Correlation and Regression

Dr. Prashant Singh Rana
psrana@gmail.com

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.

Example

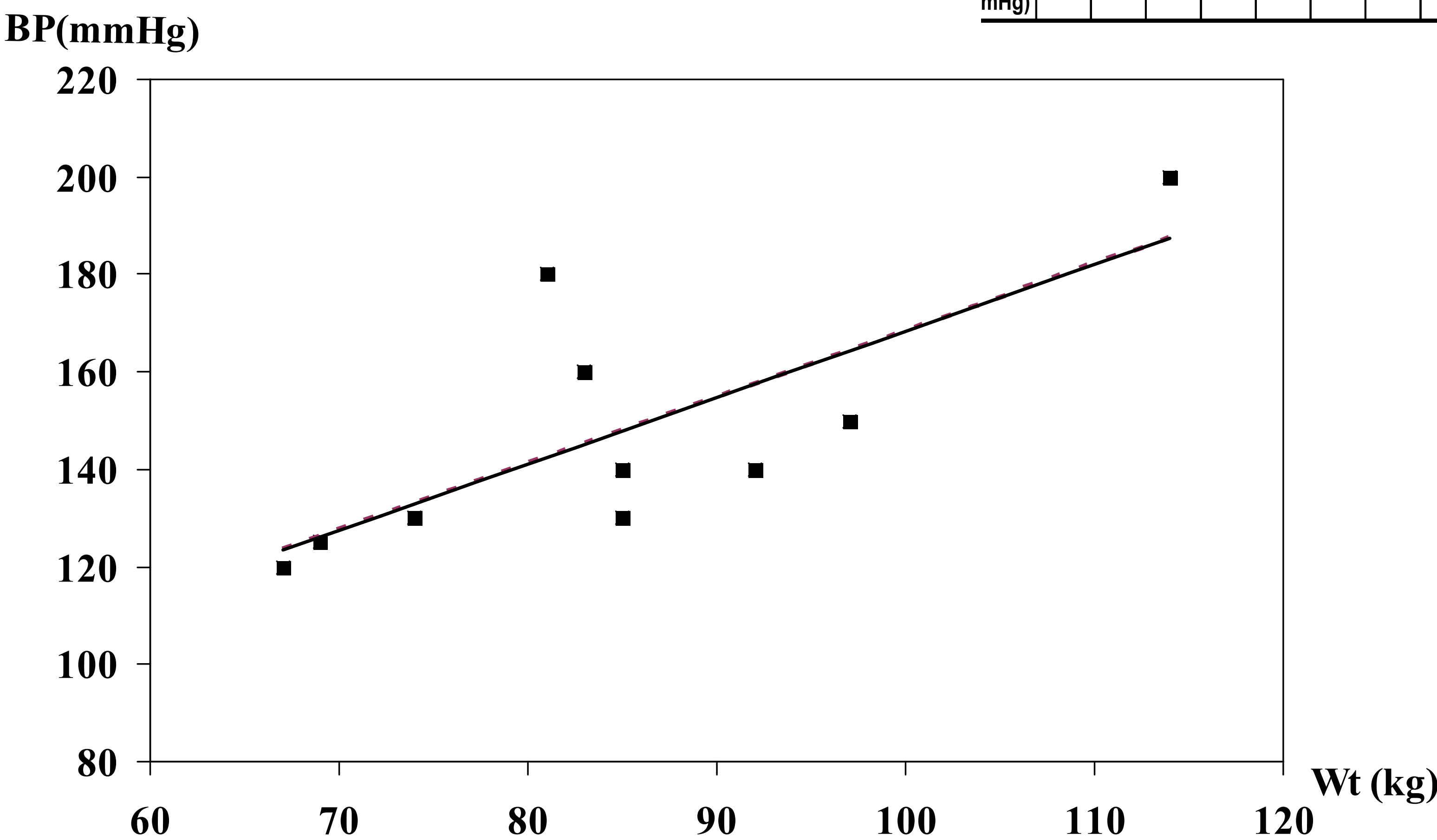
Weight of a human and its Blood Pressure

