**Regression**

The word ‘***regression’*** is used to denote estimation or prediction of the **average value** of one variable (say, ) for a specified value of the other variable (say, ).

To solve this problem it would be necessary to express the relationship between and in a mathematical form. Such an equation is known as ***regression equation*** and its geometrical representation is called a regression line/curve.

Suppose the approximate relation may be represented by a line:

(1)

where, denotes the predicted value of . To get an approximate line, it is necessary to determine and from the observed data. Let us assume that there are given pairs of values of and , the pair being denoted by (, ). The above line gives as an estimate of the value

(2)

The difference is thus the error of estimate for the pair (). Since the line is to be used for estimating purposes, it is reasonable to require that and should be such that these errors of estimates are as small as possible.

However, it will not be enough to minimize the sum of these errors, because the errors, which may be positive or negative, may even add up to zero for a line for which the individual errors are of high magnitude. Therefore, a satisfactory method of determining and would be the *method of least squares*, which consists in minimizing the sum of squares of the errors (SSE) of estimation. Thus the problem is to chose and in such a way as to minimize

(3)

(4)

The equations (3) and (4) are called ***normal equations***. The roots of these equations are

and

Therefore, the estimates of the constants and are obtained as

and .

Substituting the estimates of and in equation (1), we get the desired prediction formula:

(5)

The line represented by equation (5) is called the **regression line of on** . is the intercept of the line and is its slope. The coefficient (denoted as ) is called the **regression coefficient of on** . It is the amount by which increases for a unit increment in the value of .

Similarly, if we are interested in predicting from , we use the **regression line of on** , which has the equation

(6)

It may be noted that both the regression lines pass through the point (, ), which is their point of intersection.

**Properties of Linear Regression**

1. There are two linear regression equations - (i) Regression equation of on : and Regression equation of on : , where and are respectively regression coefficient of on and the regression coefficient of on . The values of and are obtained as follows:

and

1. The product of the two regression coefficients is equal to the square of correlation coefficient between and , i.e.

or,

Thus numerically the correlation coefficient is the geometric mean of the two regression coefficients. As regards the sign of , it is the same as the common sign of the two regression coefficients.

1. The mean of the observed values of is equal to the mean of the corresponding predicted values.

***Proof:***

Dividing both sides by and remembering that , we have

From this it follows that the mean of the errors of estimates is zero.

***Proof:***

or

or , since

Hence,

1. The residual variance,

***Proof:***

The standard deviation of , which is called the standard error of estimate of from its linear regression on , is denoted by . We have then

Since, , we have or or , a result which has already been proved in a different way.

If , then . That is, in this case for each , so that all points in the scatter diagram lie on the regression line.

On the other hand, if, then . This means that the errors of estimation are as much as variable as the original values of , and hence the linear regression equation is of no help in predicting the value of when the corresponding value of is given.

Thus, the numerical value of serves as a measure of the worth of the linear regression equation of one variable on the other as a predicting formula. The higher the numerical value of , the more efficient is the linear equation.

1. The correlation coefficient between and is zero.

***Proof:***

(From normal equation (4), )

Hence,

may be looked upon as the part of which is uncorrelated with .

1. The correlation between and its predicted value must be non-negative and must be numerically the same as the correlation coefficient between and .

***Proof:*** Since, ,

Again, and therefore,

1. The angle between the two regression lines depends on the correlation coefficient . When , the two lines are perpendicular to each other; when or , they coincide. As increases numerically from 0 to 1, the angle between the regression lines diminishes from to .

**Coefficient of determination**

We have seen earlier that

(1)

where is the predicted value of , from the linear regression equation, corresponding to . Now,

[the overall variation of the observed values of ]

[since, ]

Residual variation + Variation due to regression

Explained variation

Unexplained variation

Therefore, may be interpreted as the proportion of the total variation of which is accounted for by its linear regression on .

Thus, is called as the *coefficient of determination*.

**Correlation index**

The concept of coefficient of determination can be generalized. Suppose, the appropriate regression equation is a polynomial of degree (), e.g. . Then we can define a measure of association, similar to as

where is the predicted value of , from the degree polynomial regression equation, corresponding to .

The is called as the *correlation index* of the order.

It can be shown that as the degree of the polynomial increases, the value of the correlation index also increases i.e. .

