It is to explore and study of the **relationship** between **two variables** (x , y) with the objective of formulating an equation between the two variables and using that equation to predict one from the other.
(x is also called independent, explanatory, or predictor variable)
(y is also called dependent, response variable). So, a response variable is the variable whose value can be explained by the predictor variable.

Steps

1. To find the **nature of the relationship** (Linear or non-linear, positive, or negative relationship or no relationship) by doing a scatter diagram, y versus x

2. To measure the **strength of this relationship** by computing the correlation coefficient $= r$

3. Finding **slope** and **y-intercept** for equation of the best fitted- line (**regression equation =** $y = a\,x + b$) between x , y variables.

4. Using the regression equation to **estimate or predict** one variable from the other.

Nature of relationship:
Positive: Both variables either increasing or decreasing $x \uparrow\uparrow y$ **or** $x \downarrow\downarrow y$
Negative: When one variable increases the other one decreases or vice versa. $x \uparrow\downarrow y$ **or** $x \downarrow\uparrow y$

What do you think is the nature of relationship between x and y variables? ***Answers at the bottom***

	x **Independent, Explanatory, or Predictor** variable	y **Dependent, or response** variable	Nature of relationship Positive, Negative
1	Hours of study per week for stat class	Stat test score	+ , − , *None*
2	Mortgage rate	Number of loans refinanced	+ , − , *None*
3	Average height of the parents	Height of the sons or daughters	+ , − , *None*
4	No. of absences in a semester for stat class	Stat test scores	+ , − , *None*
5	Daily temperature in summer	Water or electric consumption	+ , − , *None*
6	$ amount spent on advertisement	Monthly sales	+ , − , *None*
7	Fat consumption	Cholesterol level	+ , − , *None*
8	Number of years of education	Monthly salary	+ , − , *None*
9	Number of hours watching TV/week	GPA	+ , − , *None*
10	Ice cream sales	Number of drownings	+ , − , *None*

1)+ 2)− 3)+ 4) − 5) + 6)+ 7)+ 8)+ 9)none 10) none (Lurking variable)

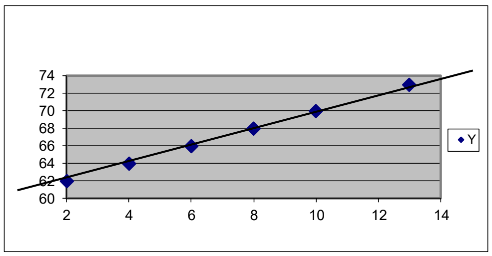
Steps to do a Correlation and Regression problem

1. Constructing a Scatter diagram and comment on its nature (linear or non-linear, positive or negative, strong or weak relationship).

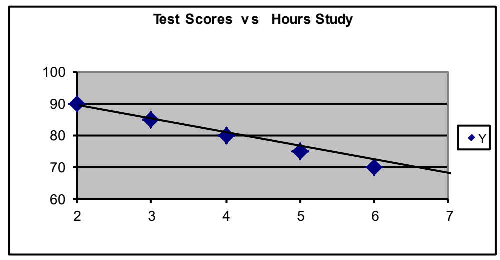
Why do we need a scatter diagram?

a) To see if data exhibit a **linear pattern** or not
b) To see if linear pattern is **positive or negative**
c) To see how closely (**strongly or perfectly**) data are **clustered around** the **a straight line.**
d) To detect any **outlier** (a point that is lying far away from the other data points).

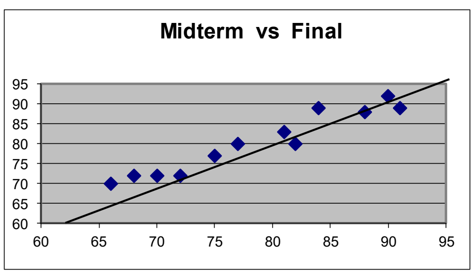
Different Possible shapes of a Scatter Diagram



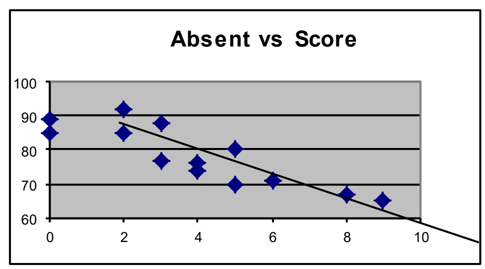
$r = 1$
Perfect Positive Linear Correlation



