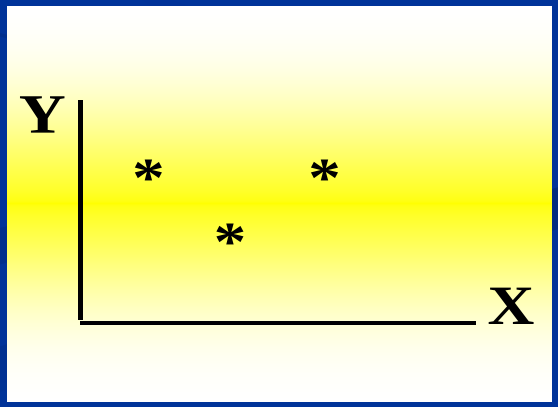
# Correlation & Regression

# Correlation

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

# Scatter diagram

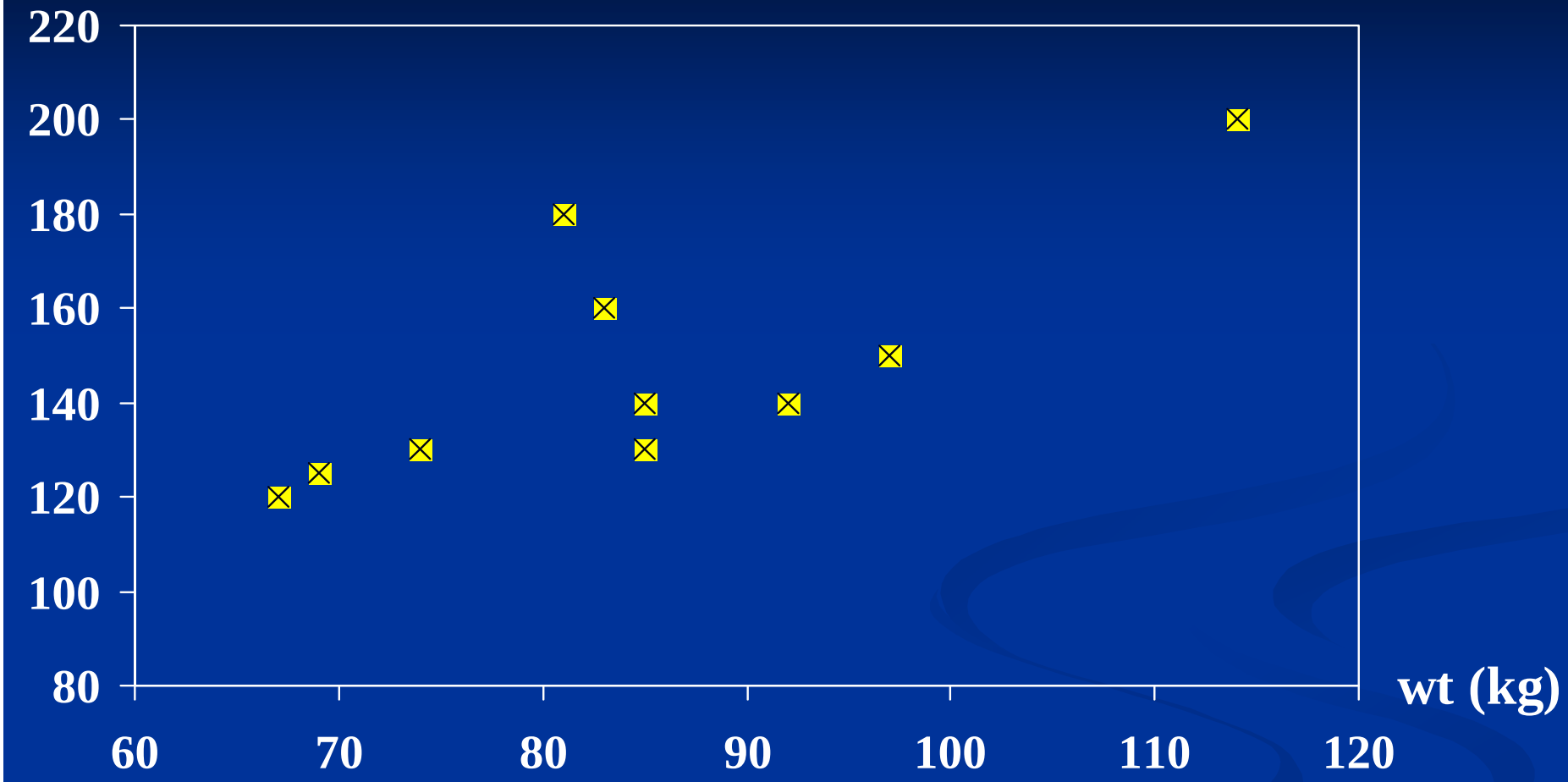
- Two quantitative variables
- One variable is called independent (X) and the second is called dependent (Y)
- Points are not joined



