

# Correlation

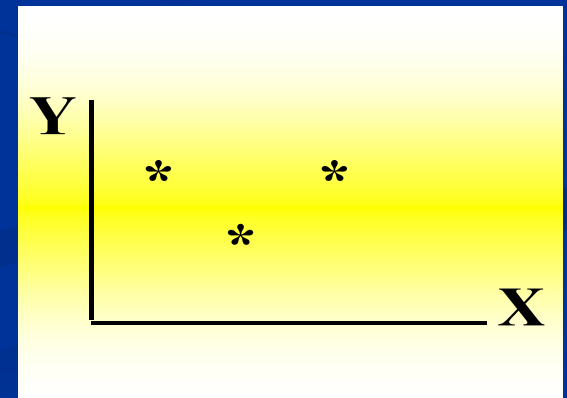
# Correlation

Finding the relationship between two quantitative variables without being able to infer causal relationships

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

# Scatter diagram

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

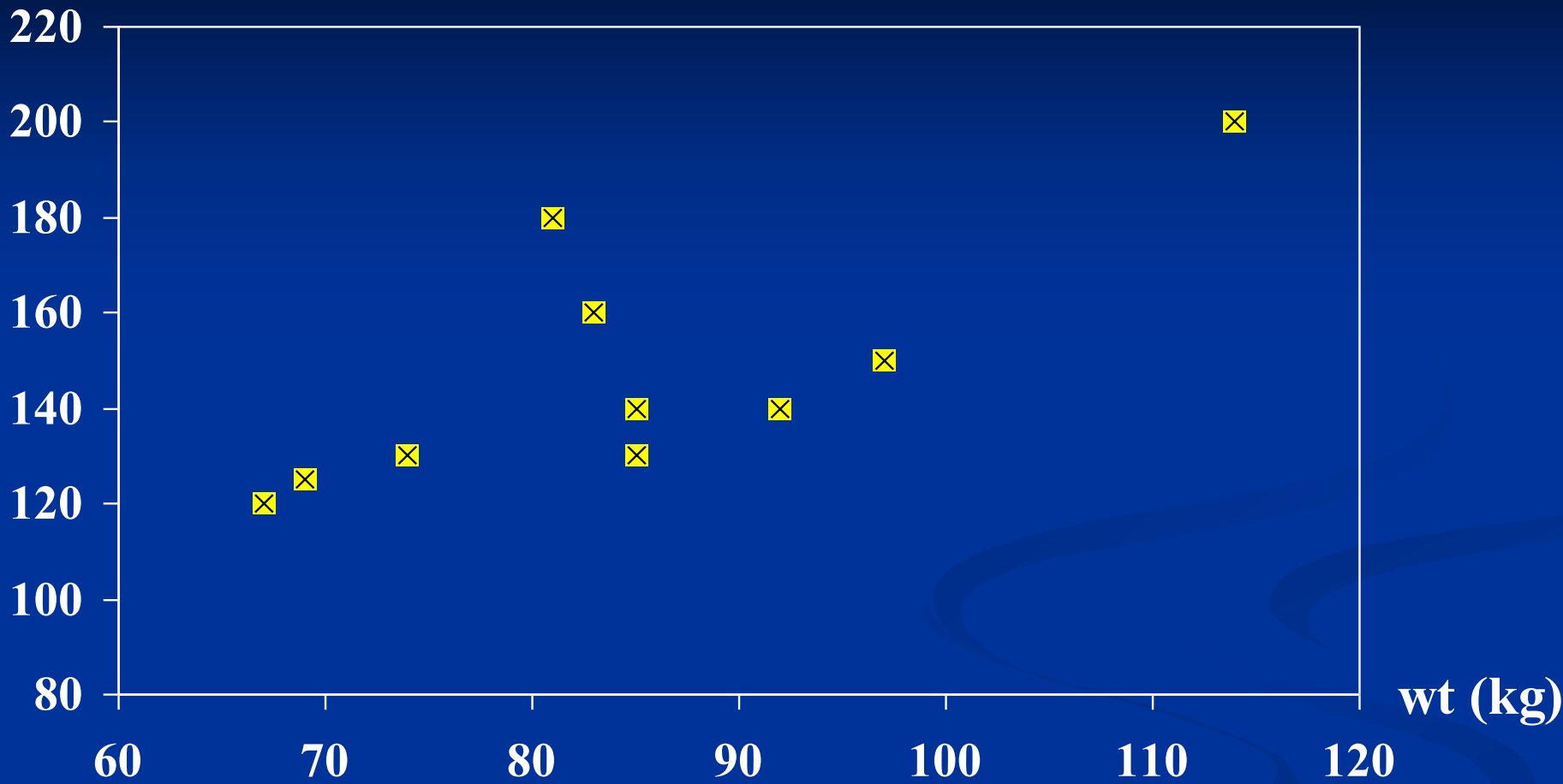


# Example

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

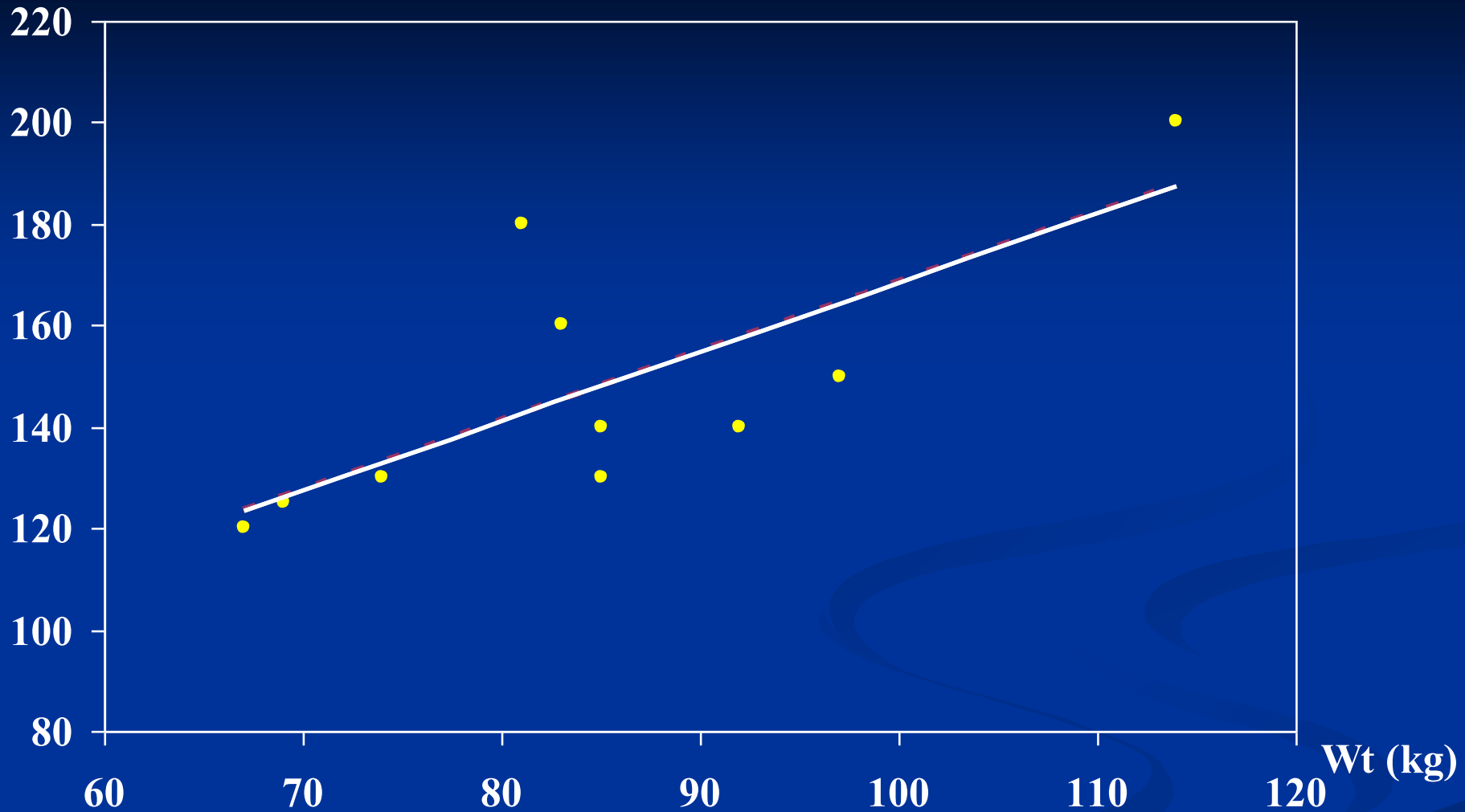
SBP(mmHg)

