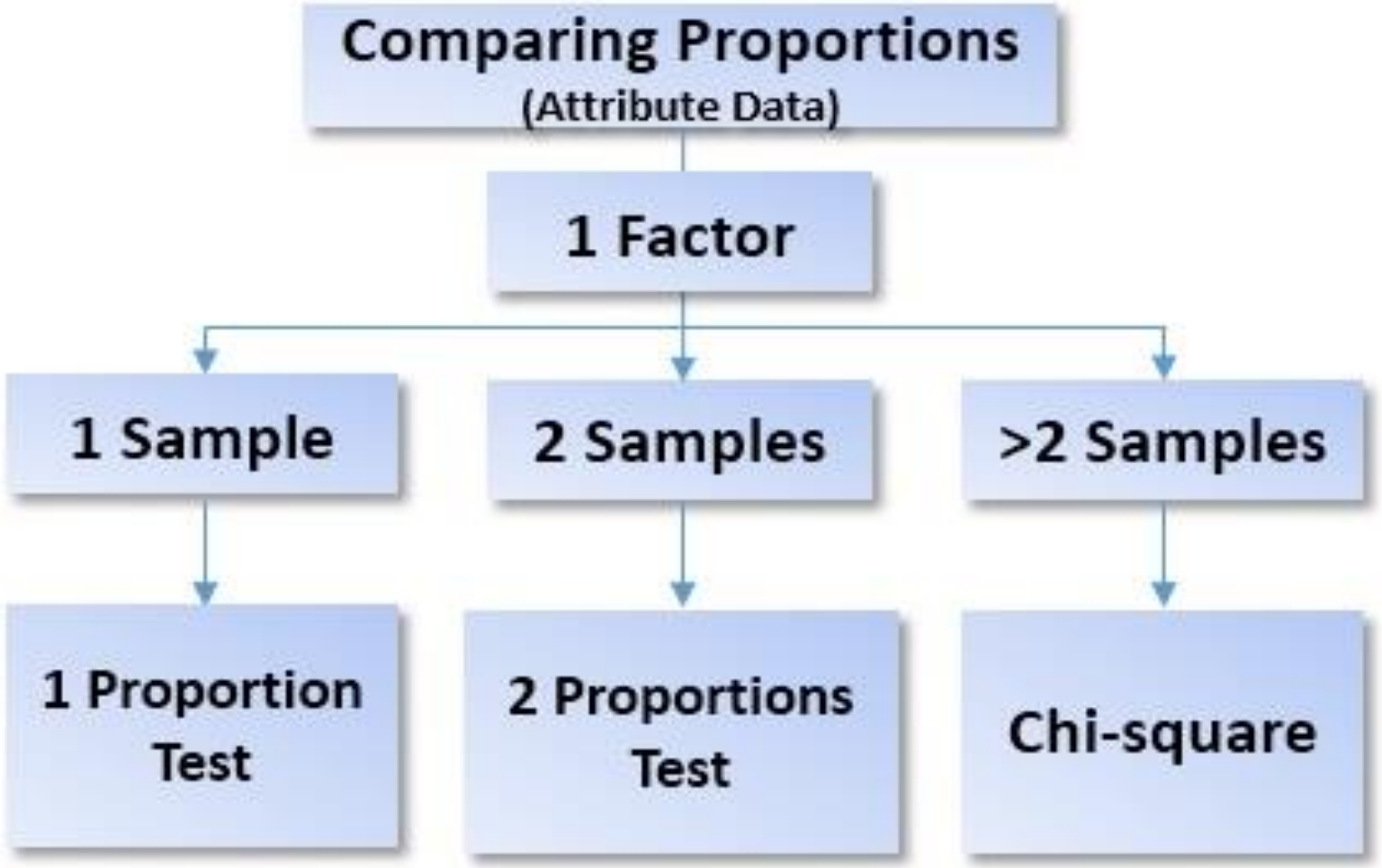
Proportions Flowchart

www.six-sigma-material.com



Regression and Correlation

PROF DR NAJLAA FAWZI


WHAT IS
CORRELATION ?
shutterstock
www.shutterstock.com · 271612295

Correlation & regression analysis are two procedures used to analyze associations involving continuous or interval/ ratio data.

1- Correlation analysis measures the strength of the association between two study variables. The term correlation analysis is refers to Pearson's Product Moment Correlation Coefficient (also known as Pearson's r).

2- Regression analysis derives a prediction equation for estimating the value of one variable given the value of the second.

Regression

The idea is similar to and sometimes confused with correlation.
It is important to clarify the difference between correlation and regression.

Correlation only indicates the strength of the relationship between two factors or parameters.

Regression measures the relationship

Regression is used only when there is cause–effect relationship.
It can quantify the relation: that is to say, once regression is applied, one parameter can be calculated if the other parameter is known.

Pearson's correlation coefficient is used for parametric data (following normal distribution curve), and **Spearman correlation coefficient** is used for nonparametric data.

Correlation is not an all-or-none phenomenon. There may be multiple factors correlating with a variable.

USES OF CORRELATION AND REGRESSION

There are three main uses for correlation and regression.

- **One is to test hypotheses about cause-and-effect relationships.**

In this case, the experimenter determines the values of the X-variable and sees whether variation in X causes variation in Y.
For example, giving people different amounts of a drug and measuring their blood pressure.

- **The second main use for correlation and regression is to see whether two variables are associated, without necessarily inferring a cause-and-effect relationship.**

In this case, neither variable is determined by the experimenter; both are naturally variable. If an association is found, the inference is that variation in X may cause variation in Y, or variation in Y may cause variation in X, or variation in some other factor may affect both X and Y.

