Regression

Regression

Regression is a mathematical measure of expressing the average of relationship between two or more variables in terms of the original units of the data. In a regression analysis there are two types of variables. The variable whose is influenced or is to be predicted is called dependent variable, regressed predicted or explained variable and the variable which influences the values or is used for prediction is called independent variable or regressor or predictor or explanator. These relationships between two variables can be considered between say rainfall and agricultural production, price of an output and the overall cost of product, consumer expenditure and disposable income.

Regression lines or Regression Equations

If the variables in a bivariate distribution are related we will find that points in the scatter diagram will cluster around some curve called the "Curve of regression". If the curve is straight line of, it is called the line of regression and there is said to be linear regression between the variables, otherwise regression is said to be curvilinear. The line of regression is the line which gives the best estimate to the value of one variable for any specific value of the other variable.

In the ***regression model*** $y = A + Bx + \epsilon$, A is called the y-intercept or constant term, B is the slope, and ϵ is the random error term. The dependent and independent variables are y and x, respectively.

In the model $\hat{y} = a + bx$, a and b, which are calculated using sample data, are called the ***estimates of A and B***.

Dr. Mohd. MuziburRahman (Mzu)
Professor, Department of Statistics, JU

SIMPLE LINEAR REGRESSION ANALYSIS cont.

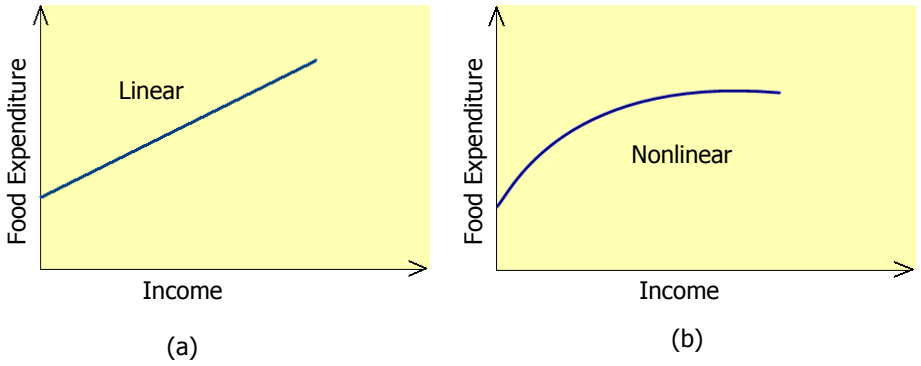
Constant term or y-intercept — Slope

$$y = A + Bx$$

Dependent variable — Independent variable

9

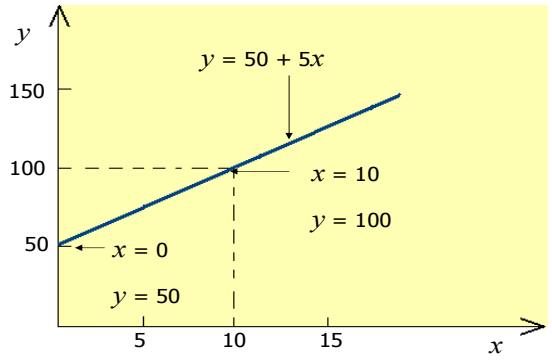
Figure_R1 Relationship between food expenditure and income. (a) Linear relationship. (b) Nonlinear relationship.



5

Dr. Mohd. MuziburRahman (Mzu)
Professor, Department of Statistics, JU

Figure_R2 Plotting a linear equation.



6

Q. What are the regression coefficients?

In the line of regression of Y on X

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

The coefficient 'b' which is the slope of the line of regression of Y on X is called the coefficient of regression of Y on X. It represents the increment in the value of the dependent variable Y for a unit change in the value of the independent variable X. For notational convenience, coefficient of regression of Y on X is denoted by b_{yx}.