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



Scatter diagram of weight and systolic blood pressure

**SBP(mmHg)**



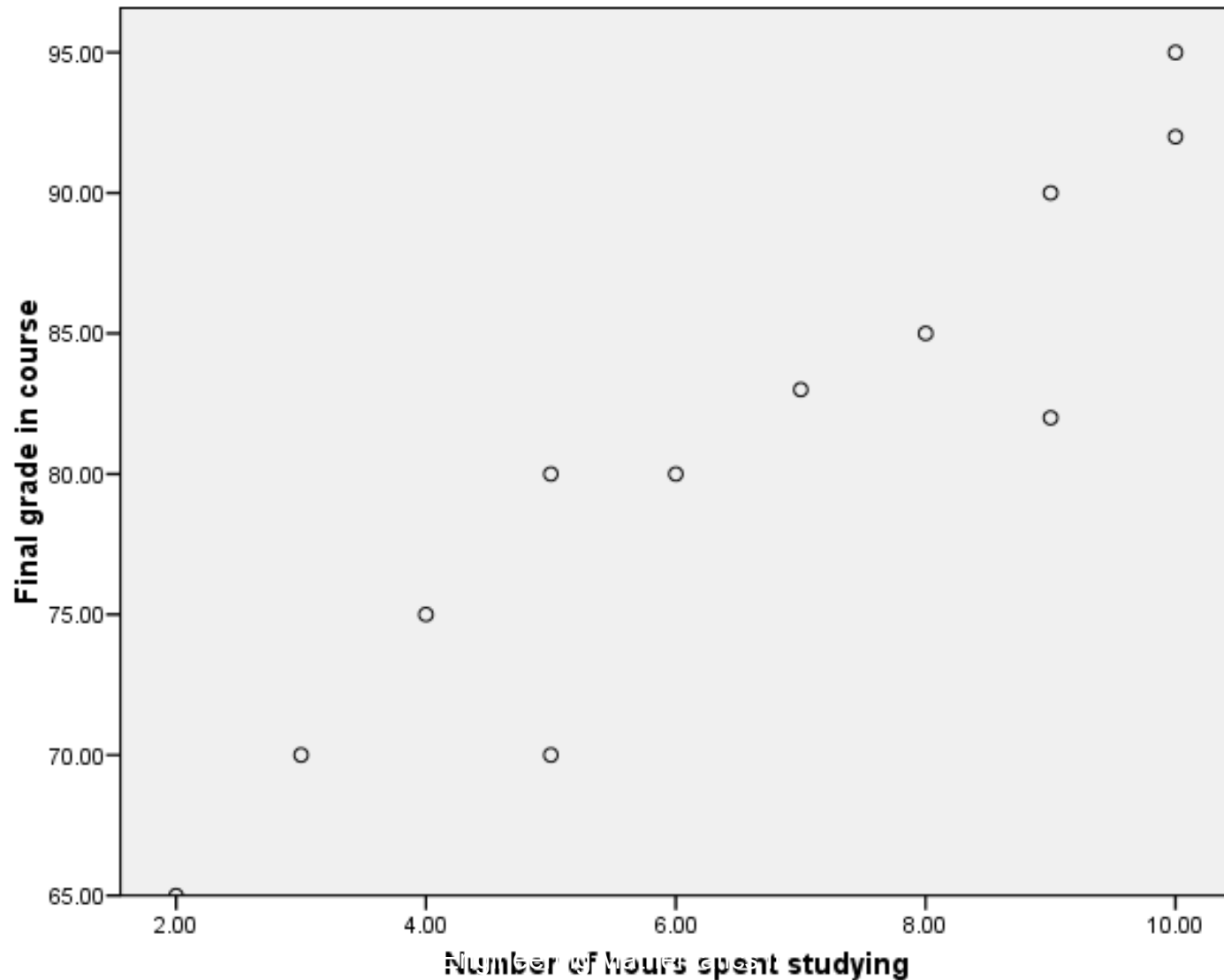
**Scatter diagram of weight and systolic blood pressure**

