Module 2

Correlation and regression

Correlation

- Correlations tell us to the degree that two variables are similar or associated with each other. It is a measure of association.
- When two variables are related in such a way that a change in the value of one is accompanied either by a direct change or by an inverse change in the values of the other, the two variables are said to be correlated.
- A greater change in one variable resulting in a corresponding greater or smaller change in the other variable is also known as correlation.

Example: Relationship exists between the price and demand of a commodity because keeping other things equal, an increase in the price of a commodity shall cause a decrease in the demand for that commodity. Relationship might exist between the heights and weights of the students and between amount of rainfall in a city and the sales of raincoats in that city.

Types of Correlation

Correlation can be categorized as one of the following:
(i) Positive and Negative,
(ii) Simple and Multiple.
(iii) Partial and Total.
(iv) Linear and Non-Linear (Curvilinear)

(i) **Positive and Negative Correlation :** Positive or direct Correlation refers to the movement of variables in the same direction. The correlation is said to be positive when the increase (decrease) in the value of one variable is accompanied by an increase (decrease) in the value of other variable also.
Negative or inverse correlation refers to the movement of the variables in opposite direction. Correlation is said to be negative, if an increase (decrease) in the value of one variable is accompanied by a decrease (increase) in the value of other.

(ii) **Simple and Multiple Correlation :** Under simple correlation, we study the relationship between two variables only i.e., between the yield of wheat and the amount of rainfall or between demand and supply of a commodity. In case of multiple correlation, the relationship is studied among three or more variables.
For example, the relationship of yield of wheat may be studied with both chemical fertilizers and the pesticides.

(iii) **Partial and Total Correlation** : There are two categories of multiple correlation analysis. Under partial correlation, the relationship of two or more variables is studied in such a way that only one dependent variable and one independent variable is considered and all others are kept constant.
For example, coefficient of correlation between yield of wheat and chemical fertilizers excluding the effects of pesticides and manures is called partial correlation. Total correlation is based upon all the variables.

(iv) **Linear and Non-Linear Correlation:** When the amount of change in one variable tends to keep a constant ratio to the amount of change in the other variable, then the correlation is said to be linear. But if the amount of change in one variable does not bear a constant ratio to the amount of change in the other variable then the correlation is said to be non-linear. The distinction between linear and non-linear is based upon the consistency of the ratio of change between the variables.

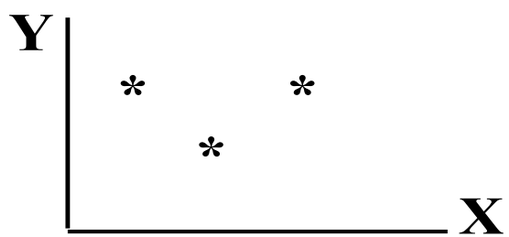
Methods of Studying Correlation:

There are different methods which helps us to find out whether the variables are related or not.

