

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. [Online calculator](#)

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** $= y = ax + b$) 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)

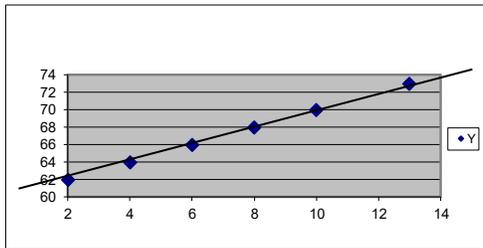
Steps to do a Correlation and Regression problem

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

Why do we need a scatter diagram?

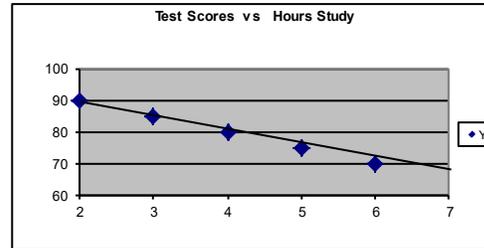
- To see if data exhibit a **linear pattern** or not
- To see if linear pattern is **positive or negative**
- To see how closely (**strongly or perfectly**) data are **clustered around the a straight line**.
- To detect any **outlier** (a point that is lying far away from the other data points).

Different Possible shapes of a Scatter Diagram



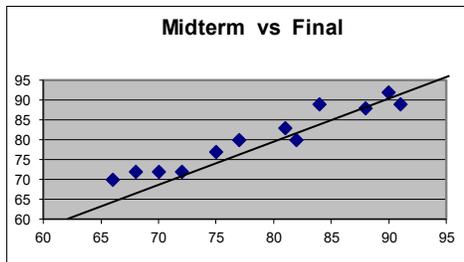
$$r = 1$$

Perfect Positive Linear Correlation

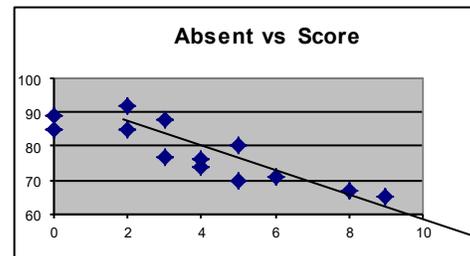


$$r = -1$$

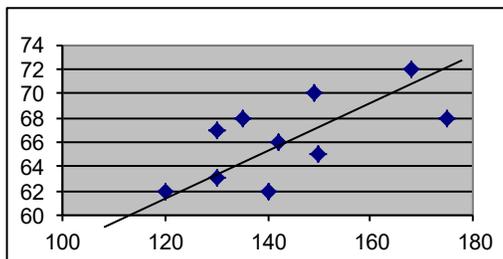
Perfect Negative Linear Correlation



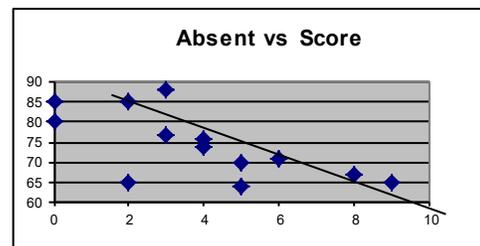
Strong Positive Linear Correlation



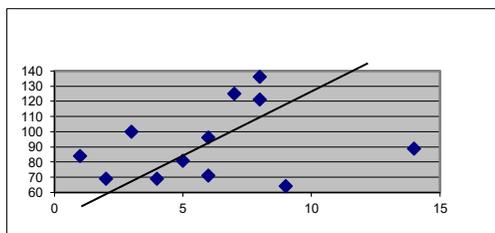
Strong Negative Linear Correlation



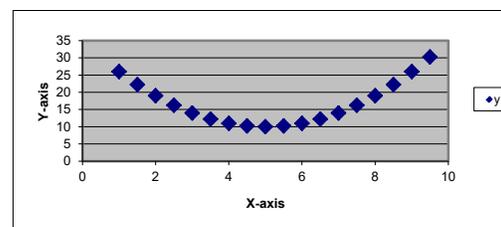
Positive Linear Correlation



Negative Linear Correlation



No Correlation



Non linear relationship

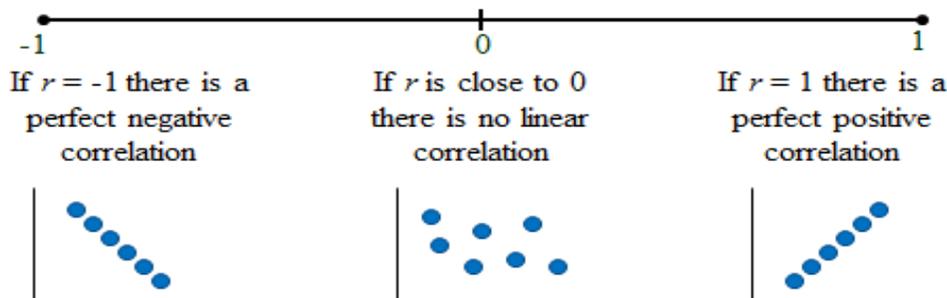
Very important: If pattern of data is not linear (looks like a curve) or it has an outlier, or it show no pattern then linear regression method is not valid and not applicable.