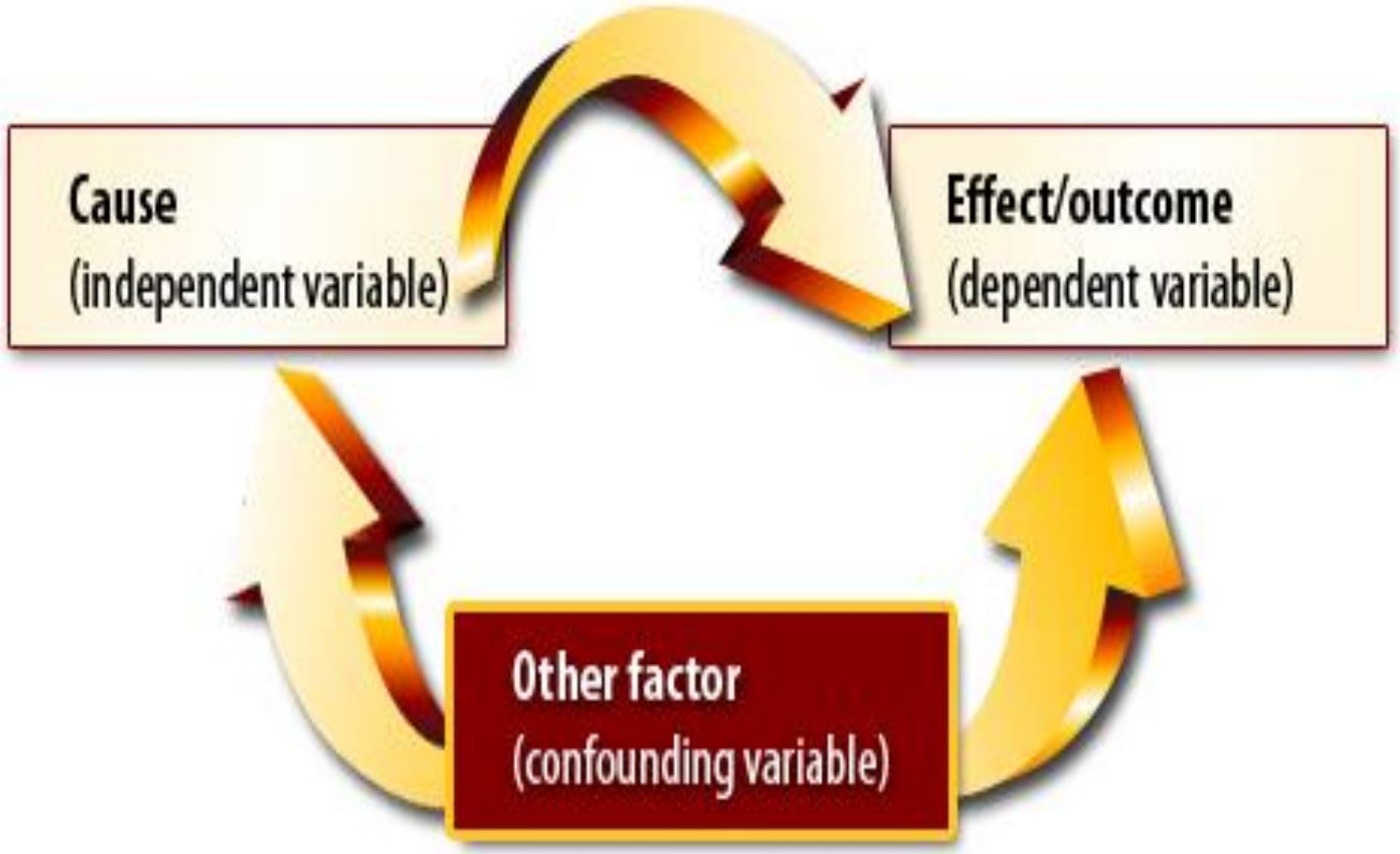
- **The third common use of linear regression is estimating the value of one variable corresponding to a particular value of the other variable.**

CORRELATION *VS* REGRESSION

BASIS OF COMPARISON	CORRELATION	REGRESSION
1. Meaning	Correlation is a statistical measure which determines co relationship or association of two variables.	Regression describes how an independent variable is numerically related to the dependent variable.
2. Usage	Correlation is used to represent linear relationship between two variables.	Regression is used to fit a best line and estimate one variable on the basis of another variable.
3. Indicates	Correlation coefficient indicates the extent to which two variables move together.	Regression indicates the impact of a unit change in the known variable (X) on the estimated variable (Y).
4. Objective	To find a numerical value expressing the relationship between variables.	To estimate values of random variable on the basis of values of fixed variables.

Difference Between correlation and Regression

Correlation	Regression
1) Relationship between two are more variables.	Average Relationship between two are more variables .
2) X and Y are random variable	X is the random variable and Y is the fixed variable.
3) It gives limited information after verifying the Relationship between variables .	It is used for the prediction of one value, relation to the other given value.
4)The range of relationship lies between -1 and +1.	Regression value is an absolute figure.
5)It studies the linear relationship between the variables.	It studies the linear and non-linear relationship between the variables.
6)If the co-efficient of correlation is positive ,then the two variables are positively correlated and vice versa.	The regression co-efficient explain that the decrease in one variable is associated with the increase in the other variable.


Cause
(independent variable)
Effect/outcome
(dependent variable)
Other factor
(confounding variable)

Data representation and organization

1- The data used in a correlation or regression analysis consist of pairs of measurements made on the same unit of observation (most often , the same study subject).

Each member of the pair corresponds to one of the two study variables.

In a study of relationship between HT and bl. Cholesterol levels ,

SBP and S. cholesterol value are the pair of measurements to be assessed for each study subject and to be represented by the two study variables.

2- The pairs are denoted symbolically (x, y) ,x typically represent the independent variable, while y represent the dependent variable.

3- Variation among the study types

A- In epidemiological studies , the independent variable x is often a suspected risk factor (low fiber diet) and the dependent variable y is the occurrence of disease or other health related out come (occurrence of colorectal cancer).

B- In experimental studies , values of the independent are fixed by the investigator rather than determined by the nature .I.e. in a study of the efficacy of a new antiviral drug , the investigator select the dosage (the independent variable).