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

Scatter Plot

Weight vs Blood Pressure

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

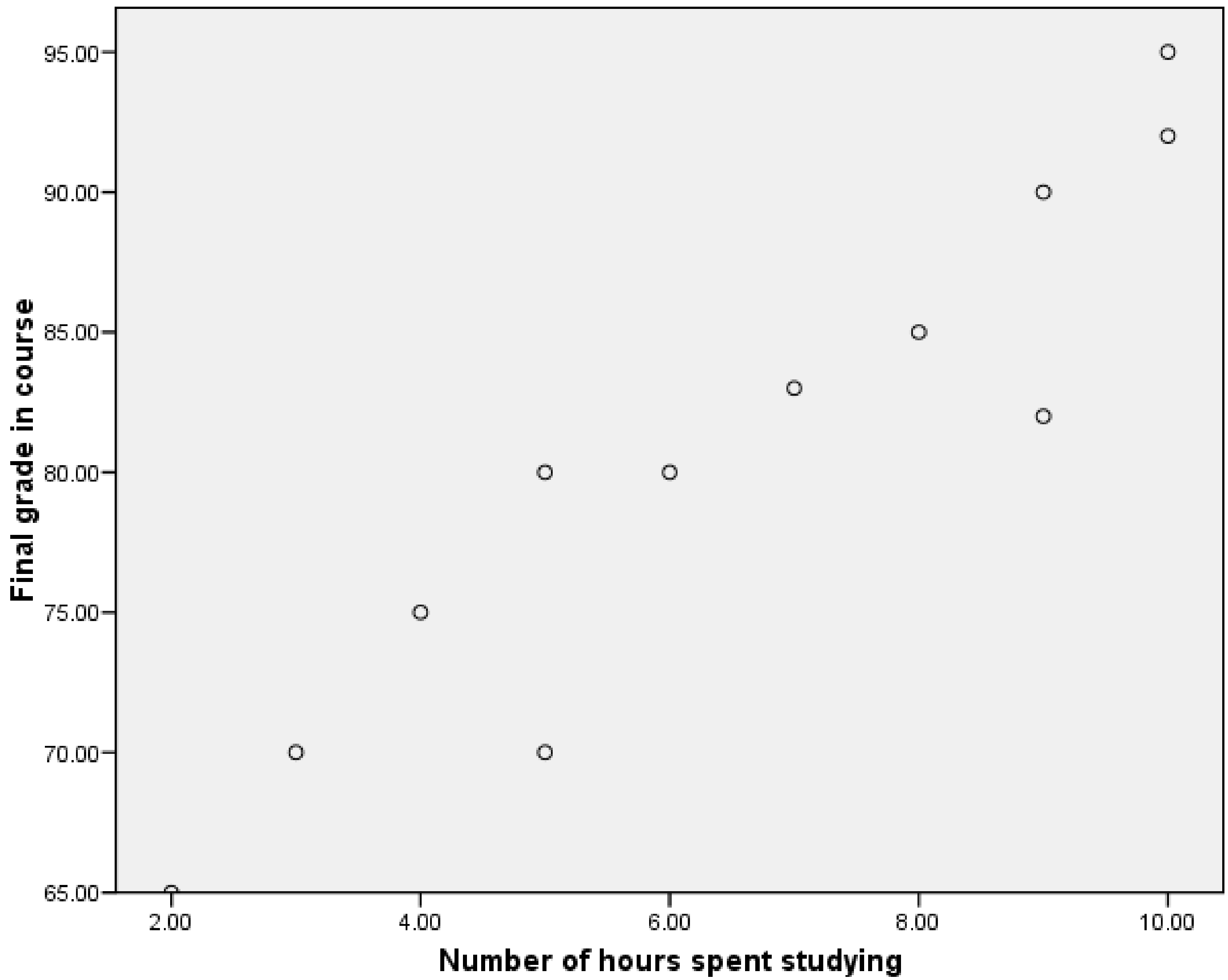


Scatter Plots

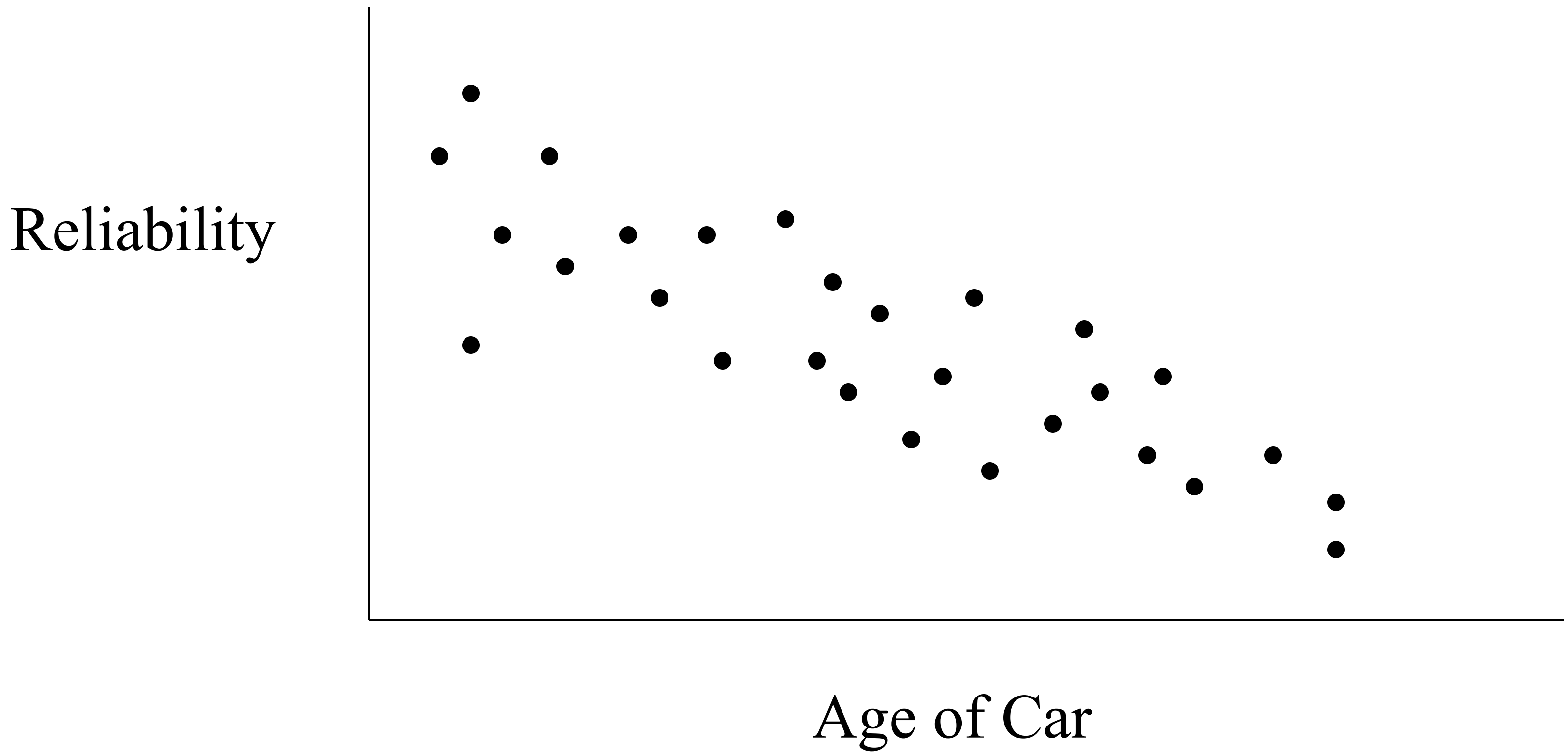
The pattern of data is indicative of the type of relationship between your two variables:

- Positive relationship
- Negative relationship
- No relationship

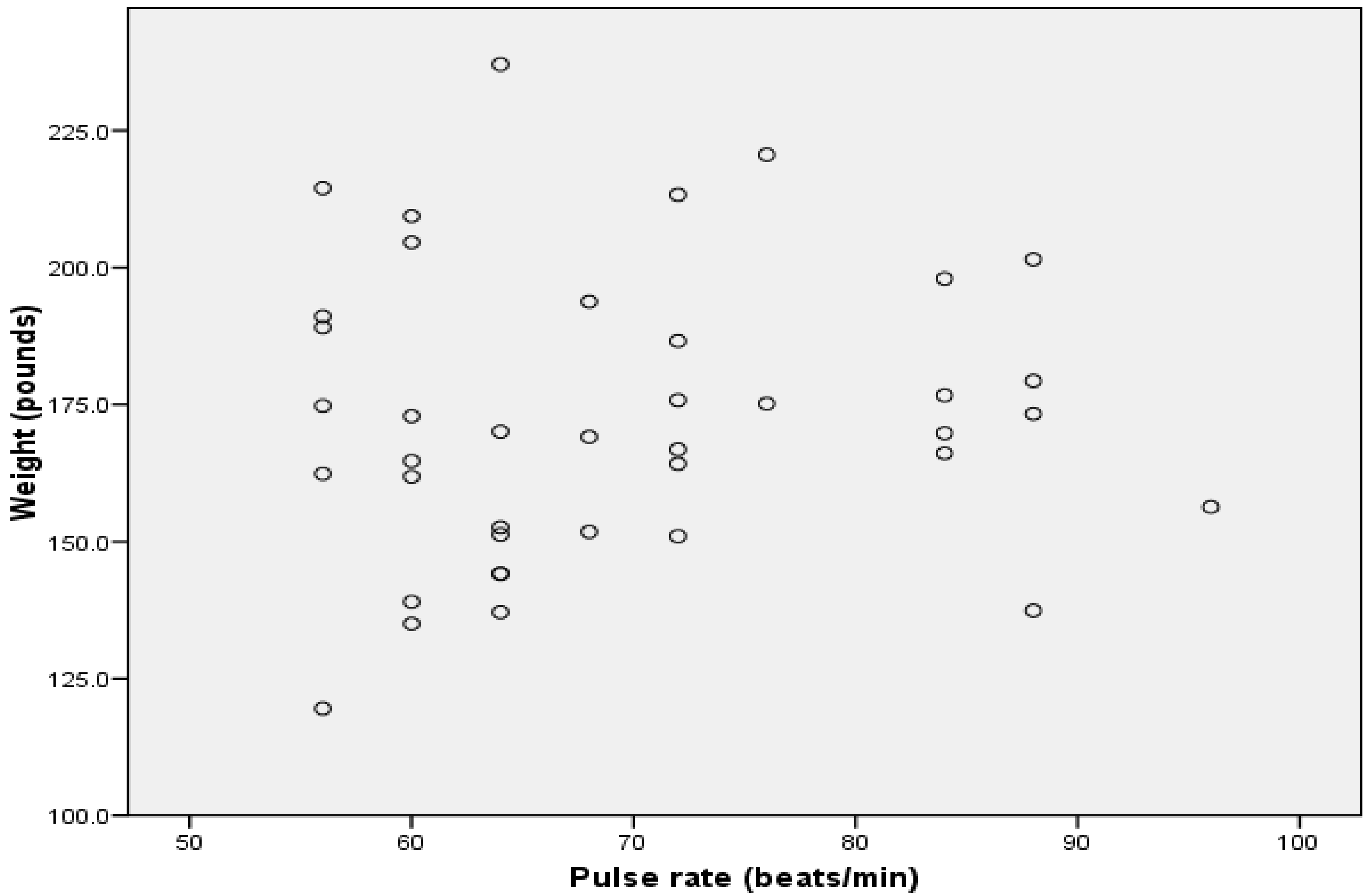
Example: Positive Relationship



Example: Negative relationship



Example: No relation



Simple Correlation Coefficient (r)

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

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

Simple Correlation Coefficient (r)

- The sign of r denotes the nature of association

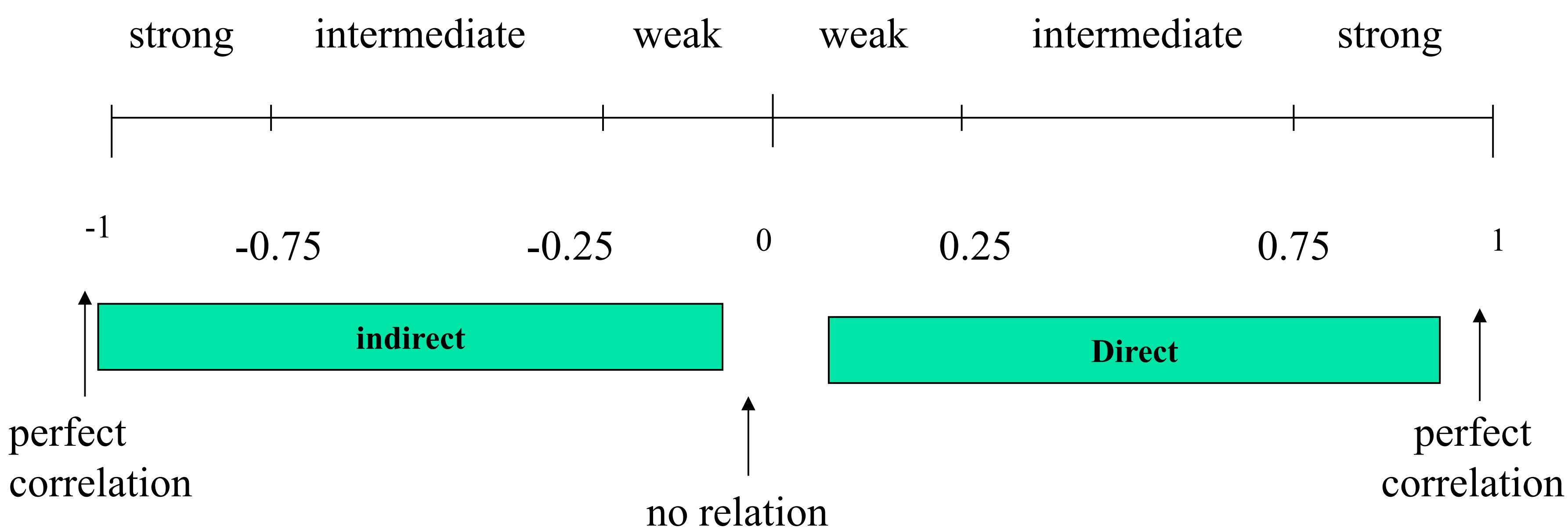
- while the value of r denotes the strength of association.