1. Scatter Diagram Method.
2. Karl Pearson's Coefficient of correlation.
3. Rank Method.

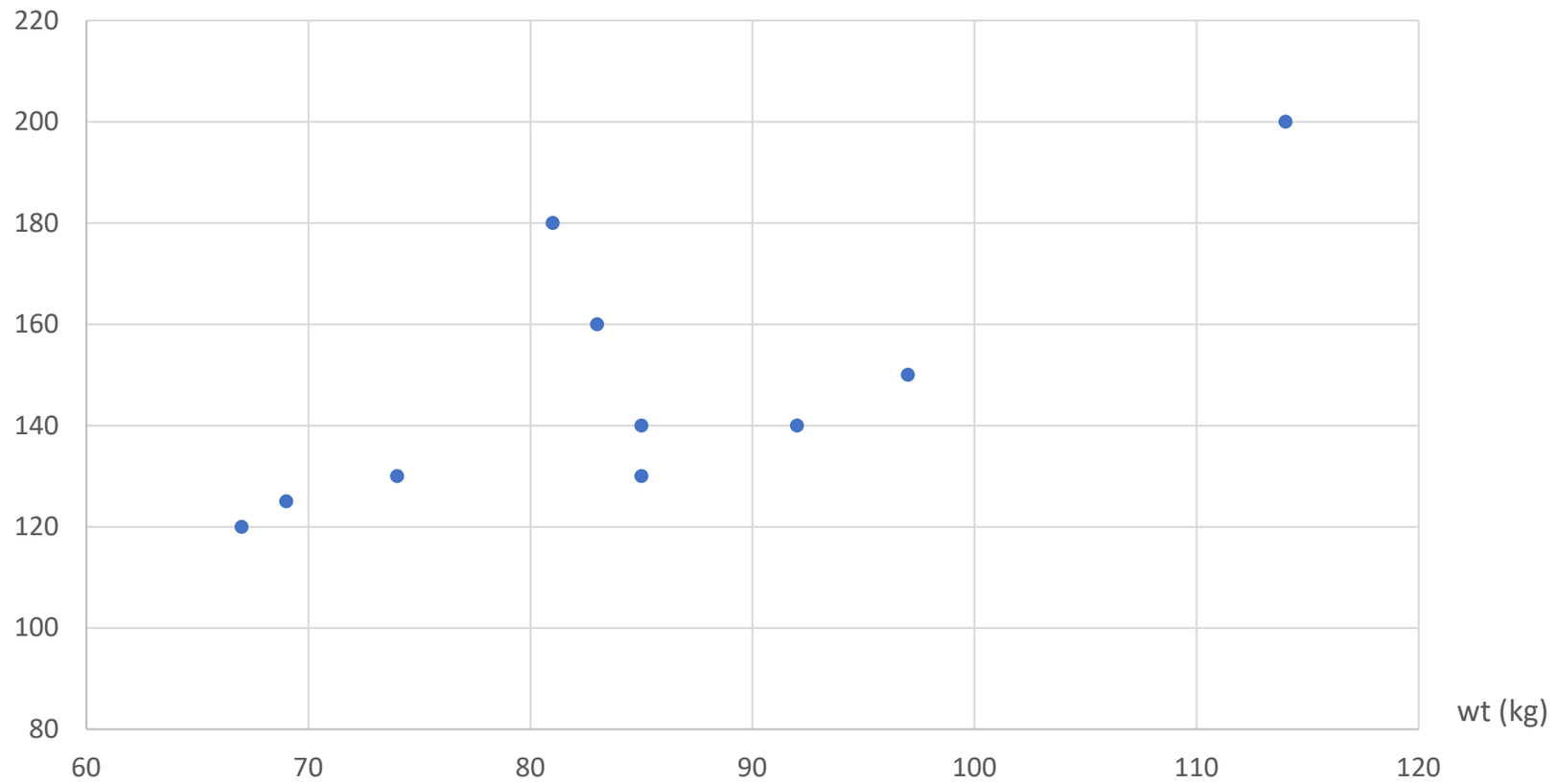
