MAT2001-Module-3: Correlation and Regression

Dr. Nalliah M

Assistant Professor

Department of Mathematics

School of Advanced Sciences

Vellore Institute of Technology

Vellore,Tamil Nadu,India.

nalliah.moviri@vit.ac.in

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

September 14, 2020

Introduction

A distribution two variables is known as bivariate distribution. If these two variables vary such that change in one variable affect the change in other variable, the variables are said to be correlated.

For example, when parents advice their children to work hard so that they may get good marks, they are correlating good marks with hard work.

Correlation refers to the relationship of two variables or more. (e-g) relation between height of father and son, yield and rainfall, wage and price index, share and debentures etc.

Intoduction

Correlation is statistical Analysis which measures and analyses the degree or extent to which the two variables fluctuate with reference to each other. The word relationship is important. It indicates that there is some connection between the variables. It measures the closeness of the relationship. Correlation does not indicate cause and effect relationship. Price and supply, income and expenditure are correlated.

Definitions

- Correlation Analysis attempts to determine the degree of relationship between variables-Ya-Kun-Chou.
- Correlation is an analysis of the covariation between two or more variables.- A.M.Tuttle.

Correlation expresses the inter-dependence of two sets of variables upon each other. One variable may be called as (subject) independent and the other relative variable (dependent). Relative variable is measured in terms of subject.

Types of Correlation

Correlation is classified into various types. The most important ones are

- Positive and negative.
- Linear and non-linear.
- Partial and total.
- Simple and Multiple.

Positive Correlation

It depends upon the direction of change of the variables. If the two variables tend to move together in the same direction (ie) an increase in the value of one variable is accompanied by an increase in the value of the other, (or) a decrease in the value of one variable is accompanied by a decrease in the value of other, then the correlation is called positive or direct correlation.

Examples: Price and supply, height and weight, yield and rainfall.

Negative Correlation

If the two variables tend to move together in opposite directions so that increase (or) decrease in the value of one variable is accompanied by a decrease or increase in the value of the other variable, then the correlation is called negative (or) inverse correlation.

Examples: Price and demand, yield of crop and price.

Linear and Non-linear correlation

If the ratio of change between the two variables is a constant then there will be linear correlation between them. Consider the following.

X	2	4	6	8	10	12
Y	3	6	9	12	15	18

Here the ratio of change between the two variables is the same. If we plot these points on a graph we get a straight line.

If the amount of change in one variable does not bear a constant ratio of the amount of change in the other. Then the relation is called Curvi-linear (or) non-linear correlation. The graph will be a curve.

Simple and Multiple correlation

When we study only two variables, the relationship is simple correlation.

For example, quantity of money and price level, demand and price.

But in a multiple correlation we study more than two variables simultaneously.

The relationship of price, demand and supply of a commodity are an example for multiple correlation.

Partial and total correlation

The study of two variables excluding some other variable is called Partial correlation.

For example, we study price and demand eliminating supply side.

In total correlation all facts are taken into account.

Methods of Studying Correlation

- Scatter Diagram
- Graphical method
- Karl Pearson's Method
- Spearman's Rank Correlation Method
- Concurrent deviation Method
- Method of Least Squares

Computation of Correlation

When there exists some relationship between two variables, we have to measure the degree of relationship. This measure is called the measure of correlation (or) correlation coefficient and it is denoted by ρ.

Co-variation

The covariation between the variables X and Y is defined as

Cov (X,Y) $= \dfrac{\sum(x-\bar{x})(y-\bar{y})}{n}$, where $\bar{x}, \bar{y}$ are respectively means of X and Y and n is the number of pairs of observations.