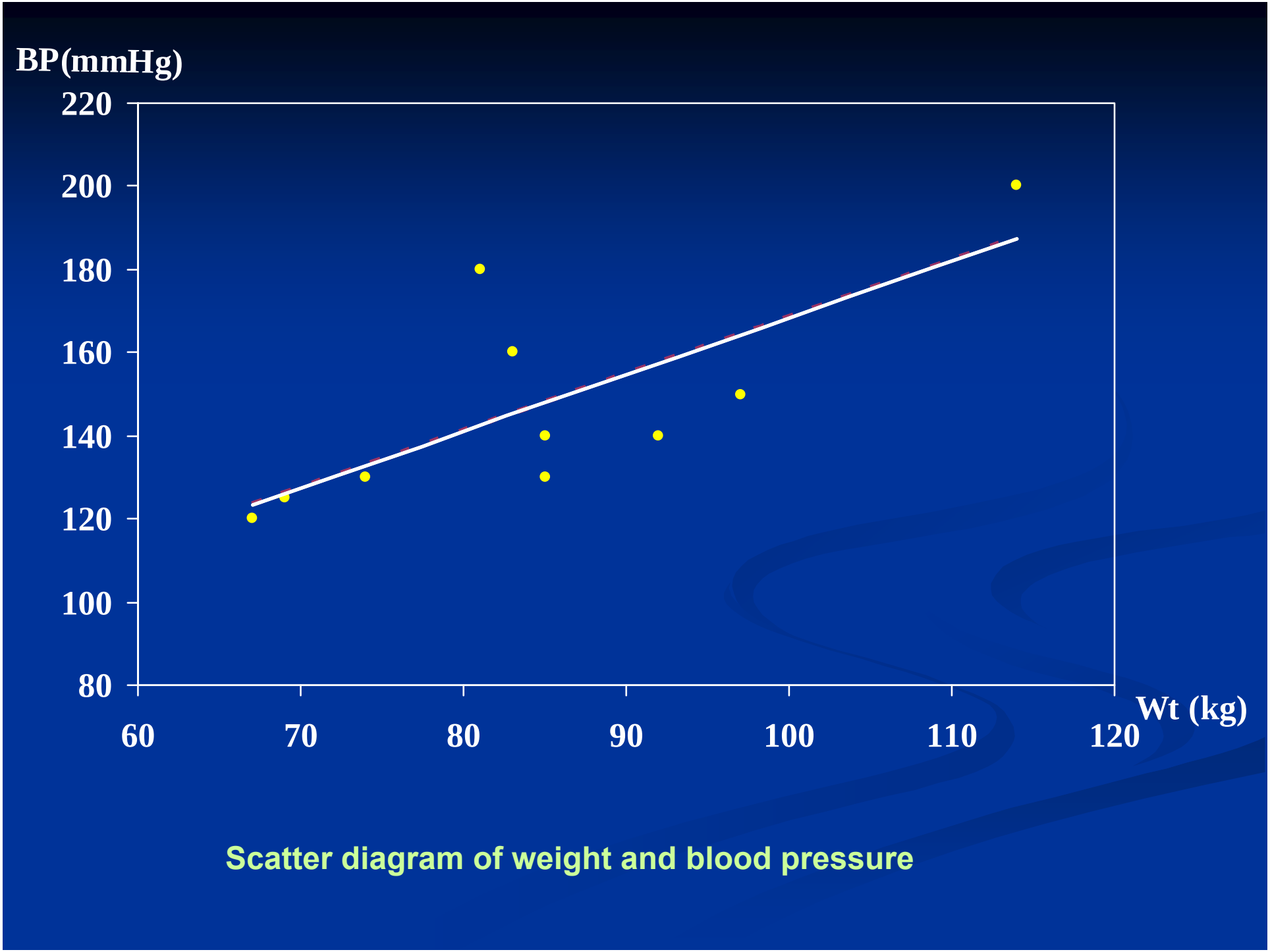
# Example

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Wt. (kg) | 67 | 69 | 85 | 83 | 74 | 81 | 97 | 92 | 114 | 85 |
| BP 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 blood pressure