Regression coefficient of Y on X is

$$b_{yx} = \frac{\sum XY - \frac{\sum X \sum Y}{n}}{\sum X^2 - \frac{(\sum X)^2}{n}}$$

and the intercept

$$a = \bar{Y} - b\bar{X} = \frac{\sum Y}{n} - b\frac{\sum X}{n}$$

Similarly in the regression equation of X on Y

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

Dr. Mohd. MuziburRahman (Mzu)
Professor, Department of Statistics, JU

Regression coefficient of Y on X is

$$b_{xy} = \frac{\sum XY - \frac{\sum X \sum Y}{n}}{\sum Y^2 - \frac{(\sum Y)^2}{n}}$$

and the intercept

$$a = \bar{X} - b\bar{Y} = \frac{\sum X}{n} - b\frac{\sum Y}{n}$$

Interpretation: The regression coefficient or slope b represents the estimated average change in Y when X increases by one unit that is when the independent variable X changes one unit then the dependent variable Y changes b unit.

Intercept: The intercept a represents the estimated average value of Y when X equals zero.

Important Properties of Regression Coefficient: The regression coefficient has the following properties:

i) Regression coefficient measures the average change in dependent variable for a unit change in independent variable.

ii) Regression coefficients are not symmetrical.

iii) Both the regression coefficient has the same sign.

iv) The correlation coefficient is the geometric mean of the two regression coefficients, that is $r = \sqrt{b_{xy} \times byx}$.

v) If one of the regression coefficients is greater than 1, then the other regression coefficient must be less than one.

vi) The sign of correlation coefficient and regression coefficients are same, since all the measures depend on the sign of the covariance appearing in the numerator.

Example 1: From the following data obtain the regression equations of Y on X :

Sales (X)	91	97	108	121	67	124	51	73	111	57
Purchase (Y)	71	75	69	97	70	91	39	61	80	47

Dr. Mohd. MuziburRahman (Mzu)
Professor, Department of Statistics, JU

Solution:

We know that,

The regression equation of Y on X is expressed as follows

$$Y = a + bX$$

Again,

Regression coefficient of Y on X is

$$b_{yx} = \frac{\sum XY - \frac{\sum X \sum Y}{n}}{\sum X^2 - \frac{(\sum X)^2}{n}}$$

Sales (X)	**Purchase** (Y)	X^2	Y^2	XY
91	71	8281	5041	6461
97	75	9409	5625	7275
108	69	11664	4761	7452
121	97	14641	9409	11737
67	70	4489	4900	4690
124	91	15376	8281	11284
51	39	2601	1521	1989
73	61	5329	3721	4453
111	80	12321	6400	8880
57	47	3249	2209	2679
$\sum X = 900$	$\sum Y = 700$	$\sum X^2 = 87360$	$\sum Y^2 = 51868$	$\sum XY = 66900$

$$\therefore b_{yx} = \frac{\sum XY - \frac{\sum X \sum Y}{n}}{\sum X^2 - \frac{(\sum X)^2}{n}} = \frac{66900 - \frac{900 \times 700}{10}}{87360 - \frac{(900)^2}{10}} = 0.613207547 = 0.613$$

$$a = \bar{Y} - b\bar{X} = \frac{\sum Y}{n} - b\frac{\sum X}{n} = \frac{700}{10} - 0.613 \times \frac{900}{10} = 14.81$$

Regression equation of Y on X is

$$Y = 14.81 + 0.613\,X$$

Dr. Mohd. MuziburRahman (Mzu)
Professor, Department of Statistics, JU

Example 2: The following data give the ages and blood pressure of 10 women

Age (X)	56	42	36	47	49	42	60	72	63	55
Blood pressure (Y)	147	125	118	128	145	140	155	160	149	150

a) Find the correlation coefficient between X and Y

b) Determine the least squares regression equation of Y on X.