Karl pearsons coefficient of correlation

Karl pearson, a great biometrician and statistician, suggested a mathematical method for measuring the magnitude of linear relationship between the two variables. It is most widely used method in practice and it is known as pearson coefficient of correlation. It is denoted by ρ. The formula for calculating ρ is $\rho = \frac{\text{Cov (X,Y)}}{\sigma_X \bullet \sigma_Y}$, where

$\text{Cov (X,Y)} = \frac{\sum(x-\bar{x})(y-\bar{y})}{n}$, $\sigma_X = \sqrt{\frac{\sum(x-\bar{x})^2}{n}}$ and $\sigma_y = \sqrt{\frac{\sum(y-\bar{y})^2}{n}}$

OR

$\rho = \dfrac{\text{Cov (X,Y)}}{\sigma_X \bullet \sigma_Y}$, where Cov (X,Y) $= \dfrac{\sum xy}{n} - \bar{x}\bar{y}$, $\sigma_X = \sqrt{\dfrac{\sum x^2}{n} - \bar{x}^2}$ and $\sigma_y = \sqrt{\dfrac{\sum y^2}{n} - \bar{y}^2}$.

OR

$$\rho = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{\left[n\sum x^2 - (\sum x)^2\right] \bullet \left[n\sum y^2 - (\sum y)^2\right]}}.$$

Properties of Correlation

- Correlation coefficient lies between -1 and +1 (i.e) $-1 \leq \rho \leq +1$.

  Note:

  $\rho = +1$ perfect positive correlation.

  $\rho = -1$ perfect negative correlation between the variables.

- ρ is independent of change of origin and scale.
- It is a pure number independent of units of measurement.
- Independent variables are uncorrelated but the converse is not true.
- Correlation coefficient is the geometric mean of two regression coefficients.
- The correlation coefficient of X and Y is symmetric $\rho_{XY} = \rho_{YX}$.

Interpretation

The following rules helps in interpreting the value of ρ.

- When $\rho = 1$, there is perfect $+$ ve relationship between the variables.
- When $\rho = -1$, there is perfect - ve relationship between the variables.
- When $\rho = 0$, there is no relationship between the variables.
- If the correlation is $+1$ or -1, it signifies that there is a high degree of correlation ($+$ve or -ve) between the two variables.

  If ρ is near to zero (ie) 0.1, - 0.1, (or) 0.2 there is less correlation.

Problem

Calculate the correlation co-efficient for the following heights (in inches) of fathers X their sons Y.

X	65	66	67	67	68	69	70	72
Y	67	68	65	68	72	72	69	71

Solution:

The correlation co-efficient between X and Y is $\rho = \dfrac{\text{Cov (X,Y)}}{\sigma_X \bullet \sigma_Y}$,

where Cov (X,Y) $= \dfrac{\sum xy}{n} - \bar{x}\bar{y}$, $\sigma_X = \sqrt{\dfrac{\sum x^2}{n} - \bar{x}^2}$ and $\sigma_y = \sqrt{\dfrac{\sum y^2}{n} - \bar{y}^2}$.

Solution Cont...

X	Y	XY	X^2	Y^2
65	67			
66	68			
67	65			
67	68			
68	72			
69	72			
70	69			
72	71			
544	552	37560	37028	38132

Solution Cont...

Here $n = 8$. Then $\bar{x} = \frac{\sum x}{n} = \frac{544}{8} = 68$ and $\bar{y} = \frac{\sum y}{n} = \frac{552}{8} = 69$.

$$\text{Cov (X,Y)} = \frac{\sum xy}{n} - \bar{x}\bar{y} = \frac{37560}{8} - (68)(69) = 4695 - 4892 = 3.$$

$$\sigma_X = \sqrt{\frac{\sum x^2}{n} - \bar{x}^2} = \sqrt{\frac{37028}{8} - (68)^2} = 2.121$$

$$\sigma_y = \sqrt{\frac{\sum y^2}{n} - \bar{y}^2} = \sqrt{\frac{38132}{8} - (69)^2} = 2.345.$$

$$\rho = \frac{\text{Cov (X,Y)}}{\sigma_X \bullet \sigma_Y} = \frac{3}{(2.121)(2.345)} = 0.6032$$

