Correlation & Regression

Md. Fazlul Karim Patwary
IIT
Jahangirnagar University

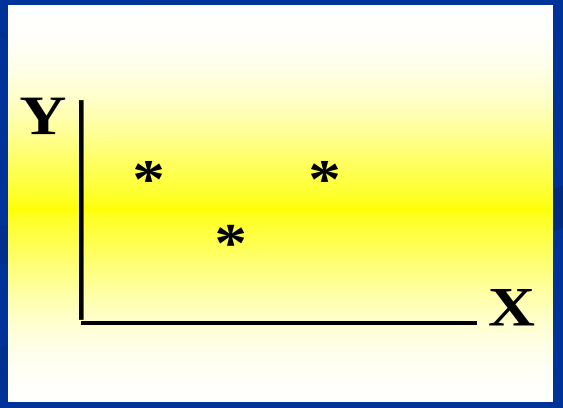
Correlation

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

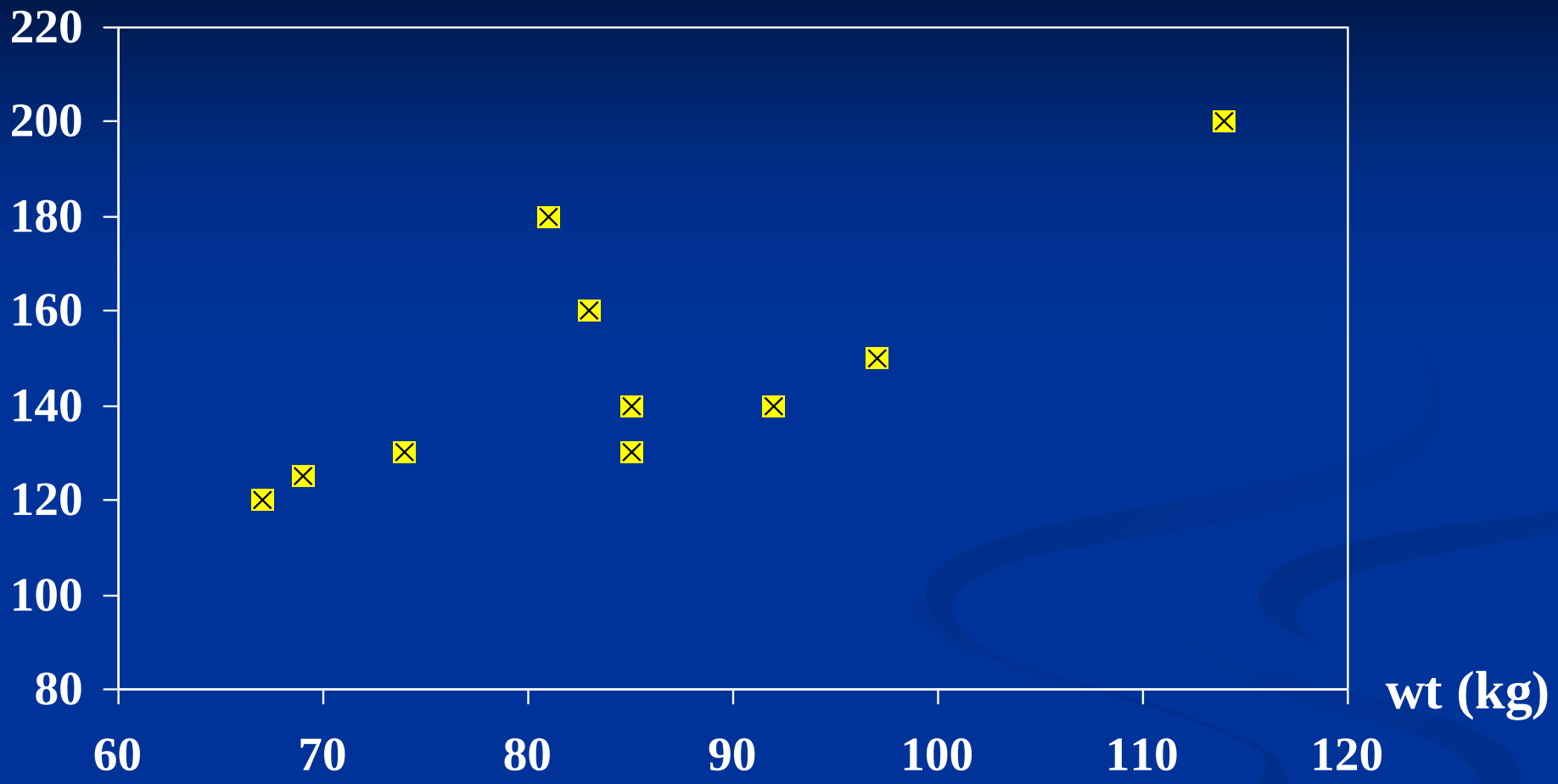
- Rectangular coordinate
- Two quantitative variables
- One variable is called independent (X) and the second is called dependent (Y)
- Points are not joined
- No frequency table