Scatter diagram

- 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

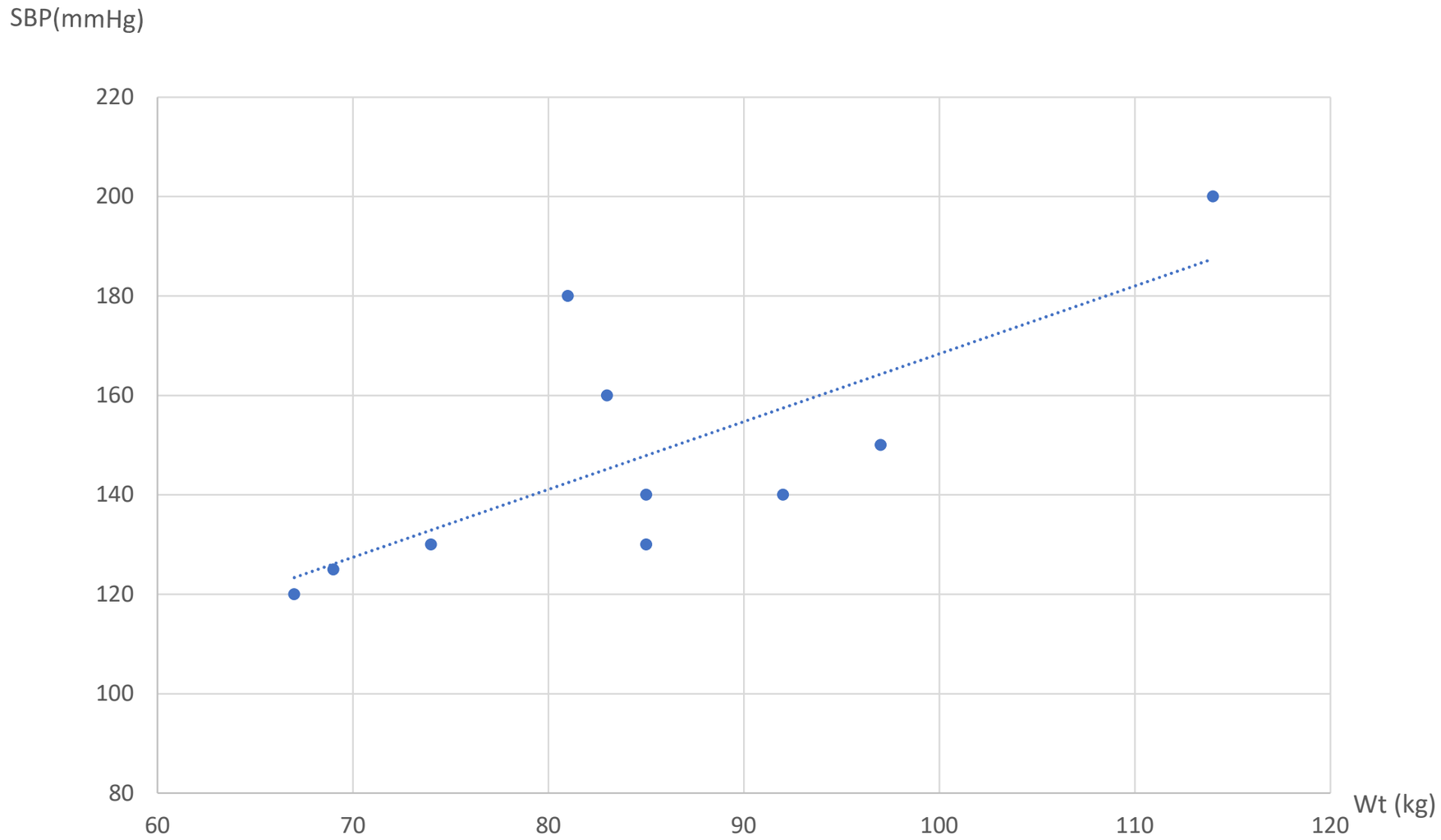


Example

