1. Introduction

Welcome to the series of E-learning modules on Fitting of Linear Regression and Related Results. In this module, we are going to study about the meaning of regression, fitting line of regression, regression coefficients, its properties and study of angle between the two regression lines.

By the end of this session, you will be able to:

- Explain the meaning of regression
- Explain the fitting linear regression – regression line of Y on X and X on Y
- Explain regression coefficients
- Explain the properties of regression coefficients
- Explain the angle between the two regression lines and using this interpretation of correlation between the two variables

If the variables in a bivariate distribution are related, we will find that the points in the scatter diagram will cluster around some curve called the "curve of regression". If the curve is a straight line, 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.

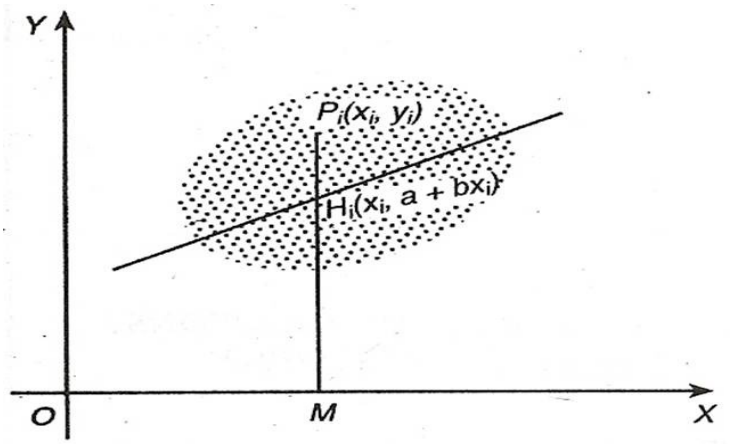
The line of regression is the line, which gives the best estimate to the value of one variable for any specific value of the other variable. Thus, the line of regression is the line of "best fit" and is obtained by Principle of Least Squares.

Let us consider a bivariate distribution x i and y i, where i take values 1, 2, up to n. Here Y is dependent variable and X is the independent variable. Let the line of regression of Y on X be,
Y is equal to a plus b into X
The above equation represents a family of straight lines for different values of the arbitrary constants 'a' and 'b'. The problem is to determine 'a' and 'b' so the above equation is the line of best fit.

The term best fit is interpreted in accordance with Legender's Principle of Least Squares, which consists in minimizing the sum of the squares of the deviations of the actual values of y from their estimated values as given by the line of best fit.

Following diagram shows the scatter plot, which is clustered around a straight line.

Figure 1



Let P i of x i, y i be any general point in the scatter diagram. Draw P i M perpendicular to X axis meeting the line in H i. Abscissa of H i is x i and its ordinate is a plus b into x i. Hence, the co-ordinates of H i are x i, and a plus b into x i.
P i H i is equal to P i M minus H i M is equal to y i minus of a plus b into x i, which is called as the error of estimate or the residual for y i.

According to the principle of least squares, we have to determine a and b so that
E is equal to summation over i from 1 to n P i H i square is equal to summation over i from 1 to n y i minus a minus b into x i whole square is minimum.

From the principle of maxima and minima, the partial derivatives of E, with respect to a and b should vanish separately.
That is, d E by d a is equal to zero is equal to minus 2 into summation y i minus a minus b into x i implies, summation y i is equal to n into a plus b into summation x i.
d E by d b is equal to zero is equal to minus 2 into summation x i into y i minus a minus b into x i implies, summation x i into y i is equal to a into summation x i plus b into summation x i square, which are the normal equations for estimating a and b.

The values of summation y_i, summation $x_i y_i$, summation x_i, summation x_i square are obtained from the given set of points (xi, yi), where i is equal to 1, 2 up to n and the two normal equations can be solved for a and b. With values of a and b so obtained, the linear regression line y is equal to a plus bx is the best fit to the given set of points.