Spearman's Rank Correlation

It is studied when no assumption about the parameters of the population is made. This method is based on ranks. It is useful to study the qualitative measure of attributes like honesty, colour, beauty, intelligence, character, morality etc.The individuals in the group can be arranged in order and there on, obtaining for each individual a number showing his/her rank in the group. This method was developed by Edward Spearman in 1904. The rank correlation coefficient is defined as

$r = 1 - \frac{6\sum D^2}{n(n^2-1)}$, where $\sum D^2$= sum of squares of differences between the pairs of ranks and n = number of pairs of observations.

Spearman's Rank Correlation

The value of r lies between -1 and $+1$. If $r = +1$, there is complete agreement in order of ranks and the direction of ranks is also same. If $r = -1$, then there is complete disagreement in order of ranks and they are in opposite directions.

Computation for tied observations: There may be two or more items having equal values. In such case the same rank is to be given. The ranking is said to be tied. In such circumstances an average rank is to be given to each individual item. For example if the value so is repeated twice at the 5^{th} rank, the common rank to be assigned to each item is $\frac{5+6}{2} = 5.5$, which is the average of 5 and 6 given as 5.5, appeared twice.

Spearman's Rank Correlation

If the ranks are tied, it is required to apply a correction factor which is $\frac{m(m^2-1)}{12}$, where m is the number of items whose ranks are common and should be repeated as many times as there are tied observations.

A slightly different formula is used when there is more than one item having the same value. The formula is

$$r = 1 - \frac{6\left[\sum D^2 + C.F\right]}{n(n^2-1)},$$

where $C.F = \sum_{i=1}^{k} c.f_i$. Now, $c.f_i = \frac{m_i(m_i^2-1)}{12}, i = 1, 2, 3, \ldots, k$, where m_1 is the numbers of times repeated the tied rank and k is the number of distinct tied ranks occurs in both X and Y.

Problem: Find the rank correlation coefficient for the following data.

X	68	64	75	50	64	80	75	40	55	64
Y	62	58	68	45	81	60	68	48	50	70

Solution: Here $n = 10$.

X	Y	rank of $X = d_1$	rank of $Y = d_2$	$D = d_1 - d_2$	D^2
68	62	4	5	-1	1
64	58	6	7	-1	1
75	68	2.5	3.5	-1	1
50	45	9	10	-1	1
64	81	6	1	5	25
80	60	1	6	-5	25
75	68	2.5	3.5	-1	1
40	48	10	9	1	1
55	50	8	8	0	0
64	70	6	2	4	16
					$\sum D^2 = 72$

Solution Cont...

The correction factor C.F is $C.F = \sum_{i=1}^{k} c.f_i$, where

$c.f_i = \frac{m_i(m_i^2-1)}{12}, i = 1, 2, 3, \ldots, k.$

Here $k = 3$ because we have 3 tied ranks(2.5,6 and 3.5) occurs in both X and Y.

In X series, the tied rank 2.5 repeated 2 times, we set $m_1 = 2$ then

$c.f_1 = \frac{m_1(m_1^2-1)}{12} = \frac{1}{2}$

The tied rank 6 repeated 3 times, set $m_2 = 3$ then $c.f_2 = \frac{m_2(m_2^2-1)}{12} = 2$

In Y series, the tied rank 3.5 repeated 2 times, we set $m_3 = 2$ then

$c.f_3 = \frac{m_3(m_3^2-1)}{12} = \frac{1}{2}$

. Therefore $C.F = \sum_{i=1}^{3} c.f_i = 3.$

Solution Cont...

Hence

$$r = 1 - \frac{6\left[\sum D^2 + C.F\right]}{n(n^2-1)}$$

$$= 0.5454$$

Problem: In a marketing survey the price of tea and coffee in a town based on quality was found as shown below. Could you find any relation between tea and coffee price.

