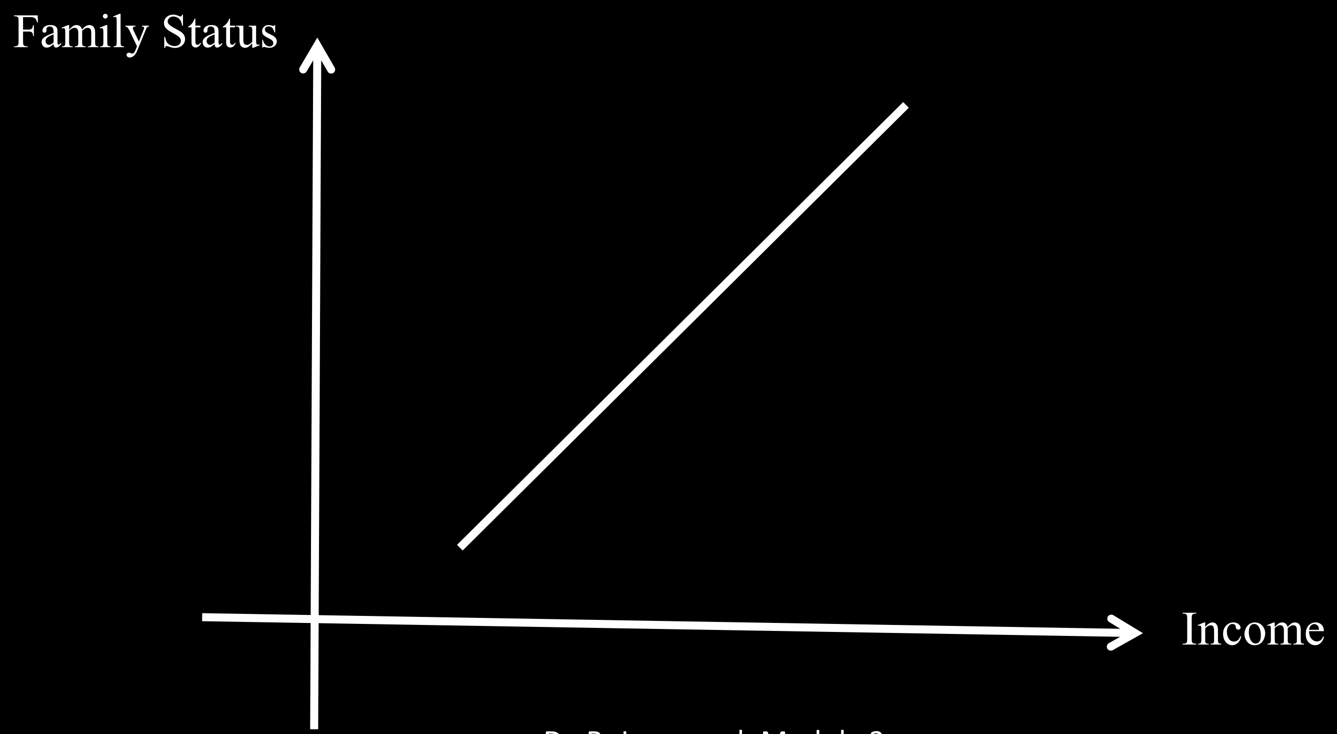
Module 3: Correlation and Regression

Topics

Correlation and Regression – Rank Correlation- Partial and Multiple correlation- Multiple regression.

Correlation

Correlation is a statistical technique used to determine the degree to which two variables are related.



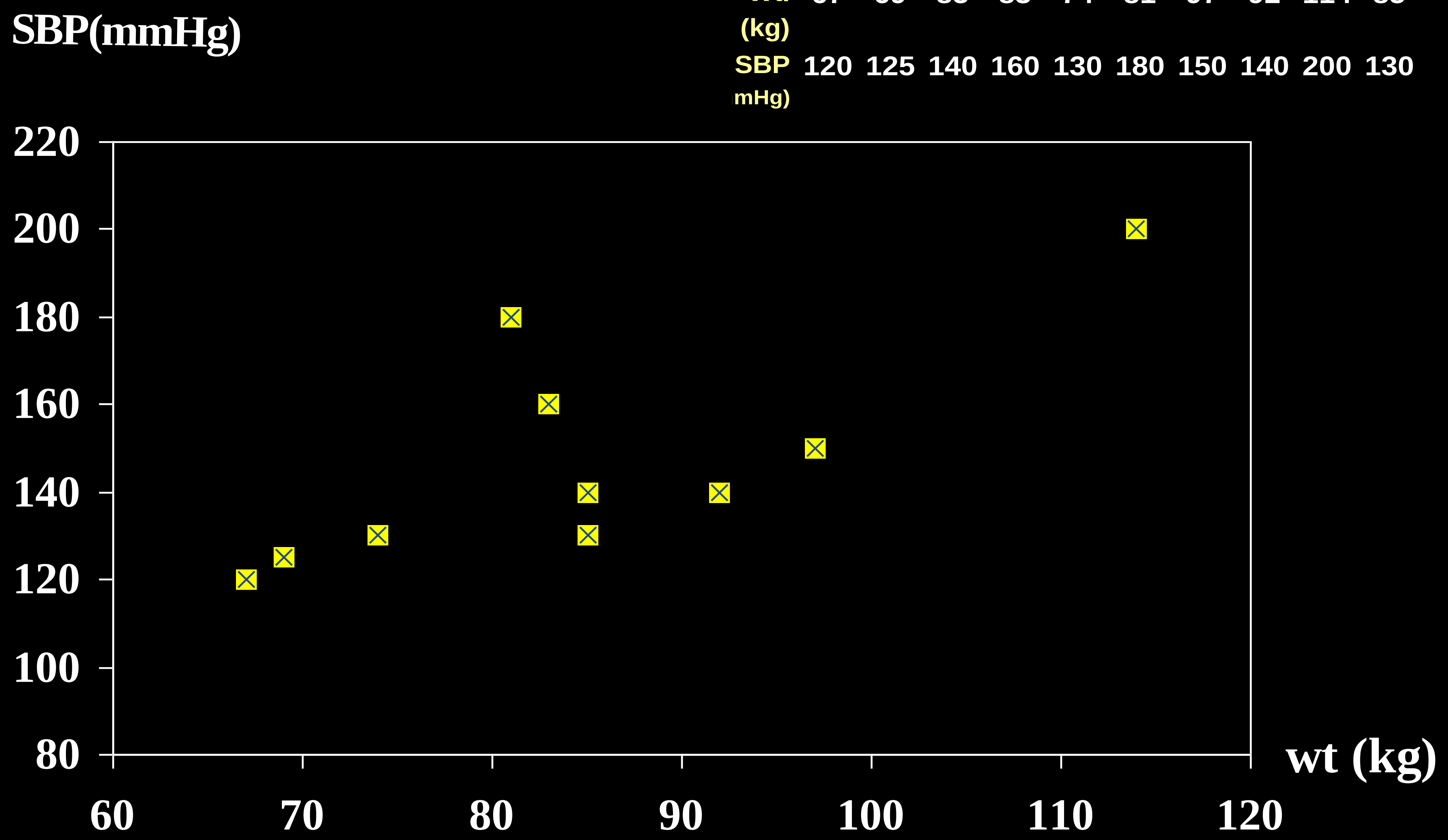
Scatter diagram

- Two quantitative variables
- One variable is called independent (X) and the second is called dependent (Y)
- Points are not joined
- No frequency table