2. Regression Line Y on X

Now, let us obtain the regression line Y on X.
Consider the equation, summation y i is equal to n into a plus b into summation x i.
On dividing by n, we get, y bar is equal to n into a plus b into x bar. Name the equation as 1.

Thus the line of regression of Y on X passes through x bar and y bar.

Now, Mew 1, 1 is equal to covariance of X and Y is equal to 1 by n into summation over i from 1 to n x i into y i minus x bar into y bar
Implies 1 by n into summation x i into y i is equal to mew 1, 1 plus x bar into y bar. Name the equation as 2.
Also, sigma x square is equal to 1 by n into summation x i square minus x bar square.
Implies, 1 by n into summation x i square is equal to sigma x square plus x bar square. Name the equation as 3.

Now consider the equation, summation x i into y i is equal to a into summation x i plus b into summation x i square. Name this equation as 4.

Dividing by n and substituting for summation x i into y i by n and summation x i square by n, we get,
Mew 1, 1 plus x bar into y bar is equal to a into x bar plus b into sigma x square plus x bar square. Name this equation as 5.

Multiplying the equation 4 by x bar and subtracting from the equation 5, we get,
Mew 1, 1 is equal to b into sigma x square implies b is equal to mew 1, 1 divided by sigma x square.

Since b is the slope of the line of regression Y on X and since the line of regression passes through the point x bar and y bar, its equation is:
Y minus y bar is equal to b into X minus x bar is equal to mew 1, 1 divided by sigma x square into X minus x bar. Name the equation as 6.
Implies, Y minus Y bar is equal to r into sigma y divided by sigma x into X minus x bar. Name the equation as 7.

Starting with the linear equation X is equal to A plus B into Y and similarly proceeding or by simply interchanging the variables X and Y in the equations, 6 and 7, the equation of the line of regression of X on Y becomes,
X minus x bar is equal to mew 1, 1 divided by sigma y square into Y minus y bar
Implies, X minus x bar is equal to r into sigma X divided by sigma Y into Y minus y bar.

b is the slope of the line of regression of Y on X is also called as the coefficient of regression of Y on X. It represents the increment in the value of dependent variable Y corresponding to a unit change in the value of independent variable X. More precisely, we write,
b Y X is equal to Regression coefficient of Y on X
That is b Y X is equal to mew 1, 1 divided by sigma X square is equal to r into sigma Y divided by sigma X.

Similarly, the coefficient of regression of X on Y indicates the change in the value of variable X corresponding to a unit change in the value of variable Y and is given by,
b X Y is equal to regression coefficient of X on Y.
That is, b X Y is equal to mew 1, 1 divided by sigma Y square is equal to r into sigma x divided by sigma y.

3. Properties of Regression Coefficients

Now, let us prove some properties of regression coefficients.
Correlation coefficient is the geometric mean between the regression coefficients.
To prove this property, multiply b X Y and b Y X. we get,

b X Y into b Y X is equal to r into sigma X by sigma Y into r into sigma Y by sigma X is equal to r square
implies, r is equal to square root of b X Y into b Y X

The second property of regression coefficient is:
If one of the regression coefficients is greater than unity, the other must be less than unity.

To prove this property, let us take one of the regression coefficient, say b Y X is greater than unity then we have to show that b X Y is less than 1
Now, b Y X is greater than 1 implies 1 by b Y X is less than 1
Also, r square is less than or equal to 1
implies, b Y X into b X Y is less than or equal to 1
That is, b X Y is less than or equal to 1 by b Y X, which is less than 1.
Hence, b X Y is less than 1.

Third property of regression coefficient is:
The modulus value of the arithmetic mean of the regression coefficient is not less than the modulus value of the correlation coefficient r.
In the above property, we have to prove that modulus of half into b Y X plus b X Y is greater than modulus r.

Substituting for b Y X and b X Y, we get,
Modulus of half into r into sigma Y divided by sigma X plus r into sigma X divided by sigma Y is greater than modulus of r square
Implies sigma Y by sigma X plus sigma X by sigma Y is greater than or equal to 2, because modulus of r is greater than zero.