Scatter diagram (Dot diagram)

- A scatter diagram is a tool for analyzing relationship between two variables.
- One variable plotted on the horizontal axis and the other is potted on the vertical axis.
- The pattern of their intersecting points can graphically show relationship patterns.
- Most often a scatter diagram is used to prove or disprove cause and effect relationships.
- While the diagram shows relationship , it does not by itself prove that one variable causes the other ,i.e., scatter diagram only show relationship between cause and effect(change in one will change other)

- **But can not prove the variable as a cause of the other.**
- **Correlation and regression are plotted on scatter diagram ⟹ also known as correlation diagram.**

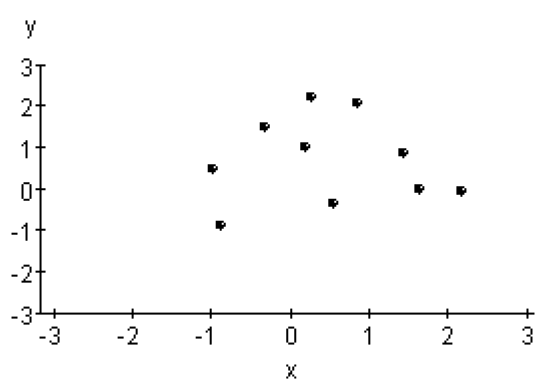
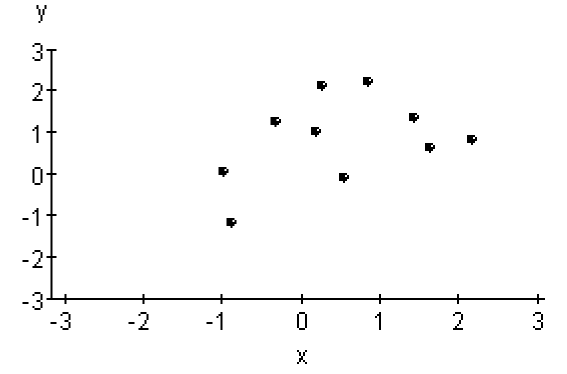
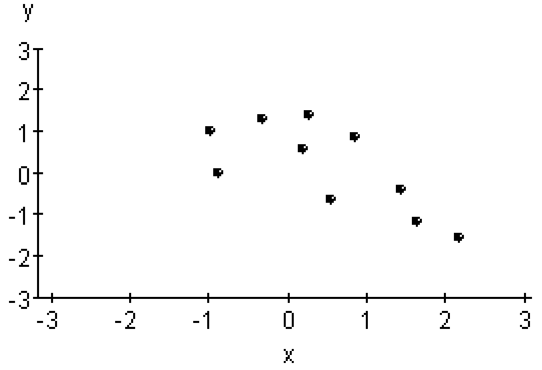
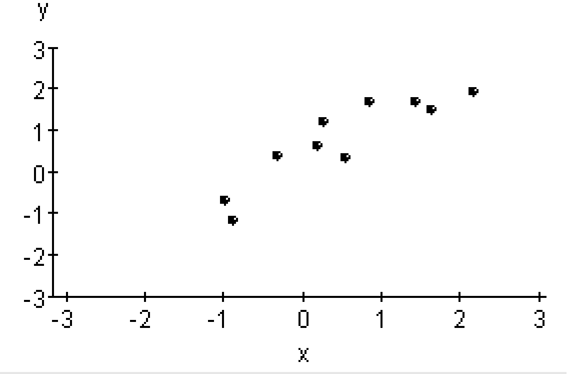
The relationship between x and y may be described by a straight line or by a more complex curvilinear relationship .

Alternatively , the scatter diagram may show that the two variables are unrelated.

1- Liner relationship may be either positive [as the values of one variable increase , the values of the other variable increase as well], or negative [as ,the value of one variable increase , the value of the other decreased].

2- The relationship between x and y may be non linear or curvilinear : the relation between age and death rate is non linear .

3- When x and y are unrelated , the data pairs are randomly distributed , i.e. either linear or non linear exists between foot size and IQ.

Scatter diagram will show one of six possible correlation:

1-Strong positive correlation (+ 1) ~ the value of one variable increase in proportionate manner as the value of the other increases.

2- Strong negative correlation(- 1) ~ the value of one variable decreases proportionally as the value of the other decreases.

3- Weak positive correlation (> 0 to< +1) ~ the value of one variable increases slightly as the value of other increases.

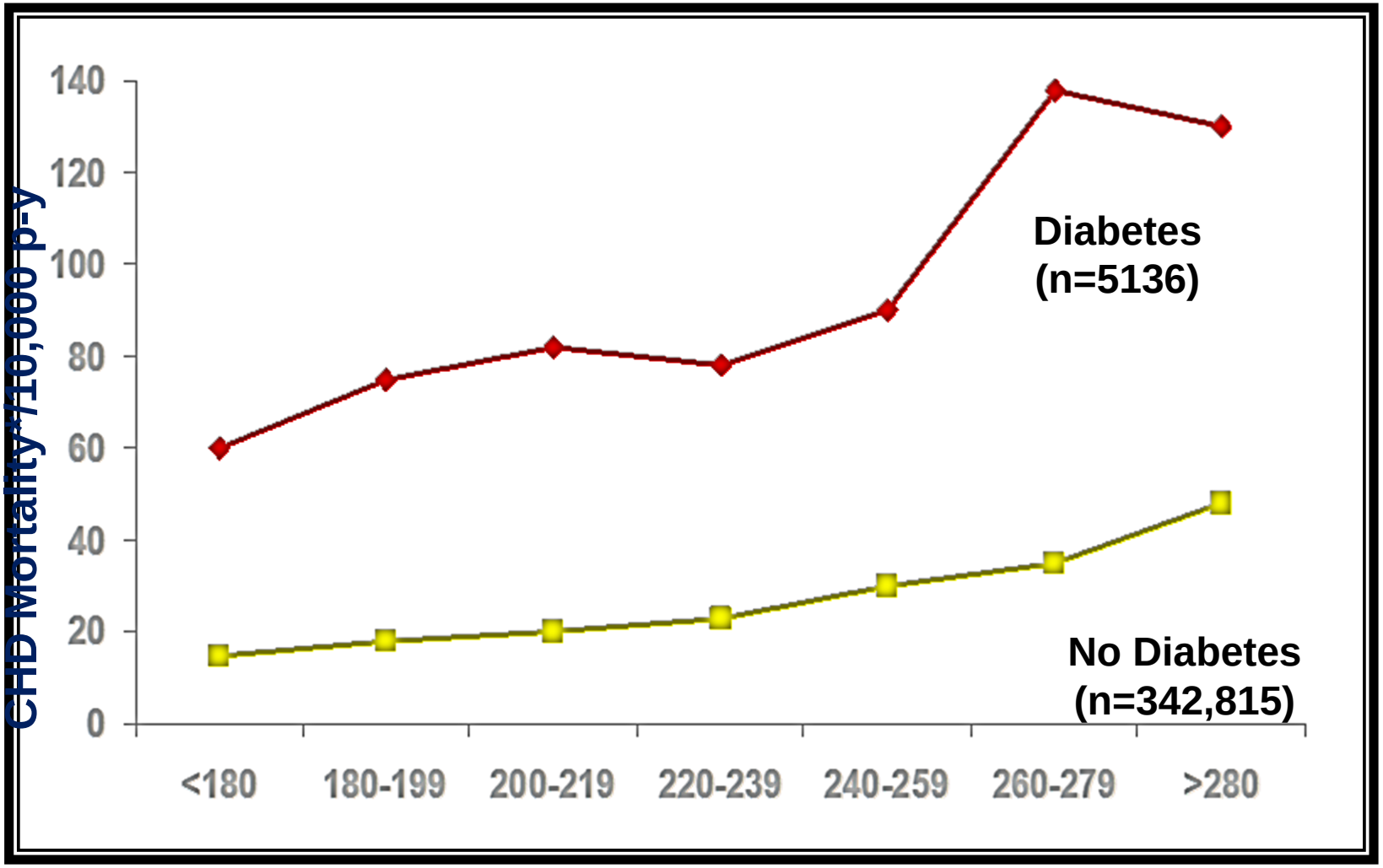
4- Weak negative correlation (between 0 to – 1) ~ the value of one variable decreases slightly as the value of other decreases.