Simple Correlation Coefficient (r)

- 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).

Simple Correlation Coefficient (r)

- The value of r ranges between (-1) and (+1)



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.

SN	Age (years)	Weight (Kg)
1	7	12
2	6	8
3	8	12
4	5	10
5	6	11
6	9	13

Example

Correlation coefficient using the following formula:

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

Example

SN	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	$\sum x=41$	$\sum y=66$	$\sum xy= 461$	$\sum x2=291$	$\sum y2=742$

Example

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

r = 0.759

strong direct correlation

Example

Relationship between Anxiety and Test Scores

Anxiety (X)	Test score (Y)
10	2
8	3
2	9
1	7
5	6
6	5

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

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 ∑di2 which is the sum of the squared values.

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

$\sum di^2 = 64$

Answer

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

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

Comment:

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

Spearman Rank Correlation Coefficient

- It is a non-parametric measure of correlation.
- Spearman Rank correlation coefficient could be computed in the following cases:
  - Both variables are quantitative.
  - Both variables are qualitative ordinal e.g.
    - Student Grade (A, A-, B, B-,C, E)
    - Product Rating (1star….. 5star).
  - One variable is quantitative and the other is qualitative ordinal.

Question: Do it yourself

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

sample numbers	Position (X)	Income (Y)
A	Teaching Assistant	25
B	Lecturer	65
C	Assistant Professor	100
D	Associate Professor	140
E	Professor	200
F	Associate Professor	140
G	Assistant Professor	110

Question: Do it yourself

Two columns are randomly defined between 1 and 10. What should be the correlation ?

X	Y
8	9
9	4
4	2
1	8
3	6
7	5
8	6
7	10
8	4
4	4

Question: Do it yourself

Two columns are randomly defined between 1 and 10. What should be the correlation ?

X	Y
8	9
9	4
4	2
1	8
3	6
7	5
8	6
7	10
8	4
4	4

Week Correlation

Question: Do it yourself

Two columns are randomly defined between 1 and 10.

Pearson correlation is: 0.013.

Find the spearman's correlation and comment

X	Y
8	9
9	4
4	2
1	8
3	6
7	5
8	6
7	10
8	4
4	4

Question: Do it yourself

Two columns are randomly defined between 1 and 10.
Pearson correlation is: 0.013.
Find the spearman's correlation and comment

X	Y
8	9
9	4
4	2
1	8
3	6
7	5
8	6
7	10
8	4
4	4

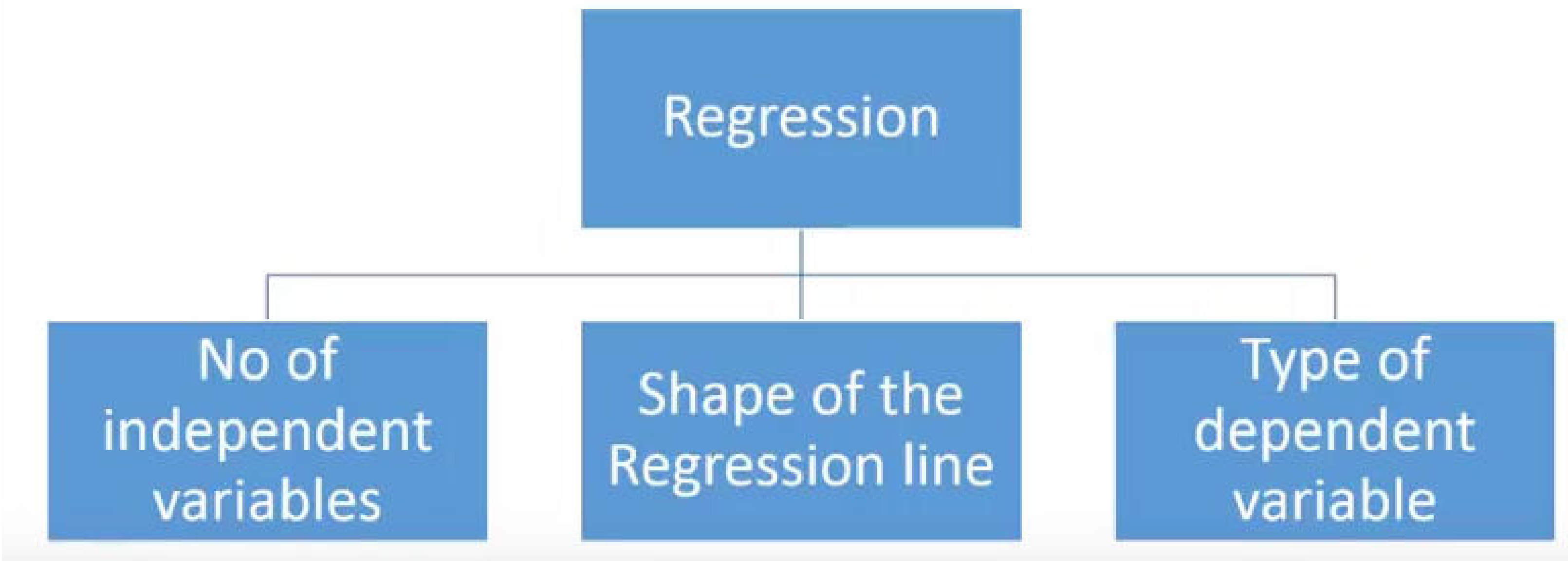
Spearman Correlation: -0.03
Both are almost (ignore sign): 0

Conclusion:
If dataset is discrete then both the correlations are almost same.