2 Computing r = Correlation Coefficient (the measurement of strength of relationship between 2

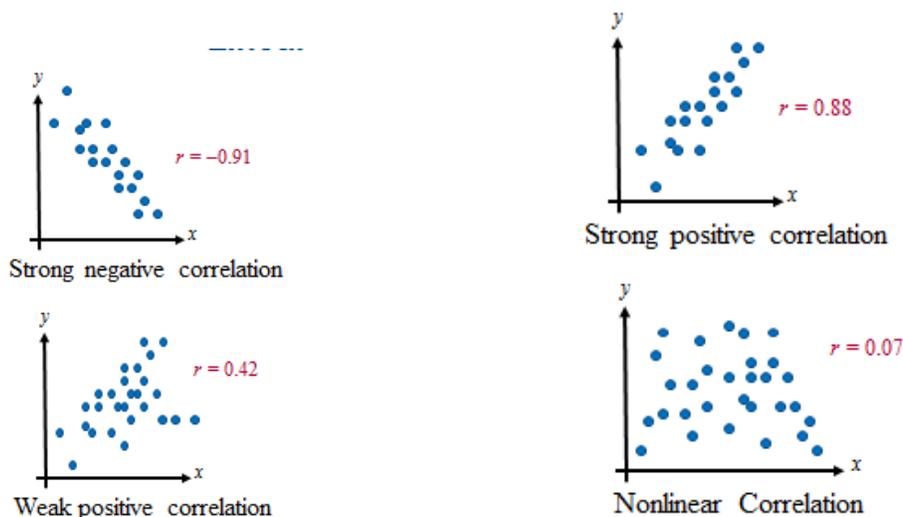
variables) by formula given by $r = \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - (\sum x)^2} \sqrt{n\sum y^2 - (\sum y)^2}}$ = or using Ti calculator and comment on its

strength. The value of r is always between $-1 \leq r \leq 1$

[Online calculator](#)



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 (a) and y-intercepts (b) for the regression equation

$y = ax + b$ by formula $Slope = a = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$ and $y - itc = b = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x)^2}$

5. Using a , b and inputting them into regression equation $y = ax + b$, 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.

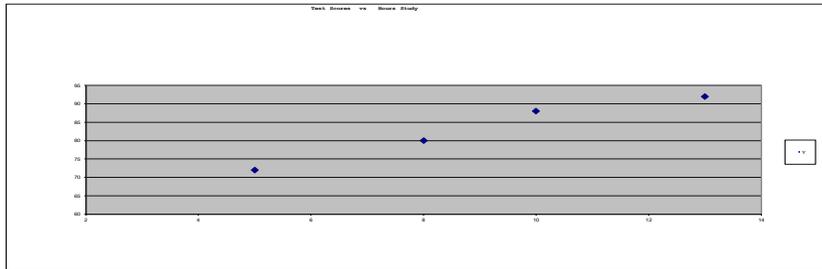
Marginal Change (Slope): in a variable is the amount that it changes in y-variable when the x-variable increases by one unit.

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? [Online calculator](#)

	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
	$\Sigma x = 36$	$\Sigma y = 332$	$\Sigma x^2 = 358$	$\Sigma y^2 = 27792$	$\Sigma xy = 3076$

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



**Strong Positive
Linear Correlation**

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

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

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

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

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

$$y = ax + b = 2.59x + 59.71$$

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.