- **Positive Correlation**
  - **Smoking and Lung Damage**

    -gestational age at birth, measured in weeks, and birth weight, measured in grams.
- **Negative Correlation**
  - **Depression & Self-esteem**
  - **Studying & test errors**

5- No correlation (r = 0) ~ there is no change in value of one variable when the value of other changes.

6- Complex correlation ; the value of one variable changes as the value of other changes, but the relationship is not easily determined



Serum Cholesterol (mg/dl)

CORRELATION

- **Establish and quantify strength and direction of relationship between 2 variables by**
  - **"Correlation coefficient" symbol : r**
  - **Range -1 or + 1**
  - **Size of coefficient: signifies strength**
  - **Sign of coefficient : signifies direction**
  - **2 types of correlation coefficients used**
    - **Pearson correlation coefficient: interval/ratio scale**
    - **Spearman correlation coefficient: ordinal scale data**

1- Characteristics of (r)

r is an index number between (-1& +1), where -1 , the two variables have a perfect negative linear relationship.

r = +1 , the study variables have a perfect positive linear relationship.

Degree of correlation:

1.Perfect: If the value is near ± 1, then it said to be a perfect correlation: as one variable increases, the other variable tends to also increase (if positive) or decrease (if negative).

2.High degree: If the coefficient value lies between ± 0.50 and ± 1, then it is said to be a strong correlation.

3.Moderate degree: If the value lies between ± 0.30 and ± 0.49, then it is said to be a medium correlation.

4.Low degree: When the value lies below + .29, then it is said to be a small correlation.

5.No correlation: When the value is zero.

Correlation Coefficient

- **The population correlation coefficient ρ (rho) measures the strength of the association between the variables**

- **The sample correlation coefficient r is an estimate of ρ and is used to measure the strength of the linear relationship in the sample observations**

- **r correlation for a sample**
  - **based on a the limited observations we have**
- **ρ actual correlation in population**
  - **the true correlation**

- **Beware Sampling Error!!**
  - **even if ρ=0 (there's no actual correlation), you might get r =.08 or r = -.26 just by chance.**
  - **We look at r, but we want to know about ρ**

2- Calculation

$$r = S_{xy} / \sqrt{(S_x)(S_y)}$$

$$S_{xy} = \sum xy - (\sum x)(\sum y)/n$$

$$S_x = \sum x2 - (\sum x)^2 / n$$

$$S_y = \sum y2 - (\sum y)^2 /n$$

Assessing the statistical significance of an association:

Is there a statistically significant linear relationship between two study variables in the population from which the sample was selected .

State the hypothesis : Ho is : x variable levels are not linearly related to y in the population from which the sample was selected .

- **Two possibilities**
  - **Ho: ρ = 0 (no actual correlation; The Null Hypothesis)**
  - **Ha: ρ ≠ 0 (there is some correlation; The Alternative Hypo.)**

•Test statistic

Testing significance of (r) is done by one of:

1-Using t-test.

$$t = r\sqrt{n-2}/1-r^2$$

$$t = \frac{r_{xy}\sqrt{n-2}}{\sqrt{1-r_{xy}^2}}$$

tabulated t = df n - 2 t 1- α

Table D.5 ***Values of Pearson's product moment correlation coefficient (r). The results are significant if the calculated value of r is higher than the table value***

df $(n-2)$	r		df $(n-2)$	r	
	$P = 0.05$	$P = 0.01$		$P = 0.05$	$P = 0.01$
1	**0.997**	1.000	16	**0.468**	0.590
2	**0.950**	0.990	17	**0.456**	0.575
3	**0.878**	0.959	18	**0.444**	0.561
4	**0.811**	0.917	19	**0.433**	0.549
5	**0.754**	0.874	20	**0.423**	0.537
6	**0.707**	0.834	21	**0.413**	0.526
7	**0.666**	0.798	22	**0.404**	0.515
8	**0.632**	0.765	23	**0.396**	0.505
9	**0.602**	0.735	24	**0.388**	0.496
10	**0.576**	0.708	25	**0.381**	0.487
11	**0.553**	0.684	26	**0.374**	0.479
12	**0.532**	0.661	27	**0.367**	0.471
13	**0.514**	0.641	28	**0.361**	0.463
14	**0.497**	0.623	29	**0.355**	0.456
15	**0.482**	0.606	30	**0.349**	0.449

2- By using table of critical values of the Pearson correlation coefficient r , for α 0.05, 0.01 levels

(n r α).

Where n = number of pairs for both testing methods.***

Correlation Assumption:

1- for each value of x there is a normally distributed subpopulation of y values .

2- for each value of y there is a normally distributed subpopulation of x values.

3- the joint distribution of x and y is a normal distribution called the bivariate normal distribution .

4- the sub pop of y values all have the same variance.

5- the sub pop of x values all have the same variance.

The significance of correlation also depends upon the sample size. If the sample size is large, even a lesser degree of correlation is also significant, and for a small sample size, even a higher degree of correlation may or may not be significant.