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

Types of Regression

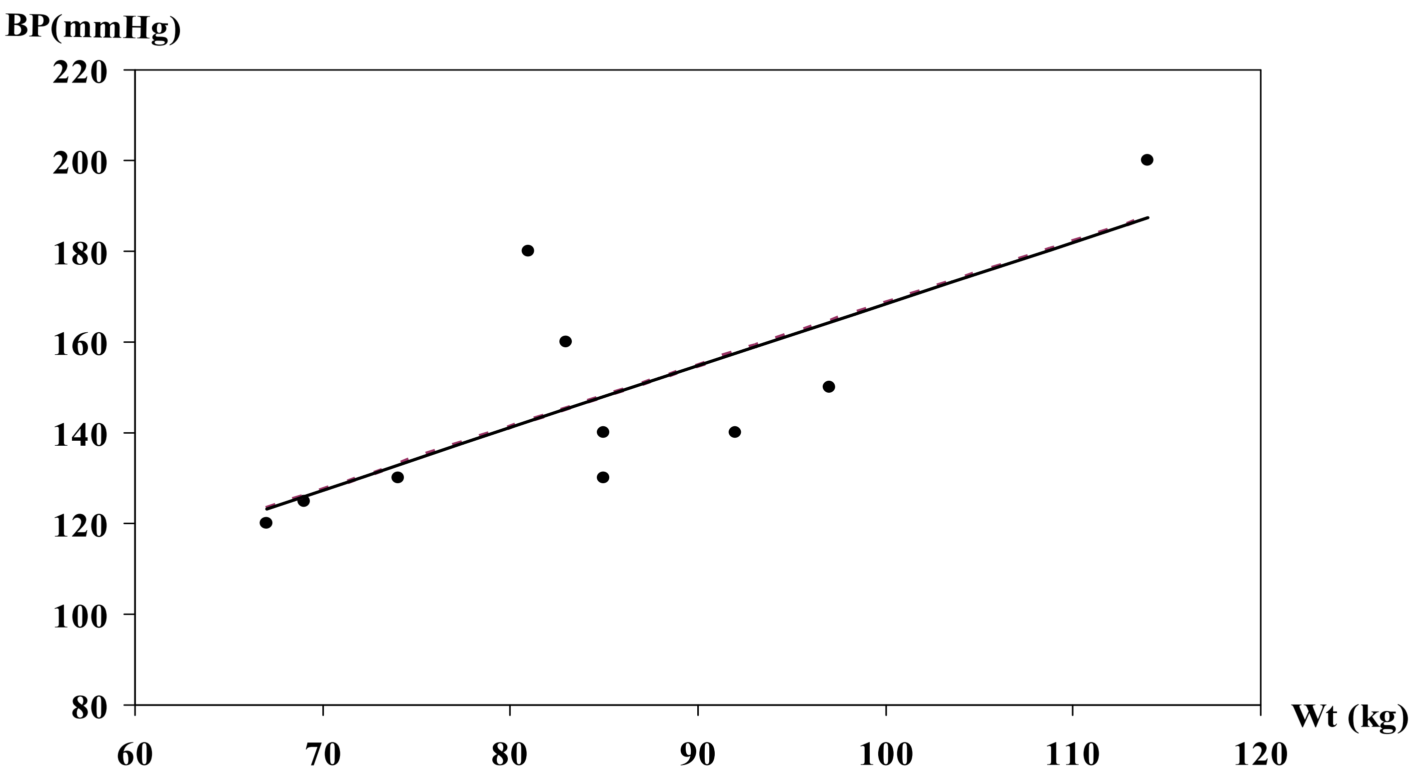


Regression

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

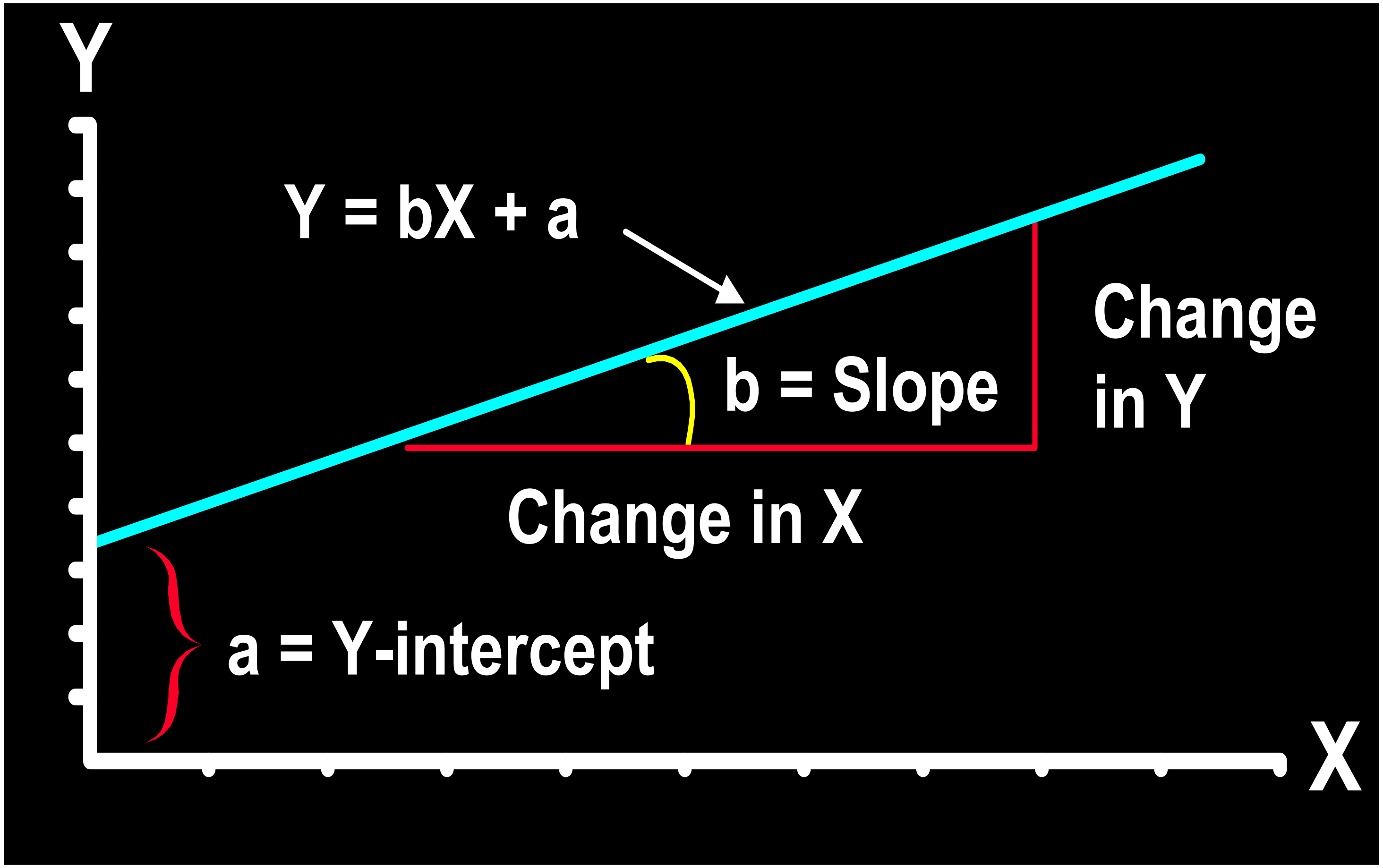
By using the least squares method (a procedure that minimizes the vertical deviations of plotted points surrounding a straight line) we are able to construct a best fitting straight line to the scatter diagram points and then formulate a regression equation in the form of:

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

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

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

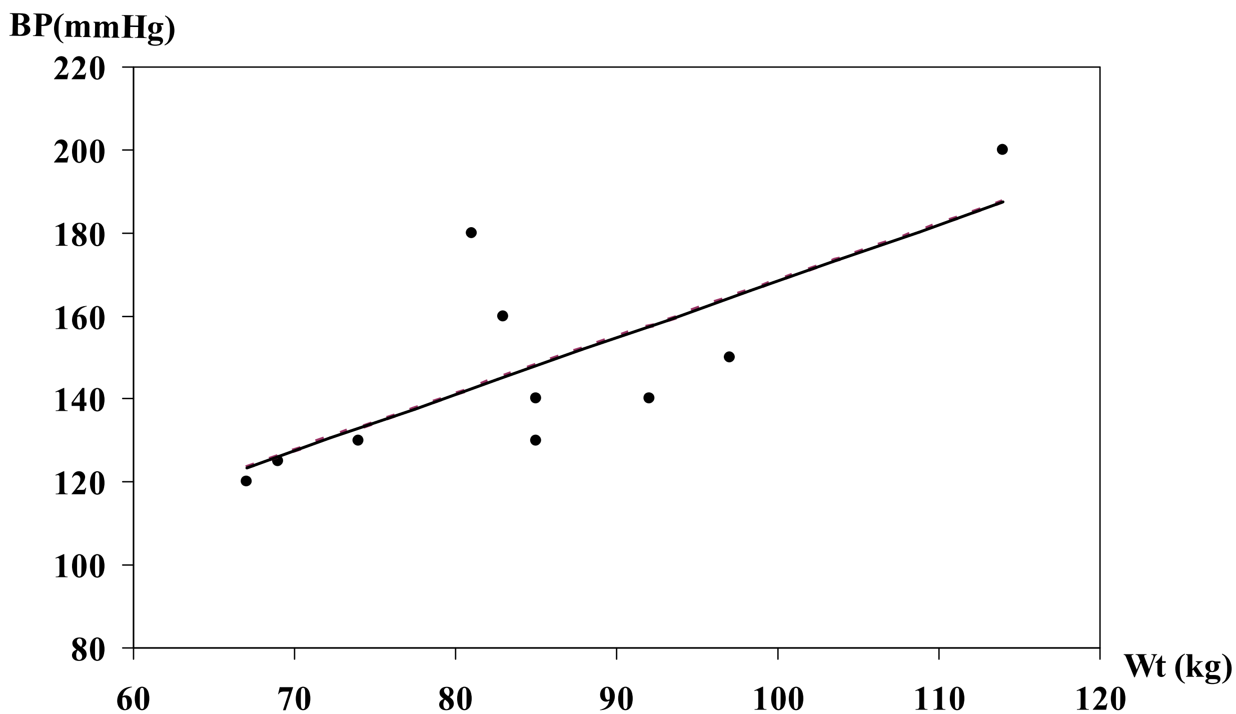
Linear Equation



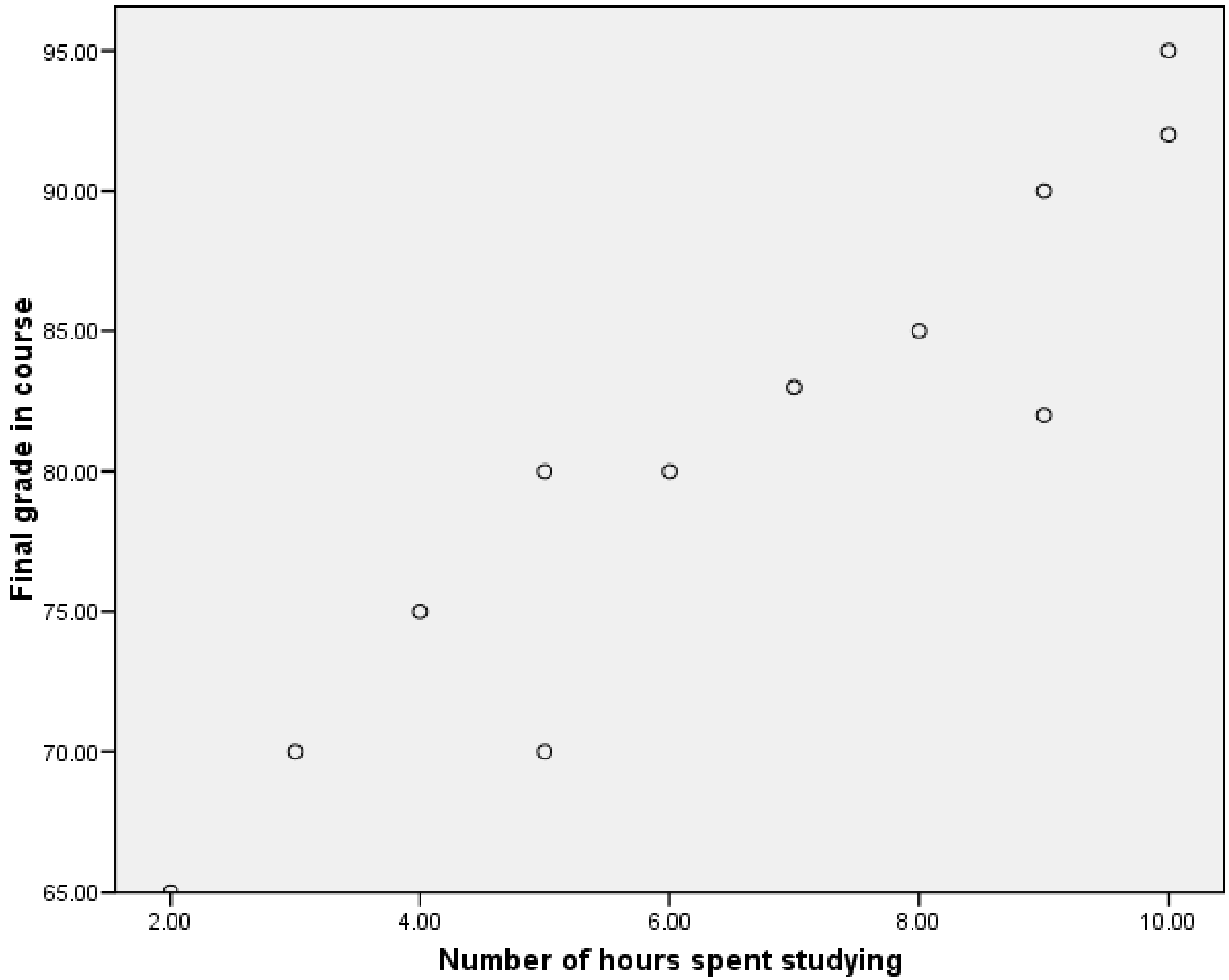
Regression Equation

Regression equation describes the regression line mathematically