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

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

$$x = 10 \text{ hrs}, \quad y' = 85.61$$

7. If one student has test score of 90, use **Reg. Equ.** to estimate number of hours he spends studying per week. and if $y = 90$, $x' = 11.69 \text{ hrs}$

Input x-values in L1 and y-values in L2

L1	L2	L3	Z
5	72	-----	
10	88		
13	92		
8	80		

L2(5) =			

2nd STAT PLOTS

```

STAT PLOTS
1:Plot1...On
  [F1] L1  1
  2:Plot2...Off
  [F2] L1  L2
  3:Plot3...Off
  [F3] L1  L2
4↓PlotsOff
    
```

for type, select the first option

```

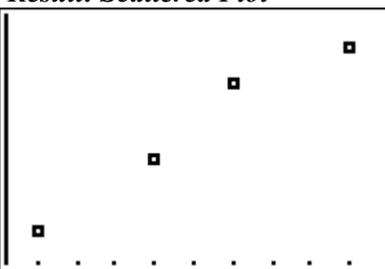
Plot1 Plot2 Plot3
Off Off Off
Type: [F1] [F2] [F3]
Xlist:L1
Ylist:L2
Mark: [F4] + .
    
```

Zoom 9

```

MEMORY
4↑ZDecimal
5:ZSquare
6:ZStandard
7:ZTrig
8:ZInteger
9:ZoomStat
0:ZoomFit
    
```

Result: Scattered Plot



2nd d → 0

select Diagnostic on → enter

enter

```

CATALOG
abs(
and
angle(
ANOVA(
Ans
Archive
Asm(
    
```

```

CATALOG
DelVar
DependAsk
DependAuto
det(
DiagnosticOff
DiagnosticOn
dim(
    
```

```

DiagnosticOn
Done
    
```

stat → calc → Option 4

```

EDIT [F1] TESTS
1:1-Var Stats
2:2-Var Stats
3:Med-Med
4:LinReg(ax+b)
5:QuadReg
6:CubicReg
7↓QuartReg
    
```

```

LinReg(ax+b)
    
```

enter → L1, L2

```

LinReg(ax+b) L1,
L2
    
```

Results

```

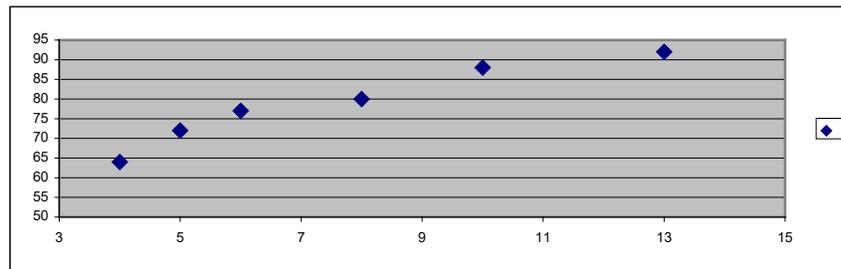
LinReg
y=ax+b
a=2.588235294
b=59.70588235
r²=.9651046859
r=.9823974175
    
```

More Practice

Online calculator

	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-ipc = $b = 56.41$

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

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. $y = 85$, $x' = 9.79$

Questions:

- a) Can Correlation coefficient be more than 1?
- b) Can Correlation coefficient be less than 1?
- c) Can correlation coefficient be close to zero? If yes what does that mean?
- d) If the slope of the regression line is negative, then what it suggests about the nature of relationship between two variables?
- e) In regression and correlation, what are the different names that we can label x and y variables?

A.

	x = Hours Study/week	y = Test Score	x²	y²	x y
1	7	68			
2	11	86			
3	16	98			
4	12	88			
5	8	79			
6	6	69			
	$\sum x = 60$	$\sum y = 488$	$\sum x^2 = 670$	$\sum y^2 = 40370$	$\sum xy = 5092$

1. Use the data and plot the data as a scattered diagram and **comment** on the pattern of the points.
2. Compute the correlation coefficient and **comment** on its value _____
3. Compute the slope and y-intercept and write the equation of regression line.
4. Explain the slope based on the regression equation and the in relation of x and y variables.
5. Compute average and standard deviation for both x and y variables.
6. If one student studies 10 hours a week, use **Reg. Equ.** to estimate her test score.
7. If one student has test score of 90, use **Reg. Equ.** to estimate number of hours he spends studying per week.

B

X = Experience(yrs)	14	3	5	6	4	9	18	5	16
Y = Monthly Salary \$(000)	42	24	33	31	29	39	47	30	43

1. Use the data and plot the data as a scattered diagram and **comment** on the pattern of the points.
2. Compute the correlation coefficient and **comment** on its value _____
3. Compute the slope and y-intercept and write the equation of regression line. _____
4. Explain the slope based on the regression equation and the in relation of x and y variables.
5. Compute average and standard deviation for both x and y variables. _____
6. If some one's experience is 10 years old, use **Reg. Equ.** to estimate his salary. _____
7. If some one's salary is \$38, 000, use **Reg. Equ.** to estimate her experience. _____

C.

