Chapter 7

Regression Analysis

Prepared by, ***Jannatul Ferdous***, Assistant Professor, Dept. of CSE, Metropolitan University,Sylhet

Regression Analysis

Linear regression is a basic and commonly used type of predictive analysis or forecasting method. The regression estimates are used to explain the relationship between one dependent variable and one or more independent variables. The simplest form of the regression equation with one dependent and one independent variable is defined by the formula,

$$y = a + bx$$

where y = estimated dependent variable score

a = constant

b = regression coefficient

x = score on the independent variable.

There are many names for a regression's dependent variable. It may be called an outcome variable, criterion variable, endogenous variable, or regressand. The independent variables can be called exogenous variables, predictor variables, or regressors.

Three major uses for regression analysis are (1) determining the strength of predictors, (2) forecasting an effect, and (3) trend forecasting.

First, the regression might be used to identify the strength of the effect that the independent variable(s) have on a dependent variable. That is, the regression analysis helps us to understand how much the dependent variable changes with a change in one or more independent variables.

In a word the definition is, "The statistical tool with the help of which we are in a position to estimate the unknown value of one variable from the known value of another variable is known as regression analysis."

From the example 1 in Lecture 6 we have

Correlation coefficient, $r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$ = 0.529809, which means the variables have a moderate positive correlation.

*** If there is correlation among the variables, then we can estimate the unknown value of variable with the value of known variable (i.e fit regression model or regression line).

From the previous math we found that there exists a moderate relationship between age & glucose level. If we assume that the relationship is linear than we express the relationship by the equation,

$$y = a + bx \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

(sign of b will depend on the sign of the r, i.e b will be negative if there exist negative correlation or b will be positive if there exist positive correlation)*

Regression co-efficient, b can be defined in two ways,

b_{yx} = Regression co-efficient when Y depends on $X = \frac{n\sum xy - \sum x \sum y}{n\sum(x^2) - (\sum x)^2}$

or, b_{xy} = Regression co-efficient when X depends on $Y = \frac{n\sum xy - \sum x \sum y}{n\sum(y^2) - (\sum y)^2}$

Equation (1) can be determined by the following equation
$y - \overline{y} = b_{yx}\,(x - \overline{x})$ ………………..(2)
or, $y - \overline{y} = b_{xy}\,(x - \overline{x})$ ………………..(3)

Where,
$\overline{Y}$ = Mean of Y series
$\overline{X}$ = Mean of X series

Example 1: Estimate the glucose level when age 30.

Subject	1	2	3	4	5	6
Age (x)	43	21	25	42	57	59
Glucose Level (y)	99	65	79	75	87	81

Solution 1:

Subject	**Age x**	**Glucose Level y**	**xy**	**x^2**	**y^2**
1	43	99	4257	1849	9801
2	21	65	1365	441	4225
3	25	79	1975	625	6241
4	42	75	3150	1764	5625
5	57	87	4959	3249	7569
6	59	81	4779	3481	6561
n = 6	Σ x = 247	Σy = 486	Σxy = 20485	Σ$\mathbf{x}^2$ = 11409	Σ$\mathbf{y}^2$ = 40022

We have,

$\Sigma x = 247$, $\Sigma y = 486$, $\Sigma xy = 20{,}485$, $\Sigma x^2 = 11{,}409$, $\Sigma y^2 = 40{,}022$, $n = 6$

Using these values we can find, $\overline{Y} = \frac{\sum Y}{n} = \frac{486}{6} = 81$

$\overline{X} = 41.167$

$$b_{yx} = \frac{\mathbf{n}\sum xy - \sum \mathbf{x} \sum \mathbf{y}}{\mathbf{n}\sum(\mathbf{x}^2) - (\mathbf{x})^2} = \frac{6\times 20485 - 247\times 486}{6\times 11409 - (247)^2} = \frac{2868}{7445} = 0.385$$

Now putting these values in equation no (2)