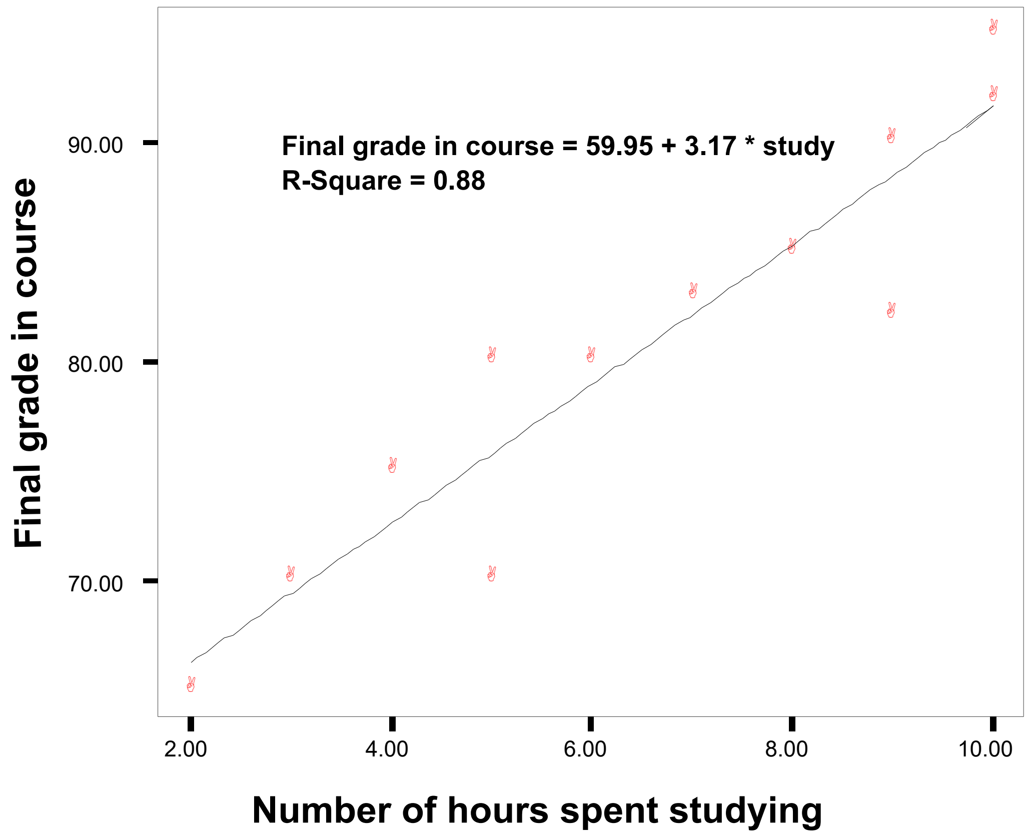
- Intercept
- Slope



Hours studying and grades



Regressing grades on hours



Predicted final grade in class = 59.95 + 3.17*(n)

n = number of hours you study per week

Results

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

Question

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.