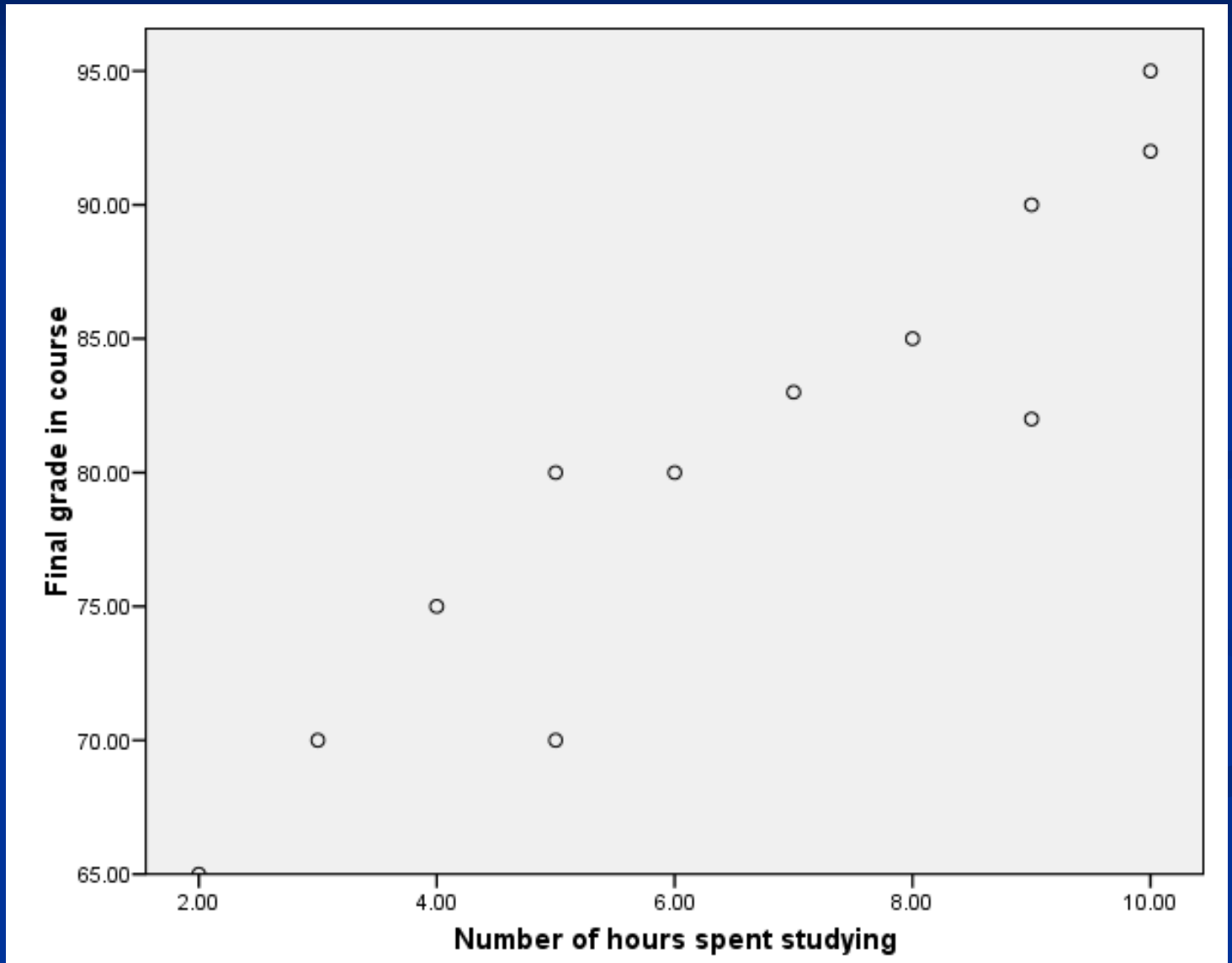

Scatter diagram of weight and blood pressure

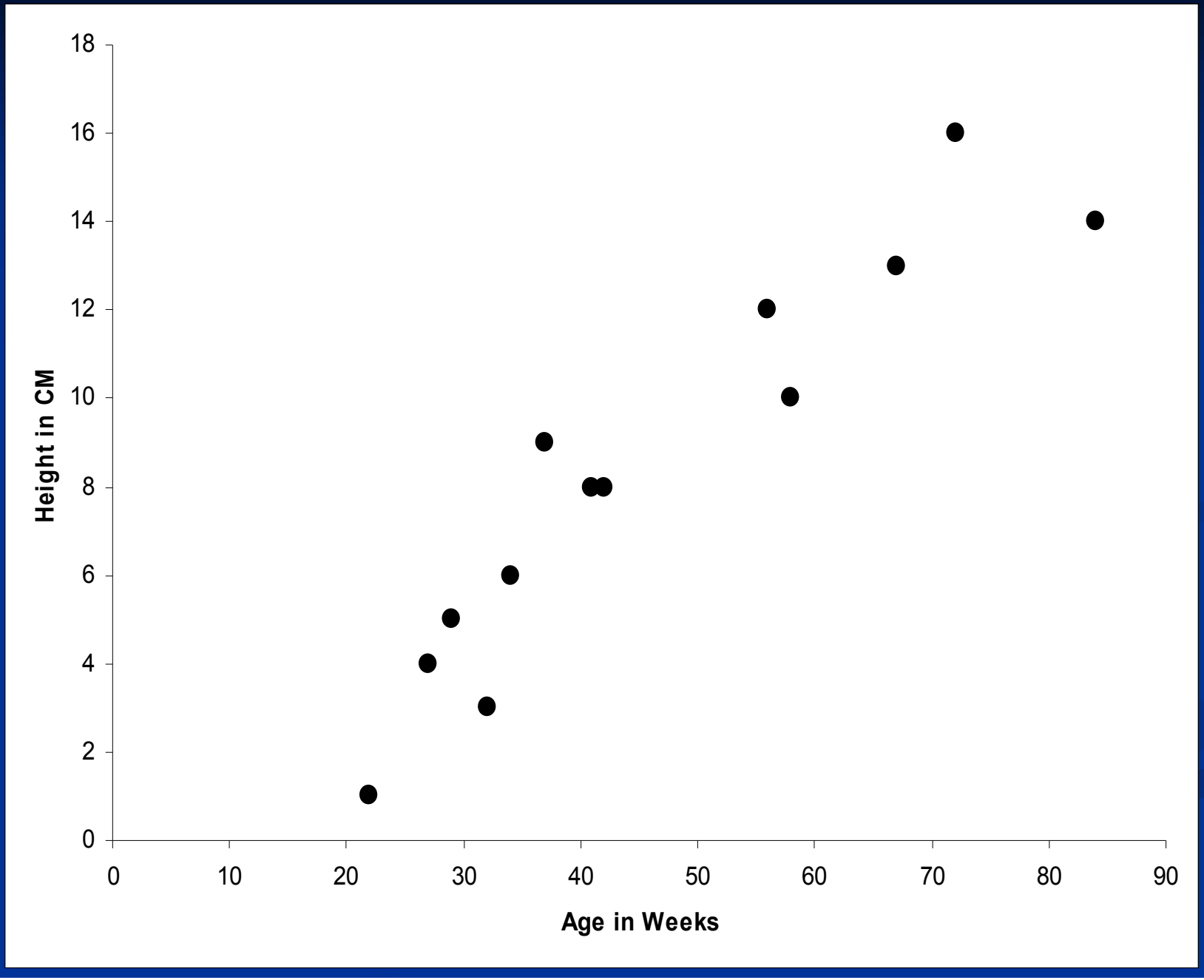
# Scatter plots

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

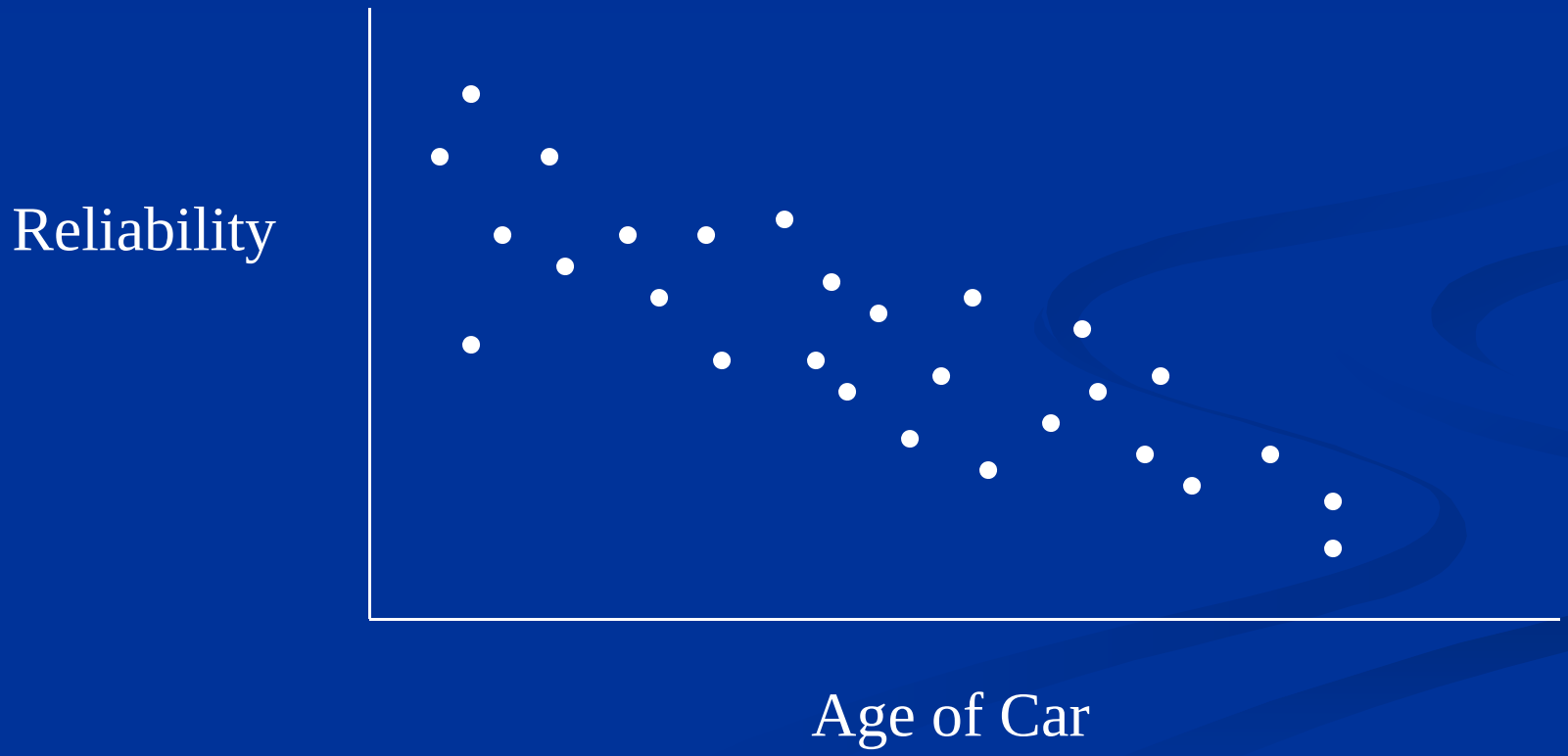
- positive relationship
- negative relationship
- no relationship