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



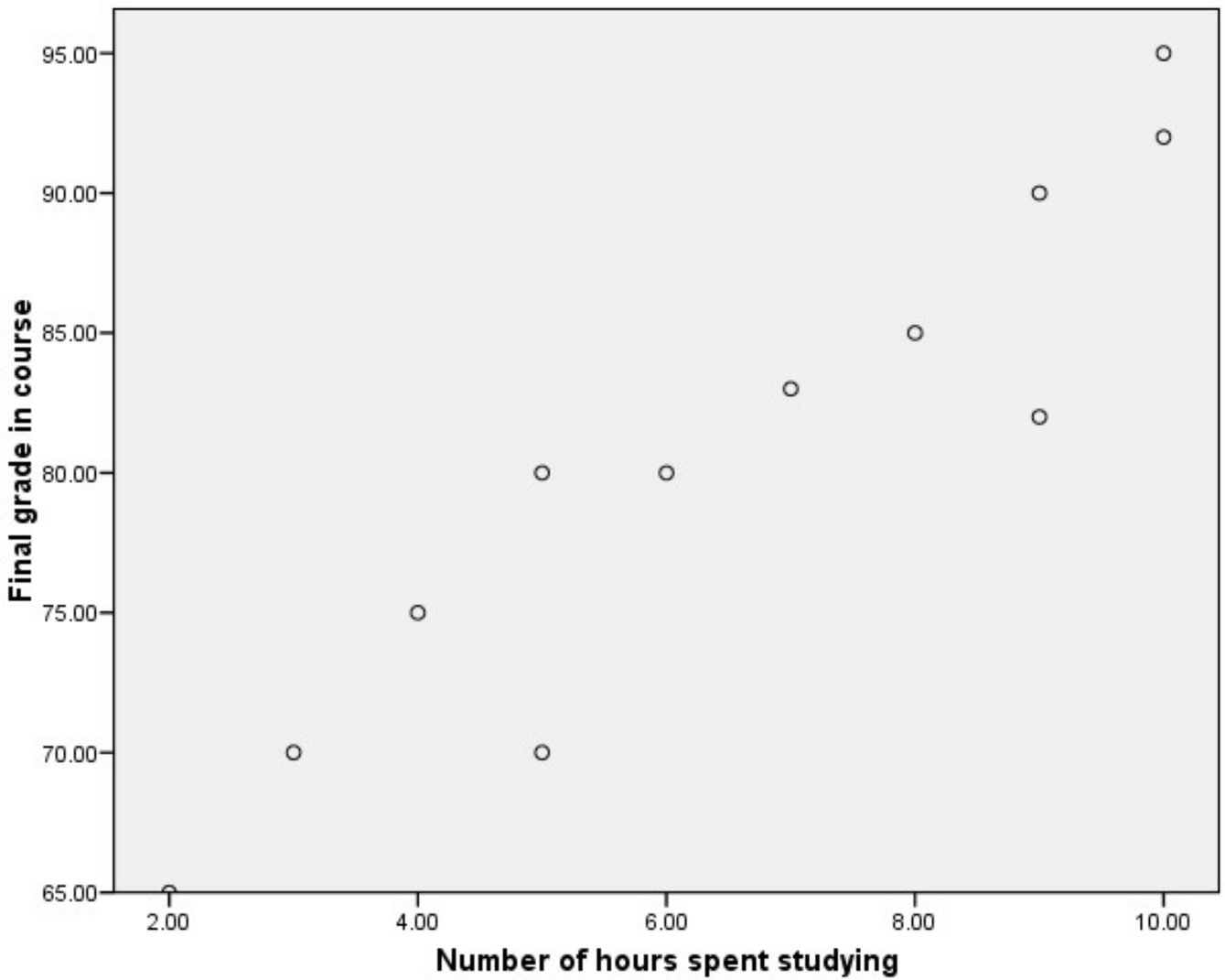
Scatter diagram of weight and systolic blood pressure