$$y - \overline{y} = b_{yx}\,(x - \overline{x})$$

$$y - 81 = 0.385(x - 41.167)$$

$$y = 81 + 0.385x - 15.849$$

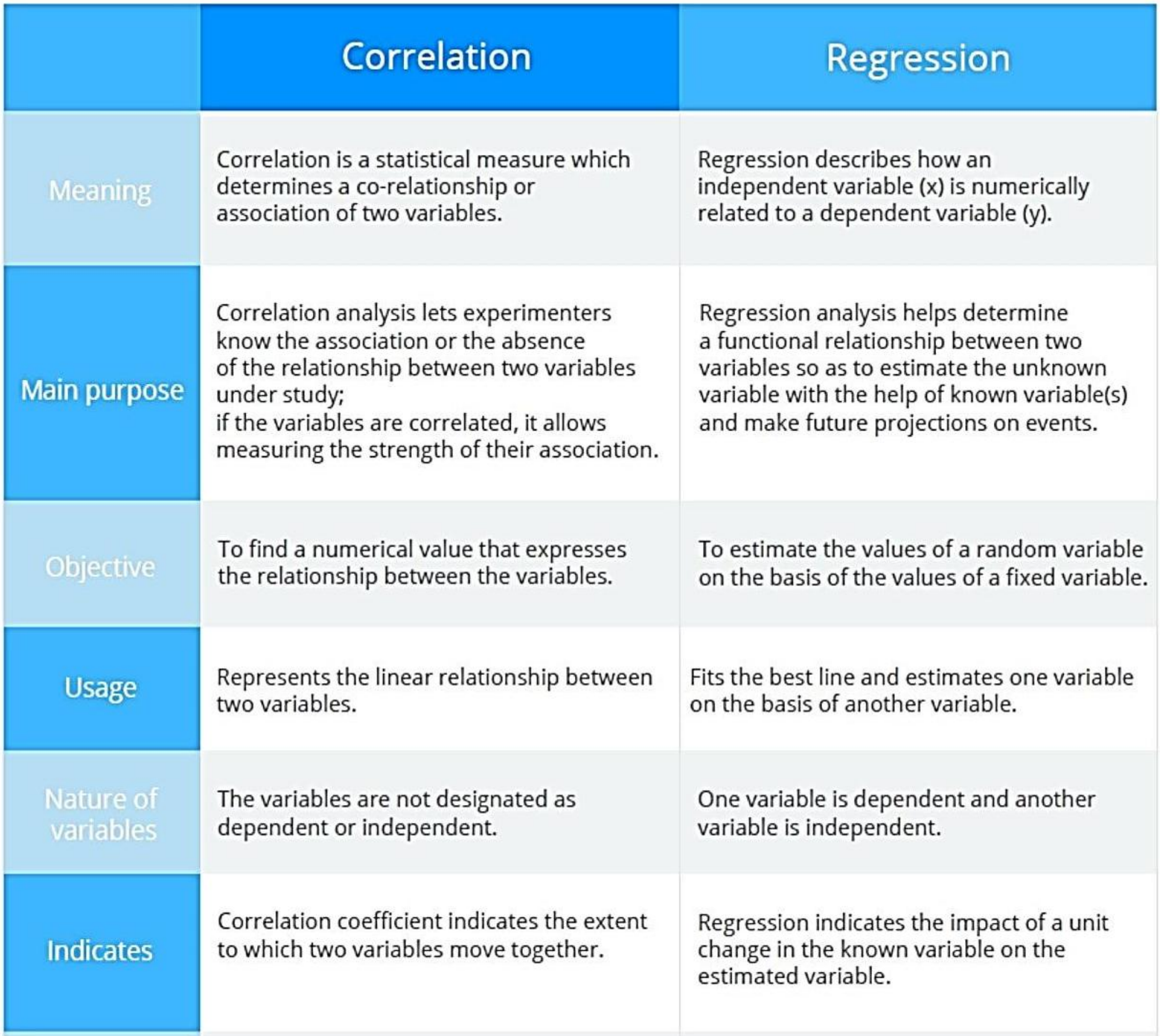
$$y = 65.151 + 0.385x$$

When, X = age 30, then Y= 65.151+ 0.385× 30 = 76.701

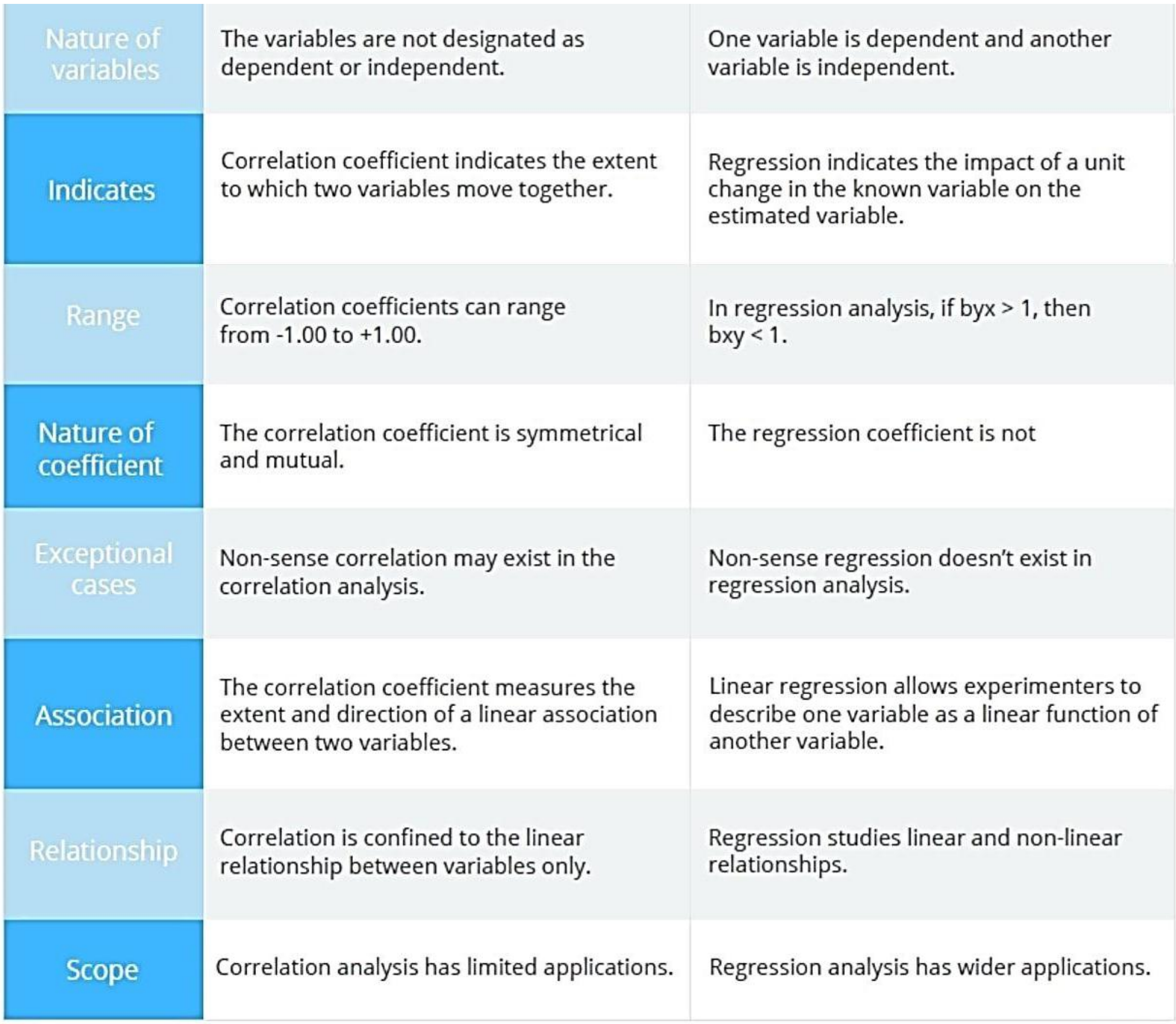
The estimated value of glucose level is 76.701 when age is 30 year.

Reversely, you may be asked to estimate the age when the glucose level is 70. In this case the dependent will be independent variable and the independent will be dependent variable. That means, here age will be dependent and glucose level will be independent variable and the whole procedure will be same.

Difference between correlation and regression analysis



	Correlation	Regression
Meaning	Correlation is a statistical measure which determines a co-relationship or association of two variables.	Regression describes how an independent variable (x) is numerically related to a dependent variable (y).
Main purpose	Correlation analysis lets experimenters know the association or the absence of the relationship between two variables under study; if the variables are correlated, it allows measuring the strength of their association.	Regression analysis helps determine a functional relationship between two variables so as to estimate the unknown variable with the help of known variable(s) and make future projections on events.
Objective	To find a numerical value that expresses the relationship between the variables.	To estimate the values of a random variable on the basis of the values of a fixed variable.
Usage	Represents the linear relationship between two variables.	Fits the best line and estimates one variable on the basis of another variable.
Nature of variables	The variables are not designated as dependent or independent.	One variable is dependent and another variable is independent.
Indicates	Correlation coefficient indicates the extent to which two variables move together.	Regression indicates the impact of a unit change in the known variable on the estimated variable.



Nature of variables	The variables are not designated as dependent or independent.	One variable is dependent and another variable is independent.
Indicates	Correlation coefficient indicates the extent to which two variables move together.	Regression indicates the impact of a unit change in the known variable on the estimated variable.
Range	Correlation coefficients can range from -1.00 to +1.00.	In regression analysis, if byx > 1, then bxy < 1.
Nature of coefficient	The correlation coefficient is symmetrical and mutual.	The regression coefficient is not
Exceptional cases	Non-sense correlation may exist in the correlation analysis.	Non-sense regression doesn't exist in regression analysis.
Association	The correlation coefficient measures the extent and direction of a linear association between two variables.	Linear regression allows experimenters to describe one variable as a linear function of another variable.
Relationship	Correlation is confined to the linear relationship between variables only.	Regression studies linear and non-linear relationships.
Scope	Correlation analysis has limited applications.	Regression analysis has wider applications.