X= Midterm	75	68	82	91	84	77	72	88	90	66	70	81	59
Y= Final	77	72	80	89	89	80	72	88	92	70	72	83	66

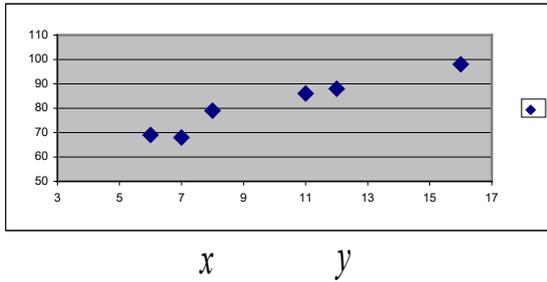
1. Use the data and plot the data as a scattered diagram and **comment** on the pattern of the points.
2. Compute the correlation coefficient and **comment** on its value _____
3. Compute the slope and y-intercept and write the equation of regression line. _____
4. Explain the slope based on the regression equation and the in relation of x and y variables.
5. Compute average and standard deviation for both x and y variables. _____
6. If some one gets 74 on the midterm estimate his final score. _____
7. If some one gets 74 on the final estimate her midterm score. _____

D.

X = Number of times absent	2	3	5	2	6	0	4	3	9	5	0	4	8
Y = Average test scores	92	88	80	85	71	85	74	77	65	70	89	76	67

1. Use the data and plot the data as a scattered diagram and **comment** on the pattern of the points.
2. Compute the correlation coefficient and **comment** on its value _____
3. Compute the slope and y-intercept and write the equation of regression line. _____
4. Explain the slope based on the regression equation and the in relation of x and y variables.
5. Compute average and standard deviation for both x and y variables. _____
6. If some one has been absent 7 times, then estimate his average test score. _____
7. If some one's average test score is 90, then estimate the number of absentees she might have _____

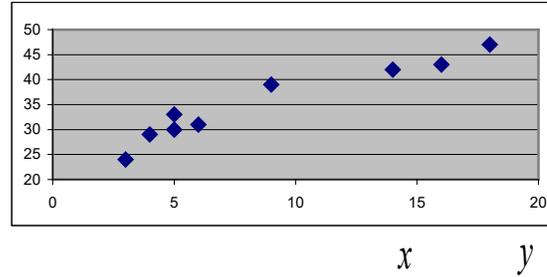
Problem A



Mean	10	81.33
St Dev.	3.74	11.66
Correl Coeff	$r = 0.972$	
Slope	3.03	
Y-itc	51.05	

$Y = 3.03 X + 51.05$		
X = 10	,	$y' = ? = 81.35$
Y = 90	,	$x' = ? = 9.79$

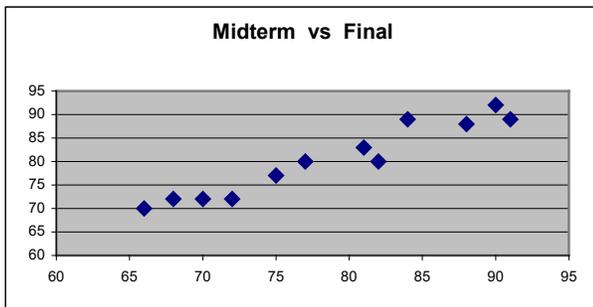
Problem B



Mean	8.889	35.333
St Dev.	5.667	7.697
Correl Coeff	$r = 0.961$	
Slope	1.305	
Y-itc	23.730	

$Y = 1.305 X + 23.73$		
X = 10	,	$y' = ? = 36.78$
Y = 38	,	$x' = ? = 10.93$

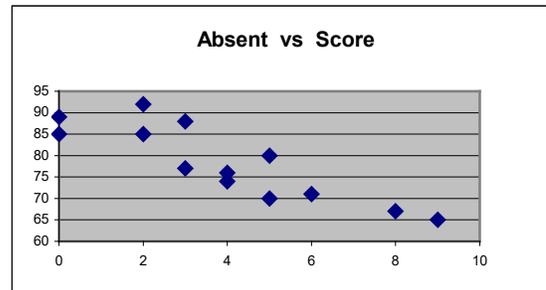
Problem C



Mean	77.154	79.231
St Dev.	9.915	8.506
Correl Coeff	$r = 0.971$	
Slope	0.833	
Y-itc	14.971	

$Y = 0.833 X + 14.971$		
X = 74	,	$y' = ? = 76.61$
Y = 74	,	$x' = ? = 70.86$

Problem D



Mean	3.923	78.385
St Dev.	2.722	8.856
Correl Coeff	$r = -0.870$	
Slope	-2.830	
Y-itc	89.485	

$Y = -2.83 X + 89.485$		
X = 7	,	$y' = ? = 69.68$
Y = 90	,	$x' = ? = -0.18$