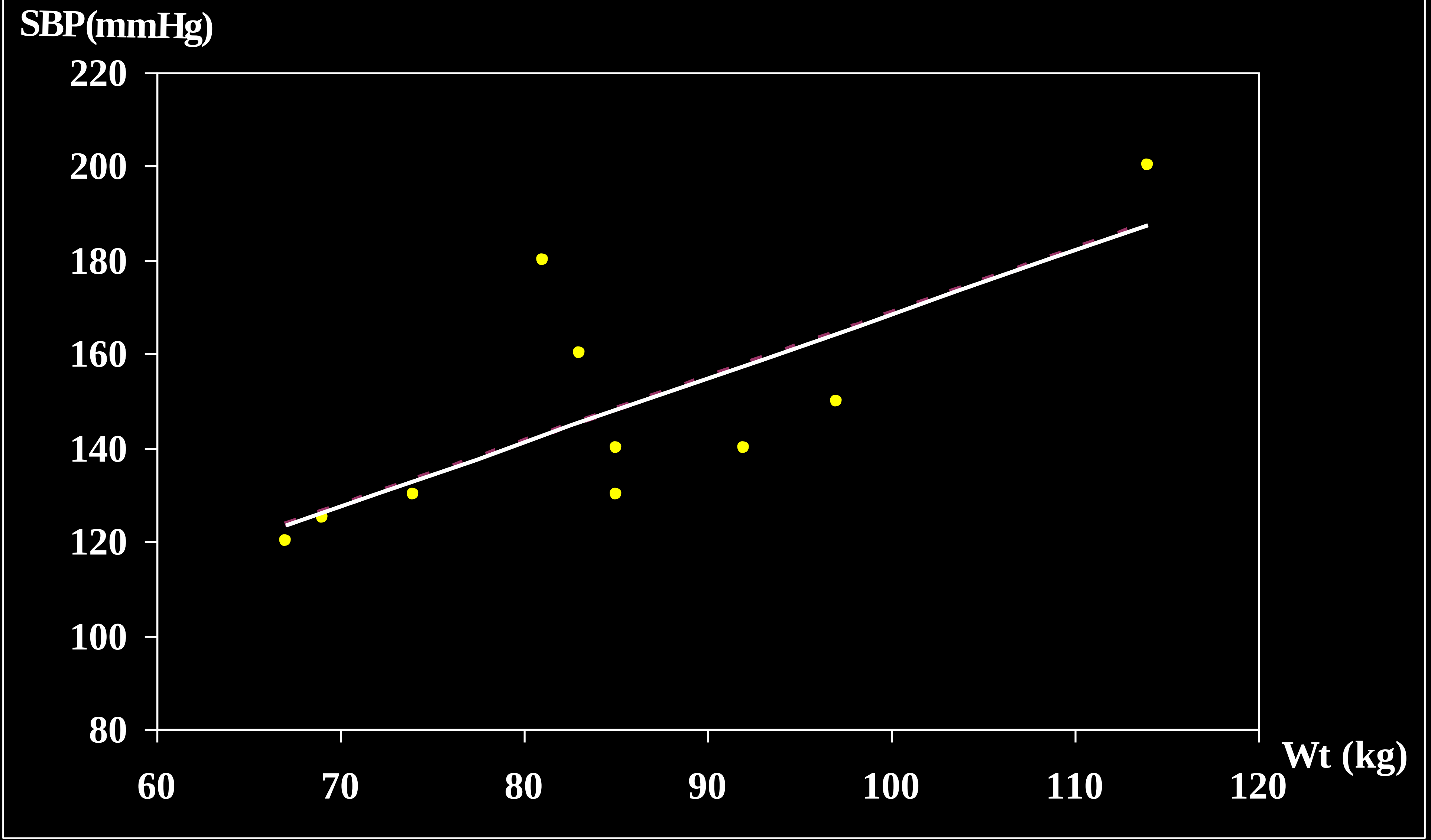
Example

Wt. (kg)	67	69	85	83	74	81	97	92	114	85
SBP mHg)	120	125	140	160	130	180	150	140	200	130

Wt. (kg)	67	69	85	83	74	81	97	92	114	85
SBP mHg)	120	125	140	160	130	180	150	140	200	130