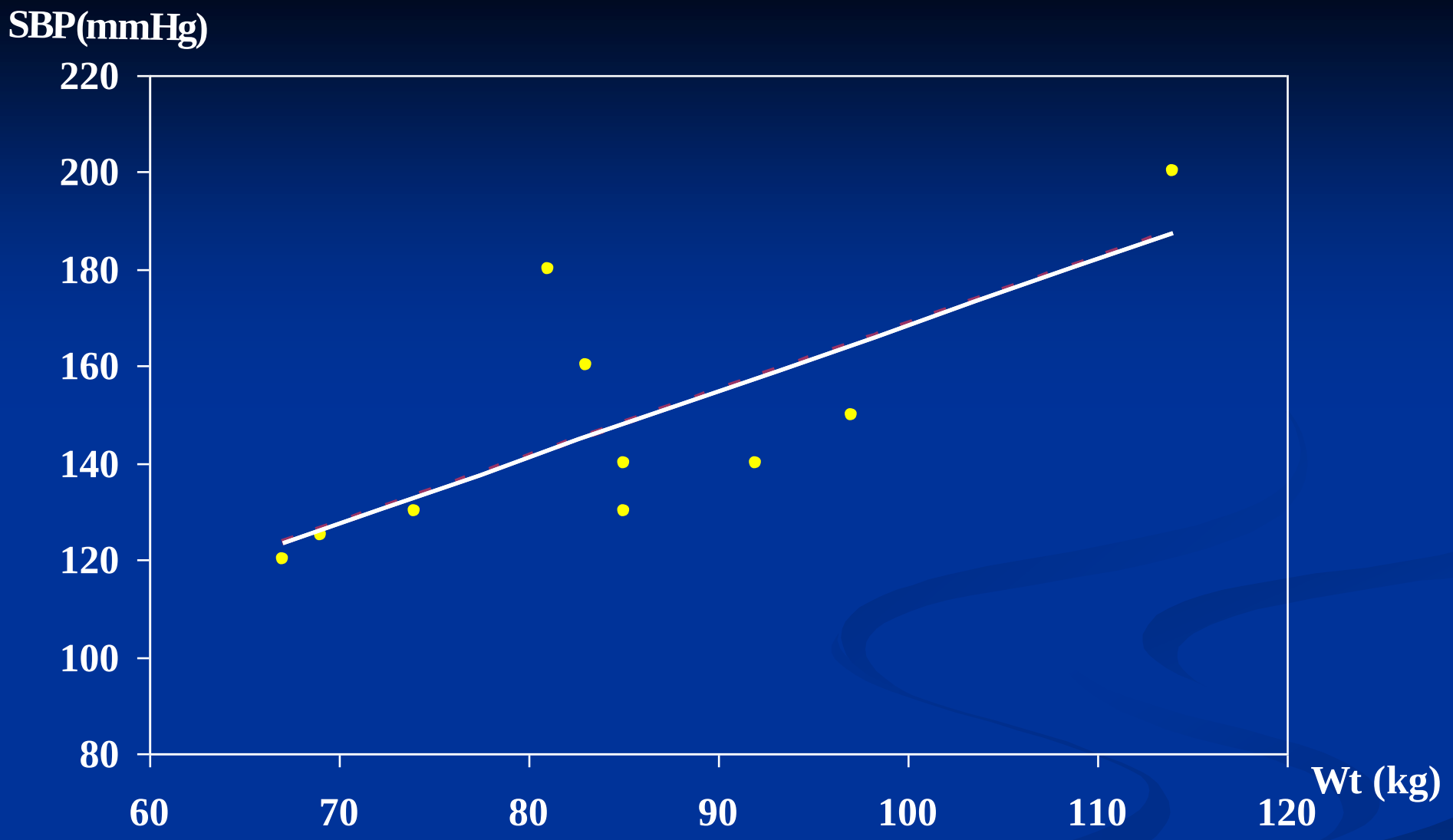
Example

Wt. (kg)	67	69	85	83	74	81	97	92	114	85
SBP mmHg)	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



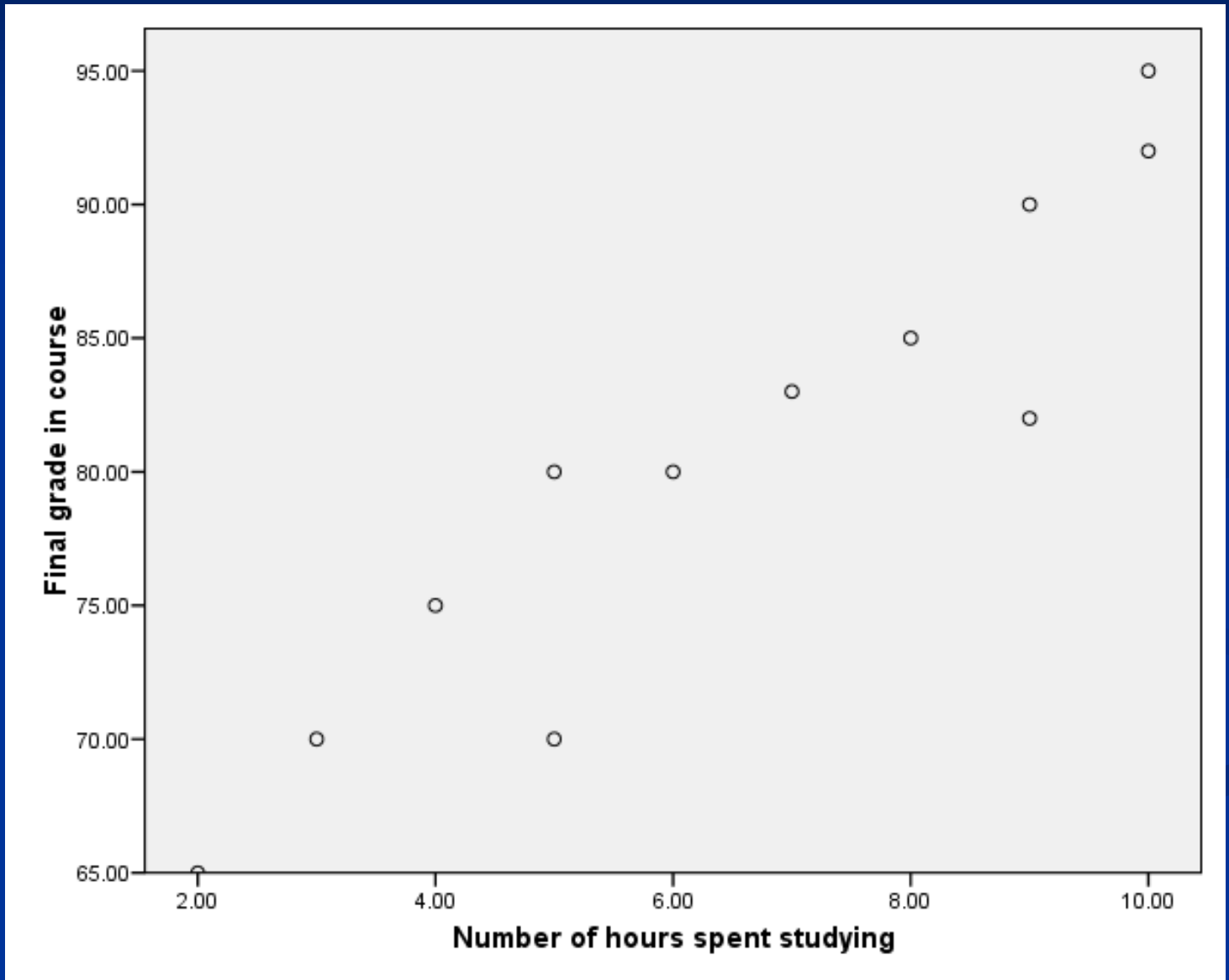
Scatter diagram of weight and systolic blood pressure

