	4	5	6	7	8
y	1.0	1.2	1.8	2.5	3.6	4.7	6.6	9.1

x_i	y_i	$Y_i = \log y_i$	$X = x$	$X_i Y_i$
		$\sum Y_i =$	$\sum X_i$	$\sum X_i Y_i$

$X_i^2 = x_i^2$

$\sum X_i^2 =$

Solution:

Let the line of best fit be $y = ab^x$

$\log y = \log a + x \log b$

$Y = A + Bx$

$Y = \log y$

$A = \log a$

$B = \log b$

$X = x$

$Y = A + B\,x$

$$\sum Y = nA + B\sum x$$

$$\sum xY = A\sum x + B\sum x^2$$

$8.608 = 8A + 36B$

$52.349 = 36A + 204B$

$$\sum Y_i = nA + B\sum X_i$$

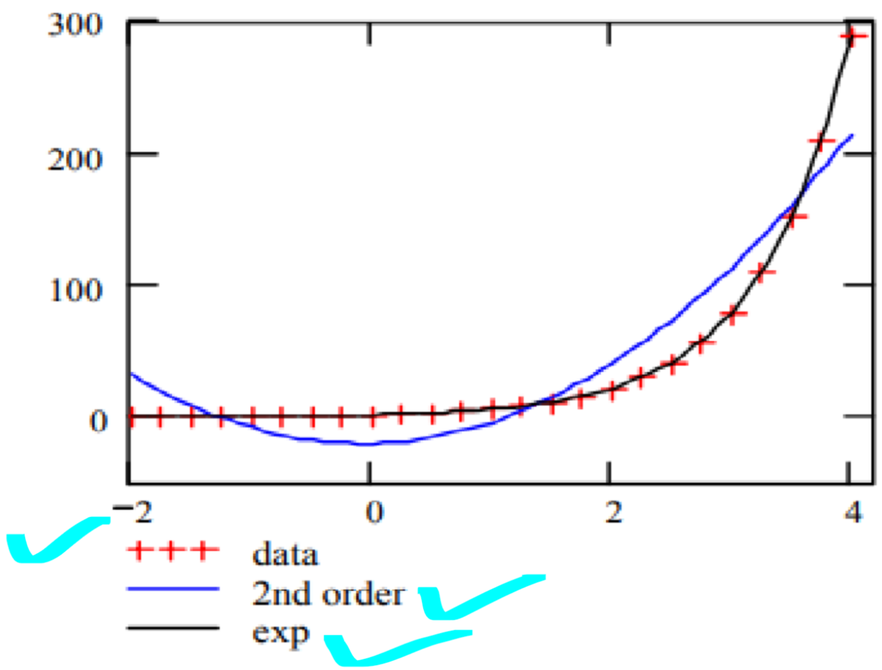
$$\sum X_i Y_i = A\sum X_i + B\sum X_i^2$$

$A = -0.382$, $B = 0.324$

$$y = 0.6825\,(1.3826)^x$$

A, B ✓

$a = e^A$, $b = e^B$



In this graph u can see for certain random datas second degree polynomial curve fit shown with blue color curve where as best fit curve is exponential curve shown with black color curve

Laws Reducible to the Linear Law

We give below some of the laws in common use, indicating the way these can be reduced to the linear form by suitable subsitutions.

Law	Transformation	Converted Law
$y = mx^n + c$	$x^n = X$ and $y = Y$	$Y = mX + c$
$y = ax^n$	$\log_{10} x = X$ and $\log_{10} y = Y$	$Y = nX + c,\ c = \log_{10} a$
$y = ax^n + b \log x$	$\frac{x^n}{\log x} = X$ and $\frac{y}{\log x} = Y$	$Y = aX + b$
$y = ae^x$	$x = X$ and $\log_{10} y = Y$	$Y = mX + c,$ $m = \log_{10} e, \log_{10} a$
$xy = ax + by$	$\frac{y}{x} = X$ and $y = Y$	$Y = bX + a$

Exercise

Convert the following equations into linear form

(a) $y = x/a+bx$ (b) $y = ax+b/x$ (c) $x a^y = b$ (d) $y = b/x(x-a)$ (e) $xy = ax + b$ (f) $y = ax + bx^2$ (g) $y = ax + bxy$.

