Kenya Medical Training College
Department of Clinical Medicine
Year Two Semester One
Scatter Diagram and Regression Analysis: Worked Examples
3rd December 2020

Willis J. Opalla

Learning Objective

- To apply the knowledge on scatter diagrams and regression analysis in calculations and make inferences on relationship between various variables.

Learning Outcomes

- By the end of this session, you should be able to
  1. Explain the concepts of scatter diagrams and regression analysis.
  2. Define the regression coefficient
  3. Construct a scatter diagram and integrate its use with other appropriate measures of relationship.
  4. Apply the regression equations in statistical analysis for relationship between independent and dependent variables.

Scatter Diagram Method

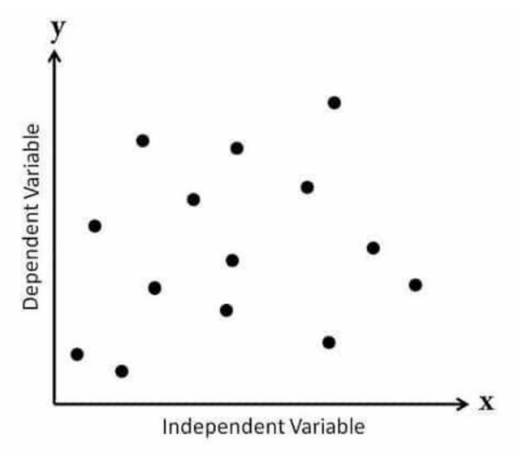
- Scatter Diagrams are convenient mathematical tools to study the correlation between two random variables.
- They are a sheet of paper upon which the data points corresponding to the variables of interest, are scattered.
- The association between the two variables is determined by the pattern that the data points form on the sheet of paper.
- This can further be coupled with a suitable correlation analysis technique.

Scatter Diagram Method

- Application
  - A quick way of confirming a hypothesis that two variables are correlated.
  - Provides a graphical representation of the strength of the relationship between two variables.
  - It also helps in understanding cause and effect relationship to evaluate whether manipulation of independent variable (cause) is producing the change in dependent variable (effect).

Construction of a Scatter Diagram

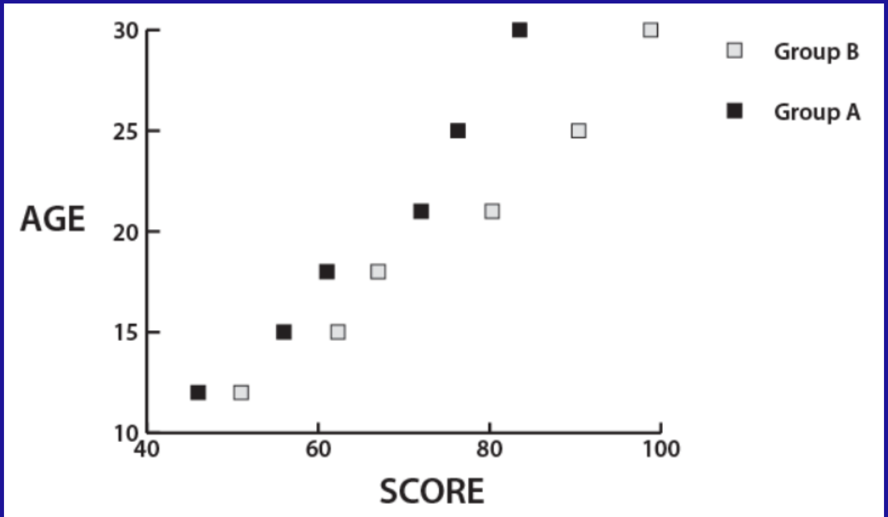
- Step 1: Draw a line "L", with the horizontal part of "L" as x axis and vertical part as y axis.



- Step 2: Make the scale units at even multiples such as 10,20,30,40 etc so as to have an even scale system.