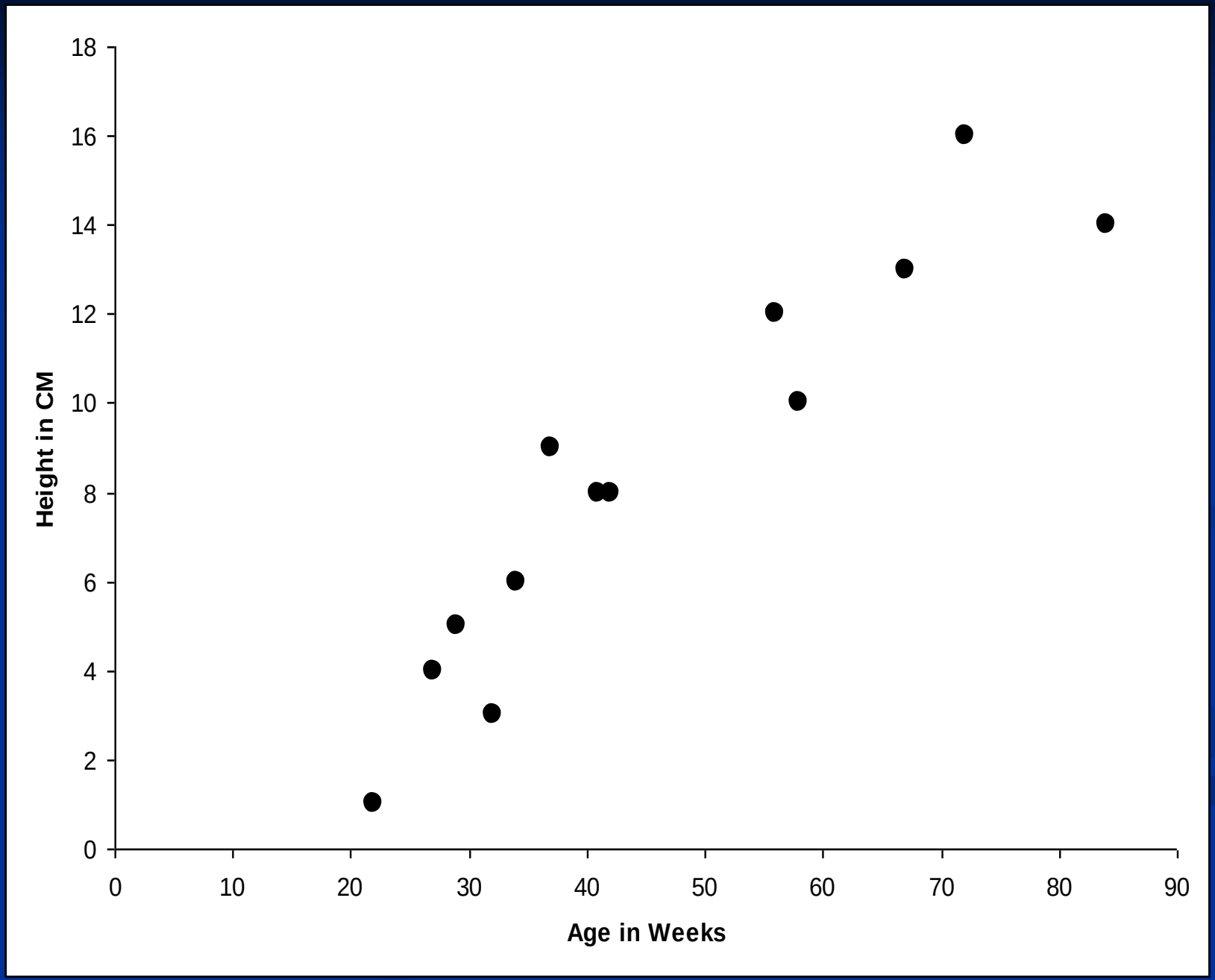
Scatter plots

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

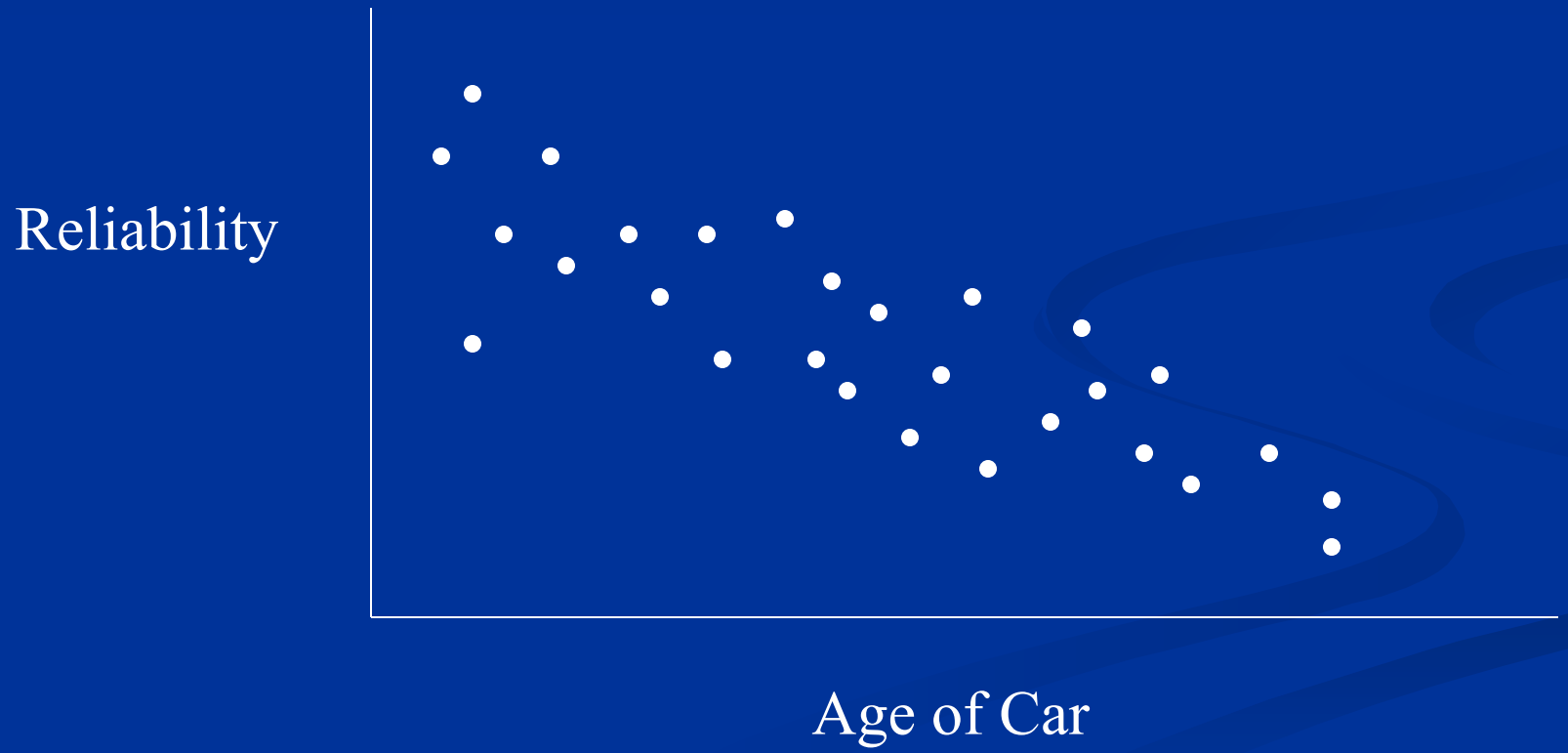
- positive relationship
- negative relationship
- no relationship