# Positive relationship

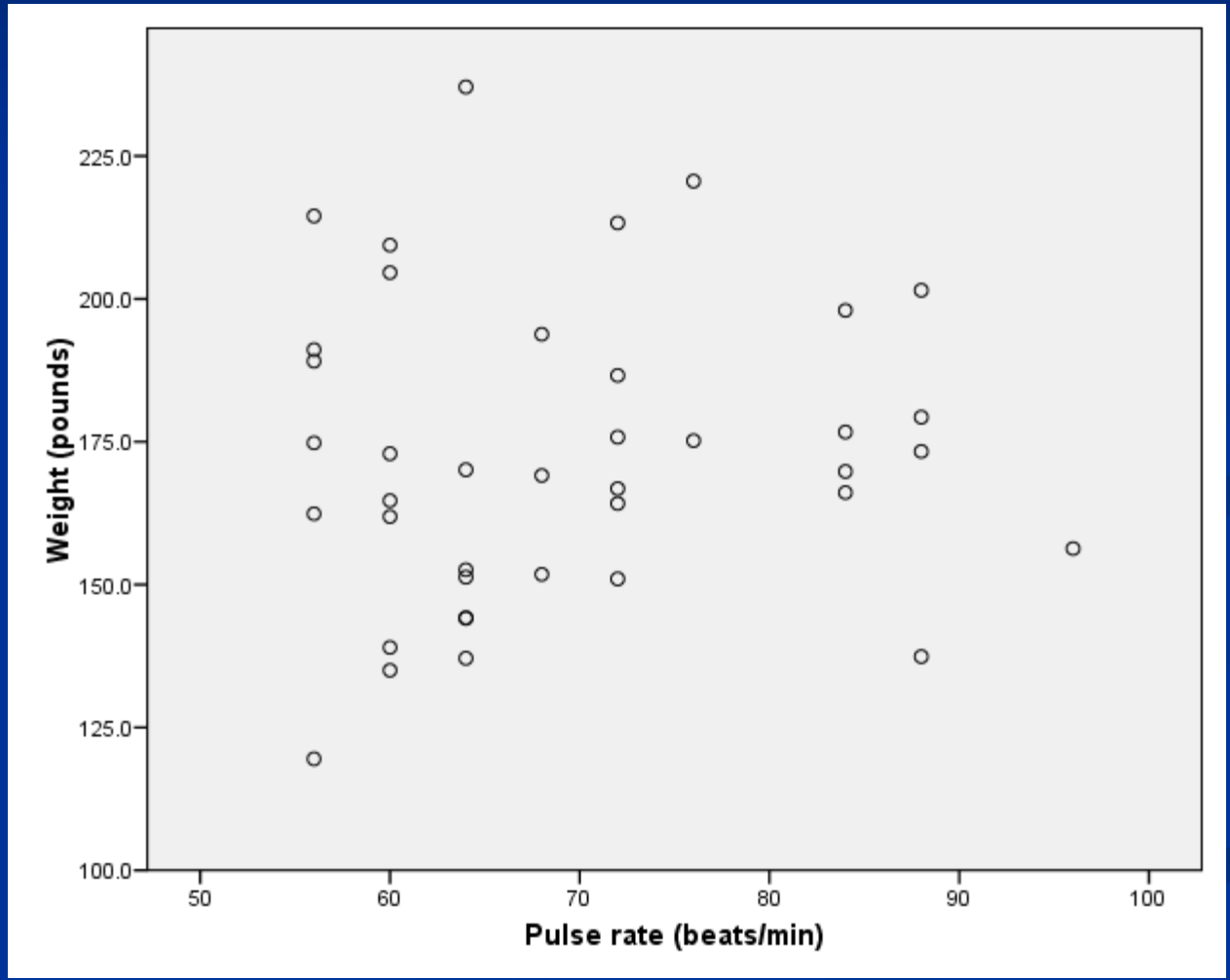



Height in CM
Age in Weeks
0
2
4
6
8
10
12
14
16
18
0
10
20
30
40
50
60
70
80
90

# Negative relationship



# No relation


Weight (pounds)
225.0
200.0
175.0
150.0
125.0
100.0
50
60
70
80
90
100
Pulse rate (beats/min)

# Correlation Coefficient

Statistic showing the degree of relation between two variables

# Simple Correlation coefficient (r)

- It is also called Pearson's correlation or product moment correlation coefficient.
- It measures the nature and strength of relationship 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 $r$ = 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)

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

## 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:

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

| Serial n. | Age Years(x) | Weight Kg(y) | xy | $X^2$ | $Y^2$ |
|---|---|---|---|---|---|
| 1 | 7 | 12 | 84 | 49 | 144 |
| 2 | 6 | 8 | 48 | 36 | 64 |
| 3 | 8 | 12 | 96 | 64 | 144 |
| 4 | 5 | 10 | 50 | 25 | 100 |
| 5 | 6 | 11 | 66 | 36 | 121 |
| 6 | 9 | 13 | 117 | 81 | 169 |
| Total | ∑x= 41 | ∑y= 66 | ∑xy= 461 | ∑x2= 291 | ∑y2= 742 |

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

| Serial n. | Age Years(x) | Weight Kg(y) | xy | $X^2$ | $Y^2$ |
|---|---|---|---|---|---|
| 1 | 7 | 12 | 84 | 49 | 144 |
| 2 | 6 | 8 | 48 | 36 | 64 |
| 3 | 8 | 12 | 96 | 64 | 144 |
| 4 | 5 | 10 | 50 | 25 | 100 |
| 5 | 6 | 11 | 66 | 36 | 121 |
| 6 | 9 | 13 | 117 | 81 | 169 |
| Total | ∑x= 41 | ∑y= 66 | ∑xy= 461 | ∑x2= 291 | ∑y2= 742 |