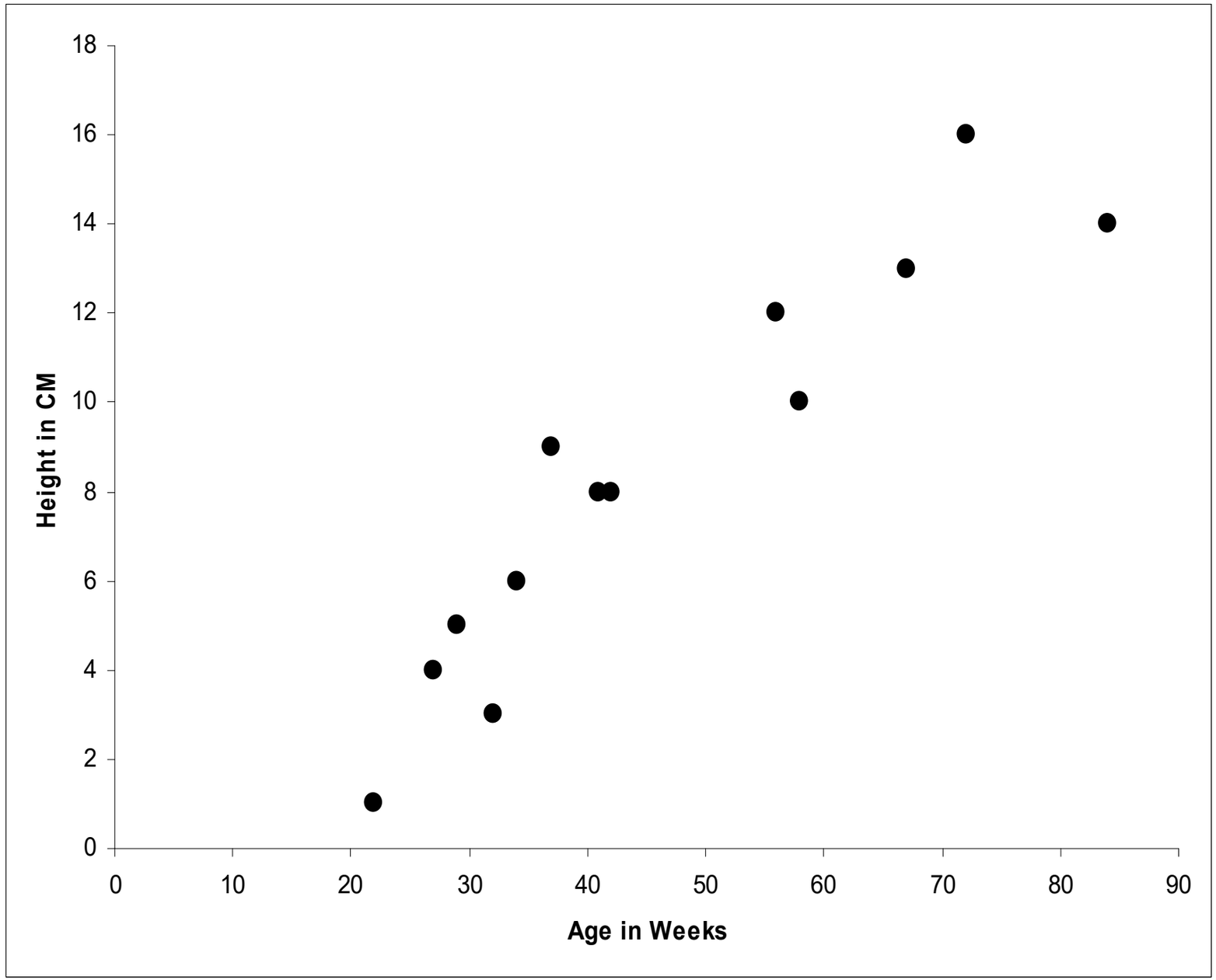
Scatter plots

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