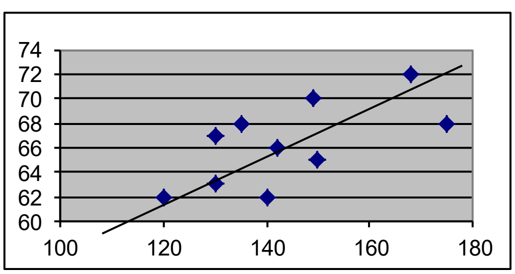
$r = -1$
Perfect Negative Linear Correlation



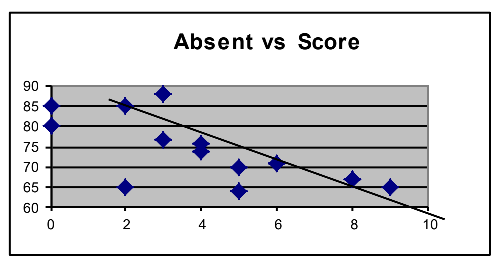
Strong Positive Linear Correlation



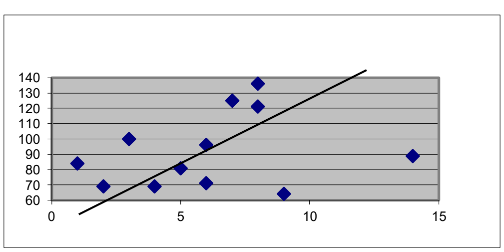
Strong Negative Linear Correlation



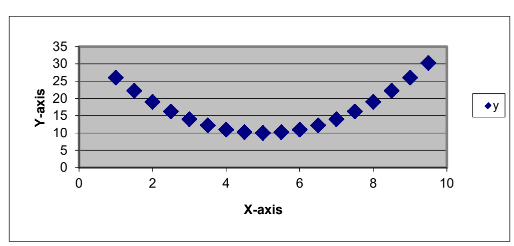
Positive Linear Correlation



Negative Linear Correlation



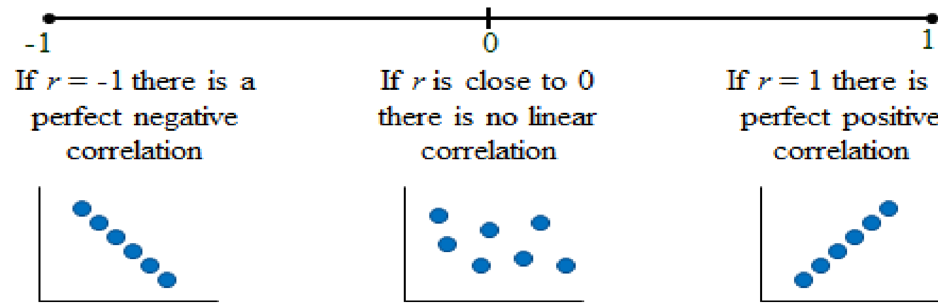
No Correlation



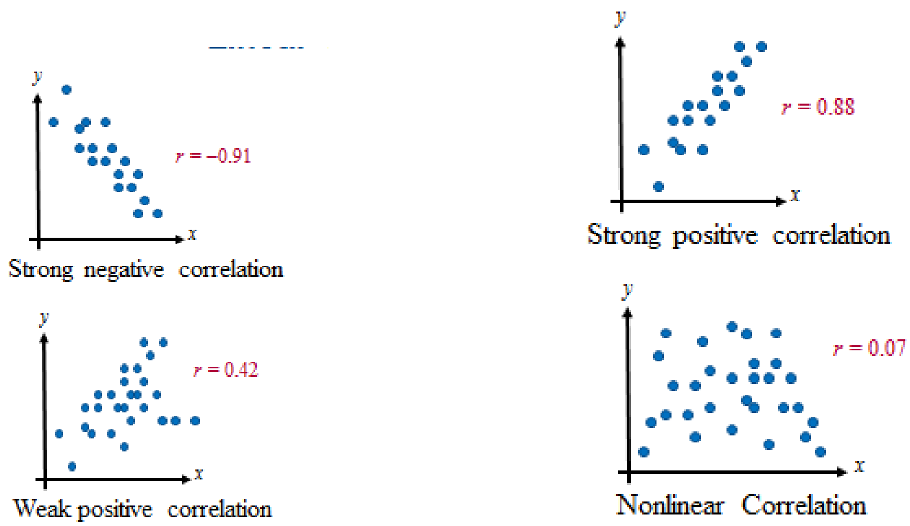
Non linear relationship

Very important: If pattern of data is not linear (looks like a **curve**) or it has an **outlier,** or it show **no pattern** then **linear regression method** is **not valid** and **not applicable.**

2 Computing $\boldsymbol{r}$ = Correlation Coefficient (**the measurement of strength of relationship between 2 variables)** by formula given by $r = \dfrac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}} =$ or using Ti calculator and comment on its strength. The value of $\boldsymbol{r}$ is always between $-1 \le r \le 1$



Linear Correlation Coefficient and scatter Diagram



3. Computing $\bar{x}, \bar{y}, s_x, s_y$,

4. Using the **formula or TI calculator** to computing Slope ($\boldsymbol{a}$) and y-intercepts ($\boldsymbol{b}$) for the regression equation $\boldsymbol{y = a\,x + b}$ by formula $\boldsymbol{Slope = a} = \dfrac{n\left(\sum xy\right) - \left(\sum x\right)\left(\sum y\right)}{n\left(\sum x^2\right) - \left(\sum x\right)^2}$ and $\boldsymbol{y\text{-}itc = b} = \dfrac{\left(\sum y\right)\left(\sum x^2\right) - \left(\sum x\right)\left(\sum xy\right)}{n\left(\sum x^2\right) - \left(\sum x\right)^2}$