$\downarrow$

$xy = ax + b$

$y = a + \frac{b}{x}$

$\boxed{Y = A + BX}$

$A = a$

$Y = y$

$X = \frac{1}{x}$

x_i	y_i	$\frac{1}{x_i} = X_i$

ΣX_i

Correlation

A statistic representing how closely two variables co-vary; it can vary from -1 (perfect negative correlation) through 0 (no correlation) to +1 (perfect positive correlation).

- The correlation coefficient, denoted by r, is a measure of the strength of the straight-line or linear relationship between two variables. The correlation coefficient takes on values ranging between +1 and -1.
- The quantity r, called the linear correlation coefficient, measures the strength and the direction of a linear relationship between two variables

Positive Correlation

When the values of two variables move same direction, correlation is said to be positive

ie; an increase in the value of one variable results into an increase in the other variable also or if decrease in the value of one variable results into a decrease in the other variable also correlation is said to be positive.

Eg. Temperature and volume

$y = f(x)$

x	y

$x \uparrow \quad y \uparrow$

$x \downarrow \quad y \downarrow$

Negative correlation

When the values of two variables move opposite direction, correlation is said to be negative.

ie; an increase in the value of one variable results into an decrease in the other variable also or if decrease in the value of one variable results into a increase in the other variable also correlation is said to be positive.

Eg. Pressure and volume

$x \uparrow \; y \downarrow$

$x \downarrow \; y \uparrow$

Linear Correlation

When the amount of change in one variable leads to a constant ratio of change in the other variable, correlation is said to be linear.

$y = ax$

Non linear Correlation

Correlation is said to be non linear (curve linear) when the amount of change in one variable is not in constant ratio to the change in the other variable.

Type of correlation coefficient

1. Perfect ***Positive correlation***
2. Perfect ***negative correlation***
3. ***Moderately*** Positive correlation
4. ***Moderate*** negative correlation
5. Absolute ***no correlation***