Scatter diagram of weight and systolic blood pressure



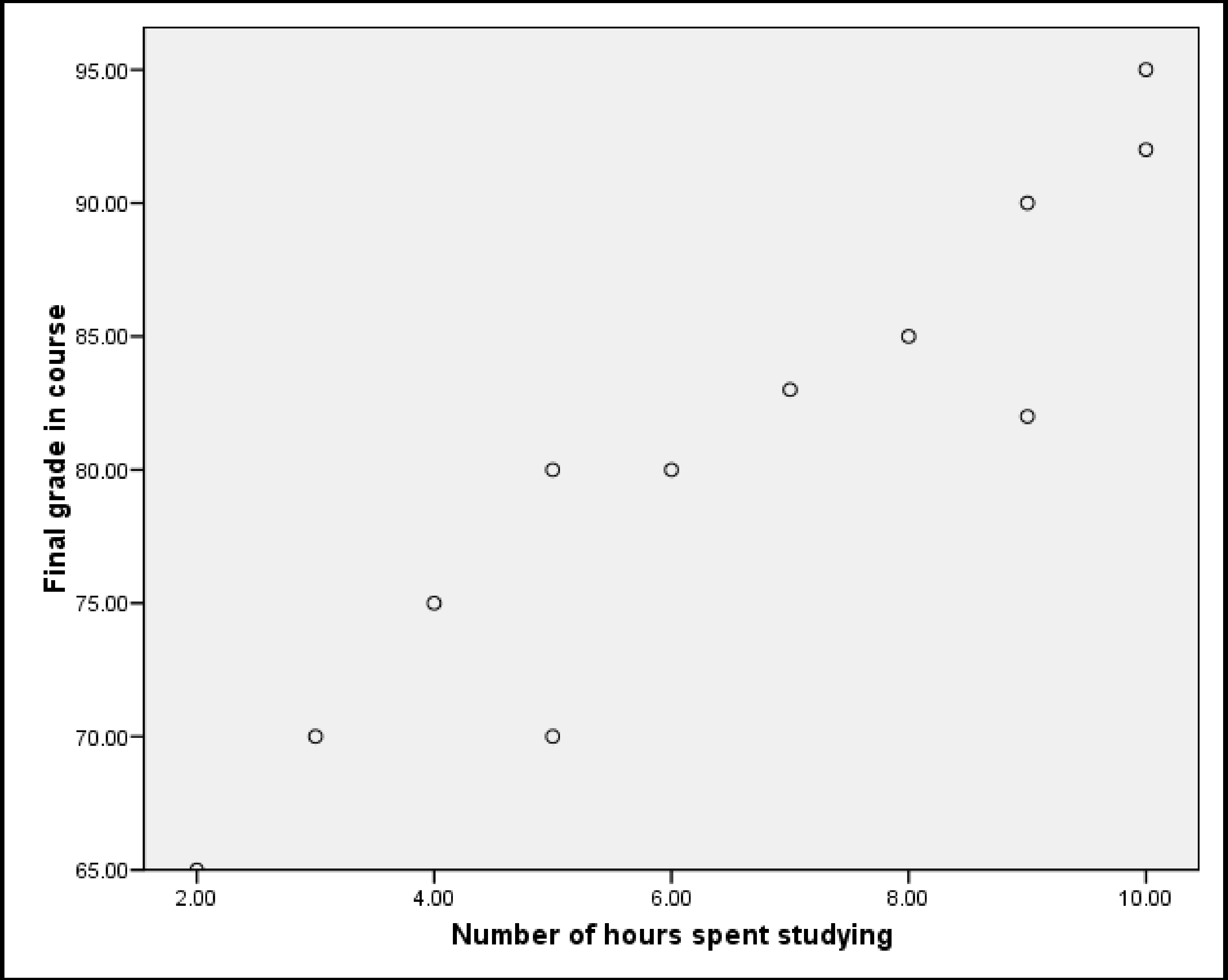
Scatter diagram of weight and systolic blood pressure

Scatter plots

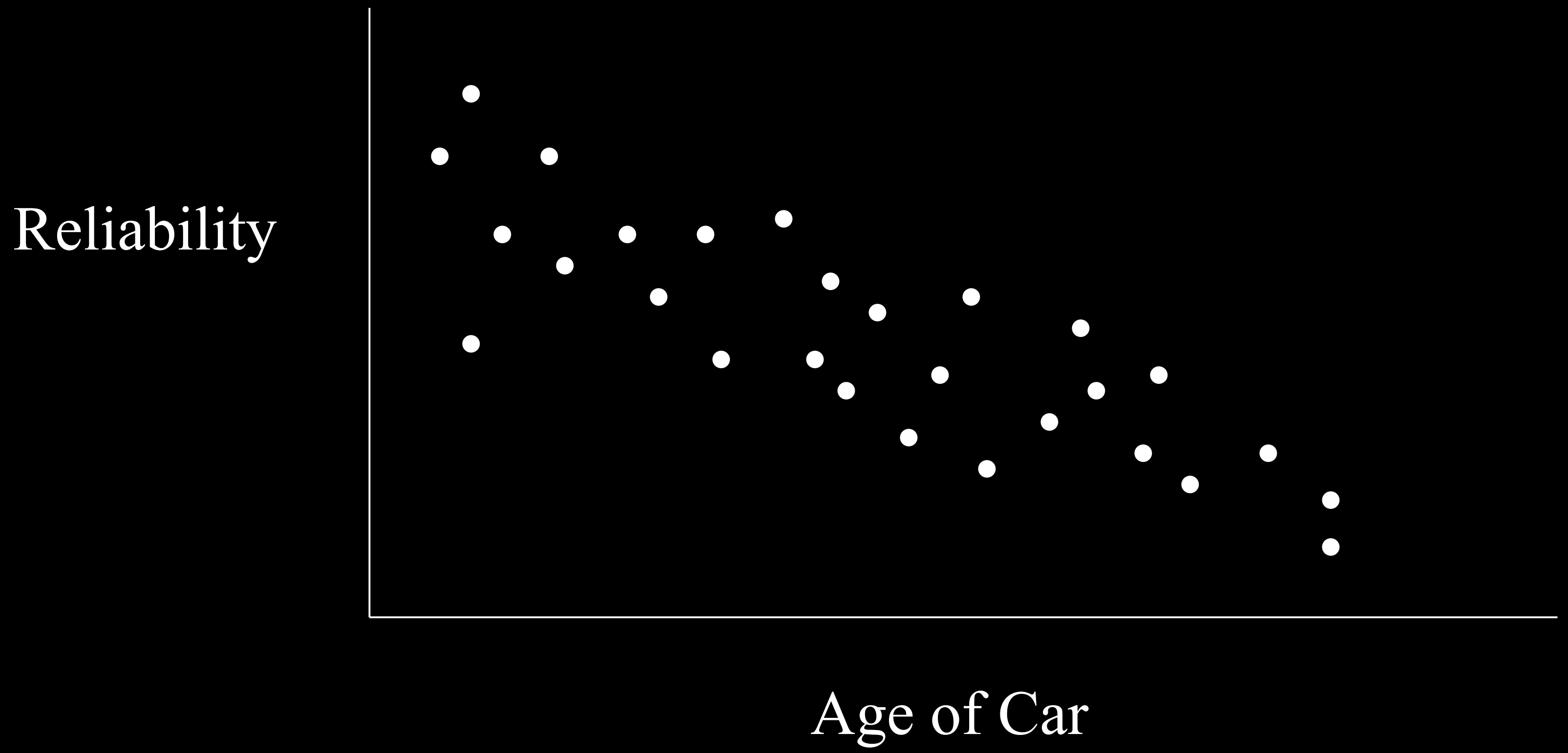
The pattern of data is indicative of the type of relationship between your two variables:

- **Positive relationship**
- **Negative relationship**
- **No relationship**

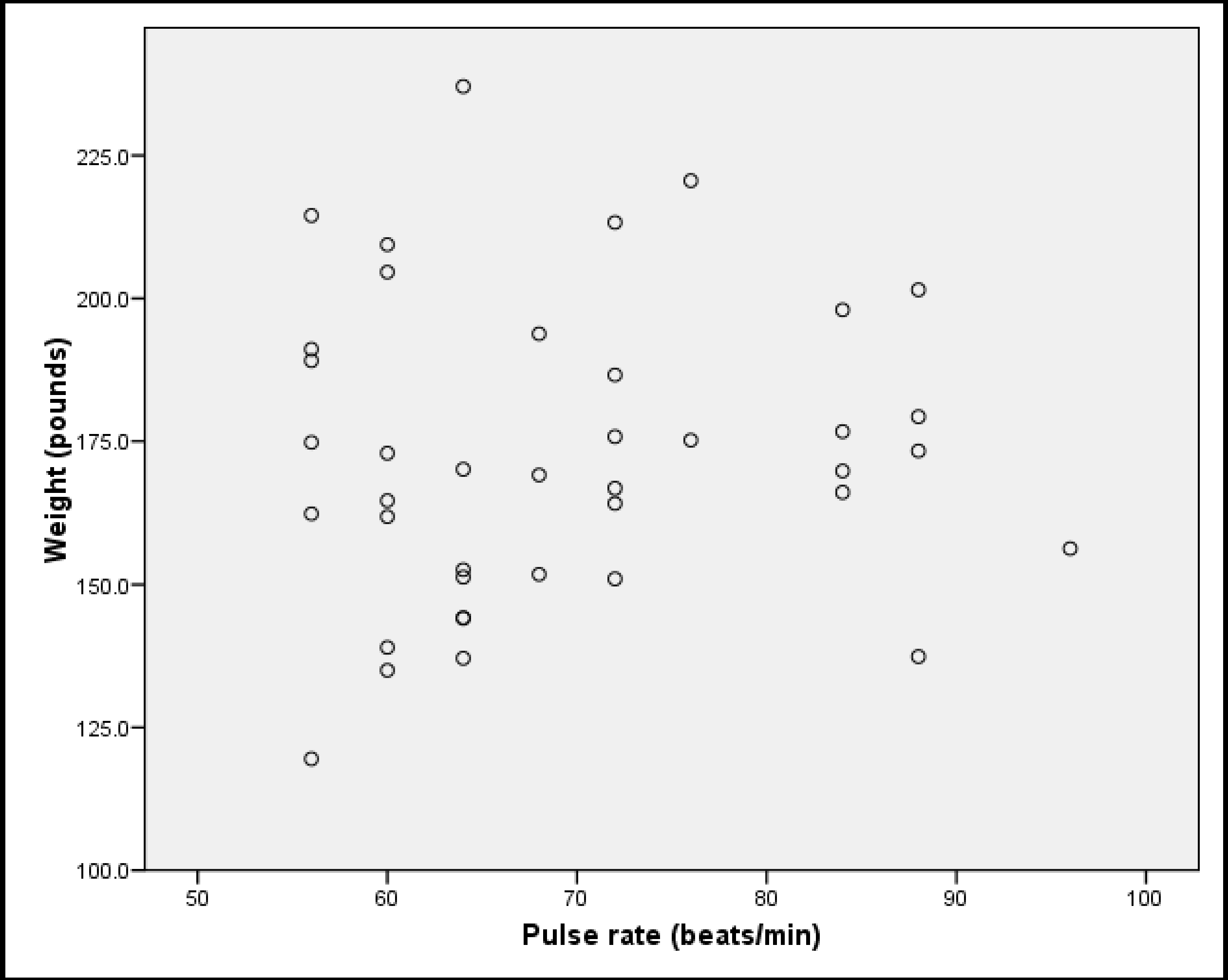
Positive Relationship



Negative Relationship



No Relationship



Simple Correlation coefficient (*r*)

- It is also called Karl Pearson's correlation coefficient.
- It measures the **nature** and **strength** between two variables of the quantitative type.

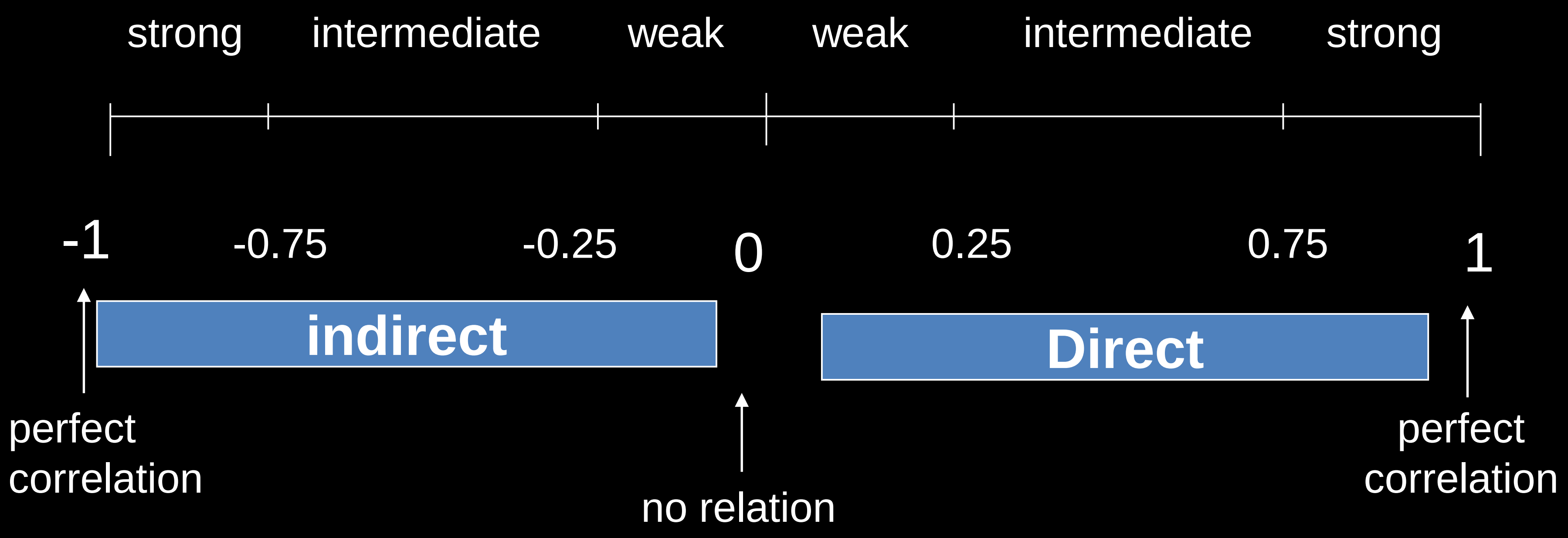
- The **sign** of r denotes the nature of association between X and Y.