Price of tea	88	90	95	70	60	75	50
Price of coffee	120	134	150	115	110	140	100

Partial Correlation

Another measure of importance in a multivariate problem is the partial correlation. Partial correlation coefficient provides a measure of the relationship between the dependent variable and other variable, with the effect of the rest of the variables eliminated.

If there are three variables X_1, X_2 and X_3, there will be three coefficients of partial correlation, each studying the relationship between two variables when the third is held constant. If we denote by $r_{12.3}$ that is, the coefficient of partial correlation X_1 and X_2 keeping X_3 constant, it is calculated as:

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\left(\sqrt{1 - r_{13}^2}\right)\left(\sqrt{1 - r_{23}^2}\right)}.$$ 'Similarly

$$r_{13.2} = \frac{r_{13} - r_{12}r_{23}}{\left(\sqrt{1 - r_{12}^2}\right)\left(\sqrt{1 - r_{23}^2}\right)}, \text{and}$$

$$r_{23.1} = \frac{r_{23} - r_{12}r_{13}}{\left(\sqrt{1 - r_{12}^2}\right)\left(\sqrt{1 - r_{13}^2}\right)},$$

where r_{12}, r_{23} and r_{13} is the correlation coefficient between respective two variables.

Problem: Suppose on the basis of observations on hens in poultry, the number of eggs laid by hens (X_1) depends on food (X_2), and types of hen (X_3). It is found that $r_{12} = 0.8, r_{13} = 0.65$ and $r_{23} = 0.7$ Find the $r_{12.3}, r_{13.2}$ and $r_{23.1}$.

Solution: Let $r_{12.3}$ be the partial correlation between the number of eggs (X_1) and food taken (X_2) excluding the effect of type of hen (X_3) is given by

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\left(\sqrt{1 - r_{13}^2}\right)\left(\sqrt{1 - r_{23}^2}\right)}$$

$$= 0.636$$

Multiple correlation

In multiple correlation, we are trying to make estimates of the value of one of the variable based on the values of all the others. The variable whose value we are trying to estimate is called the dependent variable and the other variables on which our estimates are based a known as independent variables.

The coefficient of multiple correlation with three variables X_1, X_2 and X_3 are $R_{1.23}, R_{2.13}$ and $R_{3.21}$.

$R_{1.23}$, is the coefficient of multiple correlation related to X_1 as a dependent variable and X_2, X_3 as two independent variables and it canbe expressed interms of r_{12}, r_{23} and r_{13} as below

Multiple correlation

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{23}r_{13}}{1 - r_{23}^2}}$$

$$R_{2.13} = \sqrt{\frac{r_{12}^2 + r_{23}^2 - 2r_{12}r_{23}r_{13}}{1 - r_{13}^2}}$$

$$R_{3.12} = \sqrt{\frac{r_{13}^2 + r_{32}^2 - 2r_{12}r_{23}r_{13}}{1 - r_{12}^2}}.$$

Problem:A simple correlation coefficient between yield (X_1), temperature (X_2) and rainfall (X_3) are given by $r_{12} = 0.6, r_{13} = 0.5$ and $r_{23} = 0.8$

Calculate $R_{1.23}, R_{2.13}$ and $R_{3.21}$

Regression

Introduction: After knowing the relationship between two variables we may be interested in estimating (predicting) the value of one variable given the value of another. The variable predicted on the basis of other variables is called the "dependent" or the ' explained' variable and the other the 'independent' or the 'predicting' variable. The prediction is based on average relationship derived statistically by regression analysis. The equation, linear or otherwise, is called the regression equation or the explaining equation.

For example, if we know that advertising and sales are correlated we may find out expected amount of sales for a given advertising expenditure or the required amount of expenditure for attaining a given amount of sales.