Perfect ***Positive correlation***

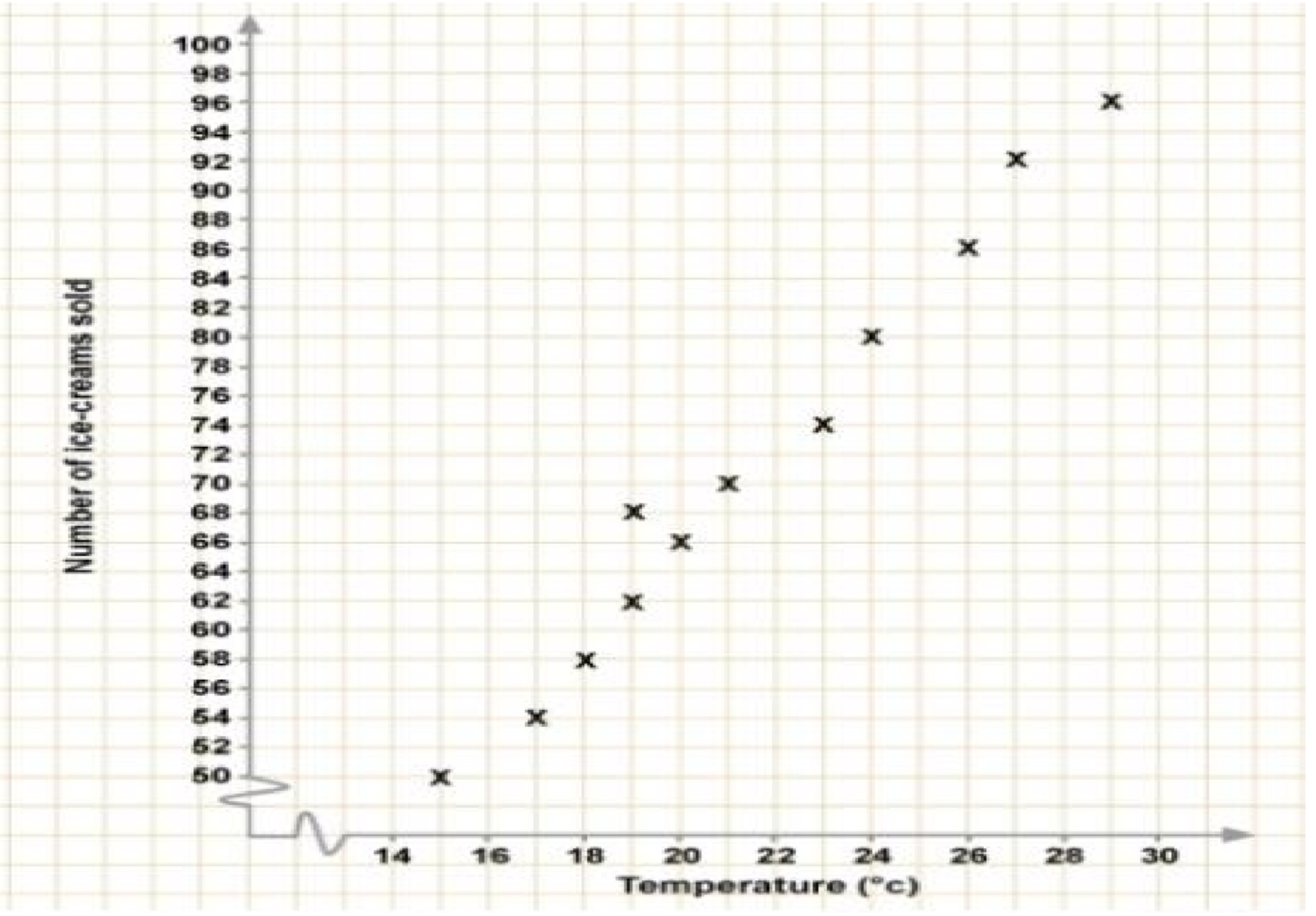
- If x and y have a strong positive linear correlation, r is close ***to +1***. An r value of exactly +1 indicates a perfect positive fit. Positive values indicate a relationship between x and y variables such that as values for x increases, values for y also increase.