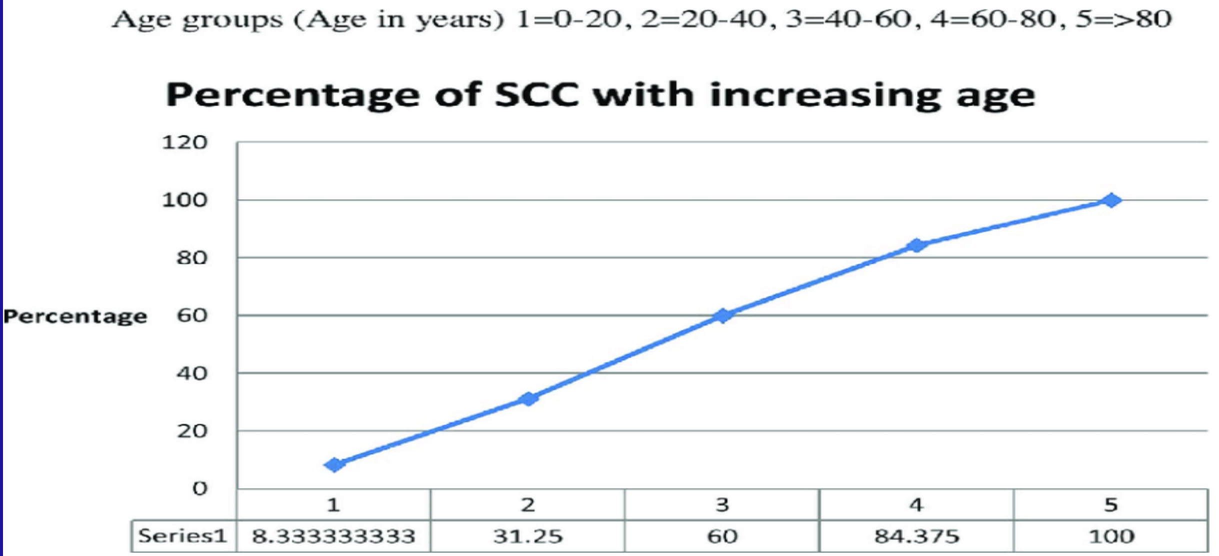
Construction of a Scatter Diagram

- Step 3: Place the independent (cause) variable on horizontal axis and dependent (effect) variable on vertical axis.
- Plot the data points at the intersection of x and y axis.



Scatter Diagram Method

- The plots on the graphs generally look scattered, hence the name scatter plot.
- Interpret the data and find the relationship.



Line graph showing increase in prevalence of SSCD (Y-axis) with increasing age (X-axis).

Interpretation of Scatter Diagram

- It suggests the degree and the direction of the correlation.
- The greater the scatter of plotted points on the chart the lesser is the relationship.
- The closer the points to the diagonal line from the left corner to the upper right corner, the perfectly positive the correlation ($r = +1$).
- If all the plots are on the diagonal line from upper left corner to the lower right corner, then the correlation is perfectly negative. ($r = -1$)

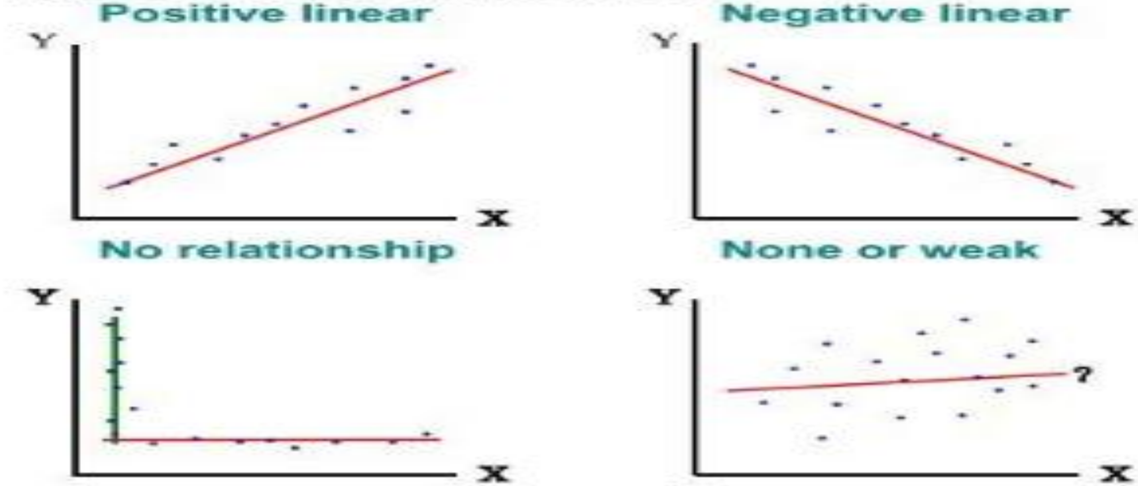
Interpretation of Scatter Diagram

Correlation and Regression

Linear correlation:

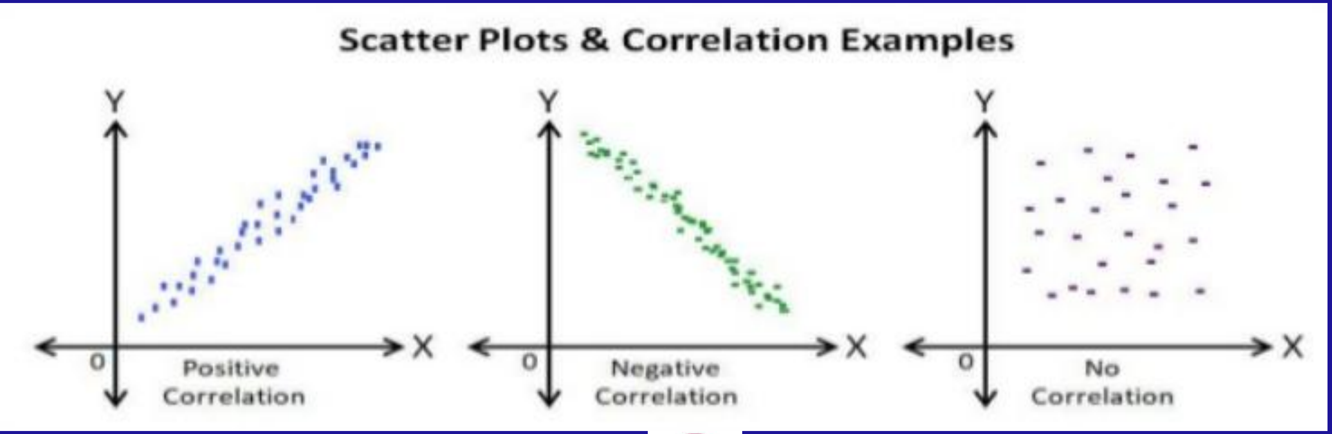
- Does one variable increase or decrease linearly with another?
- Is there a linear relationship between two or more variables?

Types of linear relationships:



Interpretation of Scatter Diagram

- If the points are widely scattered on the graph, then it indicates very little relationship, i.e. a weak positive or weak negative relationship.
- If the plotted points lie on the diagram in a disorganized manner, then it shows no correlation.



Merits and Demerits of Scatter Diagram

- Merits
  - It is a simple and non-mathematical method to study correlation.
  - Easily understood and can enable a rough idea to be formed quickly.
  - Is not influenced by the extreme values of x and y.
- Demerits
  - Cannot determine the exact degree of correlation.
  - It is not mathematical, hence less reliable.

Regression

- Regression analysis is a reliable method of identifying the variables that have an impact on a topic of interest.
- Dependent Variable:
  - This is the main factor that the study seeks to understand or predict.
- Independent Variables:
  - These are the factors that are hypothesized to have an influence on the dependent variable of the study.

Regression

- Is done by deriving a suitable equation on the basis of available bivariate data.
- The equation is called the Regression equation and its geometrical representation is called the Regression curve.
- The regression equation requires a Regression coefficient, b/b^1.

Regression Analysis

- Regression analysis seeks to determine the nature of relationship between the variables,

  i.e. to study the functional relationship between the variables and thereby provide a mechanism for prediction.

- Regression analysis describes the relationship between dependent variable (y) and independent variable (x).
- This way, unknown values of ‘y’ can be estimated for the known values of ‘x’ through the mathematical equation, $y = a+bx$.

Properties of Regression Coefficient

- The regression coefficient is denoted by b.
- Between two variables (x and y), two values of regression coefficient can be obtained:
  - One is obtained when x considered as the independent and y as dependent variable and the other when it is reversed.
  - The regression coefficient of y on x is represented as b_{yx} and that of x on y as b_{xy}.
- The *correlation coefficient* is the square root of the products of two regression coefficients ($b = b_{yx}$ and $b^1 = b_{xy}$).

Regression Equations

- Two equations:
  1. Regression Equation of y on x.
  2. Regression equation of x on y.

Regression Equation of y on x

- $y = a + bx$

  where,

  y is the dependent variable,

  x, the independent variable.

  a and b are constants.

- It is also to be noted that

  $b = b_{yx}$ (regression coefficient of y on x) •

  $$b = \frac{\Sigma xy - n\bar{x}\,\bar{y}}{\Sigma x^2 - nx^2}$$

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

Regression Equation of x on y

- $x = a^1 + b^1 x$

  where,

  x is the dependent variable,

  y, the independent variable.

  a^1 and b^1 are constants.

- It is also to be noted that

  $b^1 = b_{xy}$ (regression coefficient of y on x) •

  $$b^1 = \frac{\Sigma xy - n\bar{x}\bar{y}}{\Sigma y^2 - ny^2}$$

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

Types of Regression

- Simple linear regression:
  - It is the relationship between a scalar response or dependent variable and one or more independent variables.
- Multiple linear regression:
  - More than one explanatory variable.
- Multivariate linear regression:
  - Multiple correlated dependent variables are predicted, rather than a single scalar variable.

Types of Regression

- Positive regression:
  - A positive sign indicates that as the predictor variable increases, the response variable also increases.
- Negative regression:
  - A negative sign indicates that as the predictor variable increases, the response variable decreases.

