

Elementary Statistics (15060101)

Linear Correlation Coefficient

And

Simple Linear Regression

Independent Versus **dependent** variable:

X: independent (explanatory) variable

The independent variable is the cause. Its value is independent of other variables in your study.

Y: dependent (response) variable

The dependent variable effected by the independent variable.
Its value depends on changes in the independent variable.

Example:

Determine independent and dependent variables:

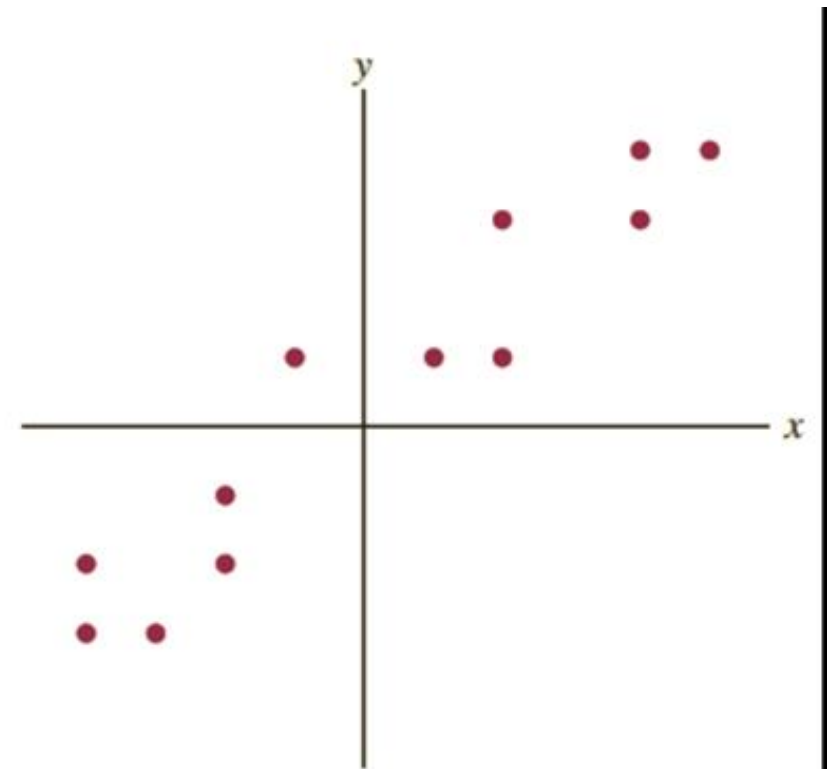
An insurance company wants to predict sales from the amount of money they spend on advertising.

Correlation Coefficient

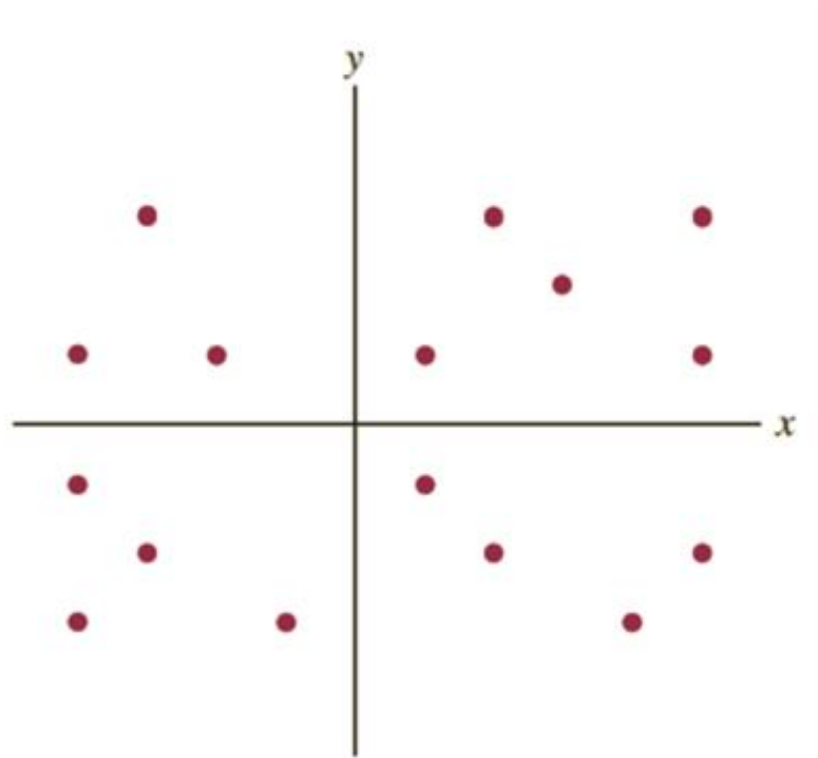
X: independent (**explanatory**) variable