Positive relationship

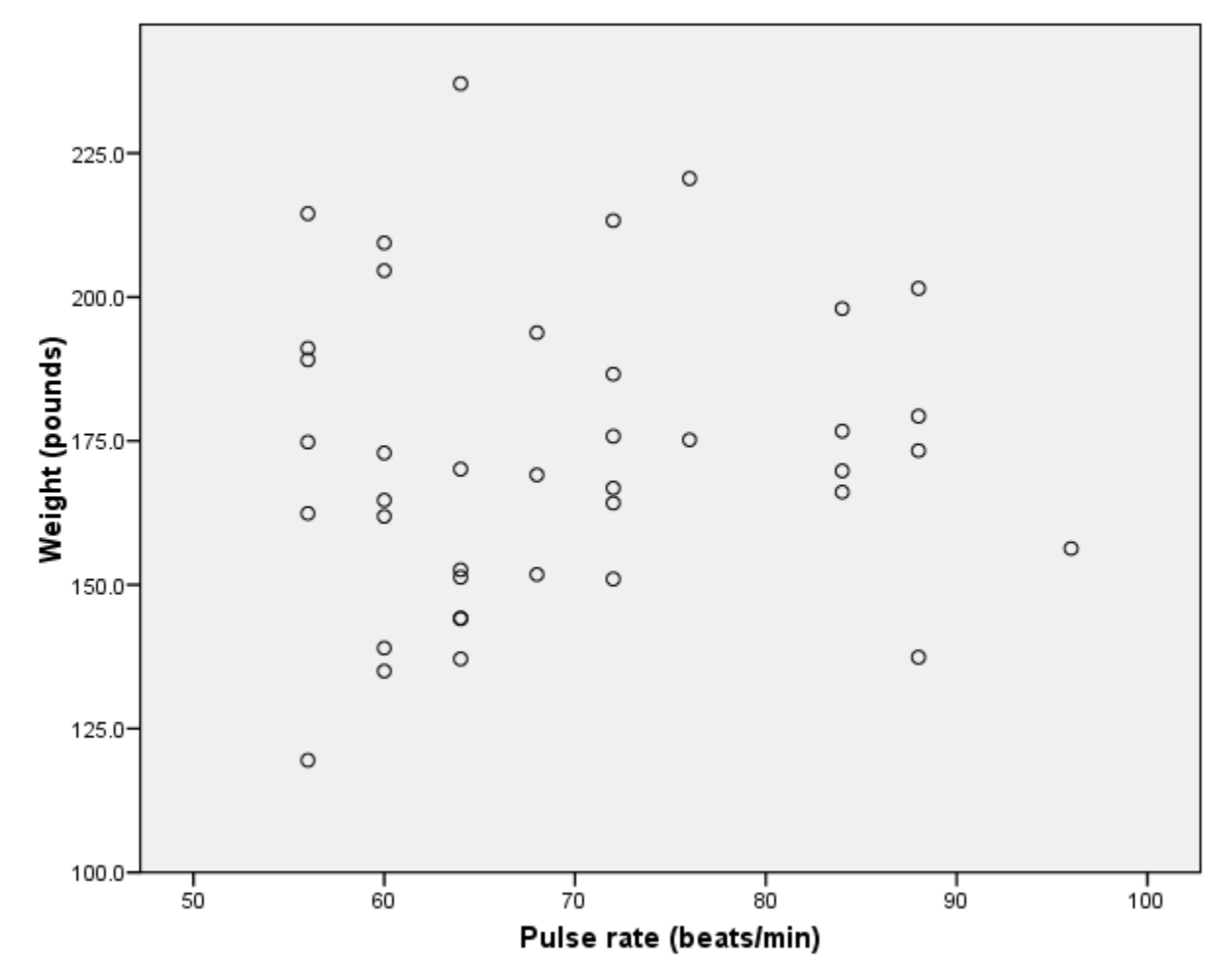



Height in CM
Age in Weeks
0
2
4
6
8
10
12
14
16
18
0
10
20
30
40
50
60
70
80
90

Negative relationship

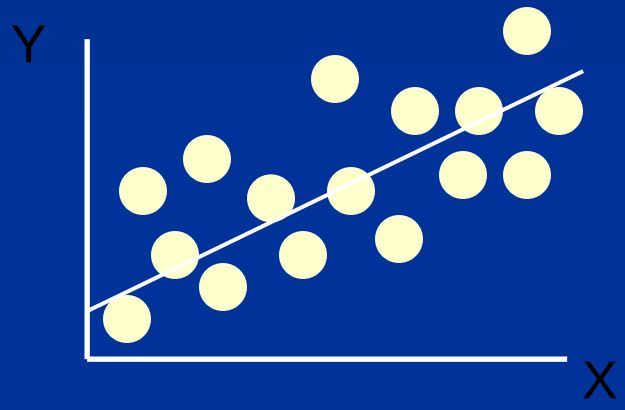


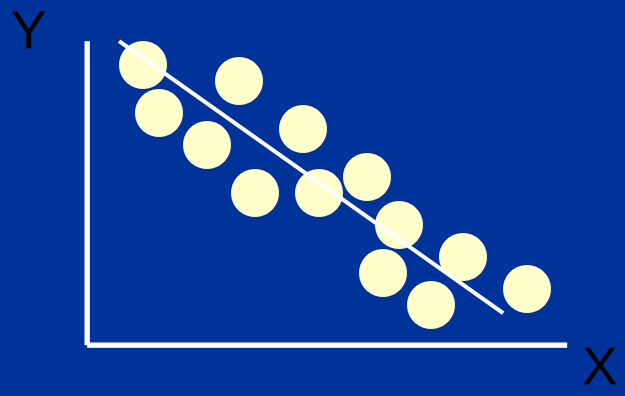
No relation


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

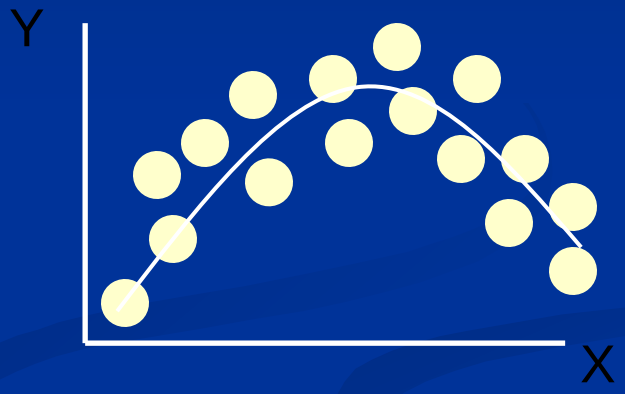
Linear Correlation

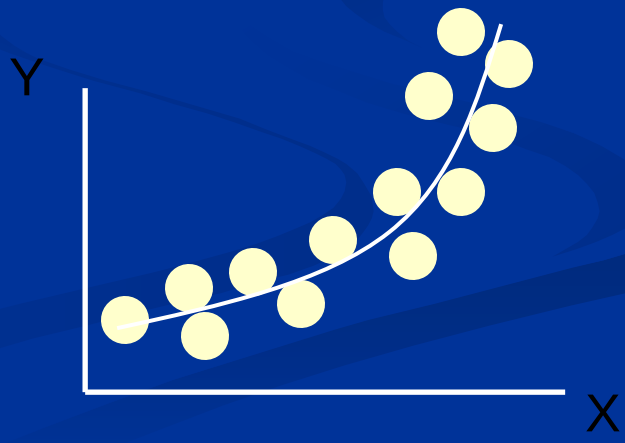
Linear relationships





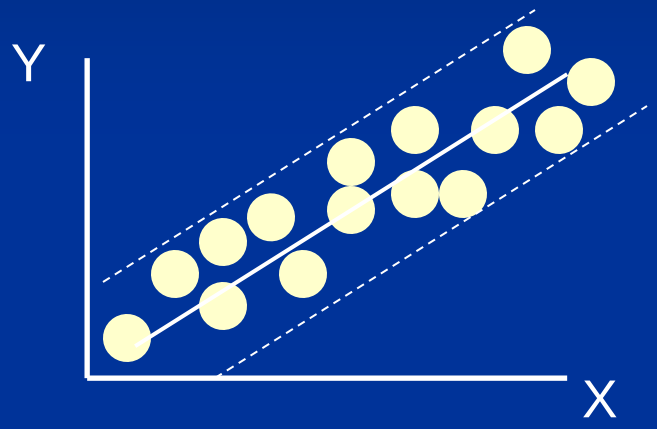
Curvilinear relationships

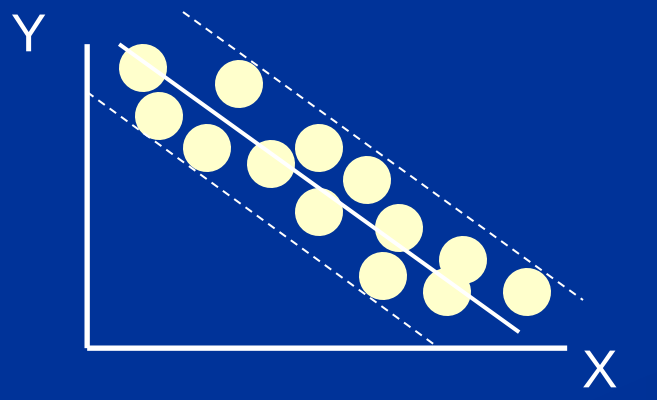




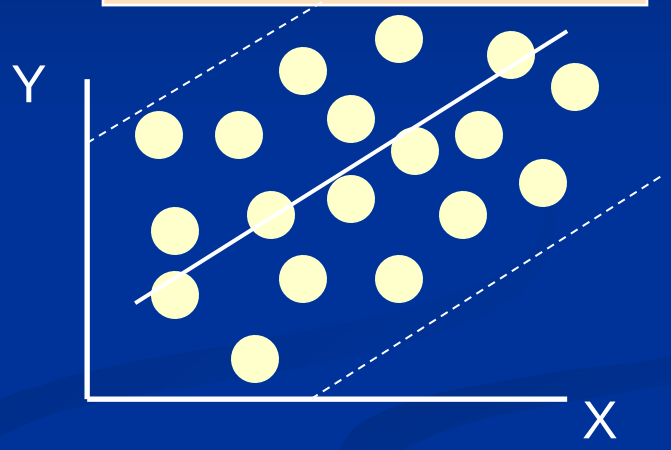
Linear Correlation

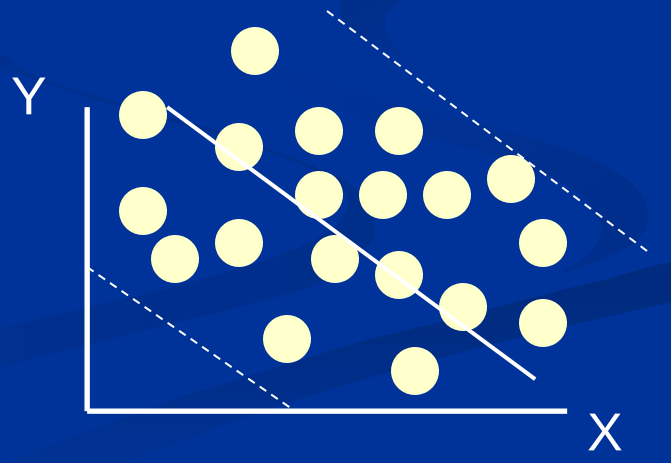
Strong relationships





Weak relationships





from: Statistics for Managers Using Microsoft® Excel 4th Edition, 2004 Prentice-Hall

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 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:

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

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

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

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

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

r = - 0.94

Indirect strong correlation

Coefficient of Determination

It is often difficult to interpret r without some familiarity with the expected values of r.

A more appropriate measure to use when interest lies in the dependence of Y on X, is the ***Coefficient of Determination, R^2***.

It measures the ***proportion of variation*** in Y that is explained by X, and is often expressed as a percentage.

Using anova to find R^2

Anova (for 93 rural female headed HHs) of log consumption expenditure versus number of persons per sleeping room is:

Source	d.f.	S.S.	M.S.	F	Prob.
Regression	1	4.890	4.890	21.9	0.000
Residual	91	20.342	0.2235		
Total	92	25.231	0.2743		

R^2 = Regre. S.S. / Total S.S.

= 4.89/25.23

= 0.194

Interpretation of R^2

From above, we can say that 19.4% of the variability in the income poverty proxy measure is accounted for by the number of persons per sleeping room.