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

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

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

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

r = - 0.94

**Indirect strong correlation**

# In R

```
X<-c(7,6,8,5,6,9)
Y<-c(12,8,12,10,11,13)
cor(X,Y)

X<-c(10,8,2,1,5,6)
Y<-c(2,3,9,7,6,5)
cor(X,Y)
```

# Regression Analyses

- Regression: technique concerned with predicting some variables by knowing others
- The process of predicting variable Y using variable X

# 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**"

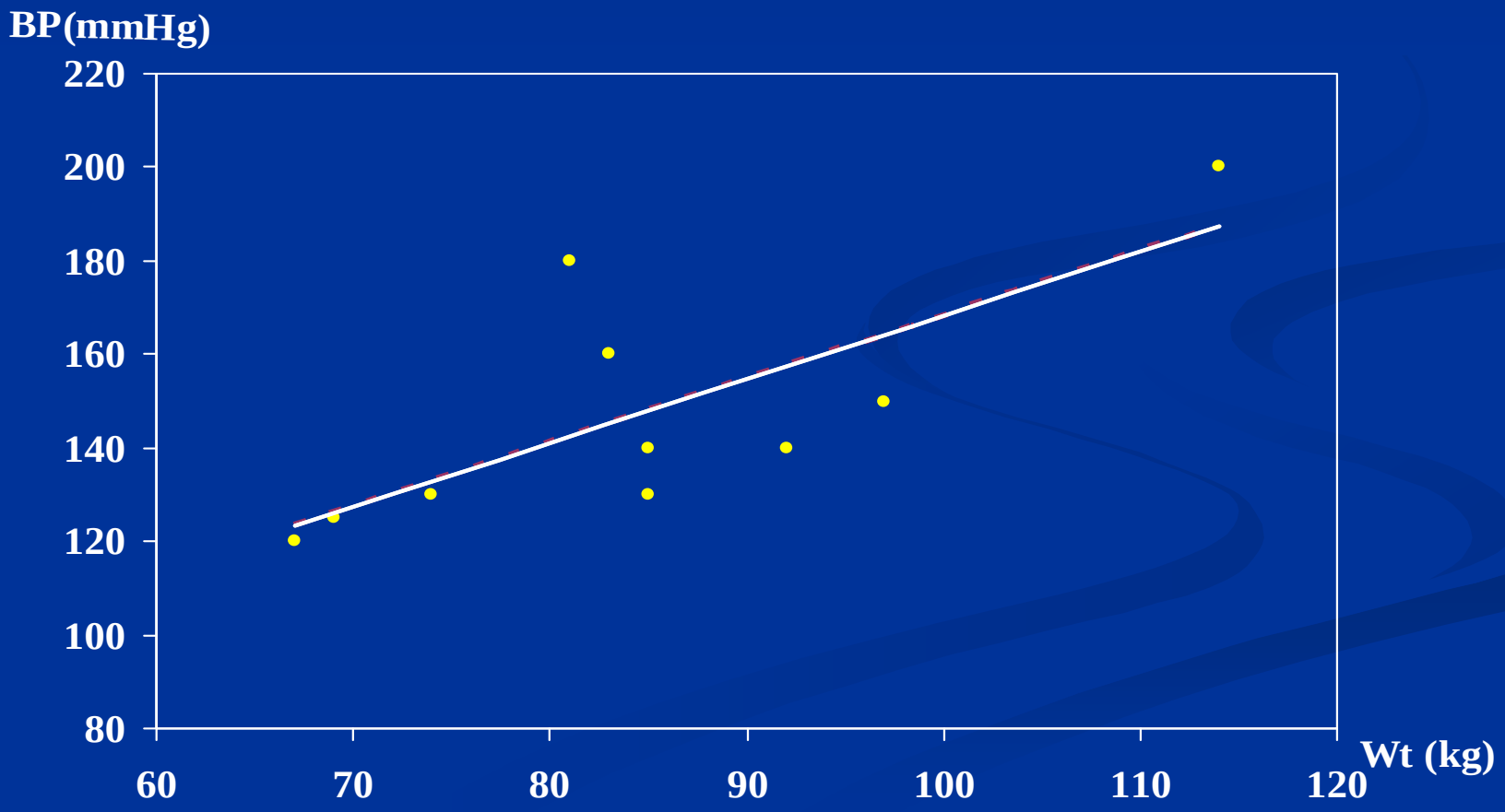
- **Regression** tells us how to draw the straight line described by the correlation

# 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**



By using the least squares method (a procedure that minimizes the vertical deviations of plotted points surrounding a straight line) we are able to construct a best fitting straight line to the scatter diagram points and then formulate a regression equation in the form of:

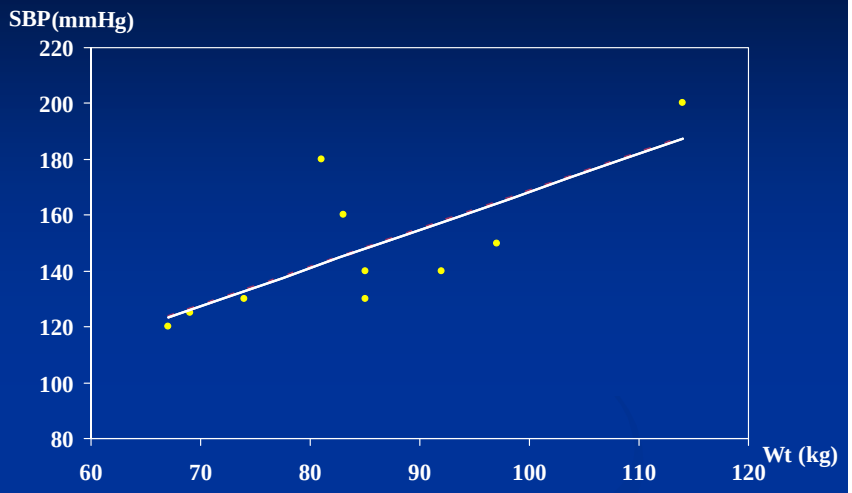
$$\hat{y} = a + bX$$

$$b = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

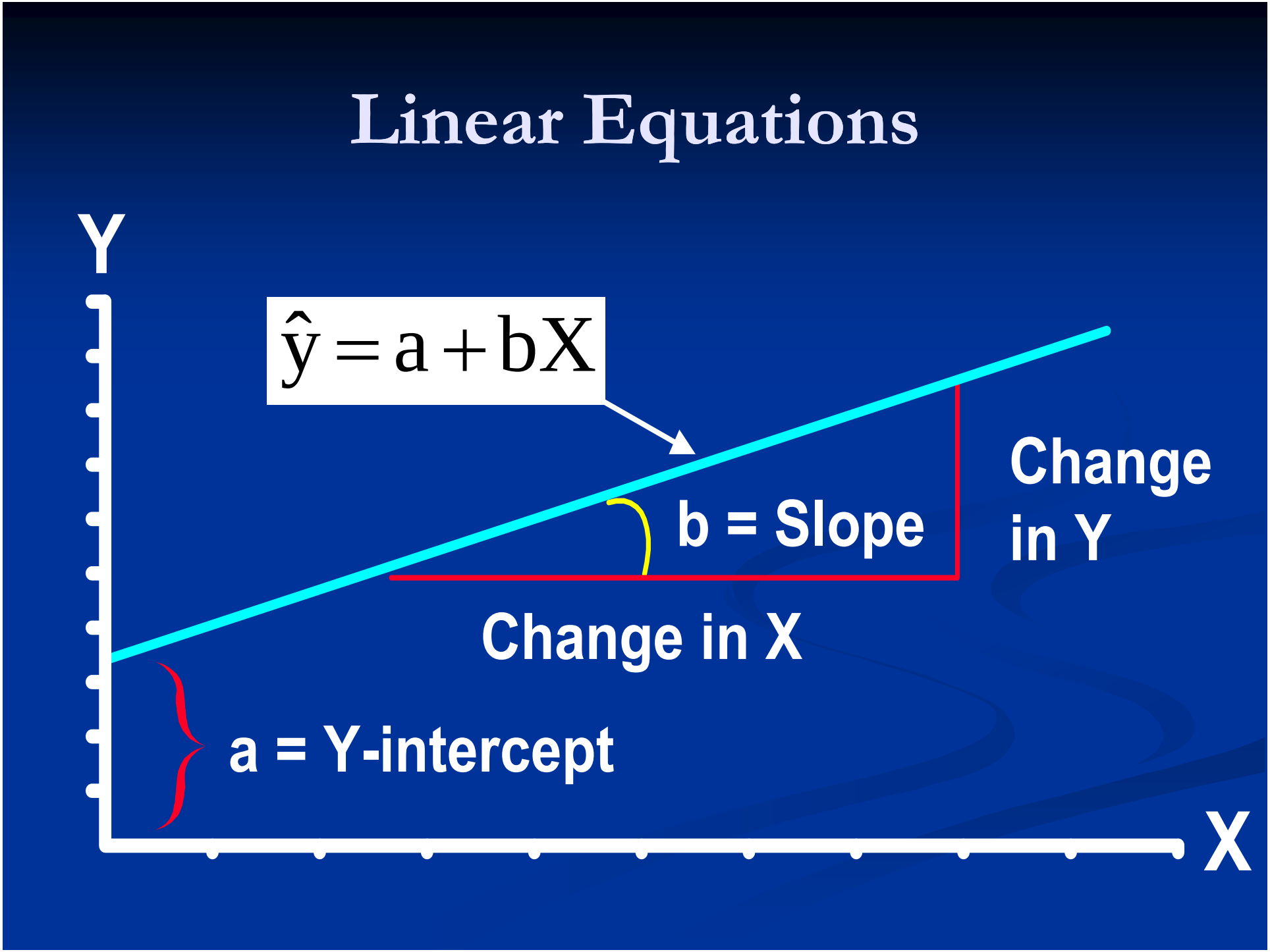
$$\hat{y} = \bar{y} + b(x - \bar{x})$$

# Regression Equation

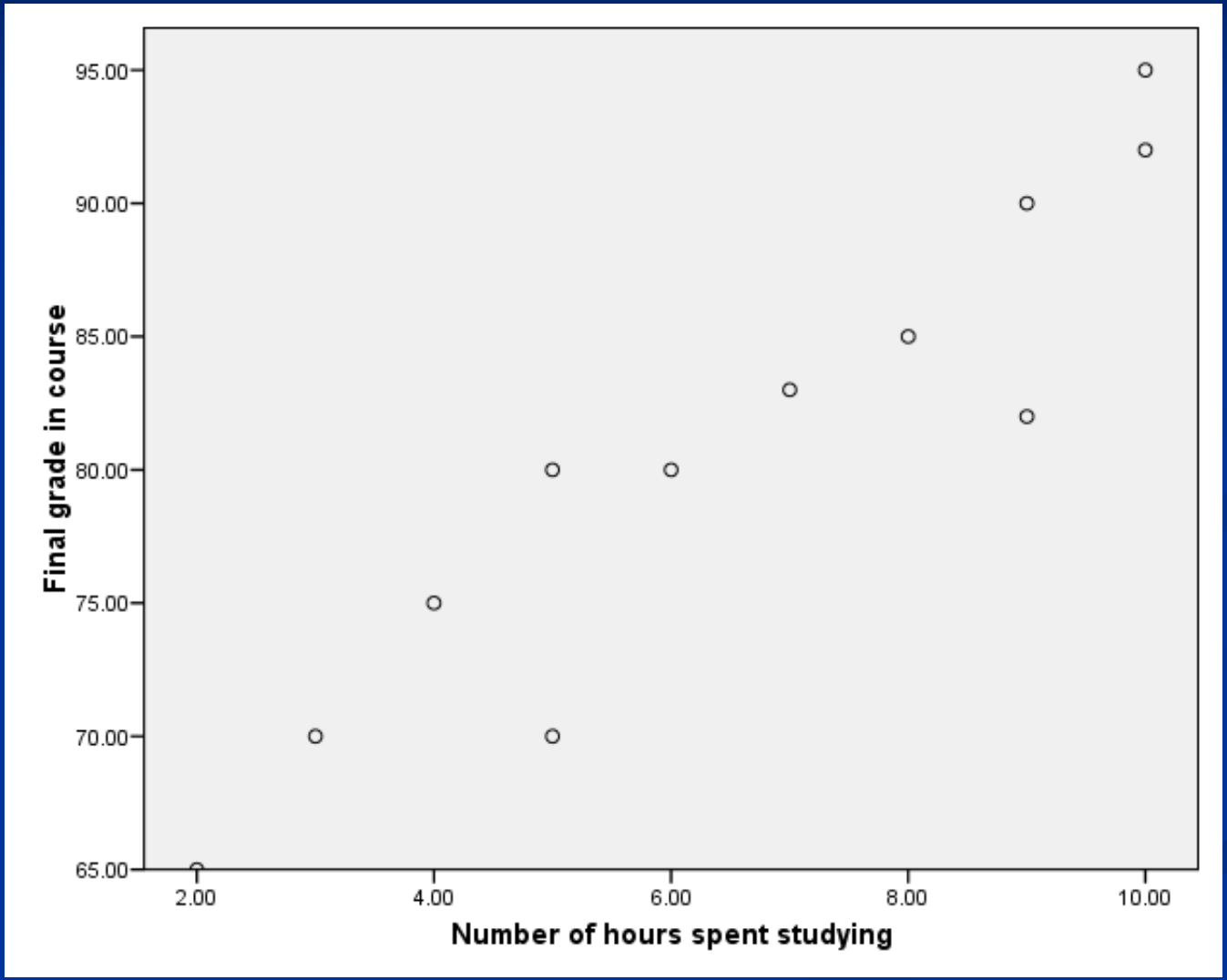
- Regression equation describes the regression line mathematically
  - Intercept
  - Slope