Y: dependent (**response**) variable

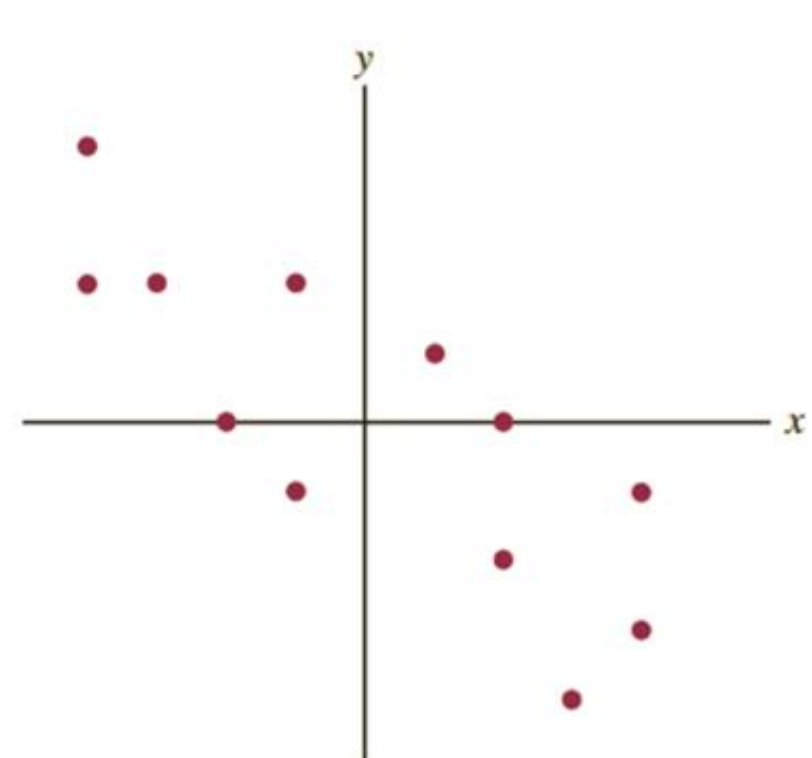
The Scatter Plot



s_{xy} **Positive:**
(\bar{x} and \bar{y} are positively linearly related)



s_{xy} **Approximately 0:**
(x and y are not linearly related)



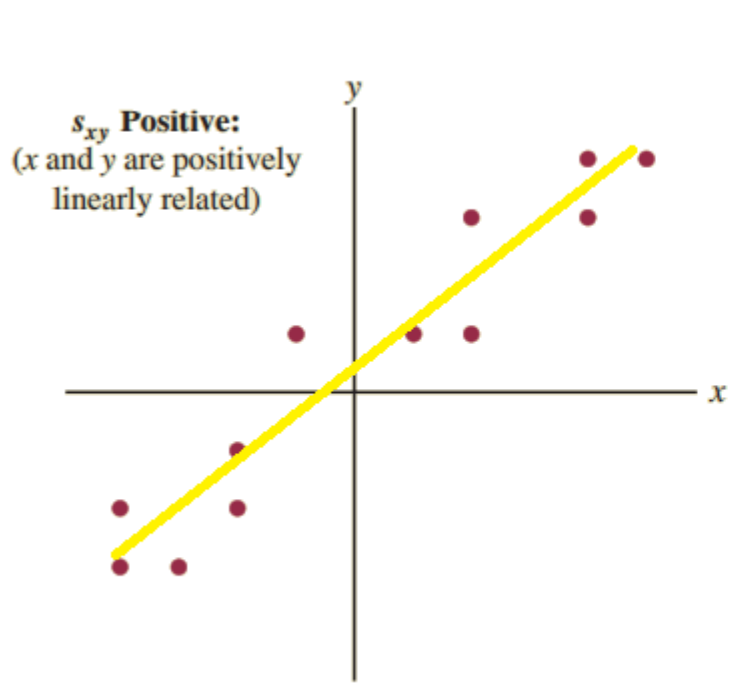
s_{xy} **Negative:**
(x and y are negatively linearly related)