Clearly there are many other factors that influence the poverty proxy since over 80% of the variability is left unexplained!

Relationship of R^2 to r

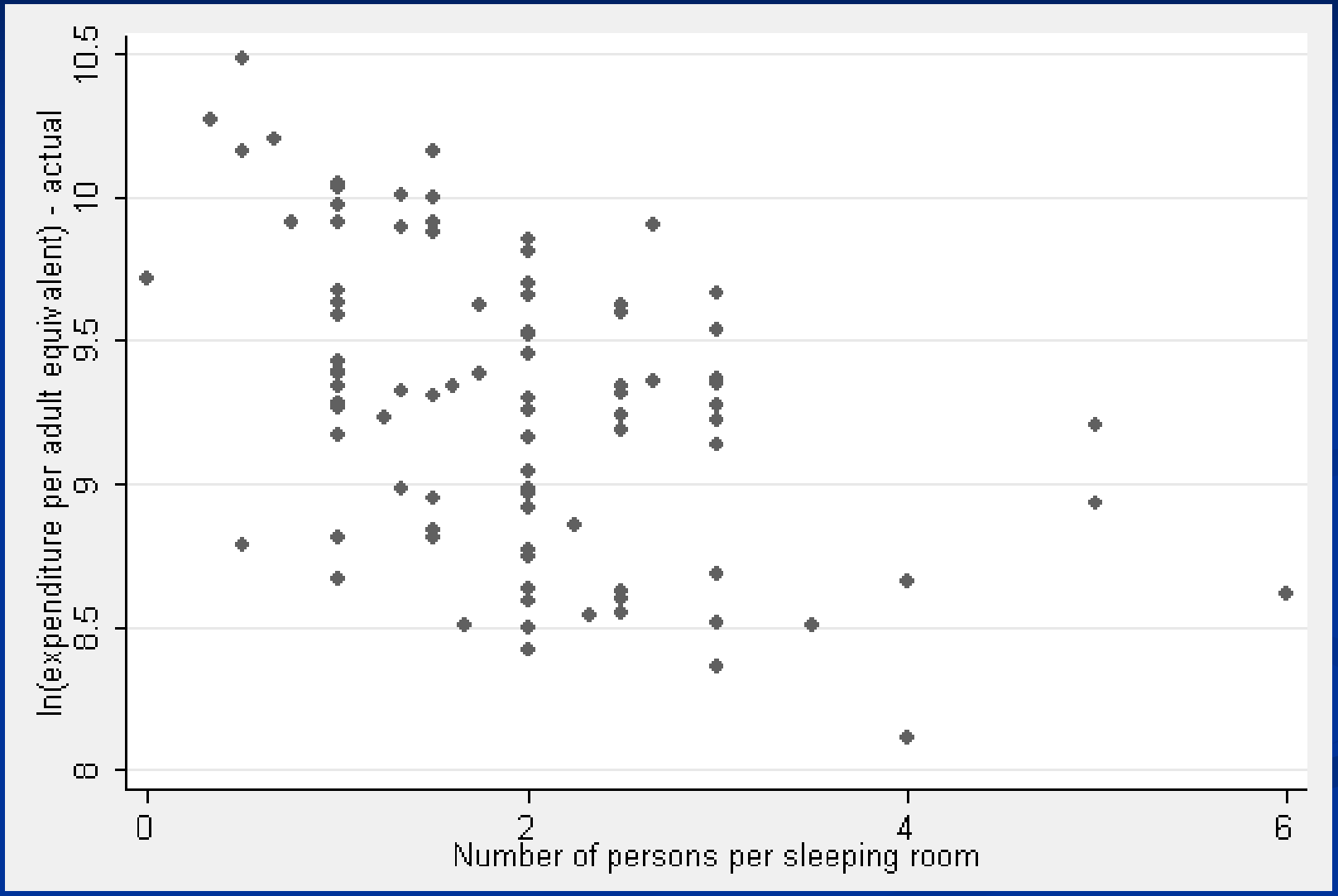
When there is just one explanatory variable being considered (as in above example), the squared value of r equals R^2.

In the above example,

value of $r = -\sqrt{0.194} = -0.44$

The negative value is used when taking the square root because the graph indicates a negative relationship (see next slide).

Plot of poverty proxy measure vs. persons per sleeping room



Benefits of R^2 and r

- r is useful as an initial exploratory tool when several variables are being considered. The sign of r gives the direction of the association.

- R^2 is useful in regression studies to check how much of the variability in the key response can be explained.

- R^2 is most valuable when there is more than one explanatory variable. High values of R^2 are particularly useful when using the model for predictions!

Limitations of r

- Observe that seemingly high values of r, e.g. r=0.70, explain only about 50% of the variability in the response variable y. So take care when interpreting correlation coefficients.

- a low value for r does not necessarily imply absence of a relationship – could be a curved relationship! So plotting the data is crucial!

- Tests exist for testing there is no association. But depending on the sample size, even low values of r, e.g. r=0.20 can give significant results – not a very useful finding!

Limitations of R^2

- Note that R^2 is only a descriptive measure to give a quick assessment of the model. Other methods exist for assessing the goodness of fit of the model.

- Adding explanatory variables to the model always increases R^2 . Hence in practice, it is more usual to look at the **adjusted R^2**

- The **adjusted R^2** is calculated as
1 – (Residual M.S./Total M.S.)
As with R^2 , the adjusted R^2 is often expressed as a percentage.