- The **value** of r denotes the strength of association between X and Y.

- If the sign is **+ve** this means the relation is **direct relationship** (an increase in one variable is associated with an increase in the other variable and a decrease in one variable is associated with a decrease in the other variable).

- While if the sign is **-ve** this means an **inverse** or **indirect relationship** (which means an increase in one variable is associated with a decrease in the other).

- The value of ***r*** ranges between **-1** and **+1**
- The value of ***r*** denotes the strength of the association as illustrated by the following diagram.



- **If r = Zero this means no association or correlation between the two variables.**

- **If $0 < r < 0.25$, weak correlation.**

- **If $0.25 \leq r < 0.75$, intermediate correlation.**

- **If $0.75 \leq r < 1$, strong correlation.**

- **If $r = 1$, perfect correlation (Direct).**
- **If $r = -1$, perfect correlation (Indirect).**

Covariance

- $Var(X) = E[X - E(X)]^2$, it measures the variations of the random variable X from its mean value $E(X)$.
- Likewise, $Covariace$ of X and Y measures the simultaneous variations of the two random variables X and Y from their respective means.
- Its is denoted by $Cov(X, Y)$.
- $Cov(X, Y) = E\big[\big(X - E(X)\big)\big(Y - E(Y)\big)\big]$

$$= E(XY) - E(X)E(Y)$$

- $Cov(X, Y) = 0$, if X and Y are independent random variables.

Computation of correlation coefficient (r)

$$r = \frac{\mathrm{Cov}(x, y)}{\sigma_x \sigma_y}.$$

$$r = \frac{E(xy) - E(x) \cdot E(y)}{\sigma_x \cdot \sigma_y}.$$

$$r = \frac{\Sigma xy/n - (\Sigma x/n)(\Sigma y/n)}{\sqrt{\frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2} \times \sqrt{\frac{\Sigma y^2}{n} - \left(\frac{\Sigma y}{n}\right)^2}}.$$

Problem No 1:

Find the coefficient of Correlation between X and Y using the following data:

X :	5	10	15	20	25
Y :	16	19	23	26	30

Correlation

X	5	10	15	20	25
y	16	19	23	26	30

Co.eff. of Correlation,

$$r = \frac{\Sigma xy/n - [(\Sigma x/n)(\Sigma y/n)]}{\sigma_x \cdot \sigma_y}$$

W.K.T, $\sigma_x = \sqrt{\frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2}$

$$\sigma_y = \sqrt{\frac{\Sigma y^2}{n} - \left(\frac{\Sigma y}{n}\right)^2}$$

x	y	x^2	y^2	xy
5	16			
10	19			
15	23			
20	26			
25	30			
Σx	Σy	Σx^2	Σy^2	Σxy

Here, $\boxed{n = 5}$

$$r = \frac{(\Sigma xy/5) - [(\Sigma x/5)(\Sigma y/5)]}{\sigma_x \cdot \sigma_y}$$

$\boxed{r = 0.9907}$, Strong relationship.

Rank Correlation Coefficient (r_s)

Spearman Rank Correlation Coefficient (r_s)

- It is a non-parametric measure of correlation.
- This procedure makes use of the two sets of ranks that may be assigned to the sample values of X and Y.
- Spearman rank correlation coefficient could be computed when both variables are qualitative or quantitative.

Procedure:

1. Rank the values of X from 1 to n.
2. Rank the values of Y from 1 to n.

Note: where n is the numbers of pairs of values of X and Y in the sample.

3. Compute the value of d for each pair of observation by subtracting the rank of Y from the rank of X.
4. Square each d and compute $\sum d^2$ which is the sum of the squared values.

5. Apply the following formula

$$r_s = 1 - \frac{6\sum d^2}{n(n^2 - 1)}$$

For non repeated ranking

$$r_s = 1 - \frac{6\left[\sum d^2 + \sum (m^3 - m)/12\right]}{n(n^2 - 1)}$$

For repeated ranking

- The value of r_s denotes the magnitude and nature of association giving the same interpretation as simple r.

Problems

Spearman's rank correlation (r_s).

Problem:

1). Ten students got the following percentage of marks in Maths and Physical sciences.

Student	1	2	3	4	5	6	7	8	9	10
Maths	78	36	98	25	75	82	90	62	65	39
Physics	84	51	91	60	68	62	86	58	63	47

Calculate the rank correlation co-eff.

Sol:

Maths (x)	78	36	98	25	75	82	90	62	65	39
R_x	4	9	1	10	5	3	2	7	6	8
Physics (y)	84	51	91	60	68	62	86	58	63	47
R_y	3	9	1	7	4	6	2	8	5	10
$d^2 = (R_x - R_y)^2$	1	0	0	9	1	9	0	1	1	4

Now, $\Sigma d^2 = 1+0+0+9+1+9+0+1+1+4$

$$\boxed{\Sigma d^2 = 26}$$

Spearman's rank correl. coeff. is,

$$r_s = 1 - \frac{6 \Sigma d^2}{n^3 - n}$$

Here, $n = 10$ → no. of students.

$$= 1 - \frac{6(26)}{10^3 - 10}$$

$$\boxed{r_s = 0.8424}$$

y.

X

$R_x \rightarrow$ Rank of x.

$R_y \rightarrow$ Rank of y.

2). Calculate the rank correlation for the following 7 observations of X and Y.

X	73.2	85.8	78.9	75.8	77.2	81.2	83.8
Y	97.8	99.2	98.8	98.3	98.3	96.7	97.8

Sol:

X	73.2	85.8	78.9	75.8	77.2	81.2	83.8
R_x	7	1	4	6	5	3	2
Y	97.8	99.2	98.8	98.3	98.3	96.7	97.8
R_y	5.5	1	2	3.5	3.5	7	5.5
d^2 $(R_x-R_y)^2$	2.25	0	4	6.25	2.25	16	12.25

Now, $\Sigma d^2 = 2.25 + 0 + 4 + 6.25 + 2.25 + 16 + 12.25$

$$\boxed{\Sigma d^2 = 43}$$

Note: For 98.3, we take the average rank b/w them.
(i.e) $\frac{3+4}{2} = 3.5$

Spearman's rank correlation coeff is,

$$r_s = 1 - \frac{6\left[\Sigma d^2 + \Sigma (m^3 - m)/12\right]}{n^3 - n}.$$

$\boxed{n = 7}$

Here, in $Y \rightarrow$ 97.8 is repeated twice.