5. Using $\boldsymbol{a}$, $\boldsymbol{b}$ and inputting them into regression equation $\boldsymbol{y = a\,x + b}$, then use this equation to **estimate or predict** one variable from the other. Estimated values are labeled as y' (y -prime) and x' (x -prime).

Guideline for using the regression line:

1. If there is no significant linear correlation, do not use the regression equation.

2. When using the regression equation for prediction, **stay** within the range of the available sample data.

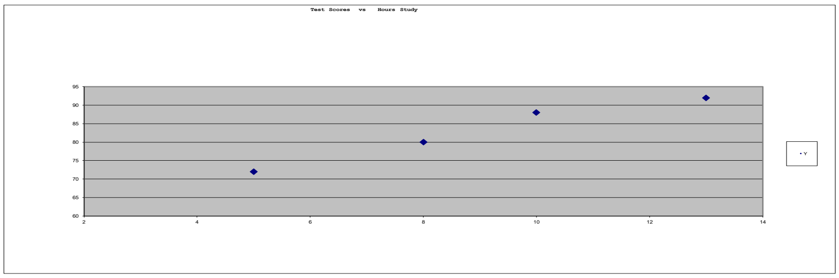
3. A Regression equation based on old data is not necessarily valid now.

Marginal Change (Slope): in a variable is the amount that it changes in y-variable when the x-variable increases by one unit.
Outlier: is a point that is lying far away from the other data points.

Is there a relationship between hours of study and test scores?

	$x =$ Hours Study/week	$y =$ Test Score	x^2	y^2	$x\,y$
1	5	72	25	5184	360
2	10	88	100	7764	880
3	13	92	169	8464	1196
4	8	80	64	6400	640
	$\Sigma x = \mathbf{36}$	$\Sigma y = \mathbf{332}$	$\Sigma x^2 = \mathbf{358}$	$\Sigma y^2 = \mathbf{27792}$	$\Sigma xy = \mathbf{3076}$

1. Use the data and plot the data as a scatter diagram and **<u>comment</u>** on the pattern of the points.



Strong Positive

Linear Correlation

2. Compute the correlation coefficient and **<u>comment</u>** on that: ***a very strong positive linear correlation.***

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - \left(\sum x\right)^2}\sqrt{n\sum y^2 - \left(\sum y\right)^2}} = \frac{(4)(3076) - (36)(332)}{\sqrt{4(358) - (36)^2}\sqrt{4(27792) - (332)^2}} = \frac{12304 - 11952}{\sqrt{136}\sqrt{944}} = \frac{352}{358.307} = 0.9824$$