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 di for each pair of observation by subtracting the rank of Yi from the rank of Xi
4. Square each di and compute $\sum di2$ 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	di	di^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

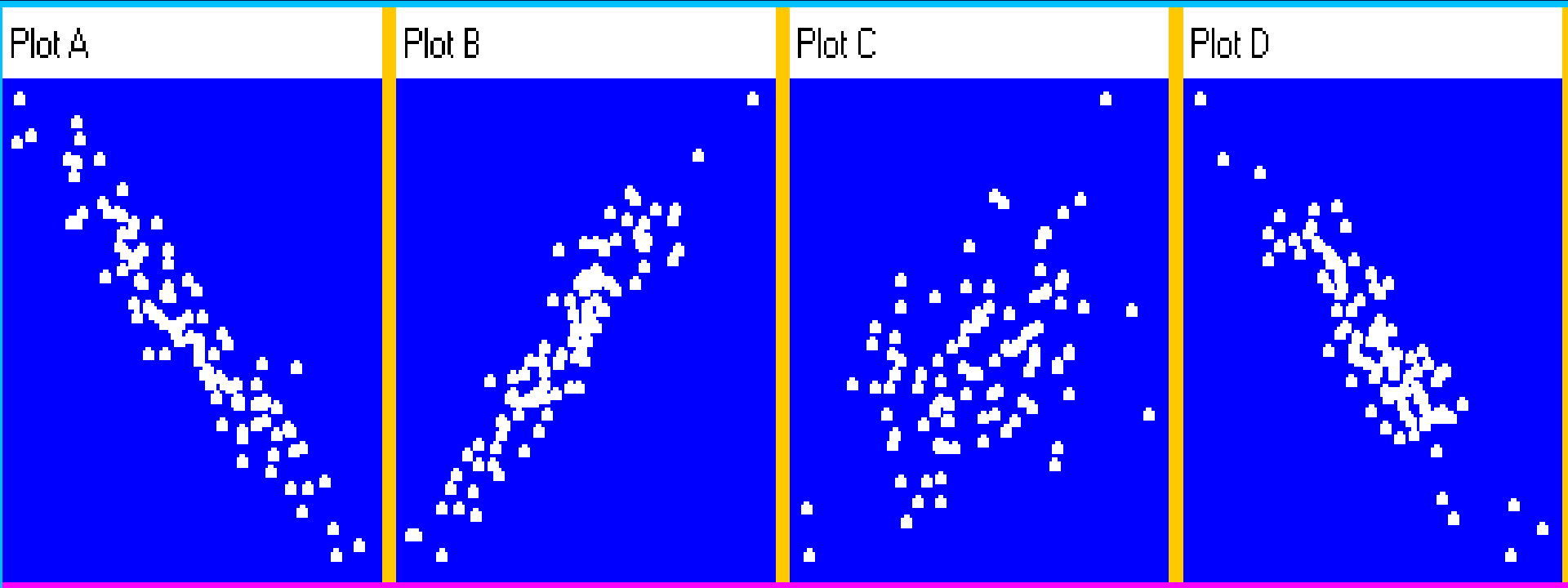
$\sum di^2=64$

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

Comment:

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

exercise



r = -0.95	A	B	C	D
r = -0.9	A	B	C	D
r = 0.49	A	B	C	D
r = 0.93	A	B	C	D

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**"

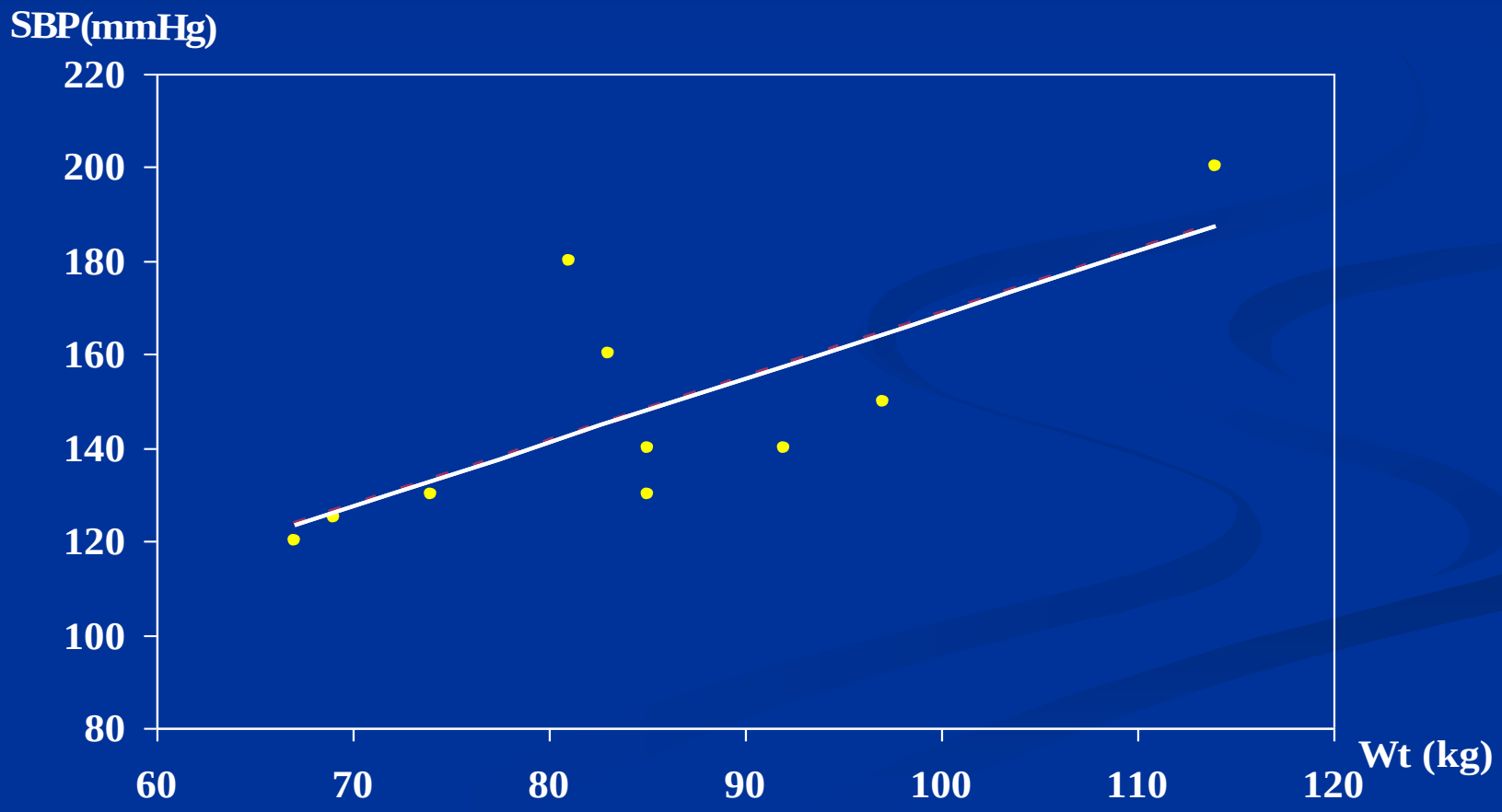
- **Regression** tells us how to draw the straight line described by the correlation

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:

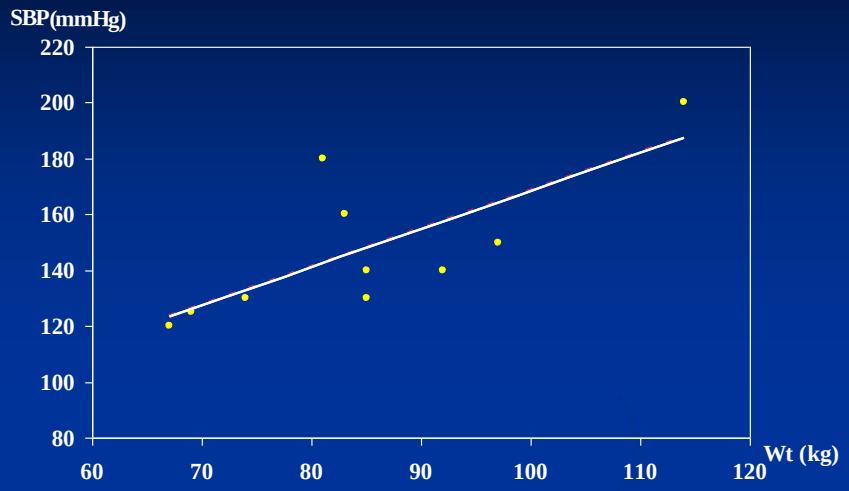
$$\hat{y} = a + bX$$

$$a = \bar{y} - b\bar{x}$$

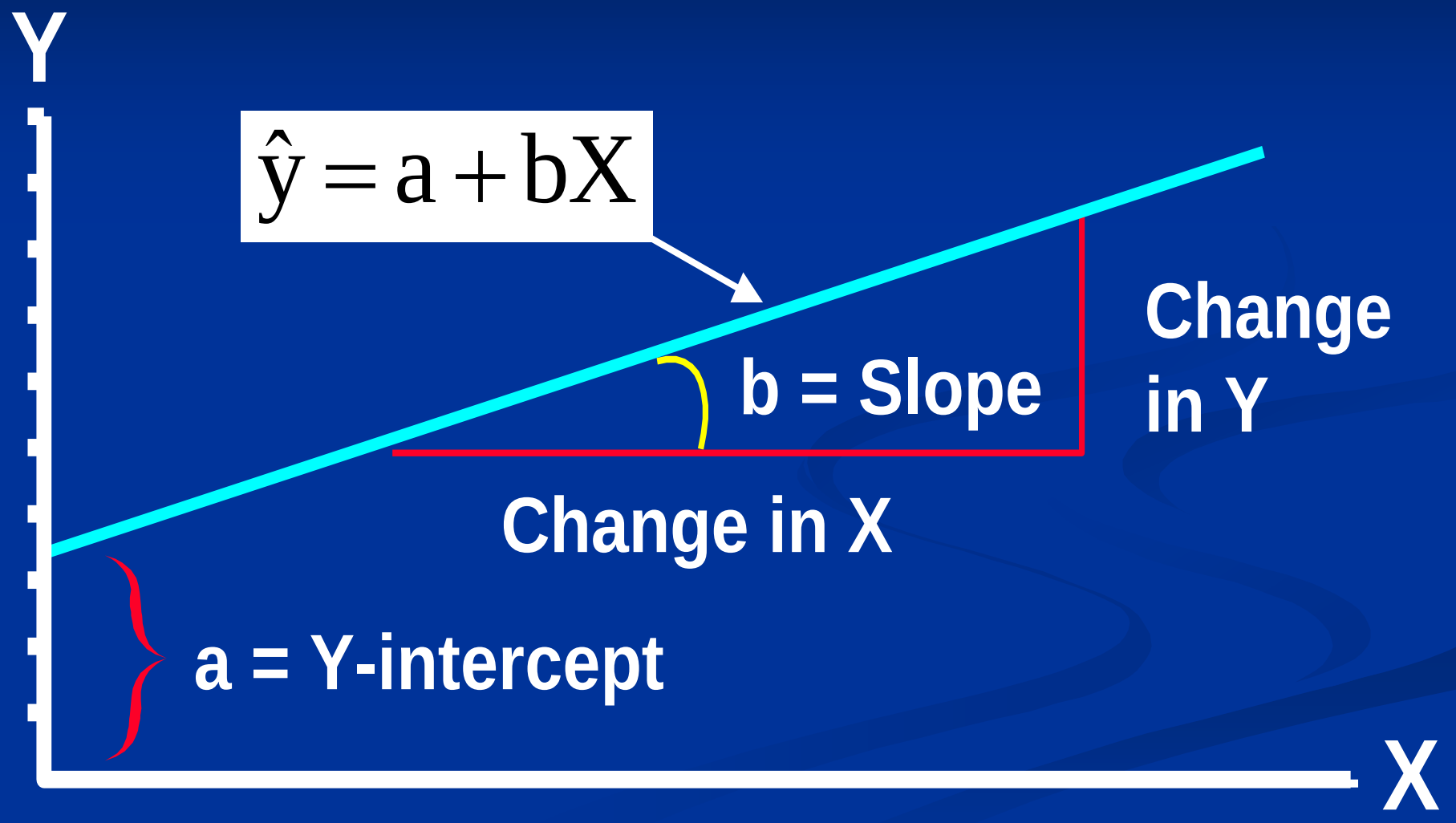
$$b = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

Regression Equation

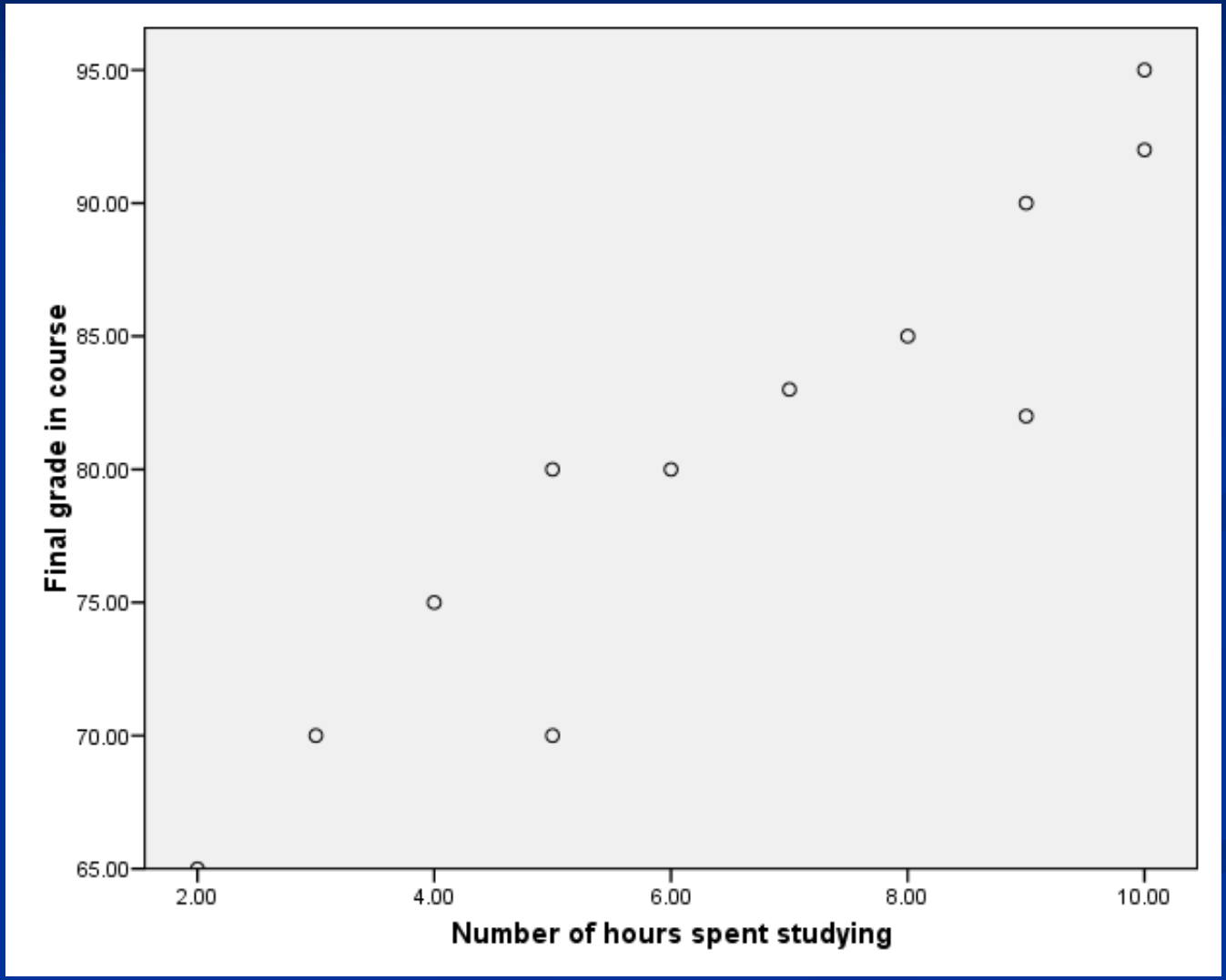
- Regression equation describes the regression line mathematically
  - Intercept
  - Slope