Regression

The relationship between two variables can be considered between, say, rainfall and agricultural production, price of an input and the overall cost of product consumer expenditure and disposable income. Thus, regression analysis reveals average relationship between two variables and this makes possible estimation or prediction.

Defintion: Regression is the measure of the average relationship between two or more variables in terms of the original units of the data.

Types of Regression

- Simple and Multiple
- Linear and Non-Linear
- Total and Partial.

Simple and Multiple

In case of simple relationship only two variables are considered, for example, the influence of advertising expenditure on sales turnover. In the case of multiple relationship, more than two variables are involved. On this while one variable is a dependent variable the remaining variables are independent ones.

For example, the turnover (y) may depend on advertising expenditure (x) and the income of the people (z). Then the functional relationship can be expressed as $y = f(x, z)$.

Linear and Non-linear

The linear relationships are based on straight-line trend, the equation of which has no-power higher than one. But, remember a linear relationship can be both simple and multiple. Normally a linear relationship is taken into account because besides its simplicity, it has a better predective value, a linear trend can be easily projected into the future. In the case of non-linear relationship curved trend lines are derived. The equations of these are parabolic.

Total and Partial

In the case of total relationships all the important variables are considered. Normally,they take the form of a multiple relationships because most economic and business phenomena are affected by multiplicity of cases. In the case of partial relationship one or more variables are considered, but not all, thus excluding the influence of those not found relevant for a given purpose.

Linear Regression Equation

If two variables have linear relationship then as the independent variable (X) changes, the dependent variable (Y) also changes. If the different values of X and Y are plotted, then the two straight lines of best fit can be made to pass through the plotted points. These two lines are known as regression lines. Again, these regression lines are based on two equations known as regression equations. These equations show best estimate of one variable for the known value of the other. The equations are linear.

Linear regression equation of Y on X is

$$Y = a + bX \tag{1}$$

and Linear regression equation of X on Y is

$$X = a + bY \tag{2}$$

where a, b are constants.

From (1), We can estimate Y for known value of X.

and (2) We can estimate X for known value of Y.

Regression Lines

For regression analysis of two variables there are two regression lines, namely Y on X and X on Y. The two regression lines show the average relationship between the two variables. For perfect correlation, positive or negative i.e., $\rho = +1$, the two lines coincide i.e., we will find only one straight line. If $\rho = 0$, i.e., both the variables are independent then the two lines will cut each other at right angle. In this case the two lines will be parallel to X and $Y-$axes. Lastly the two lines intersect at the point of means of X and Y. From this point of intersection, if a straight line is drawn on $X-$axis, it will touch at the mean value of x. Similarly, a perpendicular drawn from the point of intersection of two regression lines on $Y-$axis will touch the mean value of Y.

Principle of ' Least Squares'

Regression shows an average relationship between two variables, which is expressed by a line of regression drawn by the method of "least squares". This line of regression can be derived graphically or algebraically. Before we discuss the various methods let us understand the meaning of least squares.

A line fitted by the method of least squares is known as the line of best fit. The line adapts to the following rules:

- The algebraic sum of deviation in the individual observations with reference to the regression line may be equal to zero. i.e., $\sum(X - X_c) = 0$ or $\sum(Y - Y_c) = 0$, where X_c and Y_c are the values obtained by regression analysis.
- The sum of the squares of these deviations is less than the sum of squares of deviations from any other line. i.e., $\sum(Y - Y_c)^2 < \sum(Y - A_i)^2$ Where A_i = corresponding values of any other straight line.
- The lines of regression (best fit) intersect at the mean values of the variables X and Y, i.e., intersecting point is $(\bar{x}, \bar{y})$.

Methods of Regression Analysis

Graphical Method (through regression lines using Scatter diagram.)

Under this method the points are plotted on a graph paper representing various parts of values of the concerned variables. These points give a picture of a scatter diagram with several points spread over. A regression line may be drawn in between these points either by free hand or by a scale rule in such a way that the squares of the vertical or the horizontal distances (as the case may be) between the points and the line of regression so drawn is the least. In other words, it should be drawn faithfully as the line of best fit leaving equal number of points on both sides in such a manner that the sum of the squares of the distances is the best.