r : Pearson Correlation Coefficient $-1 \leq r \leq 1$

$$r = \frac{\sum (x_i - \bar{x}) (y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}}$$

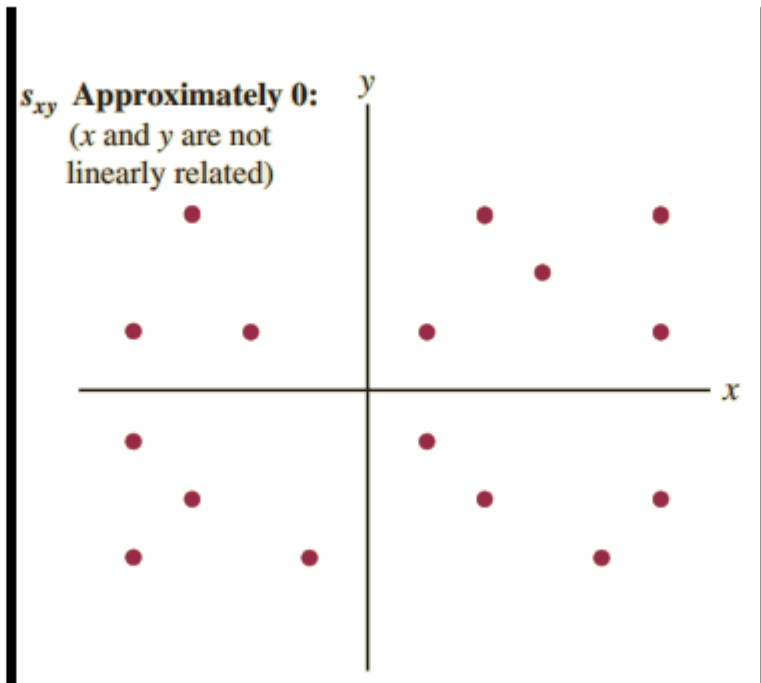
OR

$$r = \frac{\sum x_i y_i - n\bar{x} * \bar{y}}{\sqrt{\sum x_i^2 - n\bar{x}^2} \sqrt{\sum y_i^2 - n\bar{y}^2}}$$