SN	Age (x)	Weight (y)
1	7	12
2	6	8
3	8	12
4	5	10
5	6	11
6	9	13

Find regression equation

SN	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

Find regression equation

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

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

$$= 0.92$$

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

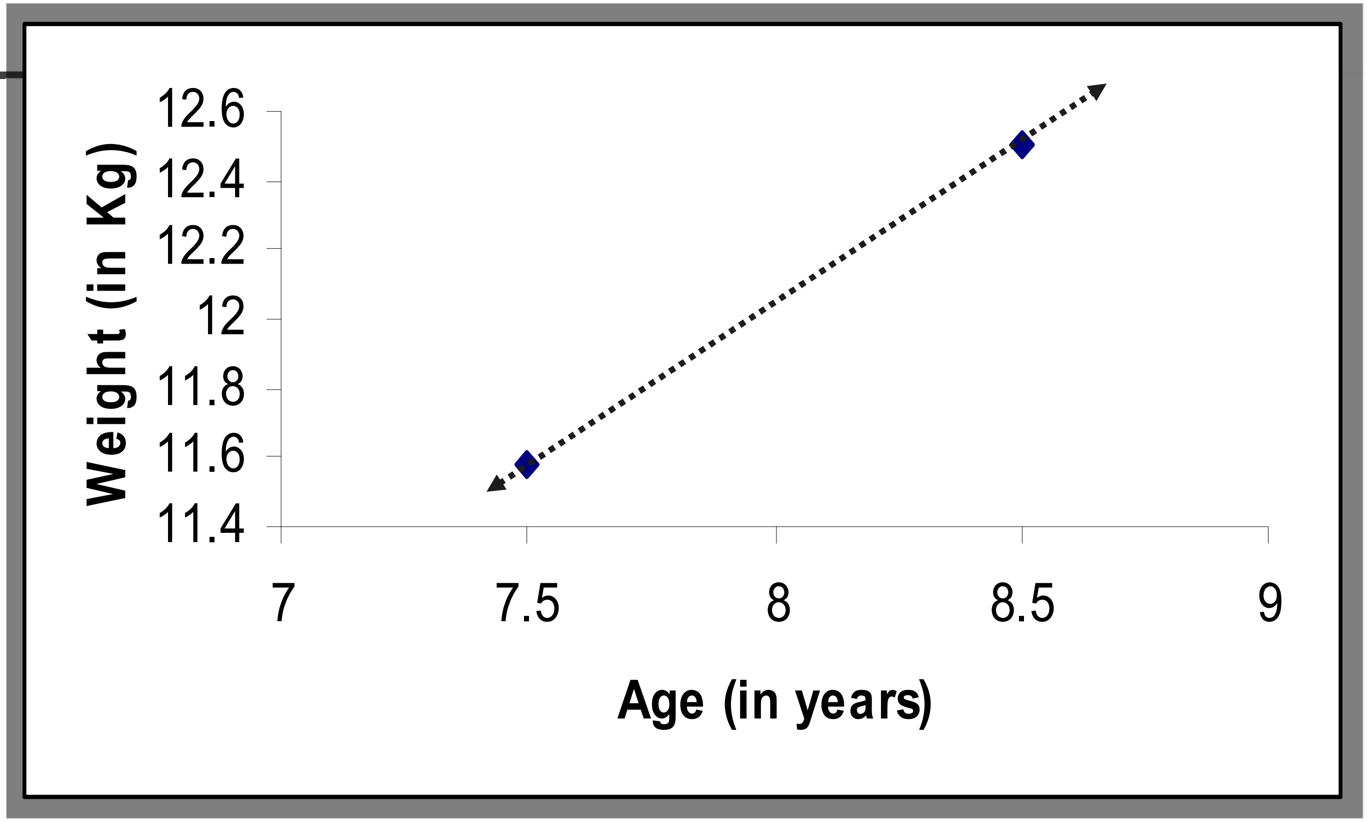
Regression equation

$$\hat{y}_{(x)} = 11 + 0.92(x - 6.83)$$

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

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

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



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

Question:

- 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?

Given Dataset

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

Solution

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

Solution

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**

Solution

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

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

for age 25

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

Regression Analysis

- Regression analysis is a form of predictive modelling technique which investigates the relationship between **dependent** (target) and **independent variable (s)**(predictor).
- This technique is used for forecasting, time series modelling and finding the causal effect relationship between the variables.
- For example, relationship between rash driving and number of road accidents by a driver is best studied through regression

Regression Analysis

- Regression: technique concerned with predicting some variables by knowing others

- The process of predicting variable Y using variable X
  - Uses a variable (x) to predict some outcome variable (y)
  - How values in y change as a function of changes in values of x

Univariate

- One input and one output
- Example:
  - OTP per transaction: Every transaction have unique OTP

Transaction ID	OTP
3424234234	9456
5653453235	9879
5909087556	4536
8797890123	2345

Multivariate

- Multiple inputs and one output
- Example:
  - Cancer Prediction
  - Cement Mixture strength

x1	x2	x3	x4	x5	Strength
17	0	-5	0.784245	37	26
12	0	-10	0.587296	25	27
18	0	-7	0.876622	40	25
11	0	-7	0.80826	24	23
18	0	-4	0.83215	37	28
10	1	-9	0.62842	27	28
19	0	7	0.522811	44	30
19	-1	4	0.548609	37	23
15	0	-6	0.177904	46	20

Multiple Regression

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

- Cover in next class

Thank

You