By taking common denominator and simplifying, we get,
Sigma Y square plus sigma X square minus 2 into sigma X into sigma Y is greater than zero
Implies, sigma Y minus sigma X whole square is greater than zero, which is always true, since the square of a real quantity is greater than or equal to zero.

The fourth property says that Regression coefficients are independent of the change of origin but not of scale.
To prove this property, let us consider the transformation, U is equal to X minus a divided by h and V is equal to Y minus b divided by k
Implies, X is equal to a plus h into U and Y is equal to b plus k into v, where a, b, h which are positive and constant k is also positive.

Then, covariance of X Y is equal to h into k into Covariance of U V
Sigma X square is equal to h square into sigma U square and Sigma Y square is equal to k square into sigma V square
Therefore, b Y X is equal to covariance of X Y divided by sigma X square is equal to h into k into covariance of U V divided by h square into sigma U square is equal to k by h into covariance of U V by sigma U square is equal to k by h into b V U.

Similarly,
B X Y is equal to covariance X Y divided by sigma Y square is equal to h into k into covariance of U V divided by k square into sigma V square
Is equal to h divided by k into covariance of U V by sigma V square is equal to h by k into b U V.

4. Angle Between Two Lines of Regression

Now, let us obtain the angle between two lines of regression.
We know that, equation of the lines of regression of Y on X and X on Y are:
Y minus y bar is equal to r into sigma Y divided by sigma X into X minus x bar and
X minus x bar is equal to r into sigma X divided by sigma Y into Y minus y bar.
Slopes of these lines are, r into sigma Y divided by sigma X and r into sigma x divided by sigma y respectively.

If θ is the angle between the two lines of regression, then,
Tan theta is equal to modulus of r into sigma Y by sigma X minus sigma Y by r into sigma X divided by,
1 plus r into sigma Y by sigma X into sigma Y by r into sigma X
Is equal to modulus of r square minus 1 divided by r into sigma X into sigma Y divided by sigma X square plus sigma Y square
Since r square is less than or equal to 1, we get,
1 minus r square divided by modulus of r into sigma X into sigma Y divided by sigma X square plus sigma Y square
Therefore, theta is equal to tan inverse of 1 minus r square divided by modulus of r into sigma X into sigma Y divided by sigma X square plus sigma Y square.

Now, let us consider the two cases, when the variables are uncorrelated and perfectly correlated.
Case i: r is equal to zero.
If r is equal to zero, tan theta is equal to infinity, implies, theta is equal to pi by 2.
Thus, if the two variables are uncorrelated, the lines of regression become perpendicular to each other.

Case ii: r is equal to plus or minus 1
If r is equal to plus or minus 1, then tan theta is equal to zero implies, theta is equal to zero or pi. In this case, either the two lines of regression coincide or they are parallel to each other. However, since both lines of regression pass through the point x bar y bar, they cannot be parallel. Hence, in the case of perfect correlation, positive or negative, the two lines of regression coincide.

5. Remarks of the Lines of Regression

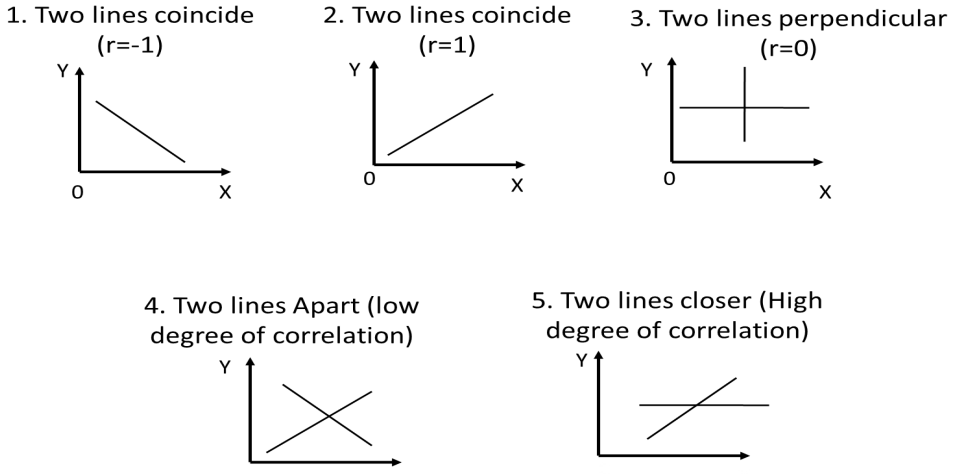
Now, let us go through some remarks.