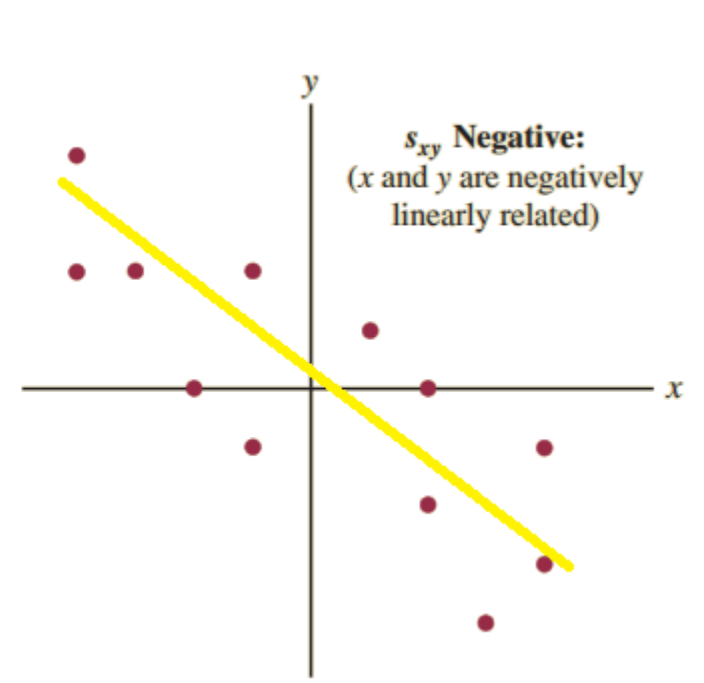
$$0 < r \leq 1$$

r Positive



$$r \approx 0$$

r : around zero



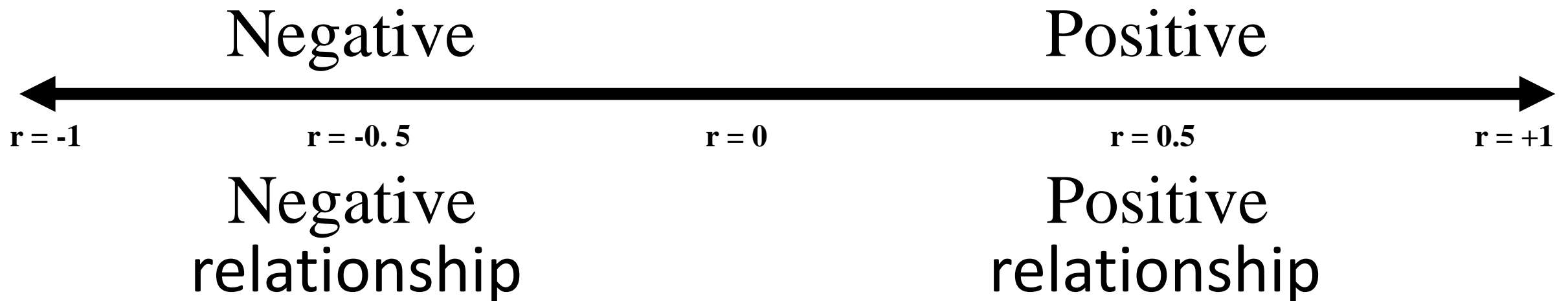
$$-1 \leq r < 0$$

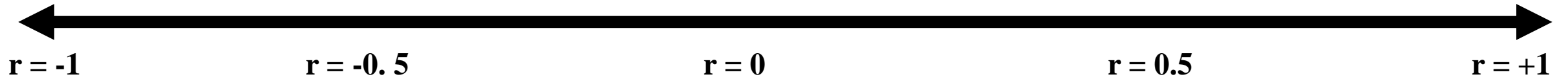
r : Negative

r : Pearson Correlation Coefficient

r : Pearson Correlation Coefficient used to measure the linear relationship between two quantitative variable

The correlation coefficient (r) ranges from -1 to $+1$. $-1 \leq r \leq 1$
Values close to -1 or $+1$ indicate a strong linear relationship.
The closer the correlation is to zero, the weaker the relationship.





Here,

Some examples of the correlation coefficient measure values

$r = 0.00$ >>Means>> **No Linear Relationship**

$r = -1$ >>>> **Complete** and **negative** linear relationship
 $r = -0.98$ >>>> **Strong** and **negative** linear relationship
 $r = -0.5$ >>> **Moderate** and **negative** linear relationship
 $r = -0.25$ >>>> **Weak** and **negative** linear relationship

$r = +1$ >>>> **Complete** and **positive** linear relationship
 $r = +0.95$ >>>> **Strong** and **positive** linear relationship
 $r = +0.5$ >>> **Moderate** and **positive** linear relationship
 $r = +0.34$ >>>> **Weak** and **positive** linear relationship

Note about the correlation coefficient measure (r):

If $r=0.00$, this implies that there is no linear relationship between variables X & Y but there may be some other non-linear relationship.

Hence, $r=0.00$ doesn't necessarily imply that the two variables are independent.