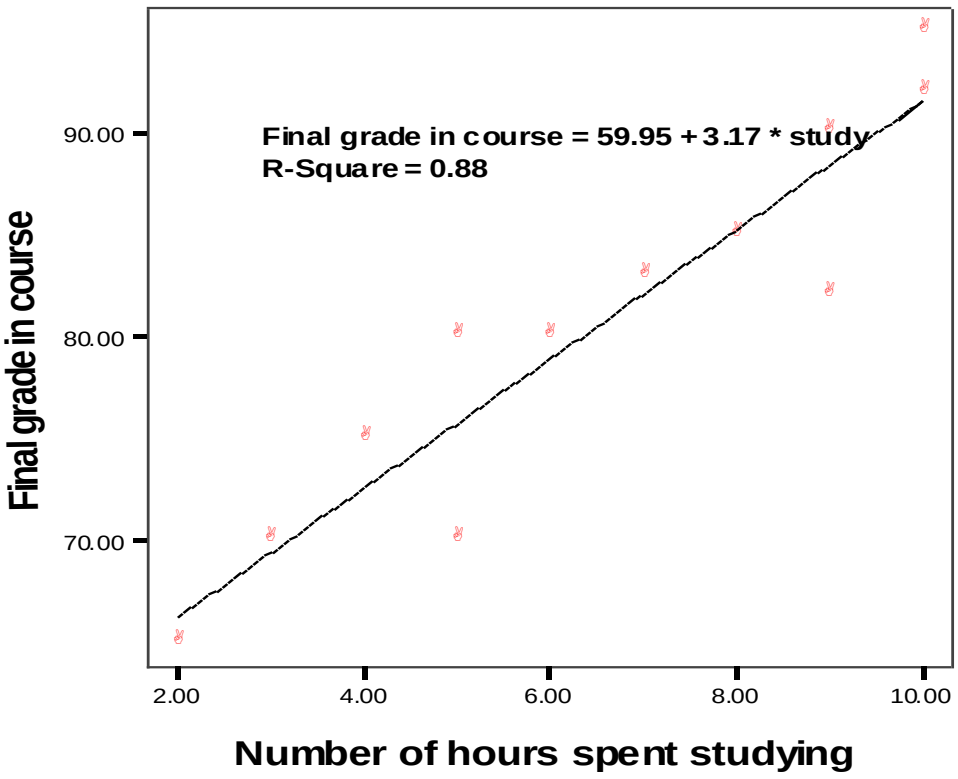
Linear Equations



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

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

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

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

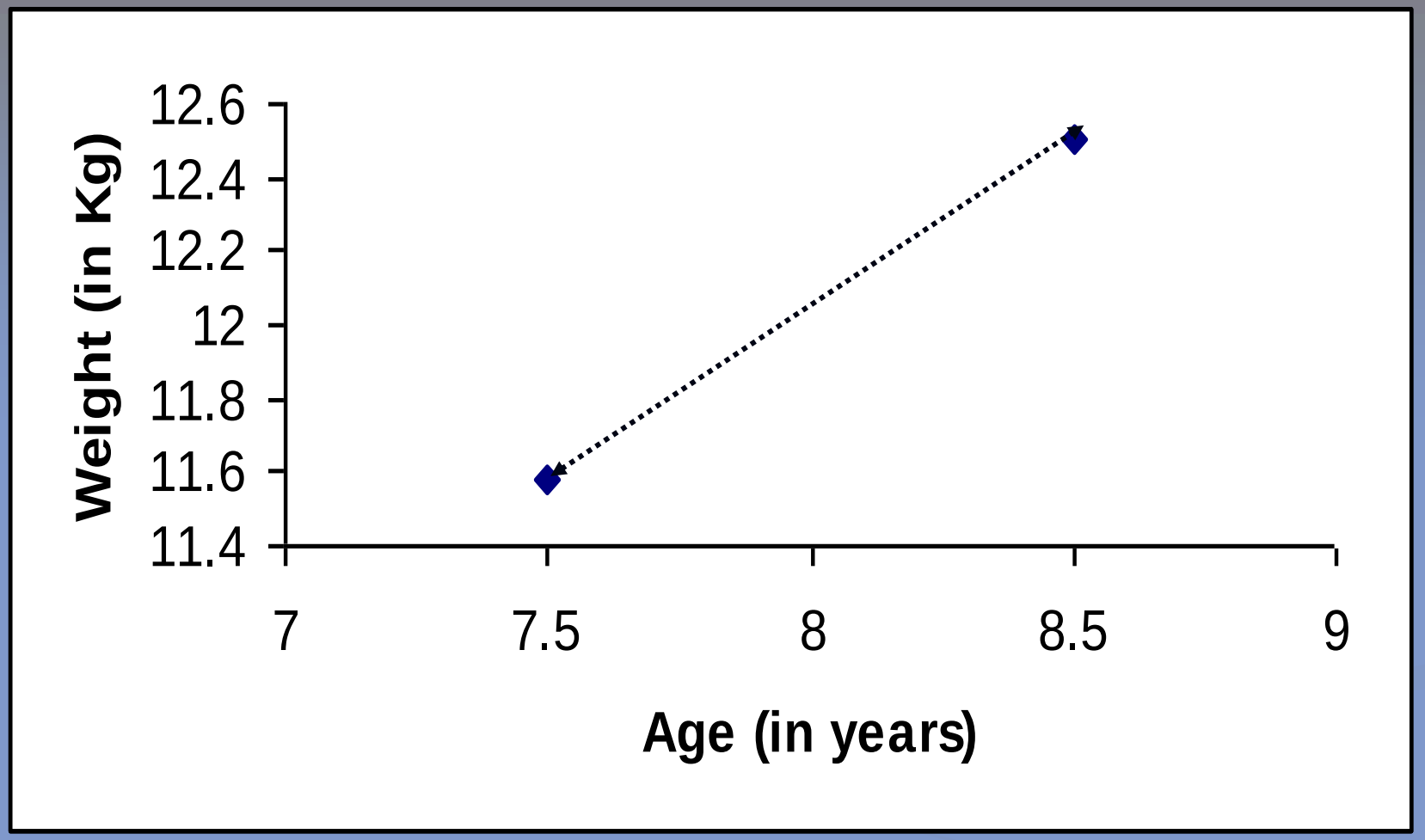
Regression equation

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

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

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

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



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 systolic 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

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

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

for age 25

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

Multiple Regression

Multiple regression analysis is a straightforward extension of simple regression analysis which allows more than one independent variable.

Thank You