There may be nonlinear correlation where correlation coefficient is small (indicating a weak relationship) but association may be strong. They are not revealed because association is not linear.

Subject	Sleeping time (hr)	Dose (Mm/kg)
1	4	3
2	6	3
3	5	3
4	9	10
5	8	10
6	7	10
7	13	15
8	11	15
9	9	15

$\sum x\, y = 780$

$\sum y = 72$

$\sum y_2 = 642$

$\sum x = 84$

$\sum x_2 = 1002$

$S\ x\ y = 108$

$S\ x = 218$

$S\ y = 66$

$r = S\ x\ y / \sqrt{(S\ x)\ (S\ y)} = 0.9$

$$t = r \sqrt{[(n-2) / (1 - r2)]}$$

$$t = 0.9 \sqrt{9-2 / 1-0.81}$$

$$t = 5.454$$

tabulated t =1.8946

From the table= 0.6664

Infant ID #	Gestational Age (wks)	Birth Weight (gm)
1	34.7	1895
2	36.0	2030
3	29.3	1440
4	40.1	2835
5	35.7	3090
6	42.4	3827
7	40.3	3260
8	37.3	2690
9	40.9	3285
10	38.3	2920
11	38.5	3430
12	41.4	3657
13	39.7	3685
14	39.7	3345
15	41.1	3260
16	38.0	2680
17	38.7	2005

Comments

Note that a relationship can be strong and yet not significant

Conversely, a relationship can be weak but significant

The key factor is the size of the sample.

For small samples, it is easy to produce a strong correlation by chance and one must pay attention to significance to keep from jumping to conclusions: i.e.,
rejecting a true null hypothesis,
which means making a Type I error.

For large samples, it is easy to achieve significance, and one must pay attention to the strength of the correlation to determine if the relationship explains very much

The proper interpretation of r

1- Correlation does not imply causality . The existence of a statistically significant correlation between study variables does not prove that a cause- and – effect relation exists between them.

2- A statistically significant correlation between two study variables does not imply that the association is clinically important.

a- statistical significance merely indicates that the calculated value of r is unlikely to have resulted from random chance when the population coefficient p=0

b- when the sample size is large , the correlation may achieve statistical significance even though the actual deviation of P from zero is small.

c- the size of p-value indicates the likelihood that an association exists , it does not specify the magnitude of that association.

The value of r or r2 is the best indicator of the magnitude of the association in the population.

3- Failure to demonstrate the statistical significance of a given value of r may be due to the absence of a linear relationship between x and y (Ho is true and the p= zero).

4- Non sensual , or spurious correlation may be obtained when average or aggregate data for groups of subjects are substituted for pairs of measurements on individual subjects (known as ecological fallacy).

Coefficient of Determination (r^2)

r^2 measures the proportion of the variation in one variable that can be attributed to, or explained by , variation in the second variable. That is (r_2) is that proportion of the variance in one variable that can be explained by its linear relationship to the other.

r_2 is the counter part of attributable risk for interval / ratio data.

Characteristics of r^2

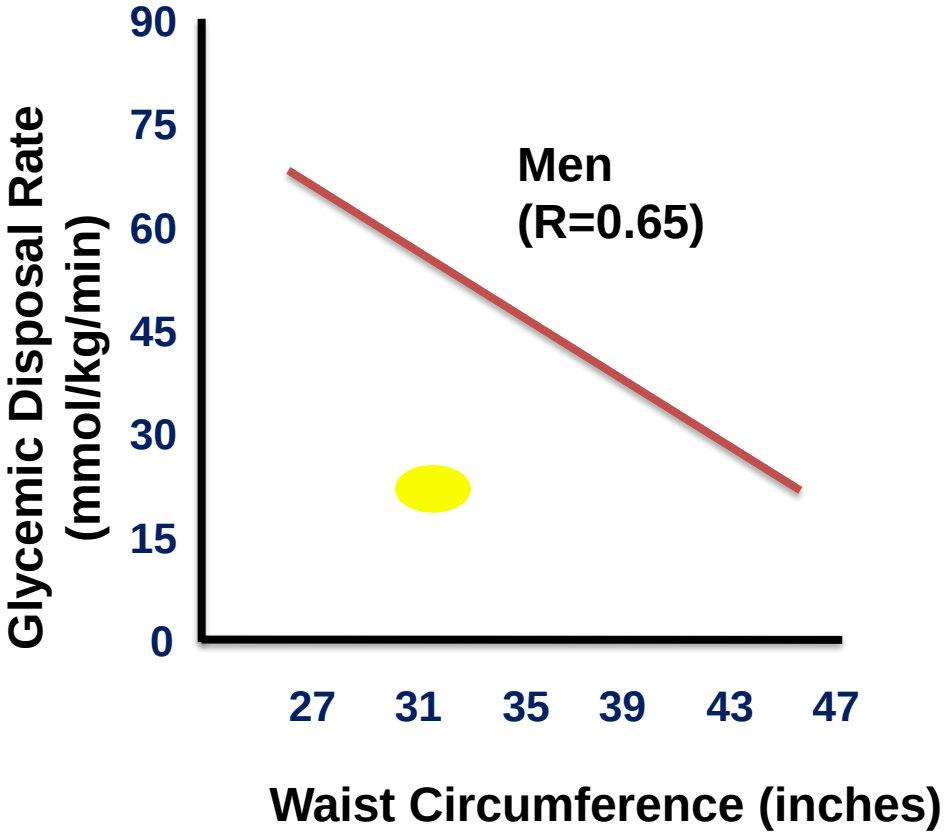
a- When r 2 = zero , (r also = zero), none of the variation in y can be attributed to changes in x.