Coeff. of correlation

$-1 \leq r \leq 1$

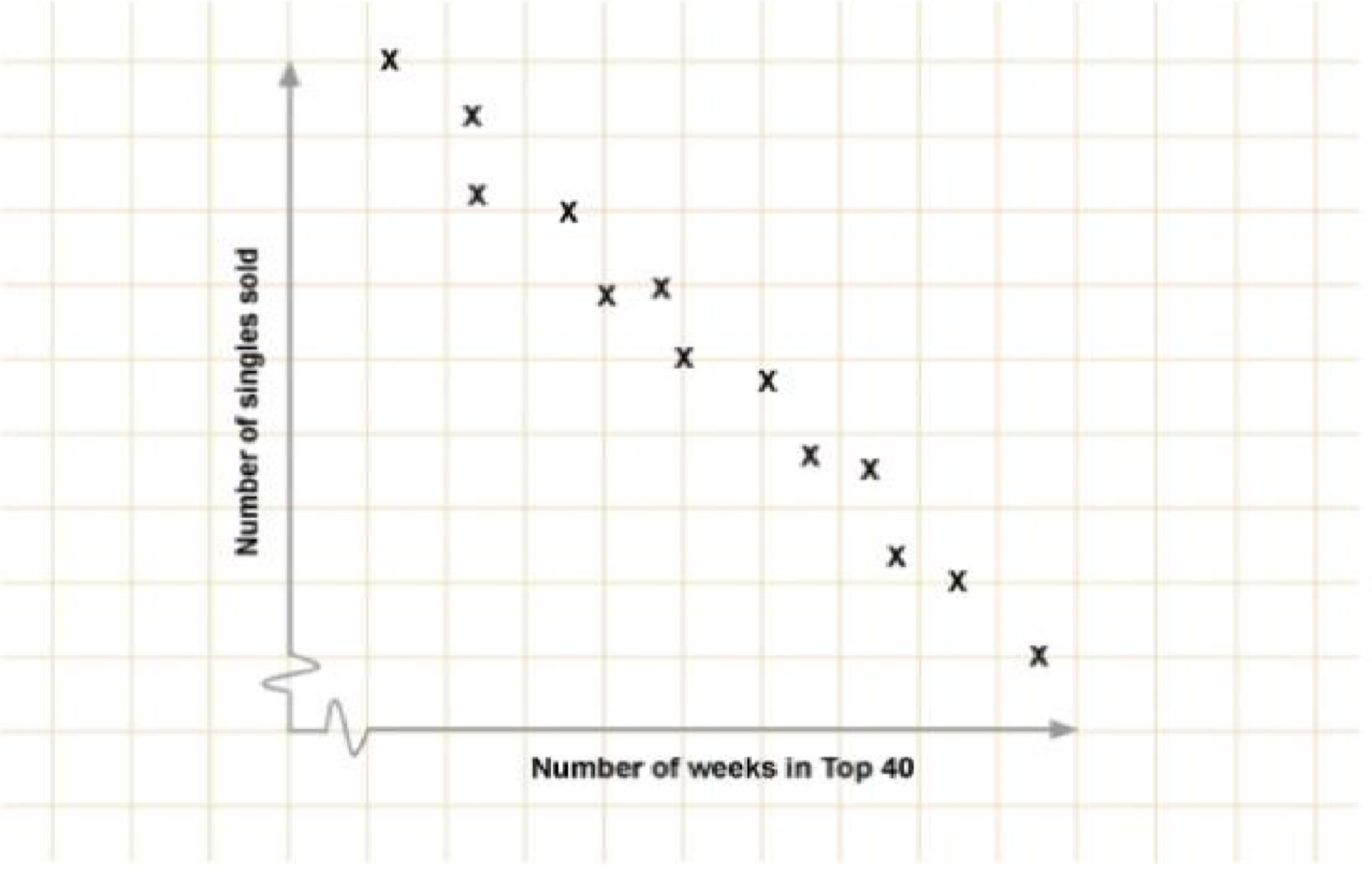
$-1 \leq \rho \leq 1$

some times ρ *is taken as* r


Number of ice-creams sold
100
98
96
94
92
90
88
86
84
82
80
78
76
74
72
70
68
66
64
62
60
58
56
54
52
50
14
16
18
20
22
24
26
28
30
Temperature (°c)

Perfect ***negative correlation***

- If x and y have a strong negative linear correlation, r is close ***to -1***. An r value of exactly -1 indicates a perfect negative fit. Negative values indicate a relationship between x and y such that as values for x increase, values for y decrease.