$\rightarrow$ 98.3 is repeated twice.

$$\therefore r_s = 1 - \frac{6\left[43 + (2^3 - 2)/12 + (2^3 - 2)/12\right]}{7^3 - 7}.$$

$$\boxed{r_s = 0.2142 \text{ (approx).}}$$

4.

Problem 3:

In a study of the relationship between level education and income the following data was obtained. Find the relationship between them and comment.

Sample number	level education (X)	Income (Y)
A	Preparatory	25
B	Primary	10
C	University	8
D	Secondary	10
E	Secondary	15
F	Illiterate	50
G	University	60

Answer:

	(X)	(Y)	Rank X	Rank Y	d	d^2
A	Preparatory	25	5	3	2	4
B	Primary	10	6	5.5	0.5	0.25
C	University	8	1.5	7	-5.5	30.25
D	Secondary	10	3.5	5.5	-2	4
E	Secondary	15	3.5	4	-0.5	0.25
F	Illiterate	50	7	2	5	25
G	University	60	1.5	1	0.5	0.25

$$r_s = 1 - \frac{6\left[\sum d^2 + \sum (m^3 - m)/12\right]}{n(n^2 - 1)} = -0.1$$

Comment: There is an indirect weak correlation between the level of education and income.

Regression Analysis

- **Regression:** Technique concerned with predicting some variables by knowing others.

- The process of predicting variable Y using variable X or vice versa.

Regression

- Uses a variable (x) to predict some outcome variable (y)
- Tells you how values in y change as a function of changes in values of x

Correlation and Regression

- Correlation describes the strength of a **linear relationship** between two variables.

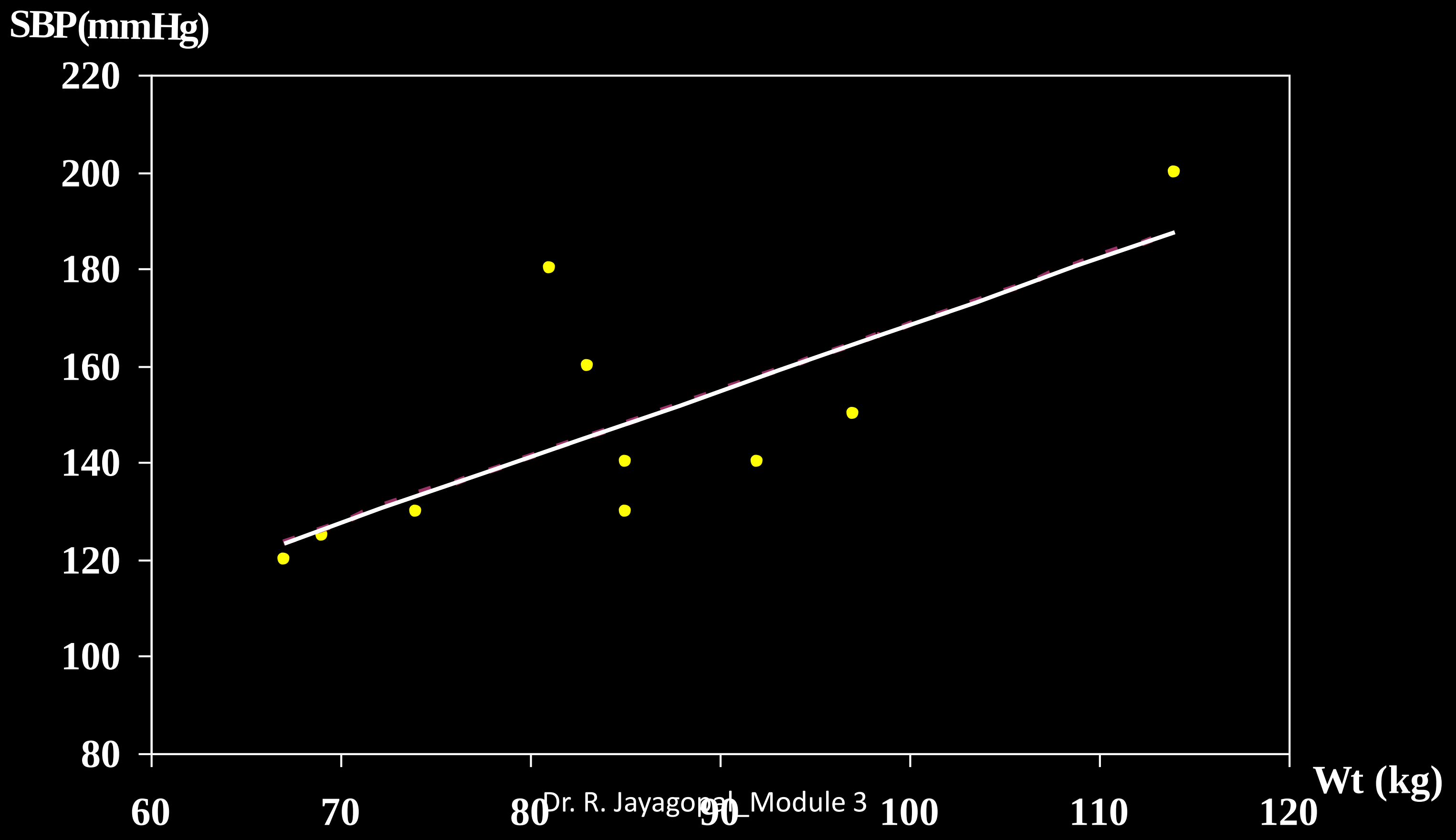
- **Linear means "straight line".**

- **Regression** tells us how to draw the straight line described by the correlation.