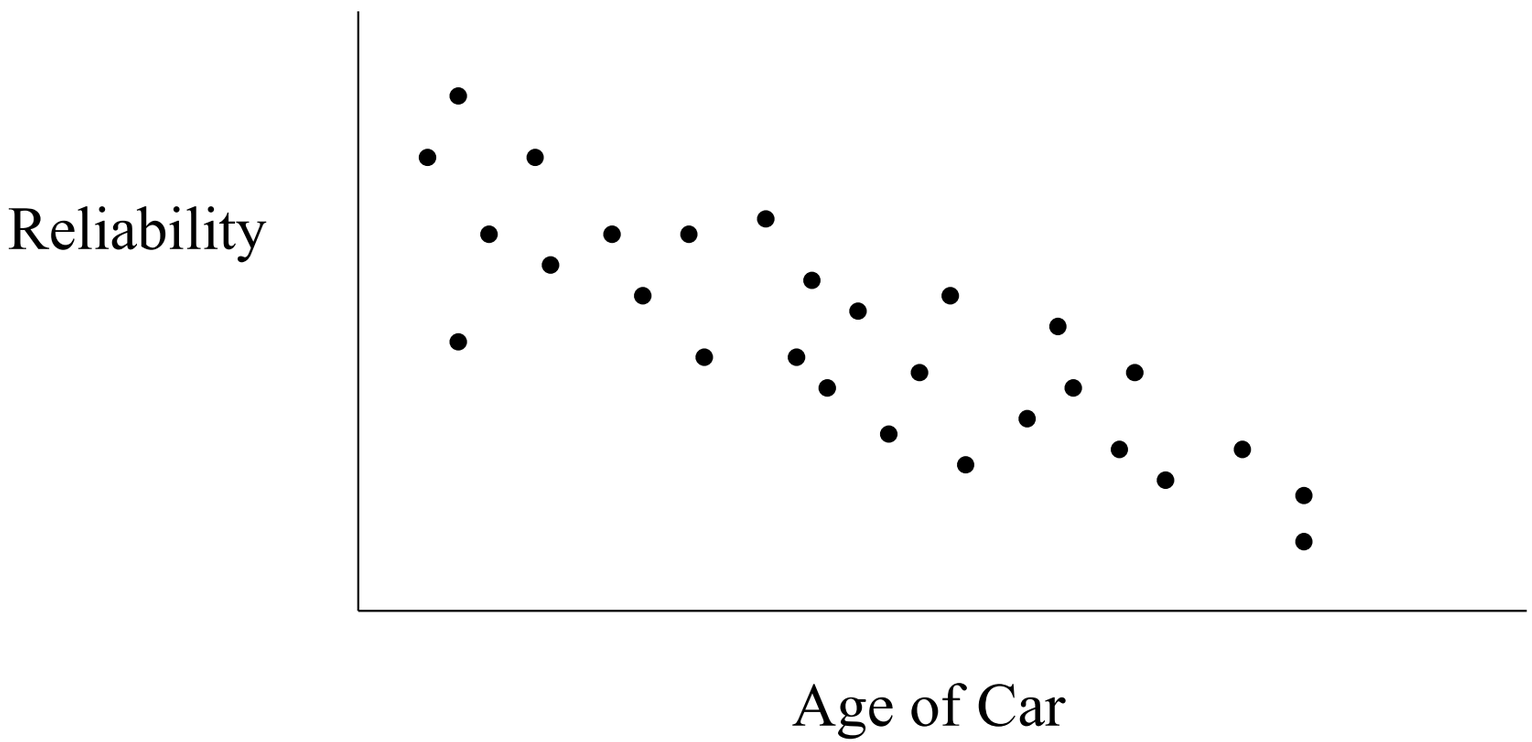
- positive relationship
- negative relationship
- no relationship