Number of singles sold
Number of weeks in Top 40

Moderately Positive correlation

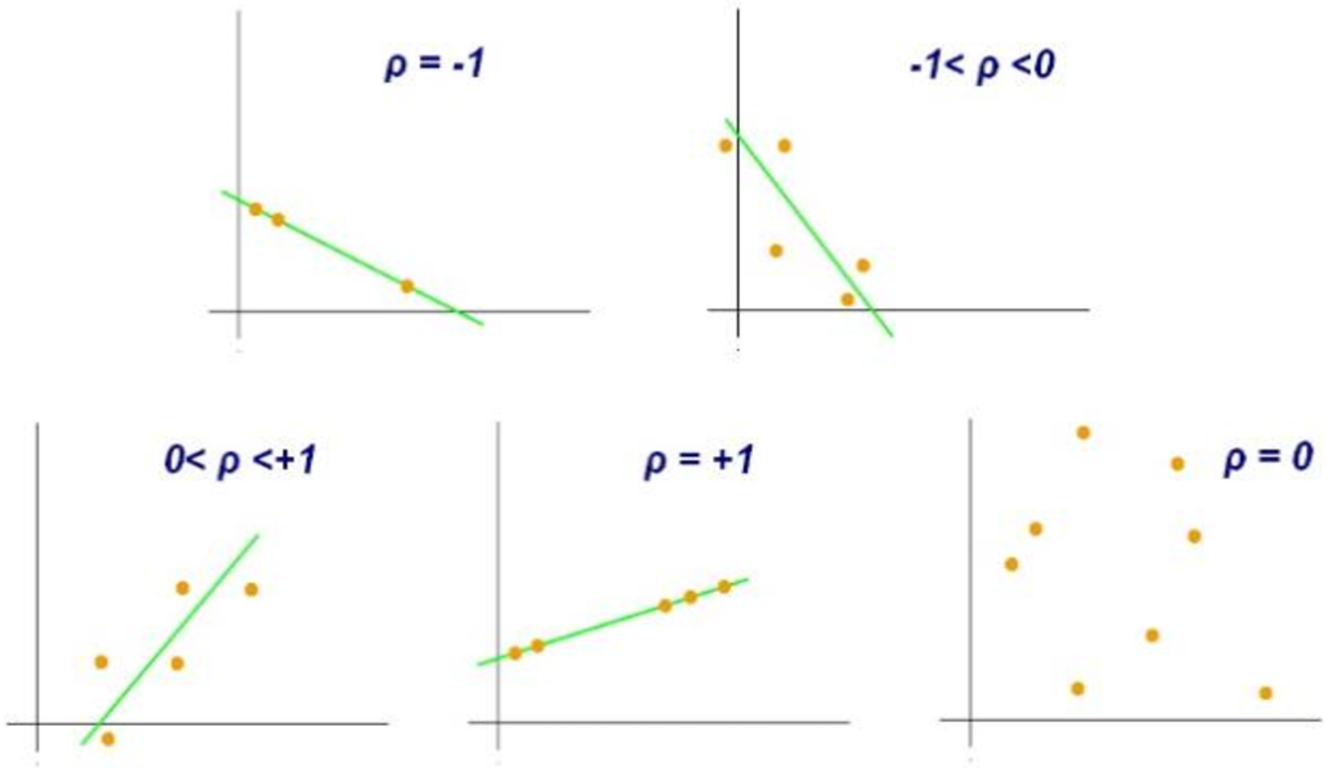
$-1 \le r \le 1$

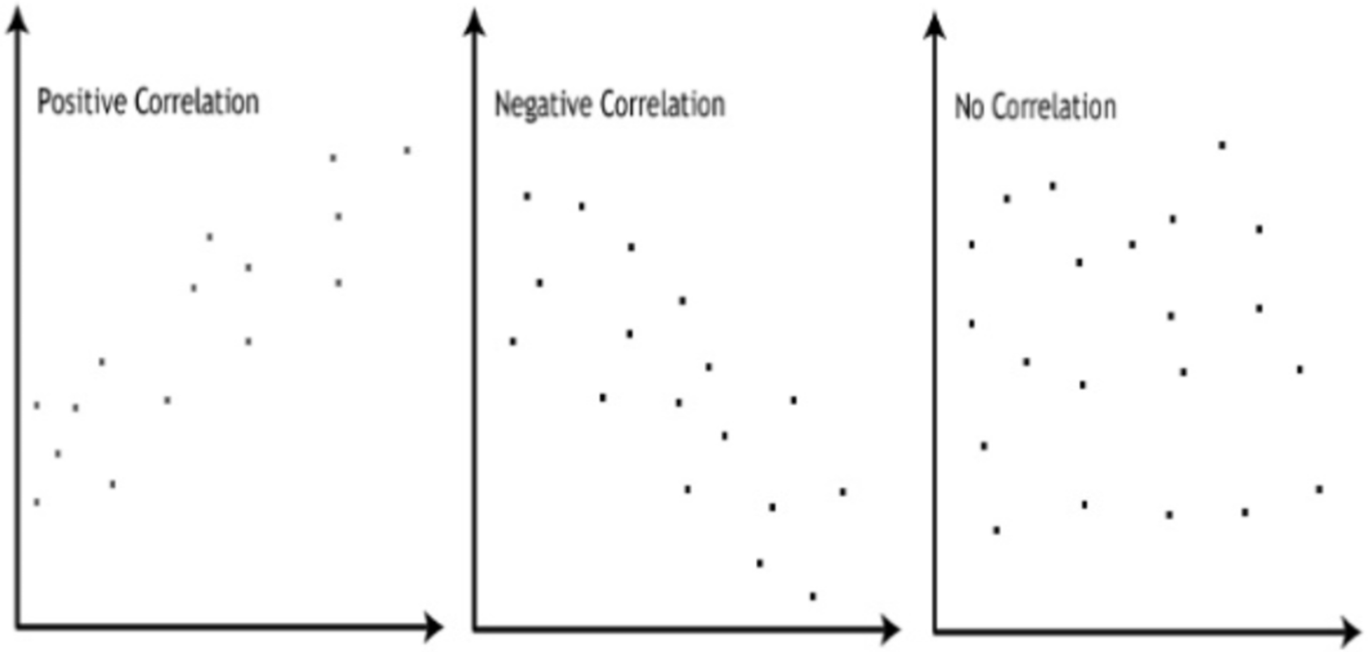
- $0<r<1$

Moderate negative correlation

~~$-1<r<1$~~

$-1 < r < 0$





Absolute *no correlation*

- If there is no linear correlation or a weak linear correlation, r is close ***to 0***. A value near zero means that there is a random, nonlinear relationship between the two variables

Coeff. of correlation = 0

Methods of computing the correlation

ρ or r

- karl pearson's correlation coefficient
- spearman's rank correlation coefficient

$$r = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

karl pearson's correlation coefficient

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

$$Cov(X,Y) = E\left[(X - E(X))(Y - E(Y))\right]$$

$$= \frac{1}{n}\sum(x_i - \bar{x})(y_i - \bar{y})$$

$$\sigma_X^2 = \frac{1}{n}\sum(x_i - \bar{x})^2$$

$$\sigma_Y^2 = \frac{1}{n}\sum(y_i - \bar{y})^2$$

Spearman's rank correlation coefficient

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)}$$

Regression

Regression means estimation or prediction of an unknown variable in terms of known variable. In regression analysis there are two types of variables. The variable whose value is to be predicted is called dependent variable or regressed variable. And the variable which is used to predict the value is called regression or independent variable.