# Scatter plots

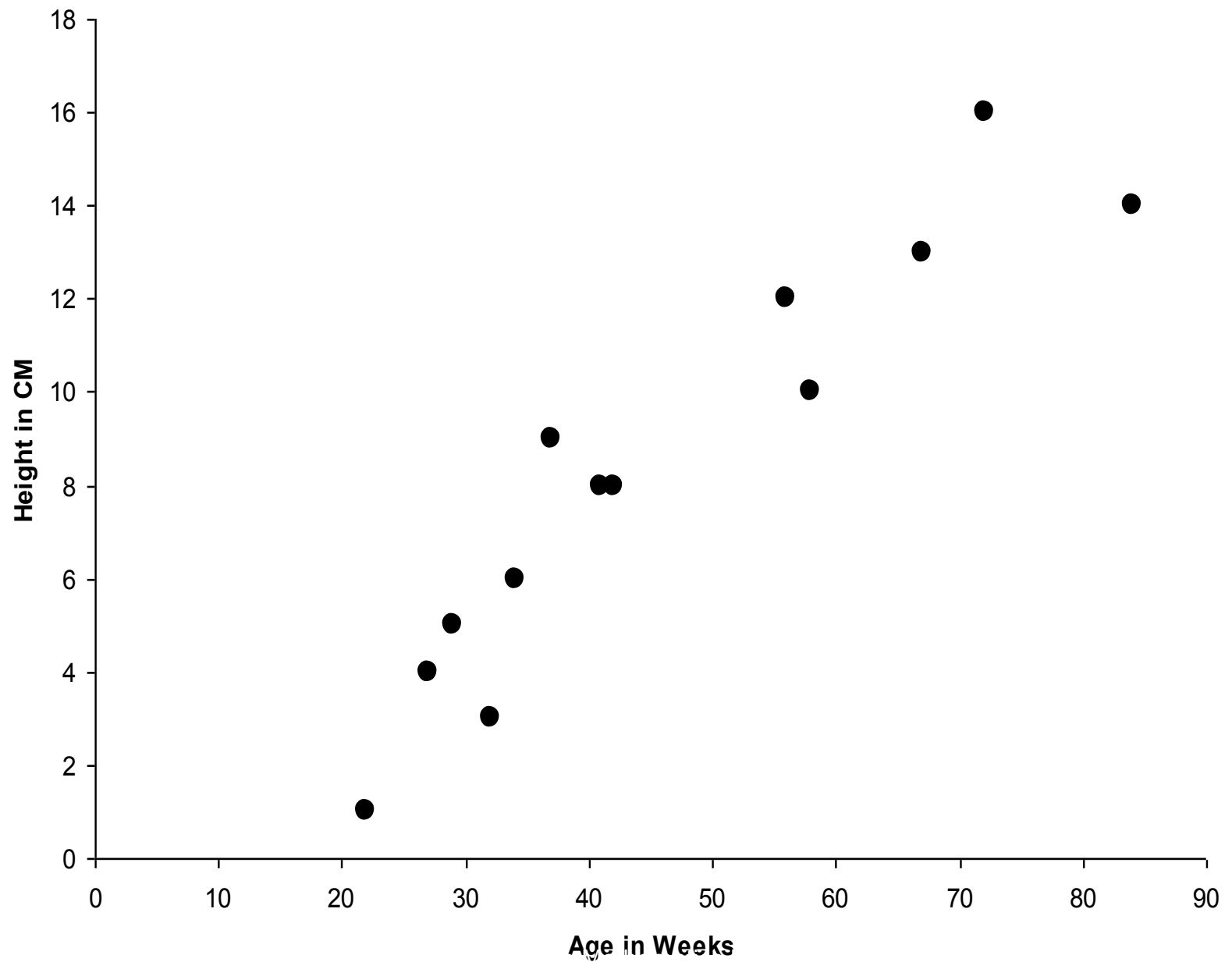
**The pattern of data is indicative of the type of relationship between your two variables:**