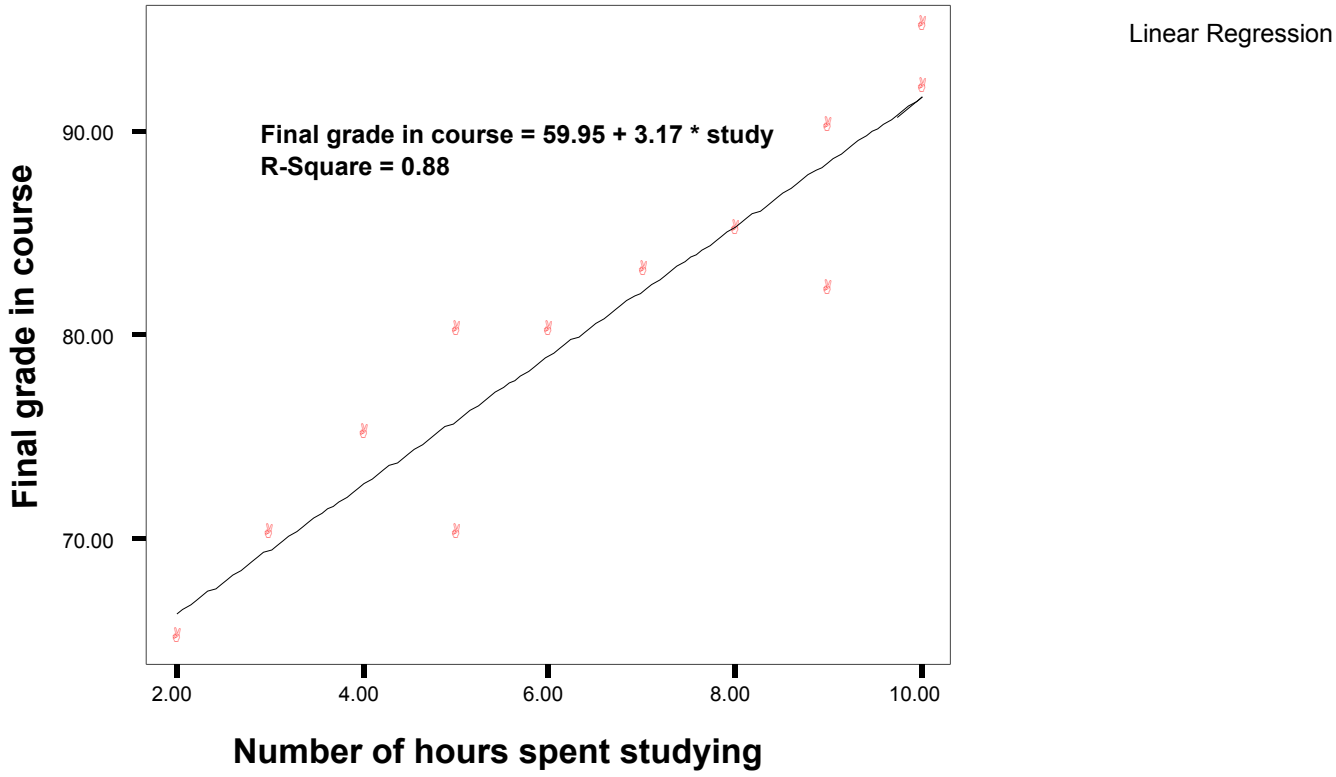
# Linear Equations



# Hours studying and grades



# Regressing grades on hours




Predicted final grade in class =

59.95 + 3.17*(number of hours you study per week)

Predicted final grade in class = 59.95 + 3.17*(hours of study)

**Predict the final grade of…**

- Someone who studies for 12 hours
- Final grade = 59.95 + (3.17*12)
- Final grade = 97.99

- Someone who studies for 1 hour:
- Final grade = 59.95 + (3.17*1)
- Final grade = 63.12

## Exercise

A sample of 6 persons was selected the value of their age ( x variable) and their weight is demonstrated in the following table. Find the regression equation and what is the predicted weight when age is 8.5 years.

| Serial no. | Age (x) | Weight (y) |
|---|---|---|
| 1 | 7 | 12 |
| 2 | 6 | 8 |
| 3 | 8 | 12 |
| 4 | 5 | 10 |
| 5 | 6 | 11 |
| 6 | 9 | 13 |

## Answer

| Serial no. | Age (x) | Weight (y) | xy | $X^2$ | $Y^2$ |
|---|---|---|---|---|---|
| 1 | 7 | 12 | 84 | 49 | 144 |
| 2 | 6 | 8 | 48 | 36 | 64 |
| 3 | 8 | 12 | 96 | 64 | 144 |
| 4 | 5 | 10 | 50 | 25 | 100 |
| 5 | 6 | 11 | 66 | 36 | 121 |
| 6 | 9 | 13 | 117 | 81 | 169 |
| Total | 41 | 66 | 461 | 291 | 742 |