Positive relationship

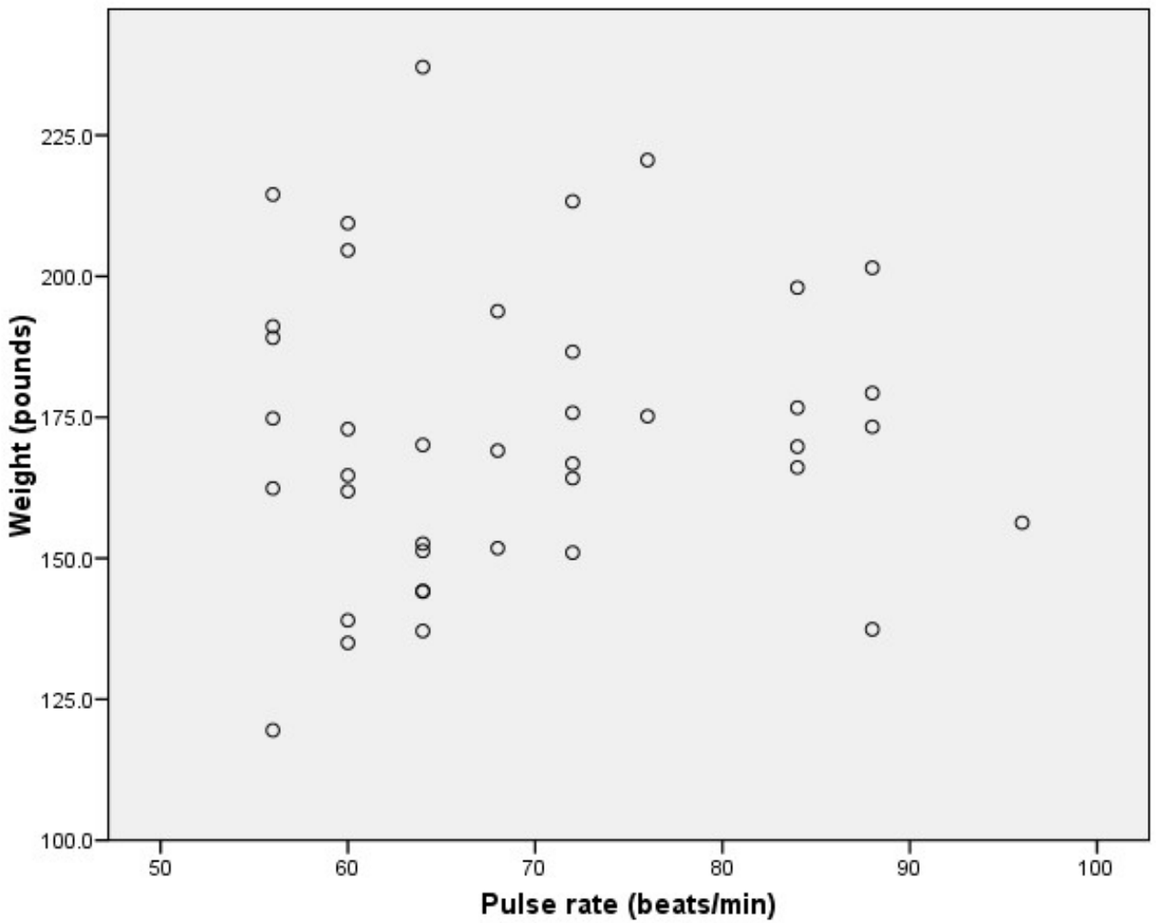



Height in CM
Age in Weeks
0
2
4
6
8
10
12
14
16
18
0
10
20
30
40
50
60
70
80
90

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)

Karl Pearson's Co-efficient of Correlation

Karl Pearson's method, popularly known as Pearsonian co-efficient of correlation, is most widely applied in practice to measure correlation.

It is denoted by r or r_{XY} or ρ_{XY}.

Variance and Covariance (in brief)

As the variance $E\{X - E(X)\}^2$ measures the variations of the RV X from its mean value $E(X)$, the quantity $E\{[X - E(X)]\,[Y - E(Y)]\}$ measures the simultaneous variation of two RV's X and Y from their respective means and hence it is called the covariance of X, Y and denoted as $Cov\,(X, Y)$.

$Cov\,(X, Y) = E\{[X - E(X)]\,[Y - E(Y)]\}$ is also called the product moment of X and Y.