There exists some association or relationship between two variables. The points on the scattered diagram will is called the curve of regression. If the curve is a straight line then we say that there is linear regression between the variables and the line is known as line of regression.

Linear Regression

Equations of line of Regression

Suppose that we have been given a set of values (x_1, y_1), (x_2, y_2),…… (x_n, y_n) of x and y. Let the equation the line of regression of y on x be

$$\boxed{y = a + b\,x} \qquad (1)$$

According to the principle of least squares the normal equations for estimating a and b are

$$\sum y = na + b\sum x \qquad (2)$$

$$\sum xy = a\sum x + b\sum x^2 \qquad (3)$$

Dividing eq.(2) by n we get

$$\frac{\sum y}{n} = a + b\frac{\sum x}{n}$$

$$\bar{y} = a + b\,\bar{x} \qquad (4)$$

Where

$\bar{y} = \frac{\sum y}{n}$, $\bar{x} = \frac{\sum x}{n}$ are means of y and x respectively

Equation (4) shows that the point $\bar{x}, \bar{y}$ lies on the line of regression of y on x.

Now shifting the origin to $\bar{x}, \bar{y}$ equation (3) becomes

$$\sum (x - \bar{x})(y - \bar{y}) = a\sum (x - \bar{x}) + b\sum (x - \bar{x})^2 \qquad (5)$$

Where $\sum (x - \bar{x}) = \sum x - \sum \bar{x} = 0$ hence equation (5) becomes

$$b = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}$$

$$b = \frac{\frac{1}{n}\sum (x - \bar{x})(y - \bar{y})}{\frac{1}{n}\sum (x - \bar{x})^2} \qquad (6)$$