3. Compute the slope and y-intercept and write the equation of regression line.

$$\textbf{\textit{Slope}} = \boldsymbol{a} = \frac{n\left(\sum xy\right) - \left(\sum x\right)\left(\sum y\right)}{n\left(\sum x^2\right) - \left(\sum x\right)^2} = \frac{4(3076) - (36)(332)}{4(358) - (36)^2} = \frac{12304 - 11952}{1432 - 1296} = \frac{352}{136} = \mathbf{2.588 = 2.59}$$

$$\textbf{\textit{y - itc}} = \boldsymbol{b} = \frac{\left(\sum y\right)\left(\sum x^2\right) - \left(\sum x\right)\left(\sum xy\right)}{n\left(\sum x^2\right) - \left(\sum x\right)^2} = \frac{(332)(358) - (36)(3076)}{4(358) - (36)^2} = \frac{118856 - 110736}{1432 - 1296} = \frac{8120}{136} = \mathbf{59.71}$$

$$\boldsymbol{y = a\,x + b = 2.59\,x + 59.71}$$

4. Explain the slope based on the regression equation and the in relation of x and y variables.

   In general for every additional hour of study per week the score goes up by 2.59 points.

5. Compute average and standard deviation for both x and y variables.

   $\bar{x} = 36/4 = \mathbf{9}$ *hrs* $s_x = 3.37$ $\quad \bar{y} = 332/4 = \mathbf{83}$ $\quad s_y = 8.87$

6. If one student studies 10 hours a week, use **Reg. Equ.** to estimate her test score. $x = 10\ hrs,\quad y' = 85.61$

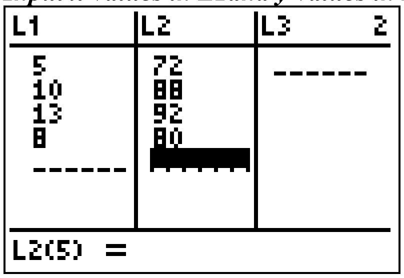
   $x = 10\ hrs,\quad y' = 85.61$

7. If one student has test score of 90, use **Reg. Equ.** to estimate number of hours he spends studying per week.

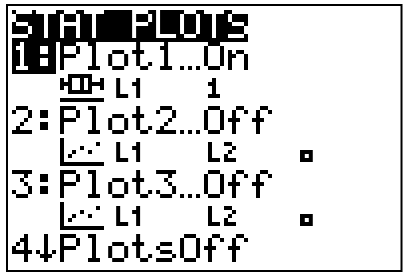
   and if $\quad y = 90, \quad x' = 11.69$ hrs