1. Whenever two lines intersect, there are two angles between them, one acute angle and other obtuse angle. Further, tan theta is greater than zero if zero less than theta less than pi by 2, that is theta is an acute angle and tan theta is less than zero if pi by 2 is less than theta less than pi, that is theta is an obtuse angle and since zero less than r square less than 1, the acute angle theta 1 and obtuse angle theta 2 between the two lines of regression are given by,
   Theta 1 is equal to acute angle is equal to tan inverse of 1 minus r square divided by modulus r into sigma X into sigma Y divided by sigma X square plus sigma Y square, and θ two is equal to pi minus theta 1.
2. When r is equal to zero, that is, variables X and Y are uncorrelated, then the lines of regressions of Y on X and X on Y are given respectively by,
   Y is equal Y bar and X is equal to X bar as shown in the adjoining figure. Hence, in this case, that is r is equal to zero, the lines of regression are perpendicular to each other and are parallel to X-axis and Y-axis respectively.
3. The fact that if r is equal to zero that is the variables uncorrelated, the two lines of regression are perpendicular to each other. If r is equal to plus or minus 1, theta is equal to zero, that is, the two lines coincide, leads us to the conclusion that for higher degree of correlation between the variables, the angle between the lines is smaller, that is, the two lines of regression is nearer to each other.

On the other hand, if the lines of regression make a larger angle, they indicate a poor degree of correlation between the variables and ultimately for theta is equal to pi by 2, and r is equal to zero, that is, the lines becomes perpendicular if no correlation exists between the variables.
Thus, by plotting the lines of regression on a graph paper, we can have an approximate idea about the degree of correlation between the two variables under study.

Consider the following illustrations:

Figure 2



The first figure shows that the line has been drawn from top to bottom, which means the two regression lines coincide when there is perfect negative correlation.
The second figure shows that the two regression lines coincide when there is perfect positive correlation.
In the third figure, observe that the two lines are perpendicular to each other which indicates that the two variables are not correlated that is r is equal to zero.
In the fourth figure, the two regression lines are apart indicating that there is low degree of correlation between the two variables.
And the last figure shows that the two regression lines are closer, which indicates that there is high degree of correlation between the two variables.

Let us consider the following illustration, where a linear regression line is used to estimate the future values.
From the following data, find the most likely price in Mumbai corresponding to the price of Rs.70 at Kolkata.
Where, we have given the average price and standard deviation in two cities Kolkata and Mumbai. Also, the correlation coefficient between the prices of commodities in the two cities is 0.8.

Let the prices (in Rupees) in Kolkata and Mumbai be denoted by X and Y respectively. Then we have given:
X bar is equal to 65, Y bar is equal to 67, sigma X is equal to 2 point 5, sigma Y is equal to 3 point 5, r is equal to zero point 8.
Line of regression Y on X is
Y minus y bar is equal to r into sigma Y divided by sigma X into X minus x bar.
On substituting and simplifying when X is equal to 70, we get, Y is equal to 72 point 6.
Hence, the most likely price in Mumbai corresponding to the price of Rs. 70 at Kolkata is Rs 72.60

Here's a summary of our learning in this session, where we have understood:

- The meaning of regression
- The fitting linear regression – regression line of Y on X and X on Y
- The regression coefficients
- The properties of regression coefficients
- The angle between the two regression lines and using this interpretation of correlation between the two variables