Algebraic Method

Regression lines (through normal equations)

The two regression equations for X on Y; $X = a + bY$

Then the normal equations are

$\sum X = na + b\sum Y$

and $\sum XY = a\sum Y + b\sum Y^2$,

where X, Y are variables, and a, b are constants whose values are to be determined.

The two regression equations for Y on X; $Y = a + bX$, Then the normal equations are

$\sum Y = na + b\sum X$ and

$\sum XY = a\sum X + b\sum X^2$

Regression lines (through Regression coefficient)

Regression coefficient Method

The regression line of Y on X is

$$y - \bar{y} = \rho \left(\frac{\sigma_y}{\sigma_x} \right) (x - \bar{x}) \tag{3}$$

Here, the regression Coefficient of Y on X is $b_{yx} = \rho \left(\frac{\sigma_y}{\sigma_x} \right)$.

The regression line of X on Y is

$$x - \bar{x} = \rho \left(\frac{\sigma_x}{\sigma_y} \right) (y - \bar{y}) \tag{4}$$

Here, the regression Coefficient of X on Y is $b_{xy} = \rho\left(\frac{\sigma_x}{\sigma_y}\right)$.

If the deviation are taken from respective means of x and y

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

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

OR

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

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

Properties of Regression Co-efficient

- Correlation coefficient is the geometric mean of the regression coefficients ie, $\rho = \pm\sqrt{b_{yx}\, b_{xy}}$
- If $\rho = 0$, the variables are uncorrelated , the lines of regression become perpendicular to each other.
- If $\rho = \pm$, the two lines of regression either coincide or parallel to each other.
- Angle between the two regression lines is $\theta = \tan^{-1}\left[\frac{m_1 - m_2}{1 + m_1 m_2}\right]$, where m_1 and, m_2 are the slopes of the regression lines X on Y and Y on X respectively.

Problem

The 10 students are scored marks in Ecomomics(X) and Statistics(Y) test is given below

X	25	28	35	32	31	36	29	38	34	32
Y	43	46	49	41	36	32	31	30	33	39

- Find the two regression equations
- Find the Coefficient of correlation between marks in Economics and Statistics
- Fine the most likely marks in statistics when marks in Economics are 30.

Solution: Here $n = 10$.

X	Y	$X-\bar{X}$	$(X-\bar{X})^2$	$Y-\bar{Y}$	$(Y-\bar{Y})^2$	$(X-\bar{X})(Y-\bar{Y})$
25	43					
28	46					
35	49					
32	41					
31	36					
36	32					
29	31					
38	30					
34	33					
32	39					
$\sum_1$	$\sum_2$		$\sum_3$		$\sum_4$	$\sum_5$

Solution Cont...

From the above table, we get $\sum_1 = \sum X = 320, \sum_2 = \sum Y = 380, \sum_3 = \sum(X - \bar{X})^2 = ?, \sum_4 = \sum(Y - \bar{Y})^2 = ?$ and $\sum_5 = (X - \bar{X})(Y - \bar{Y}) = ?$.

Therefore, $\bar{x} = 32$ and $\bar{y} = 38$.

The regression line of Y on X is

$$y - \bar{y} = b_{yx}(x - \bar{x}), \text{where} \quad b_{yx} = \rho\left(\frac{\sigma_y}{\sigma_x}\right) = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sum(X - \bar{X})^2}.$$

$$\therefore y = -0.6643x + 59.25 - - - - - - - - - - - - - - - - - (i)$$

Solution Cont..