Regression

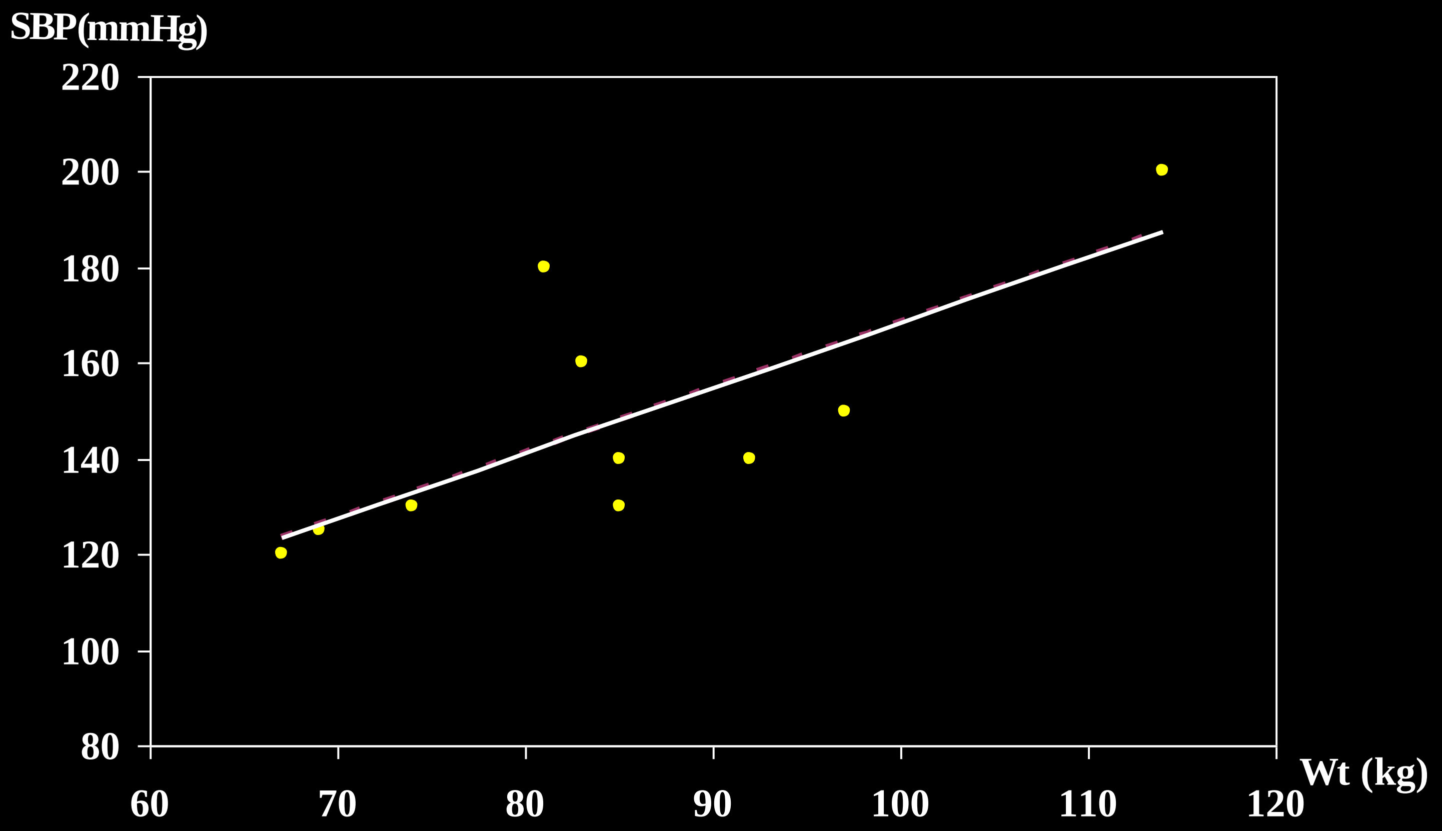
- Calculates the “best-fit” line for a certain set of data
- The regression line makes the sum of the squares of the residuals smaller than for any other line.

Regression minimizes residuals

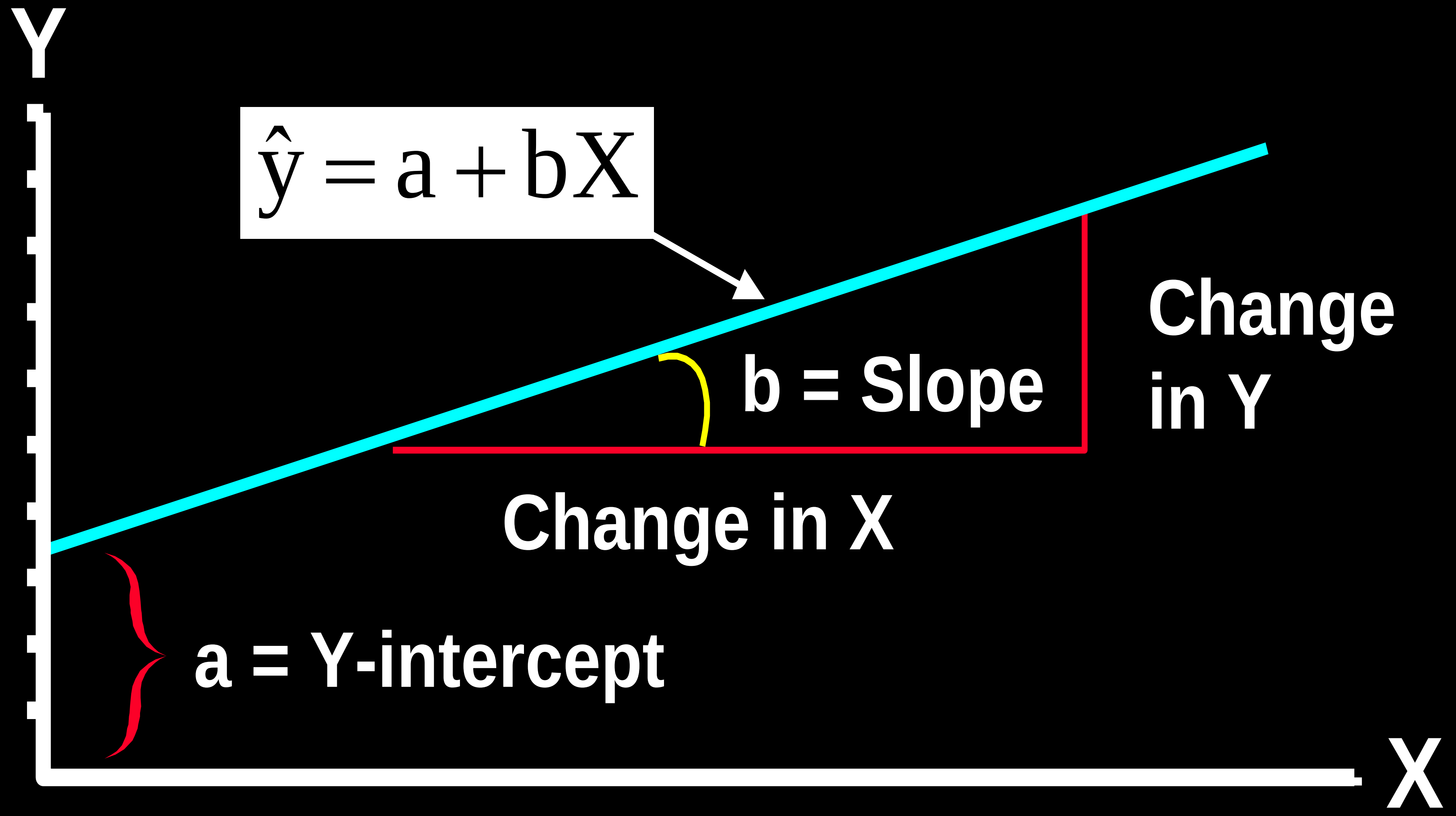


Regression Equation

- Regression equation describes the regression line mathematically – **Intercept and Slope**