- positive relationship
- negative relationship
- no relationship

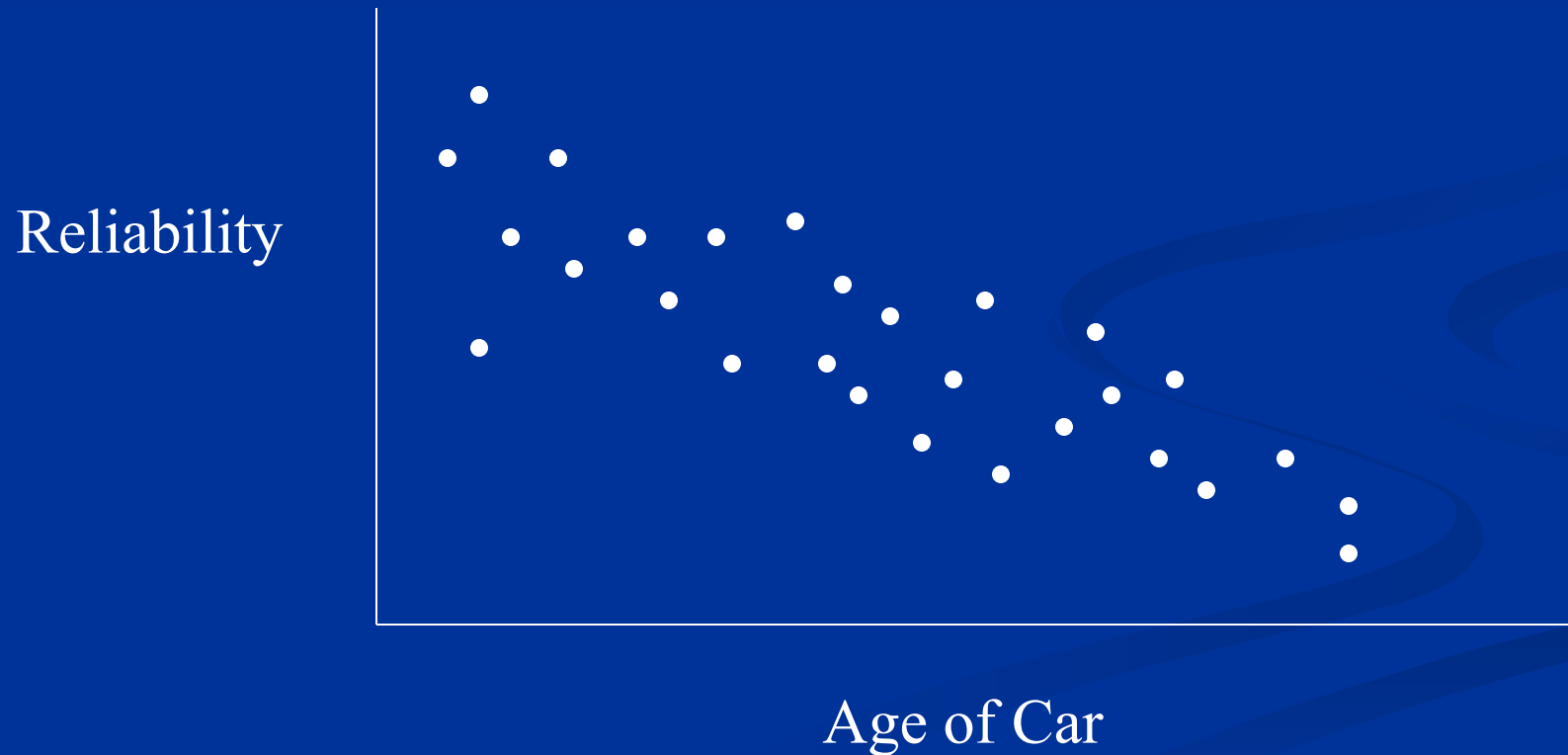
# Positive relationship



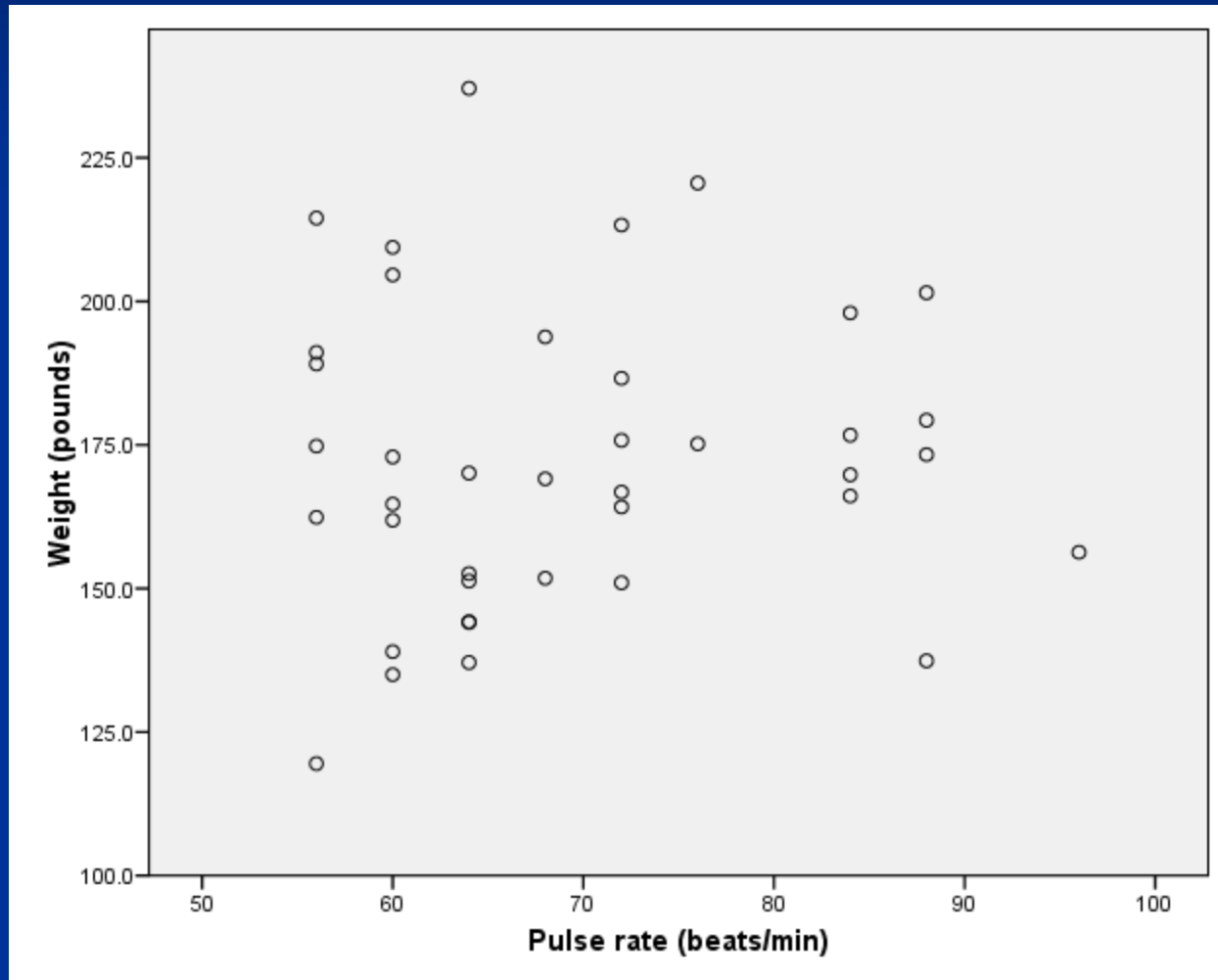




# Negative relationship



# No relation



# Correlation coefficient ( $r$ )

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

- The sign of  $r$  denotes the nature of association
- while the value of  $r$  denotes the strength of association.