The regression line of X on Y is

$$x - \bar{x} = b_{xy}(y - \bar{y}), \text{where} \quad b_{xy} = \rho\left(\frac{\sigma_x}{\sigma_y}\right) = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sum(Y - \bar{Y})^2}.$$

$$\therefore x = -0.2337y + 40.88 - \text{(ii)}$$

The value $x = 30$ substitute in the equation (i), we get $y = 39$. Therefore, the statistics marks is 39, when the marks 30 in Economics.

Problem: Heights of father(X) and sons(Y) are give in centimeters.

X	150	152	155	157	160	161	164	166
Y	154	156	158	159	160	162	161	164

Find the two lines of regression and calculate the expected average height if the son when the height of the father is 154 cm.

Problem:Given $8X - 10Y + 66 = 0$ and $40X - 18Y = 214$. Find the correlation coefficient, ρ.

Answer:± 0.6

Multiple Regression

If the number of independent variable in a regression model is more than one then the model is called as multiple regression. In fact many of the real-world applications demand the use of multiple regression models.

Example: A sample application is as stated below

$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4,$

where Y-the economic growth rate of a country,

X_1- the time period

X_2-the size of the populations of the country

X_3-the level of the employment in %

X_4-the % of literacy.

b_0-intercept, and b_1, b_2, b_3, b_4-the slopes of the variables X_1, X_2, X_3, X_4.

Regression Model with two independent variables using Normal equations

Consider the multiple regression model

$$Y = b_0 + b_1 X_1 + b_2 X_2. \tag{5}$$

Then the normal equations are

$$\sum Y = nb_0 + b_1 \sum X_1 + b_2 \sum X_2 \tag{6}$$

$$\sum YX_1 = b_0 \sum X_1 + b_1 \sum X_1^2 + b_2 \sum X_1 X_2 \tag{7}$$

$$\sum YX_2 = b_0 \sum X_2 + b_1 \sum X_1 X_2 + b_2 \sum X_2^2. \tag{8}$$

Solve the equations (6),(7) and (8), we get b_0, b_1 and b_2. Then substitute these values in (5) and hence Y is the required regression model.

Problem: The annual sales revenue (in crores of rupees) of a product as a function of sales force(number of Salesmen) and annual advertising expenditure (in lakhs of rupees) for the past 10 years are summarized in the following table. Fit a least squares regression model.

Annual sales revenue Y	20	23	25	27	21	29	22	24	27	35
Sales force X_1	8	13	8	18	23	16	10	12	14	20
Annual advertising expenditures X_2	28	23	38	16	20	28	23	30	26	32

Solution

Consider the multiple regression model

$$Y = b_0 + b_1 X_1 + b_2 X_2. \tag{9}$$

Then the normal equations are

$$\sum Y = nb_0 + b_1 \sum X_1 + b_2 \sum X_2 \tag{10}$$

$$\sum YX_1 = b_0 \sum X_1 + b_1 \sum X_1^2 + b_2 \sum X_1 X_2 \tag{11}$$

$$\sum YX_2 = b_0 \sum X_2 + b_1 \sum X_1 X_2 + b_2 \sum X_2^2. \tag{12}$$

Y	X_1	X_2	X_1^2	X_2^2	YX_1	YX_2	X_1X_2
20	8	28					
23	13	23					
25	8	38					
27	18	16					
21	23	20					
29	16	28					
22	10	23					
24	12	30					
27	14	26					
35	20	32					
253	142	264	2246	7326	3678	6751	3617

Solution Cont...

The required Normal equations are

$$253 = 10b_0 + 142b_1 + 264b_2 \tag{13}$$

$$3678 = 142b_0 + 2246b_1 + 3617b_2 \tag{14}$$

$$6751 = 264b_0 + 3617b_1 + 7326b_2. \tag{15}$$

Solve the equations (13),(14) and (15), we get $b_0 = 5.14$, $b_1 = 0.62$, and $b_2 = 0.43$

Therefore, the required answer is $Y = 5.14 + 0.62X_1 + 0.43X_2$.

Thank you