Types of Regression

- Linear and nonlinear Regression:
  - A model is linear when each term is either a constant or the product of a parameter and a predictor variable.
  - It is non linear if the equation does not meet the linear criteria.

Regression Analysis

- Worked Example
  - Fit a regression equation of BP on age based on the following data and estimate the probable BP for a 55 years old.

Age (yrs)	30	40	50	60	70
BP (mmHg)	120	130	140	150	160

$n = 5$

$\overline{X} = \Sigma x/n$

$= 250/5$

$= 50$

$\overline{Y} = \Sigma y/n = 700/5 = 140$

- The regression equation to be fitted is $y = a+bx$ where y is BP and x is the age.

Regression Equation of y on x

- Worked Example

Find b and a using the given formula.

$$b = \frac{\Sigma xy - n\bar{x}\,\bar{y}}{\Sigma x^2 - nx^2} \quad \text{and}$$

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

x	y	xy	x^2
30	120	3600	900
40	130	5200	1600
50	140	7000	2500
60	150	9000	3600
70	160	11200	4900
Σx=250	Σy=700	Σxy=36000	Σx^2=13500

Regression Equation of y on x

- Substituting,

$$b = \frac{36000 - 5 \times 50 \times 140}{13500 - 5 \times (50)^2}$$

$= (36000 - 35000)/(13500 - 12500)$

$= 1000/1000 = 1$

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

$= 140 - 1 \times 50$

$= 90$

So the fitted regression equation is $y = a+bx$.

$BP = 90 + 1 \times 55$

$= 90 +55 = 145$mmHg

Regression Analysis

- Example 2
- Fit the two line of regression equation for the following data.

X	10	20	30	40	50
Y	30	50	70	90	110

$n = 5$

$\overline{X} = \Sigma x/n = 150/5 = 30$

$\overline{Y} = \Sigma y/n = 350/5 = 70$

- The regression equation to be fitted is $y = a+bx$ and $x = a^1 + b^1 y$.

Regression Equation of y on x

- Find b^1 and a^1 using the given formula.

$$b^1 = \frac{\Sigma xy - n\bar{x}\,\bar{y}}{\Sigma x^2 - n\bar{x}^2}$$

$$a^1 = \bar{y} - b\bar{x}$$

x	y	xy	x^2	y^2
10	30	300	100	900
20	50	1000	400	2500
30	70	2100	900	4900
40	90	3600	1600	8100
50	110	5500	2500	12100
$\sum x = 150$	$\sum y = 350$	$\sum xy = 12500$	$\sum x^2 = 5500$	$\sum x^2 = 28500$

Regression Equation of y on x

- Substituting,

$$b = \frac{12500 - 5x30x70}{5500 - 5x(30)^2}$$

$$= (12500 - 10500)/(5500 - 4500)$$

$$= 2000/1000 = 2$$

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

$$= 70 - 2 \times 30$$

$$= 70 - 60 = 10$$

So the fitted regression equation is $y = 10 + 2x$.

Regression Equation of x on y

- Worked Example
  - Find b^1 and a^1 and a using the formula.
  - $$b^1 = \frac{\Sigma xy - n\bar{x}\,\bar{y}}{\Sigma y^2 - ny^2}$$

    $$a^1 = \bar{x} - b\bar{y}$$

Regression Equation of x on y

- Substituting,

$$b^1 = \frac{12500 - 5x30x70}{28500 - 5x(70)^2}$$

$b^1 = (2500 - 10500)/(28500 - 24500)$

$b^1 = 2000/4000 = 0.5$

$a^1 = \bar{x} - b1\bar{y}$

$= 30 - 0.5 \text{ x } 70$

$= 30 - 35 = -5$

- So the fitted regression equation is $x = -5 + 0.5y$.

Properties

- The square root of the products of two regression coefficients is correlation coefficient.
- In the given examples,

$b = b_{yx}$ $\quad\quad$ $b^1 = b^1_{xy}$

$= 2$ $\quad\quad$ $= 0.5$

$r = \sqrt{2 \times 0.5}$

$= \sqrt{1}$

$= 1$

Summary

- The scatter diagram informs on the degree and direction of the correlation:
- The greater the scatter of plotted points on the chart the lesser the relationship.
- The closer the points to the diagonal line from the left corner to the upper right corner, the perfectly positive the correlation.

References

- Joseph, J. K. (n.d) *Measures of Relationship*, [Online] Available: https://www.slideshare.net/JohnykuttyJoseph/measures-of-relationship, (Retrieved 26.11.2020)
- Kothari, C. R., (2004) *Research Methodology, Methods and Techniques*, 2nd ed., New Age International Publishers, New Delhi.