Though $Cov(X,Y)$ is a useful measure of the degree of correlation between X and Y, it is to be expressed in mixed units of X and Y. To avoid this difficulty and to express the degree of correlation in absolute units, we divide $Cov(X,Y)$ by $\sigma_x.\sigma_y$, so that $\rho_{XY} = \frac{Cov(X,Y)}{\sigma_x.\sigma_y} = \frac{Cov(X,Y)}{\sqrt{Var(x)}\sqrt{Var(y)}}$ is a mere number, free from the units of X and Y.

- The sign of r (or r_{XY} or ρ_{XY}) denotes the nature of association
- while the value of r denotes the strength of association.

Depending on the value of r $(-1 \leq r \leq 1)$, we can classify correlation as follows.

- If $r = 1$, both the variables X and Y increase or decrease in the same proportion. In this case we say that there is perfect positive correlation.
- If $r = -1$, both the variables X and Y are inversely proportion to each other. In this case we say that there is perfect negative correlation.
- If $r = 0$, we say that X and Y are uncorrelated.
- If $0 < r < 1$, there is moderate (partial) positive correlation between X and Y.
- If $-1 < r < 0$, there is moderate (partial) negative correlation between X and Y.

We will mainly deal with linear correlation of discrete RV's X and Y. X will take the values $x_1, x_2, \ldots, x_n$ with frequency 1 each and Y will simultaneously take the values $y_1, y_2, \ldots, y_n$ with frequency 1 each.

$$E(X) = \frac{1}{n}\sum x_i \,;\; E(X^2) = \frac{1}{n}\sum x_i^2 \,,\; E(XY) = \frac{1}{n}\sum x_i\, y_i \text{ etc.}$$

$$r_{XY} = \frac{E(XY) - E(X)\cdot E(Y)}{\sqrt{\left\{E\left(X^2\right) - E^2(X)\right\}\left\{E\left(Y^2\right) - E^2(Y)\right\}}}$$

$$r_{XY} = \frac{\frac{1}{n}\Sigma x_i\, y_i - \frac{1}{n}\Sigma x_i \cdot \frac{1}{n}\Sigma y_i}{\sqrt{\left\{\frac{1}{n}\Sigma x_i^2 - \left(\frac{1}{n}\Sigma x_i\right)^2\right\}\left\{\frac{1}{n}\Sigma y_i^2 - \left(\frac{1}{n}\Sigma y_i\right)^2\right\}}}$$

$$r_{XY} = \frac{n\,\Sigma xy - \Sigma x \cdot \Sigma y}{\sqrt{\left\{n\Sigma x^2 - (\Sigma x)^2\right\}\left\{n\Sigma y^2 - (\Sigma y)^2\right\}}}$$

Problems:

1. Find the correlation coefficient between annual advertising expenditures and annual sales revenue for the following data:

Year (i)	1	2	3	4	5	6	7	8	9	10
Annual advertising expenditure (X_i)	10	12	14	16	18	20	22	24	26	28
Annual sales (Y_i)	20	30	37	50	56	78	89	100	120	110

Solution: Now, $\bar{X} = \frac{\sum X}{n} = \frac{190}{10} = 19$, $\bar{Y} = \frac{\sum Y}{n} = \frac{690}{10} = 69$

i	X_i	Y_i	$X_i - \bar{X}$	$Y_i - \bar{Y}$	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})^2$	$(X_i - \bar{X})(Y_i - \bar{Y})$
1	10	20	-9	-49	81	2401	441
2	12	30	-7	-39	49	1521	273
3	14	37	-5	-32	25	1024	160
4	16	50	-3	-19	9	364	57
5	18	56	-1	-13	1	169	13
6	20	78	1	9	1	81	9
7	22	89	3	20	9	400	60
8	24	100	5	31	25	961	155
9	26	120	7	51	49	2601	357
10	28	110	9	41	81	1681	369
	190	**690**	**0**	**0**	**330**	**11200**	**1894**

Correlation coefficient is $r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} =$

$\frac{1894}{\sqrt{330}\sqrt{11200}} = 0.985$

The correlation coefficient between annual expenditure and annual sales revenue is 0.985.