$$\bar{x} = \frac{41}{6} = 6.83$$

$$\bar{y} = \frac{66}{6} = 11$$

Regression equation

$$\hat{y} = \bar{y} + b(x - \bar{x})$$

$$b_1 = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

$$b = \frac{461 - \frac{41 \times 66}{6}}{291 - \frac{(41)^2}{6}} = 0.92$$

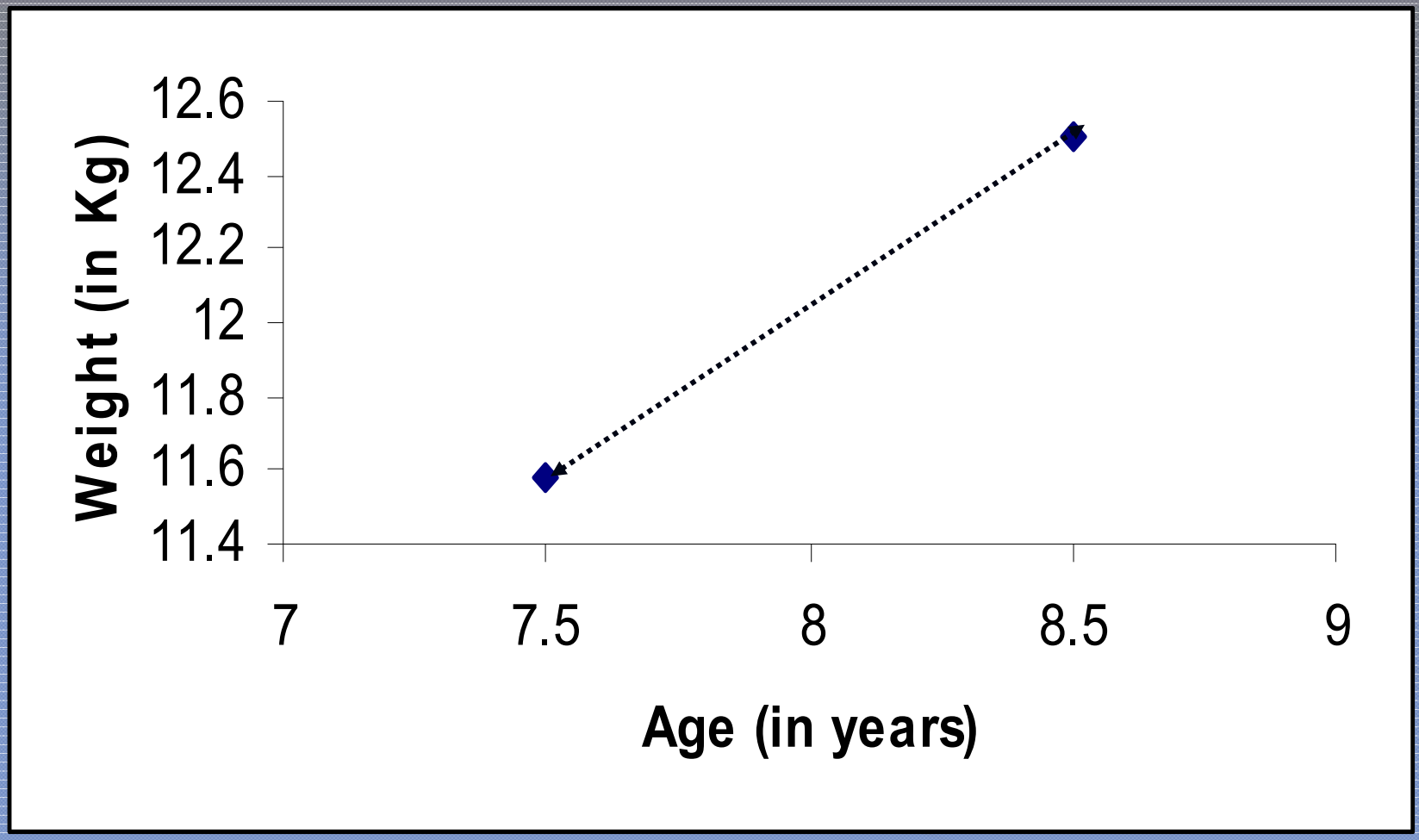
$$\hat{y}_{(x)} = 11 + 0.9(x - 6.83)$$

$$\hat{y}_{(x)} = 4.675 + 0.92x$$

$$\hat{y}_{(x)} = 4.675 + 0.92x$$

$$\hat{y}_{(8.5)} = 4.675 + 0.92 * 8.5 = 12.50 Kg$$

$$\hat{y}_{(7.5)} = 4.675 + 0.92 * 7.5 = 11.58 Kg$$




**we create a regression line by plotting two estimated values for y against their X component, then extending the line right and left.**

## Exercise 2

**The following are the age (in years) and blood pressure of 20 apparently healthy adults.**

| Age (x) | B.P (y) | Age (x) | B.P (y) |
|---|---|---|---|
| 20 | 120 | 46 | 128 |
| 43 | 128 | 53 | 136 |
| 63 | 141 | 60 | 146 |
| 26 | 126 | 20 | 124 |
| 53 | 134 | 63 | 143 |
| 31 | 128 | 43 | 130 |
| 58 | 136 | 26 | 124 |
| 46 | 132 | 19 | 121 |
| 58 | 140 | 31 | 126 |
| 70 | 144 | 23 | 123 |

- Find the correlation between age and blood pressure using simple and Spearman's correlation coefficients, and comment.
- Find the regression equation?
- What is the predicted blood pressure for a man aging 25 years?

| Serial | x | y | xy | x2 |
|---|---|---|---|---|
| 1 | 20 | 120 | 2400 | 400 |
| 2 | 43 | 128 | 5504 | 1849 |
| 3 | 63 | 141 | 8883 | 3969 |
| 4 | 26 | 126 | 3276 | 676 |
| 5 | 53 | 134 | 7102 | 2809 |
| 6 | 31 | 128 | 3968 | 961 |
| 7 | 58 | 136 | 7888 | 3364 |
| 8 | 46 | 132 | 6072 | 2116 |
| 9 | 58 | 140 | 8120 | 3364 |
| 10 | 70 | 144 | 10080 | 4900 |

| Serial | x | y | xy | x2 |
|---|---|---|---|---|
| 11 | 46 | 128 | 5888 | 2116 |
| 12 | 53 | 136 | 7208 | 2809 |
| 13 | 60 | 146 | 8760 | 3600 |
| 14 | 20 | 124 | 2480 | 400 |
| 15 | 63 | 143 | 9009 | 3969 |
| 16 | 43 | 130 | 5590 | 1849 |
| 17 | 26 | 124 | 3224 | 676 |
| 18 | 19 | 121 | 2299 | 361 |
| 19 | 31 | 126 | 3906 | 961 |
| 20 | 23 | 123 | 2829 | 529 |
| Total | 852 | 2630 | 114486 | 41678 |

$$b_1 = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}} = \frac{114486 - \frac{852 \times 2630}{20}}{41678 - \frac{852^2}{20}} = 0.4547$$

$\hat{y} = a + bX$ $\hat{y} = \bar{y} + b(x - \bar{x})$

$\hat{y}$ **=112.13 + 0.4547 x**

**for age 25**

**B.P = 112.13 + 0.4547 * 25=123.49 = 123.5 mm hg**

# Thank You