**So,
if X & Y are independent variables >> Then: $r = 0.00$**

**But
if $r = 0.00$, we can't said: X & Y are independent variables**

r : Pearson Correlation Coefficient $-1 \leq r \leq 1$

$$r = \frac{\sum (x_i - \bar{x}) (y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}}$$

OR

$$r = \frac{\sum x_i y_i - n\bar{x} * \bar{y}}{\sqrt{\sum x_i^2 - n\bar{x}^2} \sqrt{\sum y_i^2 - n\bar{y}^2}}$$

Example: Compute Pearson Correlation Coefficient

X	4	6	28	6
Y	5	8	14	9

Answer

	X	Y	(X-\bar{x})	(X - \bar{x})²	(Y-\bar{y})	(Y - \bar{y})²	(X-\bar{x})(Y-\bar{y})
	4	5	-7	49	-4	16	28
	6	8	-5	25	-1	1	5
	28	14	17	289	5	25	85
	6	9	-5	25	0	0	0
Total	44	36	0	388	0	42	118

$$r = \frac{\sum (x_i - \bar{x}) (y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}}$$

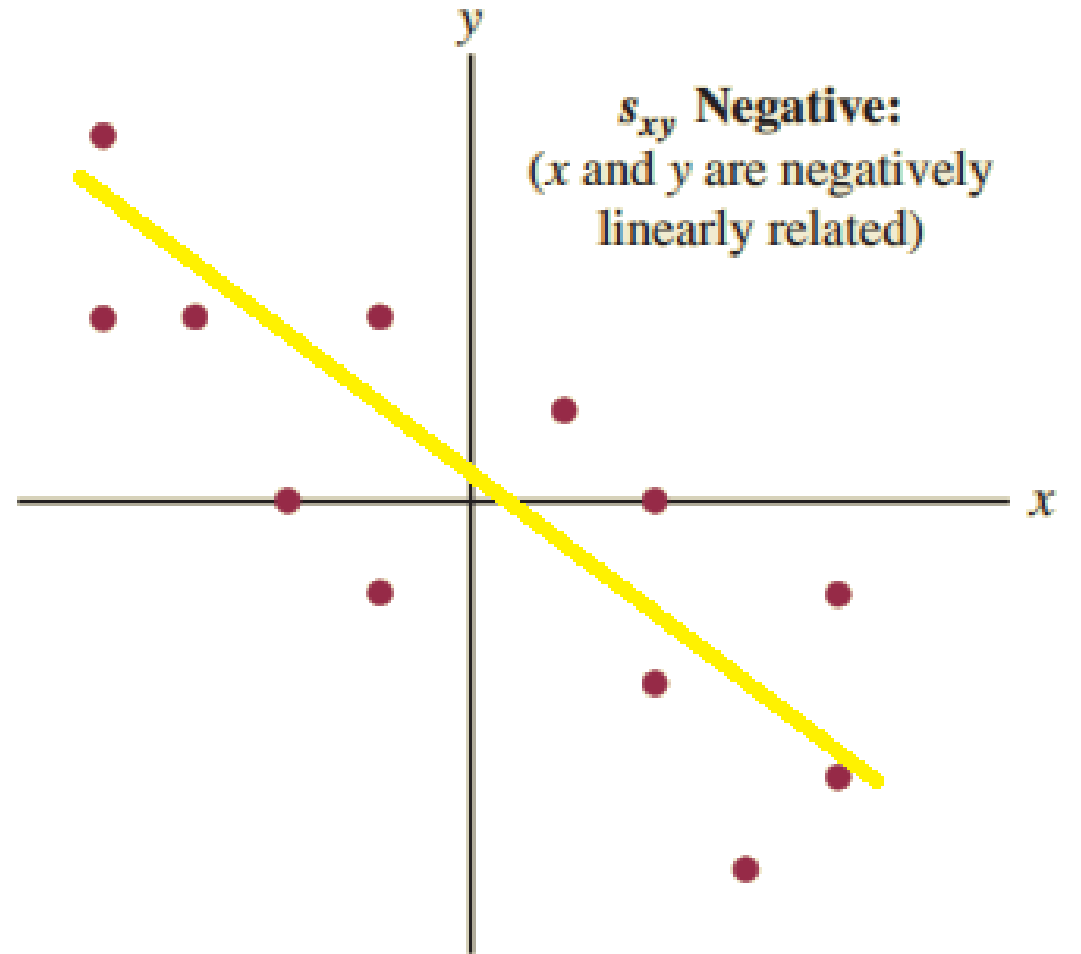
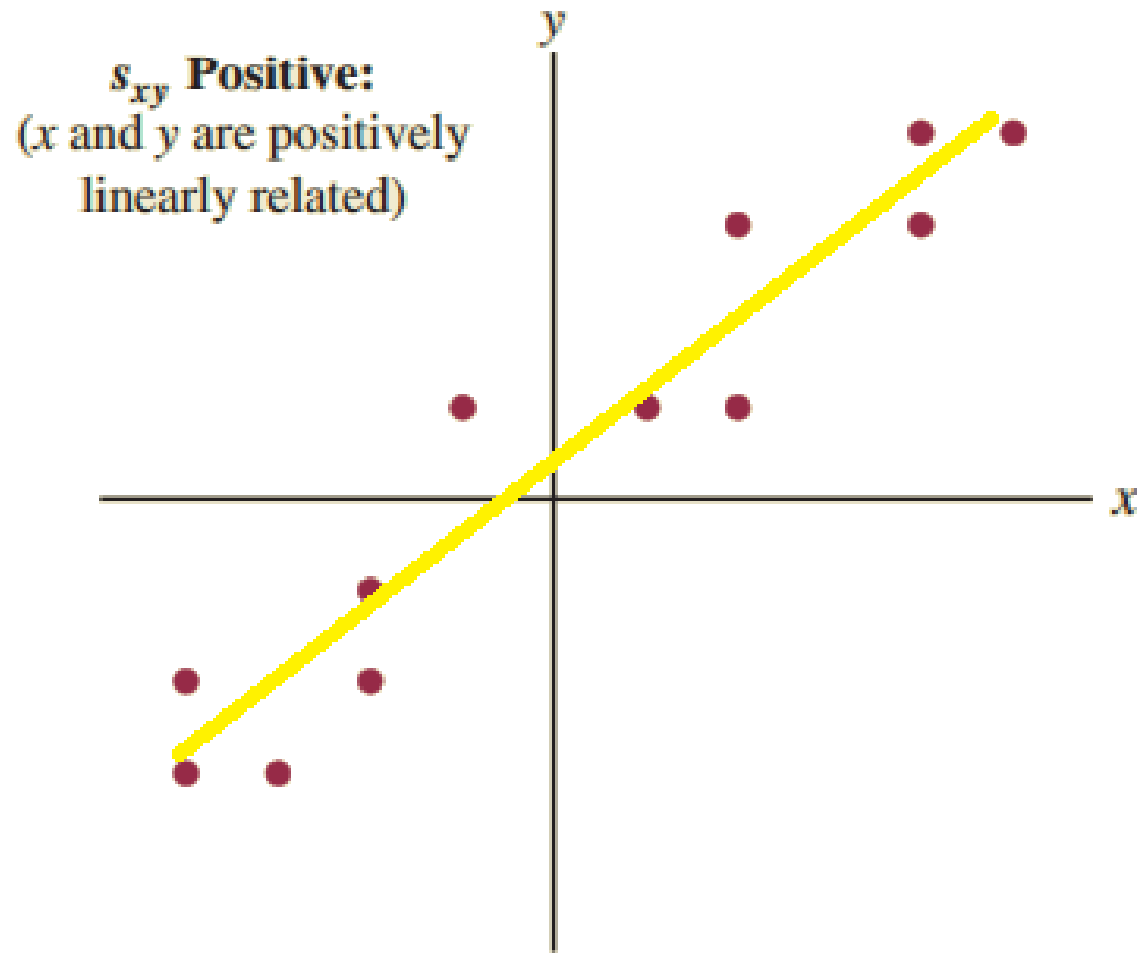
$$r = \frac{118}{\sqrt{388}\sqrt{42}} = 0.9243608$$

$$= 0.92$$

r = 0.92 >>>> **Strong and positive** linear relationship

Simple Linear Regression

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$



Linear Regression

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Estimation Error (**residual**): $e = y - \hat{y}$