**Correlation ratio**

* Correlation ratio is a coefficient of non-linear association.
* In the case of linear relationships, the correlation ratio that is denoted by etabecomes the correlation coefficient.
* In the case of non-linear relationships, the value of the correlation ratio is greater, and therefore the difference between the correlation ratio and the correlation coefficient refers to the degree of the extent of the non-linearity of relationship.
* *Whereas the correlation coefficient is symmetrical in and , the correlation ratio is not, and so generally and will not be equal.*

**Exercises**

**Exercise 1:** Marks obtained by 12 students in the college test (x) and the university test (y) are as follows:

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 41 | 45 | 50 | 68 | 47 | 77 | 90 | 100 | 80 | 100 | 40 | 43 |
|  | 60 | 63 | 60 | 48 | 85 | 56 | 53 | 91 | 74 | 98 | 65 | 43 |

What is your estimate of the marks a student could have obtained in the university test if he obtained 60 in the college test but was ill at the time of the university test?

**Exercise 2:** Let the lines of regression concerning two variables x and y be given by and . Obtain the values of the means and the correlation coefficient. [Ans. =6.7, =25.3, ]

**Exercise 3:** Regression of savings (s) of a family on income (y) may be expressed as , where and are constants. In a random sample of 100 families the variance of savings is one-quarter of the variance of incomes and the correlation is found to be 0.4. Obtain the estimate of . [Ans. ]

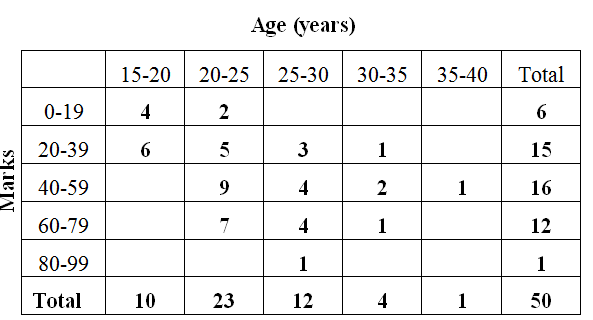
**Exercise 4:** The following data relate to the stature () and sitting height (), both in cm., for each of 30 people of a particular Indian caste:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
| 172.8 | 83.9 | 157.7 | 77.7 | 170.2 | 83.4 |
| 166.0 | 83.6 | 146.7 | 76.4 | 153.3 | 79.1 |
| 164.1 | 81.3 | 153.2 | 77.2 | 173.7 | 86.1 |
| 164.4 | 85.4 | 155.3 | 80.1 | 155.8 | 78.6 |
| 168.8 | 83.9 | 151.5 | 76.9 | 158.0 | 80.1 |
| 165.2 | 81.1 | 161.1 | 81.5 | 157.2 | 81.6 |
| 170.0 | 84.9 | 156.3 | 80.9 | 156.2 | 78.6 |
| 163.5 | 81.1 | 169.4 | 83.1 | 168.2 | 82.5 |
| 169.4 | 84.9 | 159.9 | 84.2 | 164.4 | 84.1 |
| 159.1 | 79.6 | 161.7 | 80.3 | 165.5 | 87.1 |

1. Represent the data by means of a scatter diagram
2. Compute the correlation coefficient of and

[Ans. 0.839]

**Exercise 5**: From the following bivariate frequency distribution, calculate (i) the coefficient of correlation and (ii) the regression equation of marks () on age .



**Exercise 6**: During an investigation in an agricultural farm in Bengal, the length (in cm.) of green jute plant and the weight (in gm.) of dry jute fibre were observed for 350 plants. With these data, the following bivariate frequency table was obtained:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | Length of green plant (in cm.)  Class mark | | | | | | | Total |
| 111.5 | 127.5 | 143.5 | 159.5 | 175.5 | 191.5 | 207.5 |
| Weight of dry jute fibre (gm.)  Class mark | 1.175 | 12 | 25 | 15 | 1 |  |  |  | 53 |
| 2.775 | 1 | 4 | 33 | 59 | 29 | 3 |  | 129 |
| 4.375 | 1 |  | 4 | 28 | 35 | 14 | 2 | 84 |
| 5.975 |  |  |  | 2 | 20 | 18 | 1 | 41 |
| 7.575 |  |  |  | 1 | 1 | 14 | 5 | 21 |
| 9.175 |  |  |  |  | 4 | 8 | 2 | 14 |
| 10.775 |  |  |  |  |  | 3 | 2 | 5 |
| 12.375 |  |  |  |  |  |  | 3 | 3 |
| Total | | 14 | 29 | 52 | 91 | 89 | 60 | 15 | 350 |

From the above table, compute the coefficient of correlation of the two variables. Also, find the linear regression equation of weight of dry fibre on length of green plant.

[Ans. = 0.755; regression eq.: = 8.379 + 0.0756]