- If the sign is +ve this means the relation is direct (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  $r = \text{Zero}$  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.



# How to compute the simple correlation coefficient (r)

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left(\sum x^2 - \frac{(\sum x)^2}{n}\right) \left(\sum y^2 - \frac{(\sum y)^2}{n}\right)}}$$

## Example:

A sample of 6 children was selected, data about their age in years and weight in kilograms was recorded as shown in the following table . It is required to find the correlation between age and weight.

serial No	Age (years)	Weight (Kg)
1	7	12
2	6	8
3	8	12
4	5	10
5	6	11
6	9	13

These 2 variables are of the quantitative type, one variable (Age) is called the independent and denoted as (X) variable and the other (weight) is called the dependent and denoted as (Y) variables to find the relation between age and weight compute the simple correlation coefficient using the following formula:

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left(\sum x^2 - \frac{(\sum x)^2}{n}\right) \left(\sum y^2 - \frac{(\sum y)^2}{n}\right)}}$$

<b>Serial n.</b>	<b>Age (years) (x)</b>	<b>Weight (Kg) (y)</b>	<b>xy</b>	<b>x<sup>2</sup></b>	<b>y<sup>2</sup></b>
<b>1</b>	<b>7</b>	<b>12</b>	<b>84</b>	<b>49</b>	<b>144</b>
<b>2</b>	<b>6</b>	<b>8</b>	<b>48</b>	<b>36</b>	<b>64</b>
<b>3</b>	<b>8</b>	<b>12</b>	<b>96</b>	<b>64</b>	<b>144</b>
<b>4</b>	<b>5</b>	<b>10</b>	<b>50</b>	<b>25</b>	<b>100</b>
<b>5</b>	<b>6</b>	<b>11</b>	<b>66</b>	<b>36</b>	<b>121</b>
<b>6</b>	<b>9</b>	<b>13</b>	<b>117</b>	<b>81</b>	<b>169</b>
<b>Total</b>	<b><math>\sum x =</math> 41</b>	<b><math>\sum y =</math> 66</b>	<b><math>\sum xy =</math> 461</b>	<b><math>\sum x^2 =</math> 291</b>	<b><math>\sum y^2 =</math> 742</b>

$$r = \frac{461 - \frac{41 \times 66}{6}}{\sqrt{\left[291 - \frac{(41)^2}{6}\right] \cdot \left[742 - \frac{(66)^2}{6}\right]}}$$

$$r = 0.759$$

strong direct correlation

## EXAMPLE: Relationship between Anxiety and Test Scores

Anxiety (X)	Test score (Y)	$X^2$	$Y^2$	$XY$
10	2	100	4	20
8	3	64	9	24
2	9	4	81	18
1	7	1	49	7
5	6	25	36	30
6	5	36	25	30
$\sum X = 32$	$\sum Y = 32$	$\sum X^2 = 230$	$\sum Y^2 = 204$	$\sum XY = 129$

## Calculating Correlation Coefficient

$$r = \frac{(6)(129) - (32)(32)}{\sqrt{(6(230) - 32^2)(6(204) - 32^2)}} = \frac{774 - 1024}{\sqrt{(356)(200)}} = -.94$$

$$r = -0.94$$

**Indirect strong correlation**

# ***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 in the following cases:
  - Both variables are quantitative.
  - Both variables are qualitative ordinal.
  - One variable is quantitative and the other is qualitative ordinal.



## Procedure:

1. Rank the values of  $X$  from 1 to  $n$  where  $n$  is the numbers of pairs of values of  $X$  and  $Y$  in the sample.
2. Rank the values of  $Y$  from 1 to  $n$ .
3. Compute the value of  $d_i$  for each pair of observation by subtracting the rank of  $Y_i$  from the rank of  $X_i$
4. Square each  $d_i$  and compute  $\sum d_i^2$  which is the sum of the squared values.

5. Apply the following formula

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

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

## Example

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

sample numbers	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	$d_i$	$d_i^2$
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

$$\sum d_i^2 = 64$$

$$r_s = 1 - \frac{6 \times 64}{7(48)} = -0.1$$

Comment:

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