$\hat{\beta}_1$: Slope of the Regression Line

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

OR
$$\hat{\beta}_1 = \frac{\sum x_i y_i - n\bar{x} * \bar{y}}{\sum x_i^2 - n\bar{x}^2}$$

$\hat{\beta}_0$: intercept of the Regression Line

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

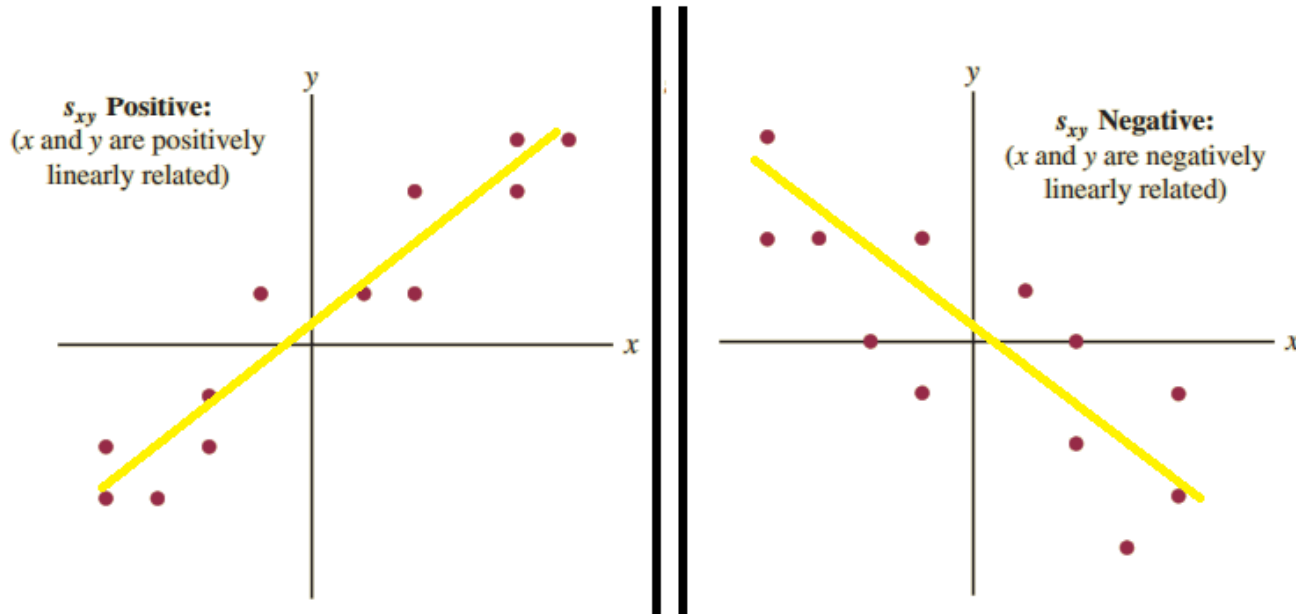
Note:

The **sign** of the linear regression **Slope** AND the **sign** of **Pearson correlation** coefficient are same

Exercise:

Based on the following two cases :

Determine when the error of estimation (residual) is a positive and when it is a negative.



Estimation Error (**residual**):

$$e = y - \hat{y}$$

Example:

For the following values

Find:

X	4	6	28	6
Y	5	8	14	9

- a) Pearson correlation coefficient (r)
- b) The relation type and strength
- c) The linear regression equation $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
- d) The regression estimate at $x = 0$
- e) The regression estimate at $x = 6$
- f) The regression estimate at $x = 4$
- g) The error of regression estimate at $x = 4$

Answer

	X	Y	(X-\bar{x})	(X - \bar{x})²	(Y-\bar{y})	(X-\bar{x})(Y-\bar{y})
	4	5	-7	49	-4	28
	6	8	-5	25	-1	5
	28	14	17	289	5	85
	6	9	-5	25	0	0
Total	44	36	0	388	0	118

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{118}{388} = 0.304$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 9 - (0.308)(11) = 5.612$$

$$\hat{y}_i = 5.61 + 0.304x_i$$

c) The linear regression equation is:

Answer

a) Pearson correlation coefficient ($r = 0.92$)

b) The relation type and strength : is **strong** and **positive** linear relationship

c) The linear regression equation $\hat{y}_i = 5.61 + 0.304x_i$

d) The regression estimate at $x = 0$ is $\hat{y} = 5.61 + 0.304(0) = 5.61$

e) The regression estimate at $x = 6$ is $\hat{y} = 5.61 + 0.304(6) = 7.434$

f) The regression estimate at $x = 4$ is $\hat{y} = 5.61 + 0.304(4) = 6.862$

g) The error of regression estimate at $x = 4$ is $e = y - \hat{y} = 5 - 6.862 = -1.826$

Example:

Based on the following data values

Find:

- a) Pearson correlation coefficient (r)
- b) The relation type and strength
- c) The linear regression equation
- d) The regression estimate at x = 0
- e) The regression estimate at x = 6
- f) The regression estimate at x = 3
- g) The error of regression estimate at x = 3, if y=4

$$n = 6, \bar{x} = 9, \bar{y} = 8, \sum x_i^2 = 930, \sum y_i^2 = 446, \sum x_i y_i = 582$$

X	4	3	7	6	28	6
Y	5	4	8	8	14	9

$$r = \frac{\sum x_i y_i - n\bar{x} * \bar{y}}{\sqrt{\sum x_i^2 - n\bar{x}^2} \sqrt{\sum y_i^2 - n\bar{y}^2}} = \frac{582 - (6)(9)(8)}{\sqrt{930 - (6)(9^2)} \sqrt{446 - (6)(8^2)}} = 0.904$$

$$\hat{\beta}_1 = \frac{\sum x_i y_i - n\bar{x} * \bar{y}}{\sum x_i^2 - n\bar{x}^2} = \frac{582 - (6)(9)(8)}{930 - (6)(9^2)} = \frac{150}{444} = 0.338$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 8 - (0.338)(9) = 4.96$$

$$\hat{y}_i = 4.96 + 0.338x_i$$

Answer:

- a) Pearson correlation coefficient ($r = 0.904$)
- b) The relation type and strength : is **strong** and **positive** linear relationship
- c) The linear regression equation $\hat{y}_i = 4.96 + 0.338x_i$
- d) The regression estimate at $x = 0$ $\hat{y}_i = 4.96$
- e) The regression estimate at $x = 6$ $\hat{y}_i = 4.96 + 0.338(6) = 6.99$
- f) The regression estimate at $x = 3$ $\hat{y}_i = 4.96 + 0.338(3) = 5.974$
- g) The error of regression estimate at $x = 3$, if $y=4$ $e = y - \hat{y} = 4 - 5.974 = -1.974$

The coefficient of determination (R^2):

is the square of the Pearson correlation (r) it ranges

$$0 \leq R^2 \leq 1$$

The coefficient of determination (R^2):

is a measurement used to **explain** how much variability of dependent variable **Y** can be caused by its relationship to independent variable **X**.

Ex. If the Pearson correlation ($r=0.9$) then $R^2 = (0.9)^2 = 0.81 = 81\%$

this means that:

The independent variable is explained about 81% of the variation of the dependent variable Y.

And 19% of the dependent variable variation comes from another factor(s).

Ex. Assuming the regression equation is $\hat{y}_i = 0.2 - 1.5x_i$,

If the coefficient of determination is ($R^2=0.64$). Then Pearson correlation (r) is:

- a) 0.8 b) -0.8 c) 0.64 d) -0.64 e) 0.2 f) -0.2

Example:

Based on the following data values

Find:

X	20	8	12	14	24
Y	2	5	7	8	2

- a) **Pearson correlation coefficient (r)**
- b) **The relation type and strength**
- c) **The linear regression equation**
- d) **What is the slope of the linear regression line**
- e) **The regression estimate at $x = 0$**
- f) **What is the value of Y that the linear regression line intercept Y-axis**
- g) **What is the value of X that the linear regression line intercept X-axis**
- h) **The regression estimate at $x = 8$**
- i) **The error of regression estimate at $x = 8$,**
- j) **What is the percent of X variable interpret the variation of variable Y (How much the coefficient of determination)?**