S C Agarkar

V N BRIMS

Thane

Regression Analysis

regression

- The term regression was first used as a statistical concept in 1877 by Sir Francis Galton. He designated the word regression as the name of a general process of predicting one variable from another.
- In regression analysis we will develop an estimating equation, that is a mathematical formula that relates the known variable to the unknown variable.

Correlation and Regression

- Correlation measures the degree of relationships between two variables whereas regression studies about the nature of relationships.
- Correlation helps in finding the degree of relationship between two variables, no dependency is established. In regression one variable is taken as dependent while other is taken as independent.
- The value of correlation coefficient (r) is symmetric while regressioncoefficient by x andbxyarenot symmetrical.
- We can have non sense correlation but not the non sense regression.
- The value of r depends on the scale but the value of regression coefficients are independent of scales.

Types of relations

- Direct relationship between X and Y, for example GDP and the number of cars per thousand persons.
- Inverse relationships between X and Y for example pressure and volume.
- Curvilinear relationship between X and Y, for example, manufacturing time per unit for new aircraft.

Scatter diagrams

- In order to see if there is any relationship in two variables we must plot the data on a graph paper.
- Such a plot will give us a scatter diagram. It can show whether there is any relationship or not. If such a relation exists we can find out an estimating equation.
- Students scores on entrance examinations and cumulative grade point average at graduation are given below. Plot them to get scatter diagram.
- Scores: 74 69 85 63 82 60 79 91
- GPA: 2.6 2.2 3.4 2.3 3.1 2.1 3.2 3.8

Regression line

- In a scatter diagram we can draw a straight line by fitting it with as many data points as possible. The same task can be done more precisely by using the equation of astraight line.
- We have studied that the equation of a straight line where the dependent variable Y is determinedby independentvariable X is given by $Y = a + bX$.Inthis equation a is the Y intercept and b is slope of the line.

Estimating equation of a line

- In order to get the equation from the data we need to estimate the values of a and b. Let us first estimate b. It is given by the equation
- $b = Y_2 – Y_1 / X_2 – X_1$
- We have to pick up two points and find out their coordinates. The value of b can be obtained by substituting the values of these coordinates.
- Once we have the value of b known we can calculate the value of a using the equation of a straight line $Y = a + bX$.Alternativelyone can find the value of a by looking at the Y intercept if a plot exists.

Problem to solve

- For the following set of data
- A) Plot the scatter diagram
- B) Develop the estimating equation that best describes the data
- C) predict Y for X = 4, 9, 12
- X: 7 10 8 5 11 3 7 11 12 6
- Y: 2.0 3.0 2.4 1.8 3.2 1.5 2.1 3.8 4.0 2.2

The method of least squares

- How can we fit the line mathematically if none of the points lie on the line. To a statistician the line will have a good fit if it minimizes the error between estimated points on the line and actual observed points. In such cases we need to useequation $\check{Y}$= a + b X.Here $\check{Y}$ isusedto symbolize estimatedvalues.
- It seems reasonable that the farther away a point is from estimated line, the more serious is the error. We would rather have several small errors than one large error. We accomplish this by squaring the individual errors. It magnifies the large errors and cancels the effect of positive and negative values. Since we are looking for estimating line that minimizes the sum of the squares of errors, we call this method the least square method.

Least Square Regression Line

- Statisticians have derived two equations to find the slope and y intercept of the best fitting regression line. The first formula is
- $b= \sum XY - n\ddot{X}\bar{Y} / \sum X^2 - n\ddot{X}^2$
- Where b is the slope of best fitting estimating line
- X is value of independent variable
- Y is he value of dependent variable
- $\ddot{X}$ is mean of values of independent variable
- $\bar{Y}$ isthe mean value of the dependent variable
- n is the number of data points that is number of pairs

- The second formula is
- $a = \bar{Y} - b\ddot{X}$
- Where a is Y intercept and b is the slope from equation.

Using the leastsquare method

- Suppose the Director of a Chapel Hill Sanitation Department is interested in the relationship between the age of a garbage truck and the annual repair expense he should expect to incur. In order to determine this relationship the Director has accumulated information concerning four trucks the city currently owns. The data is given below.

Truck Number	Age of truck in years	Repair expenses last year in hundred of USD
101	5	7
102	3	7
103	3	6
104	1	4

Calculation

- $\sum X = 12$ Hence $\ddot{X} = 12/4 = 3$
- $\sum Y = 24$,Hence $\bar{Y}= 24/4 = 6$
- $\sum XY = 78$ and $\sum X^2= 44$
- $b = \sum XY – n\ddot{X}\bar{Y}/ \sum X^2- n\ddot{X}^2$
- $= 78 – 4*3*6/ 44- 4*3^2$
- $= 78-72/44-36 = 6/8 = 0.75$
- Y intercept $a =\bar{Y}-b\ddot{X}$
- $= 6 - 0.75 * 3 = 3.75$
- Henceequation $\check{Y}= a + b X$
- $= 3.75 + 0.75 X$
- Using this equation we can estimate the expenses incurred on a truck that is four years old by putting the value of X as 4 as get
- $Y = 3.75 + 0.75 (4) = 6.75$
- It means the annual repair expense for the tuck would be USD 675.

Problem to solve

The vice president for research and development of a large chemical and fiber manufacturing company believes that the firm’s annual profits depend on the amount spent on R & D. The new chief executive officer does not agree and asked for evidence. If thedata obtained are givenin the table can you please help the vice president to provideevidence?

year	Millions spent on R & D	Annual Profit in millions
1978	2	20
1979	3	25
1980	5	34
1981	4	30
1982	11	40
1983	5	31

Standard error of estimate

- To measure the reliability of the estimating equation, statisticians have developed standard error of estimate. It is given by the formula
- $S_e= \surd\ (Y -\check{Y})^2/n-2$ where
- Y is the value of dependent variable
- $\check{Y}$ is the estimated values from estimating equation
- N =number of data points
- Calculating the valuesof $\check{Y}$ fromthe equation and substituting them in the above formula we get
- S_e= 0.866 that is USD 86.6

Short cut method of calculating standard error of estimation

- The formula for standard error of estimation can be modified as follows:
- $S_e = \sqrt{}\sum Y^2 – a\sum Y – b\sum XY/n-2$ where
- X: values of independent variable
- Y: Values of dependent variable
- a = Y intercept from equation
- b= slope of the line
- n = number of data points
- Substituting the values we once again get
- $S_e = 0.866$

Thank you, the tide regresses to its mean value

that enables you to judge when to go to the fort.