c) Estimate the blood pressure of a women whose age is 45 years

Solution:

a) We know that, Correlation coefficient between X and Y is given by

Age (X)	**Blood Pressure** (Y)	X^2	Y^2	XY
56	147	3136	21609	8232
42	125	1764	15625	5250
36	118	1296	13924	4248
47	128	2209	16384	6016
49	145	2401	21025	7105
42	140	1764	19600	5880
60	155	3600	24025	9300
72	160	5184	25600	11520
63	149	3969	22201	9387
55	150	3025	22500	8250
$\sum X = 522$	$\sum Y = 1417$	$\sum X^2 = 28348$	$\sum Y^2 = 202493$	$\sum XY = 75188$

$$\therefore r = \frac{75188 - \frac{522 \times 1417}{10}}{\sqrt{\left[\left\{28348 - \frac{(522)^2}{10}\right\}\left\{202493 - \frac{(1417)^2}{10}\right\}\right]}} = 0.891678842$$

b) We know that, the regression equation of Y on X is expressed as follows

$$Y = a + bX$$

Dr. Mohd. MuziburRahman (Mzu)
Professor, Department of Statistics, JU

Again, Regression coefficient of Y on X is

$$b_{yx} = \frac{\sum XY - \frac{\sum X \sum Y}{n}}{\sum X^2 - \frac{(\sum X)^2}{n}} = \frac{75188 - \frac{522 \times 1417}{10}}{28348 - \frac{(522)^2}{10}} = 1.110040015 = 1.11$$

$$a = \bar{Y} - b\bar{X} = \frac{\sum Y}{n} - b\frac{\sum X}{n} = \frac{1417}{10} - 1.11 \times \frac{522}{10} = 83.75591124$$

Regression equation of Y on X is

$$Y = 83.756 + 1.11\,X$$

c) When $X = 45$ then $Y = 83.756 + 1.11\ \ 45\ = 133.706$

Hence the most likely blood pressure of women of 45 years is 134.

Compare the correlation analysis with regression analysis.

Correlation	Regression
1. Simple correlation measures the direction and strength of linear relationship between two variables	1. Regression measures the effect of independent variable on dependent variable
2. Correlation does not measure the cause and effect relationship between the variables under study	2. Regression analysis measures the cause and effects relationship between the variable.
3. Correlation analysis has limited applications as it is confined only to the study of linear relationship between the variables	3. Regression analysis studies linear as well as non-linear relationship between the variables and therefore has much wider applications
4. Question of dependent and independent variables do not arise in case of correlation analysis.	4. Dependent variable is regressed on the independent variable in regression analysis.
5. Correlation coefficient is symmetric i.e. $r_{xy} = r_{yx}$	5. Regression co-efficients are not symmetric in X and Y i.e. $b_{xy} \neq b_{yx}$
6. The value of the correlation coefficient lies between -1 to +1.	6. Regression coefficient can take any real value between $-\alpha$ to $+\alpha$.
7. Correlation coefficient is a pure number. It is a relative measurement.	7. Regression coefficient is an absolute measurement. It depends on the units of measurement of the variable
8. Correlation coefficient is independent of shift of origin and scale of measurement.	8. Regression coefficient is independent of shift of origin but depends on scale.

Dr. Mohd. MuziburRahman (Mzu)
Professor, Department of Statistics, JU

Uses of regression analysis**:**

- The relation can be used for predictive purpose.
- Regression analysis is widely used in statistical estimation of demand curves, supply curves, production functions; cost functions, consumption function etc.

Example: A study was made by a retail merchant to determine the relation between weekly advertizing expenditure and sales. The following data were recorded:

Expenditure (in $)	40	20	25	20	30	50	40	20	50
Sales (in $)	385	400	395	365	475	440	490	420	560

(i) Plot a scatter diagram.
(ii) Find the equation of regression line to predict weekly sales from advertizing expenditures.
(iii) Estimate the weekly sales when advertizing costs are 570.

Dr. Mohd. MuziburRahman (Mzu)
Professor, Department of Statistics, JU