Linear Equations



Regression Lines – $(\bar{x}, \bar{y})$ and Slope

1. Equation of straight line ***y* on *x***:

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

where, $b_{yx} = r\frac{\sigma_y}{\sigma_x}$

2. Equation of straight line ***x* on *y***:

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

where, $b_{xy} = r\frac{\sigma_x}{\sigma_y}$

Note: $b_{xy} \neq b_{yx}$

Regression Lines – Least square method

Consider the equation of a straight line ***y* on *x***:

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

Normal equations are

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

$$\sum yx = \boldsymbol{b}\sum x^2 + \boldsymbol{a}\sum x$$

Solving the above two equations we get the values of $\boldsymbol{a}$ and $\boldsymbol{b}$

$$\boxed{y = \boldsymbol{b}x + \boldsymbol{a}}$$

Regression Lines – Least square method

Consider the equation of a straight line ***x* on *y***:

$$x = \boldsymbol{d}y + \boldsymbol{c}$$

Normal equations are

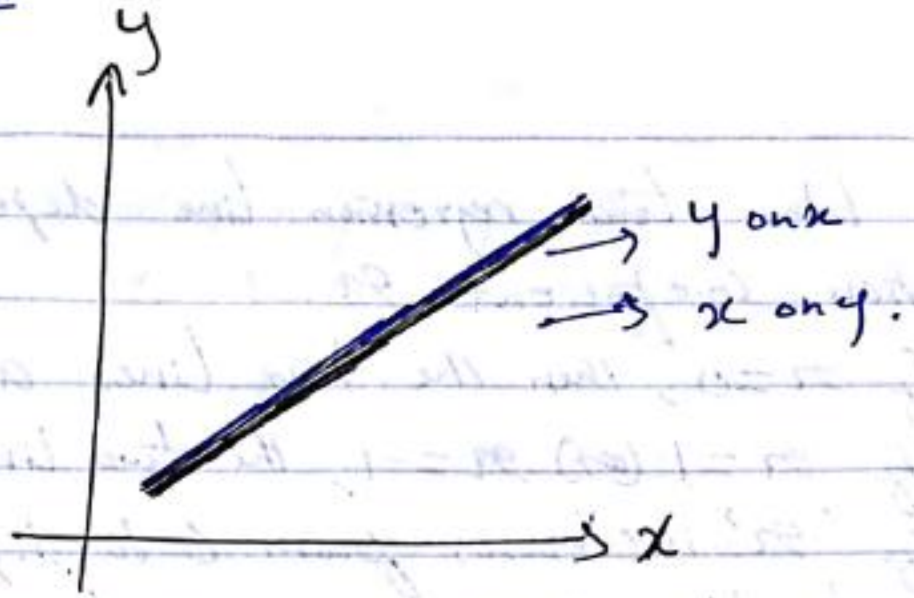
$$\sum x = \boldsymbol{d}\sum y + \boldsymbol{c}n$$

$$\sum xy = \boldsymbol{d}\sum y^2 + \boldsymbol{c}\sum y$$

Solving the above two equations we get the values of $\boldsymbol{c}$ and $\boldsymbol{d}$

$$\boxed{x = \boldsymbol{d}y + \boldsymbol{c}}$$

(8) Two lines coincide:

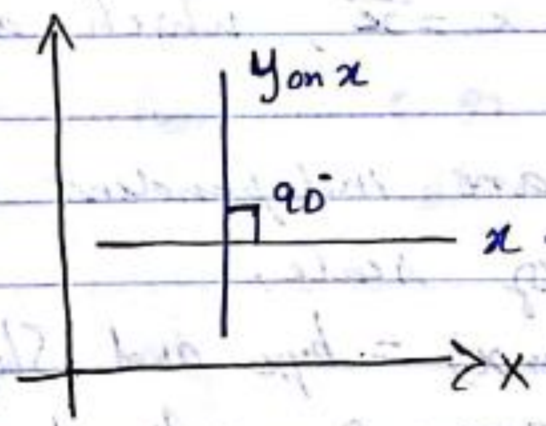


Hence, $r = 1 \Rightarrow \tan\theta = 0.$

(or) $\therefore \theta = \tan^{-1}(0)$

$r = -1$ $\boxed{\theta = 0.}$

(4). Two lines are at right angle (=90°):



Here, $r = 0 \Rightarrow \tan\theta = \infty$

$\theta = \tan^{-1}(\infty)$

$\boxed{\theta = 90^\circ}$

Note:

If θ approches from 90° to 0° then it means two regr. lines are getting closer and closer.

Note:-

1) $m \rightarrow$ Slope of Str. Line.
(i-e) It gives the ratio btwn the change in y and change in x.

2) Here, b_{yx} is the slope and hence it gives the ratio (or) relationship btwn the change in y and change in x.

$\therefore$ We call, b_{yx} to be the regr. Co-eff.

Properties:

1) $r = \sqrt{b_{yx} \cdot b_{xy}}$. (i.e) r is the G.M btwn b_{yx} & b_{xy}.

(Correlation coefficient)

2) $\frac{b_{yx}}{b_{xy}} = \frac{\sigma_y^2}{\sigma_x^2}$.

3). If $b_{yx} > 1$ then $b_{xy} < 1$. (vice-versa).

4). $\left| \frac{b_{yx} + b_{xy}}{2} \right| \geq |r|$.

5) Points of intersection of the two regression lines is the point whose co-ordinates are $(\bar{x}, \bar{y})$.

6) Angle btwn two regression line depends on the Correlation co-efficient r.

→ If $r=0$, then the two lines are $\perp^{er}$.

→ If $r=1$ (or) $r=-1$, the two lines coincide.

→ If 'r' increases from 0 to 1, then the angle btwn the regression lines diminishes from 90° to 0°.

→ when $r=\pm 1$, two lines becomes identical. Thus there is an exact linear relationship btwn the variables.

→ when $r=0$, regr. eqn reduces to $y=\bar{y}$ and $x=\bar{x}$, which are $\perp^{er}$ to each other.

7). They are independent of origin, but not of scale.