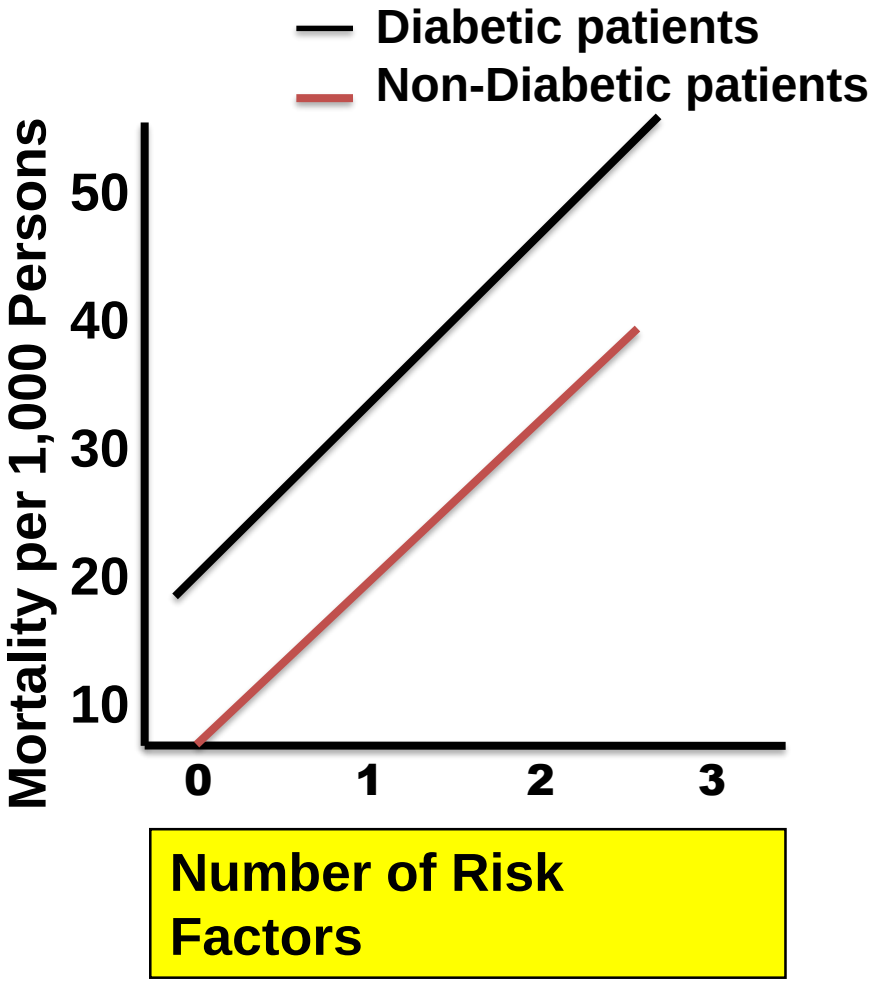
b- When r 2 =1 , all of the variation in y is attributed to its linear relationship with x.

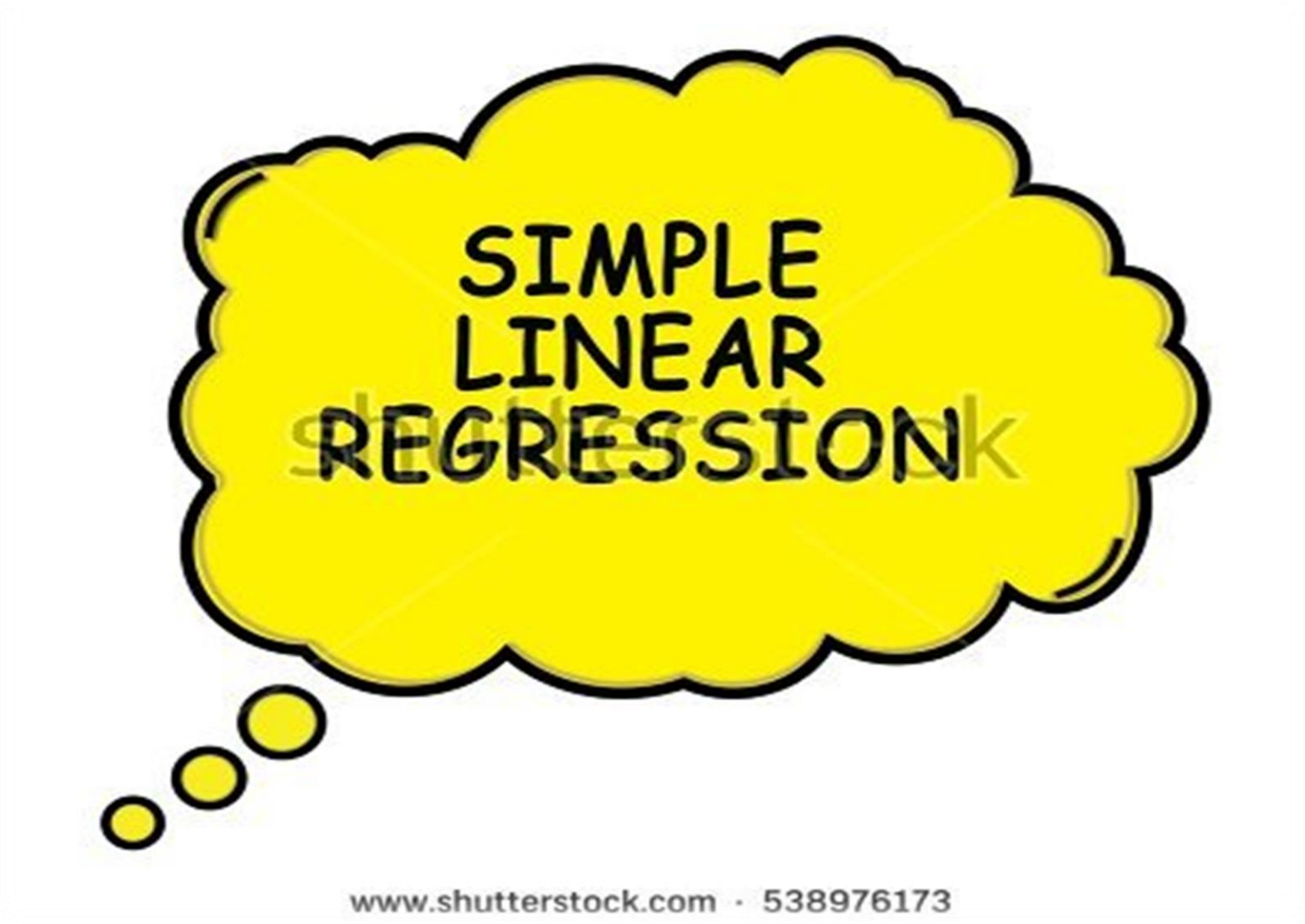
The variation in y may be attributed in part to its linear relationship with x. Other factors including systematic variation resulting from relationship between y and other unknown variables , and random subject to subject variation account for the remainder of the variation. By specifying the proportion of the total variation in y that is attributed to its linear relationship with x , r 2 provides a mathematical tool for separating these sources of variation.