TI-83/84

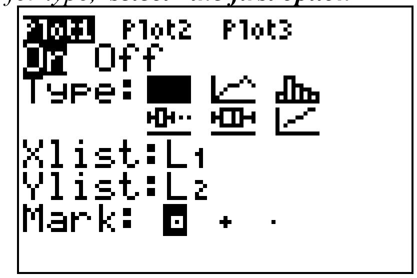
Input x-values in L1and y-values in L2



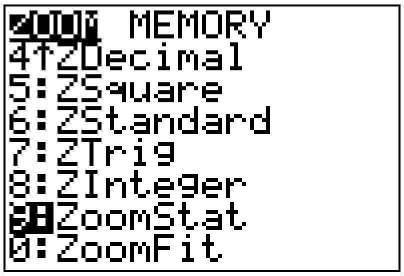
2nd STAT PlOTS



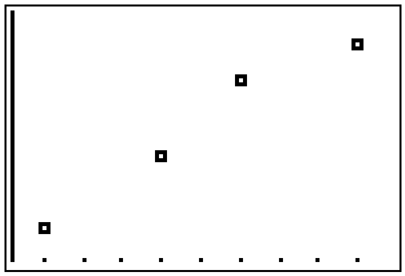
for type, ***select the first option***



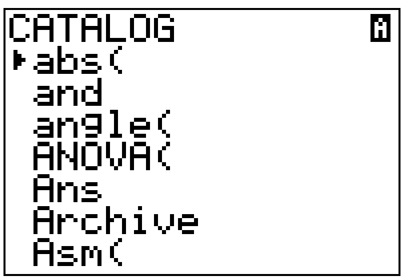
Zoom 9



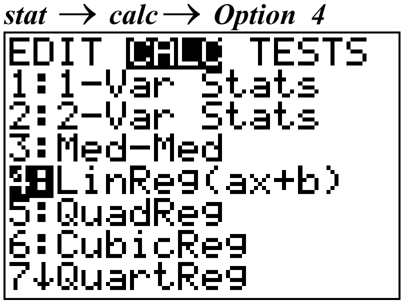
Result: Scattered Plot

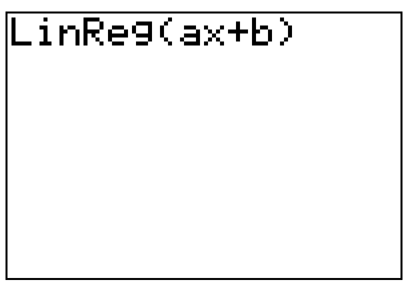


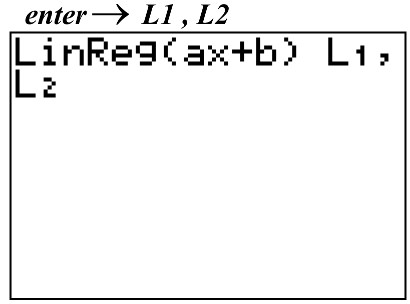
2n d → 0 ***select Diagnostic on → enter*** ***enter***



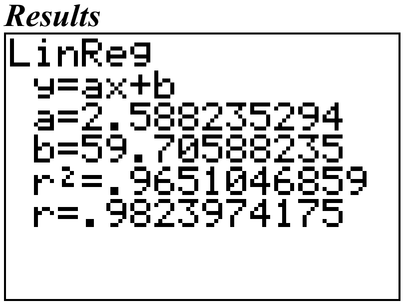
stat → calc → Option 4 ***enter → L1 , L2***







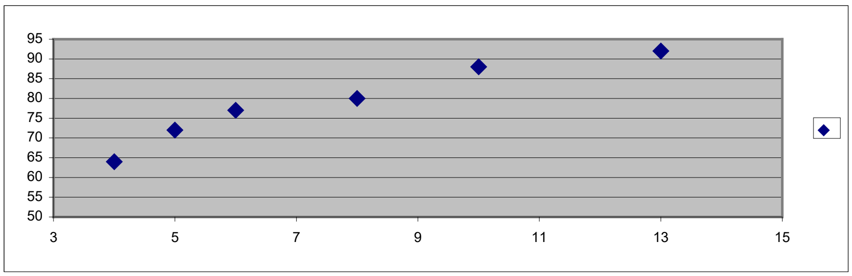
Results



More Practice

	x = Hours Study/week	y = Test Score	x^2	y^2	$x\,y$
1	5	72			
2	10	88			
3	13	92			
4	8	80			
5	6	77			
6	4	64			
	$\sum x = 46$	$\sum y = 473$	$\sum x^2 = 410$	$\sum y^2 = 37817$	$\sum x\,y = 3794$

1. Use the data and plot the data as a scatter diagram and **comment** on the pattern of the points.



Comment: A very strong positive linear correlation.

2. Compute the correlation coefficient and **comment** on that $r = 0.963$ Very strong…?

3. Compute the slope and y-intercept and write the equation of regression line. Slope = a = 2.92, y-itc = b = 56.41

$$y = a\,x + b = 2.92\,x + 56.41$$

4. Explain the slope based on the regression equation and the in relation of x and y variables.
   In general for every additional hour of study per week the score goes up by 2.92 points.

5. Compute average and standard deviation for both x and y variables. $\bar{x} = 7.67$, $\bar{y} = 78.83$, $s_x = 3.386$, $s_y = 10.28$

6. If one student studies 6 hours a week, use **Reg. Equ.** to estimate her test score. $x = 6$, $y' = 73.93$

7. If one student has test score of 85, use **Reg. Equ.** to estimate number of hours he spends studying per week. $y = 85$, $x' = 9.79$