Where $\frac{1}{n}\sum(x-\bar{x})(y-\bar{y}) = E\left((x-\bar{x})(y-\bar{y})\right) = \mathrm{cov}(x,y)$

$$\boxed{b = \frac{\mathrm{cov}(x,y)}{\mathrm{variance}(x)} = \frac{\rho\sigma_x\sigma_y}{\sigma_x^2}}$$

$$\boxed{b = \frac{\rho\sigma_y}{\sigma_x}} \rightarrow b_{yx} = b \qquad (7)$$

substitute equation (7) in (4)

$$a = \bar{y} - b\,\bar{x} = \bar{y} - \frac{\rho\sigma_y}{\sigma_x}\,\bar{x}$$

$$y - \bar{y} = \frac{\rho\sigma_y}{\sigma_y}(x - \bar{x})$$

Substituting the value of a and b in equation (1)

$$\boxed{y - \bar{y} = \frac{\rho\sigma_y}{\sigma_x}(x - \bar{x})} \qquad (8)$$

This equation of line of regression of y on x. the coefficient $\frac{\rho\sigma_y}{\sigma_x}$ is called the regression coefficient of y on x and is denoted by

b_{yx}

Similarly equation of line of regression of x on y is

$x = a + by$

$$\boxed{(x - \bar{x}) = \frac{\rho\sigma_x}{\sigma_y}(y - \bar{y})} \quad (9)$$

$b_{xy} = b$

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

The coefficient $\frac{\rho\sigma_x}{\sigma_y}$ is called the regression coefficient of x on y and is denoted by b_{xy}.

$\delta \quad r$

Note: The correlation coefficient ρ is the geometric mean between the two regression coefficient. That is $\boxed{\rho = \sqrt{b_{xy} b_{yx}}}$ and the sign of coefficient correlation is same as that of the regression coefficient.

Problem

1. Calculate the coefficient of correlation for the following data & obtain the lines of regression

x	1	3	4	6	8
y	1	2	4	4	5

Solution

$$\bar{x} = \frac{\sum x}{n} = \frac{22}{5} = 4.4 \ , \quad \bar{y} = \frac{\sum y}{n} = \frac{16}{5} = 3.2$$

$$\sum x^2 = 126, \quad \sum y^2 = 126, \quad \sum xy = 87$$

$$\sigma_x^2 = \frac{\sum x^2}{n} - \bar{x}^2 = \frac{126}{5} - (4.4)^2 = 5.84$$

$$\sigma_y^2 = \frac{\sum y^2}{n} - \bar{y}^2 = \frac{62}{5} - (3.2)^2 = 2.16$$

$$\text{cov}(x, y) = \frac{\sum xy}{n} - \bar{x}\ \bar{y} = \frac{87}{5} - (4.4)(3.2) = 3.32$$

$$\rho = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y} = \frac{3.32}{2.46)(1.469)} = 0.935$$

Line of regression of y on x

$$\boxed{(y-\bar{y}) = \frac{\rho\sigma_y}{\sigma_x}(x-\bar{x})} \rightarrow y = \bar{y} + \rho\frac{\sigma_y}{\sigma_x}(x-\bar{x})$$

$$(y-3.2) = 0.935.\frac{1.469}{2.416}(x-4.4)$$

$$y = 0.5685\,x + 0.6986$$

Line of regression of x on y

$$(x-\bar{x}) = \frac{\rho\sigma_x}{\sigma_y}(y-\bar{y})$$

$$(x-4.4) = 0.935.\frac{2.416}{1.469}(y-3.2)$$

$$x = 1.536\,y - 0.517$$

2. Following table gives the data on rainfall and discharge in a river. Obtain the line of regression of y on x and correlation coefficient.