LIMITATIONS OF CORRELATION

- **linearity:**
  - **can't describe non-linear relationships**
  - **e.g., relation between anxiety & performance**
- **truncation of range:**
  - **underestimate strength of relationship if you can't see full range of x value**
- **no proof of causation**
  - **third variable problem:**
    - **could be 3rd variable causing change in both variables**
    - **directionality: can't be sure which way causality "flows"**


Glycemic Disposal Rate
(mmol/kg/min)
90
75
60
45
30
15
0
Men
(R=0.65)
27
31
35
39
43
47
Waist Circumference (inches)


Diabetic patients
Non-Diabetic patients
Mortality per 1,000 Persons
50
40
30
20
10
0
1
2
3
Number of Risk Factors


SIMPLE
LINEAR
REGRESSION
www.shutterstock.com · 538976173

REGRESSION

- **Means change in measurements of a variable character on the positive or negative side beyond the mean**

- **For strongly correlated variables , value of dependent variable can be predicted from the values of the independent variable**

  - **Simple linear regression**
  - **Multiple linear regression**
  - **Logistic regression**

Regression Analysis: the goal of regression analysis is to derive a linear equation that best fit a set of data pairs (x, y) represented as points on a scatter diagram .

The equation of a straight line relationship between variable x and y is

$$Y = a + b\,x$$

a the intercept distance of regression line cuts the y axis . It is the value of y when x is zero.

$$E(y) = \beta_0 + \beta_1 x$$

b slop or gradient of the line. It is the increase in y corresponding to increase of one unit in x (regression coefficient)