Rain fall x (inches)	1.53	1.78	2.60	2.95	3.42
Discharge y (1000cc)	33.5	36.3	40.0	45.8	53.5

Solution:

$$\bar{x} = \frac{\sum x}{n} = 2.456 \ , \quad \bar{y} = \frac{\sum y}{n} = 41.82$$

$$\sigma_x^2 = 0{,}7083$$

$$\sigma_y^2 = 7.145 \qquad \text{cov}(x, y) = 4.8796$$

$$\rho = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y} = \frac{4.8796}{(0.7083)(7.145)} = 0.9642$$

Line of regression of y on x

$$(y - \bar{y}) = \frac{\rho \sigma_y}{\sigma_x}(x - \bar{x})$$

$$(y - 41.82) = 0.9642 \cdot \frac{7.145}{0.7083}(x - 2.456)$$

$$\boxed{y = 9.7264\,x + 17.932}$$

Curvilinear regression

Fit the regression line

$$\underbrace{Y = a + b_1 X + b_2 X^2}$$

Method of least squares $\Downarrow$ a, b_1, b_2

Let (X, Y) — two-dimensional r.v

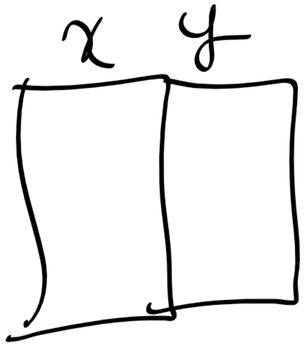
$E(X) = \bar{X}$, $E(Y) = \bar{Y}$

$V(X) = \sigma_X^2$, $V(Y) = \sigma_Y^2$

coeff. of correlation $= \rho$

The regression of Y on X

$$E(Y|X) = \bar{Y} + \rho \frac{\sigma_Y}{\sigma_X}(X - \bar{X})$$



nly

the regression of X on Y

$$E(X|Y) = \bar{X} + \rho \frac{\sigma_X}{\sigma_Y}(Y - \bar{Y})$$

Problem: Let (X,Y) have the joint pdf given by

$$f(x,y) = \begin{cases} 1 & \text{if } |y| < x, \quad 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

Show that the regression of Y on X is linear but regression of X on Y is nonlinear

$E(Y|X)$ & $E(X|Y)$

$|y| < x$
$\Rightarrow -x < y < x$

$$g(x) = \int_{y=-x}^{x} f(x,y)\,dy = \int_{y=-x}^{x} 1\,dy = 2x \; ; \; 0 < x < 1$$

$$h(y) = \int_{x=|y|}^{1} f(x,y)\,dx = \int_{x=|y|}^{1} 1 \cdot dx = 1 - |y| \; ; \; -1 < y < 1$$

$$g(x|y) = \frac{f(x,y)}{h(y)} = \frac{1}{1-|y|} \; ; \; -1 \le y \le 1 \; , \; 0 < x < 1$$

$$= \begin{cases} \dfrac{1}{1-y} \; ; \; 0 < y < 1 \; , \; 0 < x < 1 \\[2ex] \dfrac{1}{1+y} \; ; \; -1 < y < 0 \; ; \; 0 < x < 1 \end{cases}$$

$$h(y|x) = \frac{f(x,y)}{g(x)} = \frac{1}{2x} \; ; \; 0 < x < 1 \; ; \; |y| < x$$

$$E(Y|X=x) = \int_{y=-x}^{x} y\, h(y|x)\, dy = \int_{y=-x}^{x} \frac{y}{2x}\, dy = \frac{1}{4x} y^2 \Big|_{-x}^{x}$$

$$= 0$$

$\therefore \boxed{y=0} \Rightarrow$ The curve of regression of Y on X is $y=0$] which is a st. line

$$E(X|Y=y) = \int_{x=0}^{1} x\, g(x|y)\, dx = \int_{x=0}^{1} x \times \frac{1}{1-y}\, dx$$

$$= \frac{1}{2(1-y)}, \quad 0<y<1$$

$$E(X|Y=y) = \int_{x=0}^{1} x\, g(x|y)\, dx = \int_{x=0}^{1} x \left(\frac{1}{1+y}\right) dx$$

$$= \frac{1}{2(1+y)}, \quad -1<y<0$$

$\therefore$ The curve of regression of X on Y is

$$x = \begin{cases} \dfrac{1}{2(1-y)} ; & 0<y<1 \\ \dfrac{1}{2(1+y)} ; & -1<y<0 \end{cases}$$

which is not a st. line [non linear]