$a = my - b\ m\ x$

my= mean of y

m x=mean of x

$$b = \frac{\sum x\ y) - (\sum x)\ (\sum y)\ /n}{[\sum x2 - (\sum x)2/n\ ]}$$

$$E(y) = \beta_0 + \beta_1 x$$

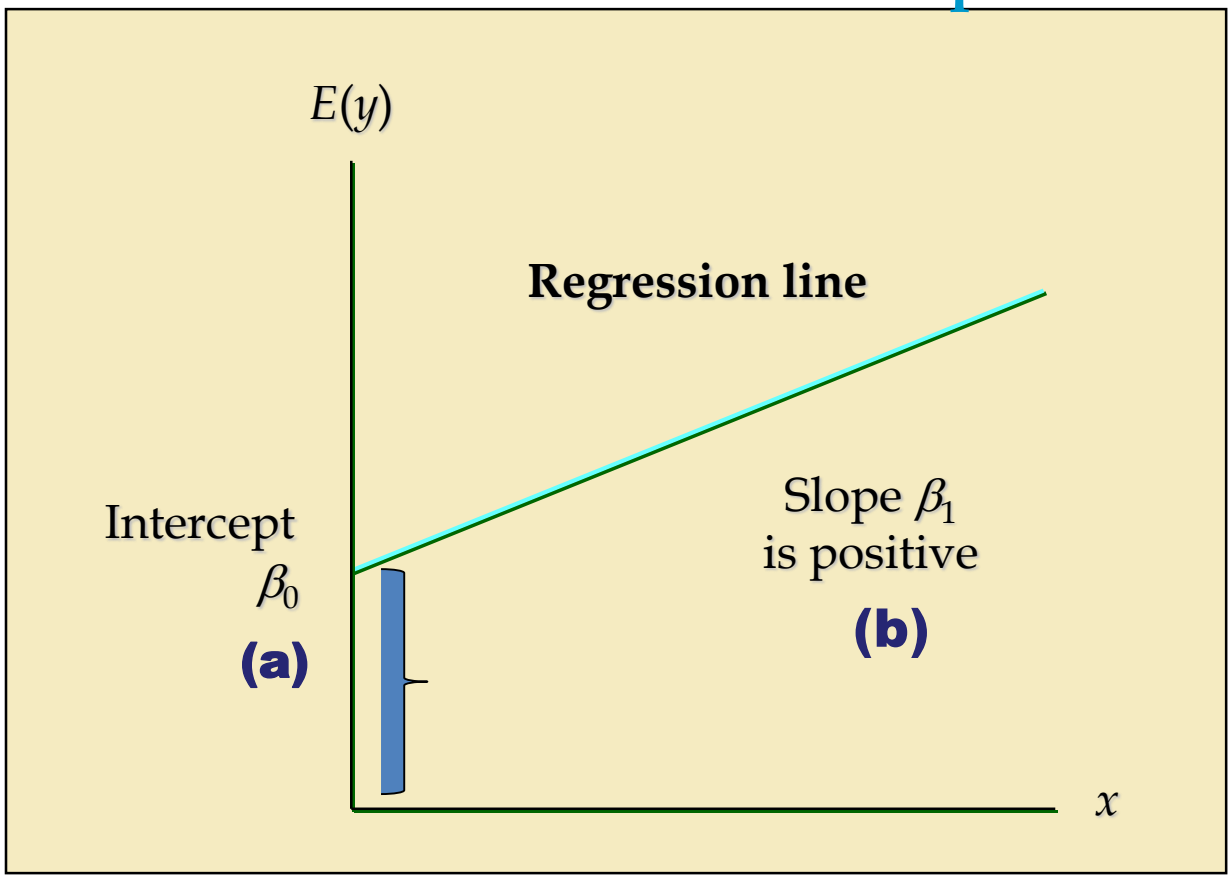
For population

- **The <u>estimated simple linear regression equation</u> is:**

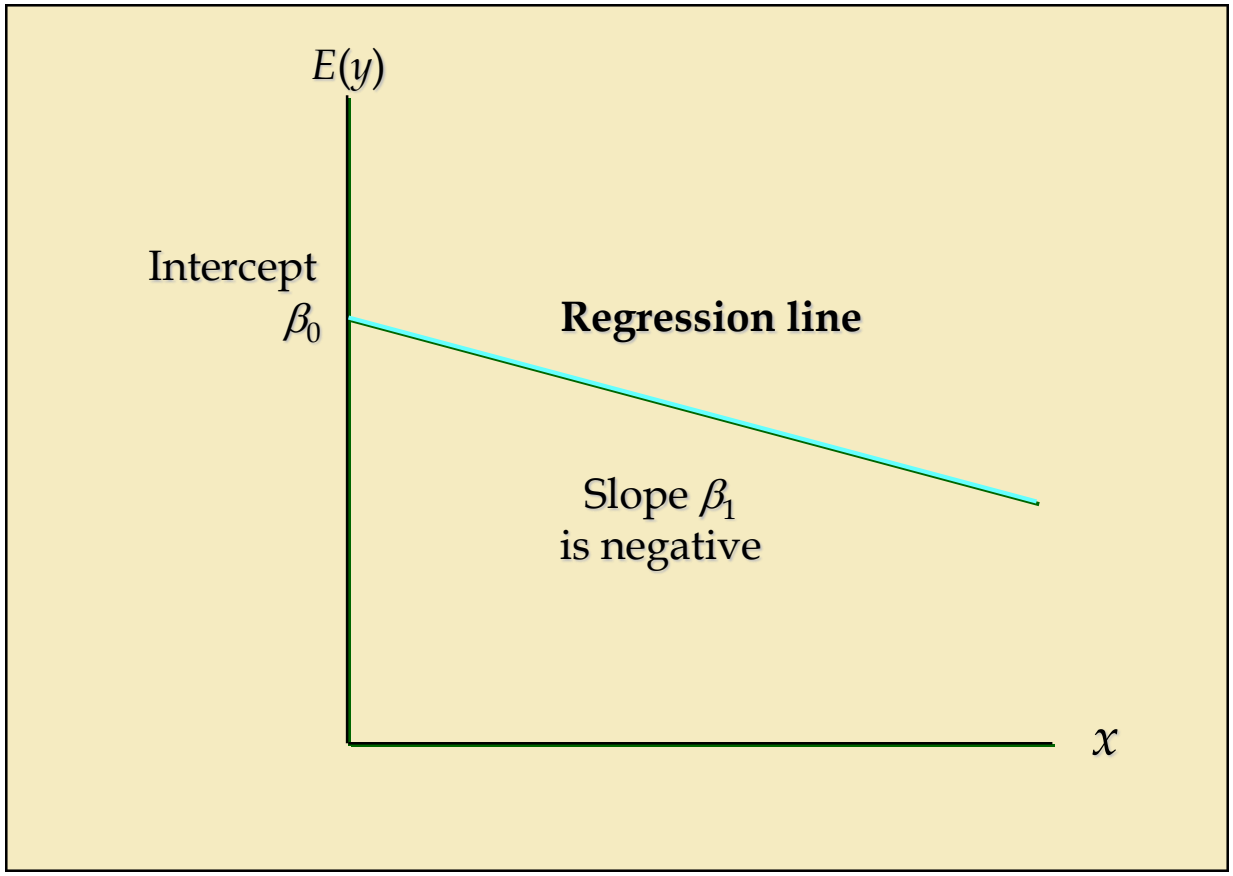
$$\hat{y} = b_0 + b_1 x$$

  - **The graph is called the estimated regression line.**
  - **b_0 is the *y* intercept of the line.**
  - **b_1 is the slope of the line.**
  - **$\hat{y}$ is the estimated value of *y* for a given *x* value.**

Positive Linear Relationship



Negative Linear Relationship



Regression Analysis

Regression analysis examines associative relationships between a dependent variable and one or more independent variables in the following ways:

- **Determine whether the independent variables explain a significant variation in the dependent variable: whether a relationship exists.**

- **Determine how much of the variation in the dependent variable can be explained by the independent variables: strength of the relationship.**

- **Determine the structure or form of the relationship: the mathematical equation relating the independent and dependent variables.**

- **Predict the values of the dependent variable.**

- **Control for other independent variables when evaluating the contributions of a specific variable or set of variables.**

- **Regression analysis is concerned with the nature and degree of association between variables and does not imply or assume any causality.**

Evaluating the adequacy of the regression line

Two criteria determine the suitability of the regression equation for predicting values of the dependent variable.

a- A statistically significant linear relationship between x& y can be demonstrated in the population from which the sample was draw.

b- A sufficiently large proportion of the variation in y can be accounted for by its linear relationship to x .

The value of r^2 , provides this information.

Using regression analysis correctly :

The sample regression equation should not be used to predict values of y out side the relationship between the two variables is unknown.

Systolic blood pressure readings in mmHg by 2 methods in 25 patients with essential hypertension as in the table below:

No.	Method I (X)	Method II (Y)	X^2	Y^2	XY
1	132	130	17424	16900	17160
2	138	134	19044	17956	18492
3	144	132	20736	17424	19008
4	146	140	21316	19600	20440
5	148	150	21904	22500	22200
6	152	144	23104	20736	21888
7	158	150	24964	22500	23700
8	130	122	16900	14881	15860
9	162	160	26244	25600	25920
10	168	150	28224	22500	25200
11	172	160	29584	25600	27520
12	174	178	30276	31684	30972
13	180	168	32400	28224	30240
14	180	174	32400	30276	34320
15	188	186	35344	34596	34968
16	194	172	37636	29584	33368
17	194	182	37636	33124	35308
18	200	178	40000	31684	35600
19	200	196	40000	38446	39200
20	204	188	41616	35344	38352
21	210	180	44100	32400	37800
22	210	196	44100	38416	41160
23	216	210	46656	44100	45360
24	220	190	48400	36100	41300
25	220	202	48400	40804	44140
totals	4440	4172	808408	710952	757276

r = 0.995 → strong direct relationship

$t = r \sqrt{[(n-2) / (1 - r2)]} = 0.955 \sqrt{[(25-2) / (1 - (0.955)2)]} = 16.17$

Calculated t (16.17) > tabulated t (n - 2 t 1-α = 23 t 0.95 = 2.0687) →

reject H0

$b = [(\sum x\ y) - (\sum x)(\sum y)/ n] / [\sum x2 - (\sum x)2/ n] = 0.822$

$a = (\sum y - b \sum x) / n = [(4172) - (0.822)(4440)] / 25 = 20.89$

y = a + b x = 20.89 + 0.822 (x)

SERUM CHOLESTEROL	Mean daily calorie intake
162.2	1990
158	1450
157	2385
155	1850
156	1750
154.1	1950
169.1	1860
181	1740
174.9	2260
180.2	2240
174	2170
182.5	2820

r = 0.5890