SIMPLE LINEAR REGRESSION Part 1

DR NISHITA GUPTA
DEPARTMENT OF COMMERCE
SRI AUROBINDO COLLEGE

LEARNING OUTCOMES:

- Understand the concept of linear regression
- Compute regression coefficients and regression line
- Apply regression analysis to predict the values of a dependent variable from known values of independent variable
- Distinguish between regression and correlation

MEANING OF REGRESSION

- Regression analysis means the estimation or prediction of the unknown value of one variable from the known values of one or more variables. It attempts to establish the "nature of relationship" between variables.
- Simple regression means study of only two variables, a dependent and an independent variable.
- Dependent/Explained variable -The variable whose value is to be predicted is called dependent or explained variable.
- Independent/Explanatory variable – The variables which are used to predict the values of a dependent variable are called independent or explanatory variables.

MEANING OF REGRESSION

- Simple linear regression means when the relationship between the dependent and independent variable is linear.
- Regression analysis plays a very important role in the field of every human activity.
- A businessman may be keen to know what would be his estimated profit for a given level of investment on the basis of the past records.
- y is known as dependent variable and x is known as independent variable.
- In the previous examples since profit depends on investment ***profit*** is the ***dependent variable*** and ***investment*** is the ***In-dependent variable***.

LINES OF REGRESSION-THE LEAST SQUARES APPROACH

- The **least square line of regression of Y on X**: This equation is used to estimate
- value of Y for a given value of X.
- Y = a + bX

Where, a and b are constants

The value of constant a and b can be find out with the help of two normal equations. The two normal equations are as follows:

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

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

LINES OF REGRESSION-THE LEAST SQUARES APPROACH

- b = it is called the **regression coefficient of Y on X** and is denoted by b_{yx}.

It measures the change in Y corresponding to a unit change in X. Thus

b_{yx} = Slope of the line of regression of Y on X and given by

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

And $a = \overline{Y} - b\overline{X}$

LINES OF REGRESSION-THE LEAST SQUARES APPROACH

- The line of regression of Y on X passes through the point $(\bar{X}, \bar{Y})$ and hence the equation of the line of regression of Y on X (Y = a + bX) can also be written as
- $Y - \bar{Y} = b_{yx} (X - \bar{X})$

LINES OF REGRESSION-THE LEAST SQUARES APPROACH

- The **least square line of regression of X on Y**: This equation is used to estimate a value of X for a given value of Y.
- $X = c + dY$

Where, c and d are constants

- The two normal equations for estimating c and d are given by

$\sum X = nc + d\sum Y$

$\sum XY = c\sum Y + d\sum Y^2$

LINES OF REGRESSION-THE LEAST SQUARES APPROACH

- d = it is called the **regression coefficient of X on Y** and is denoted by b_{xy}. It measures the change in X corresponding to a unit change in Y. Thus
- b_{xy}= Slope of the line of regression of X on Y and given by

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

And $c = \overline{X} - d\overline{Y}$

LINES OF REGRESSION-THE LEAST SQUARES APPROACH

- The line of regression of X on Y passes through the point $(\bar{X}, \bar{Y})$ and hence the equation of the line of regression of X on Y (X = c + dY) can also be written as
- $X - \bar{X} = b_{XY}(Y - \bar{Y})$

- **Remarks:**
- **It may be remarked that there are always two lines of regression, one of Y on X and the other X on Y.**
- **Y on X = to predict value of Y from known values of X**
- **X on Y = to predict value of X from known values of Y**

LINES OF REGRESSION-THE LEAST SQUARES APPROACH

- **Since the two lines of regression pass through the point (X,Y), the mean values (X,Y) can be obtained as the point of intersection of the two regression lines.**
- **Regression Coefficient – Some Formulas**
- 1. Formulas for regression Coefficients in terms of Covariance and Variance:
- $b_{YX} = \frac{Cov(X,Y)}{\sigma^2{}_X}$
- $b_{XY} = \frac{Cov(X,Y)}{\sigma^2{}_Y}$

LINES OF REGRESSION-THE LEAST SQUARES APPROACH

- Formulas for Regression Coefficients in terms of Deviations of X and Y values from their respective means:
- $b_{YX} = \dfrac{\sum(X-\overline{X})(Y-\overline{Y})}{\sum(X-\overline{X})^2}$
- $b_{XY} = \dfrac{\sum(X-\overline{X})(Y-\overline{Y})}{\sum(Y-\overline{Y})^2}$

LINES OF REGRESSION-THE LEAST SQUARES APPROACH

- Formulas for Regression Coefficient in terms of r, σ_x and σ_y
- $b_{YX} = \dfrac{r\,\sigma_Y}{\sigma_X}$
- $b_{XY} = \dfrac{r\,\sigma_X}{\sigma_Y}$

Example:

X	Y
3	11
4	12
8	9
7	3
2	5

a) What likely is the value of X, if Y=10.
b) What likely is the value of Y, if X=10.

So here the question is asking us to find regression equation Y on X for b) part and regression equation X on Y for a) part.

- Regression Equation:
- Y on X
- $Y - \bar{Y} = b_{yx} (X - \bar{X})$

- So let us calculate mean first:
- $\sum X = (3+4+8+7+2) = 24$
- $N = 5$
- $\bar{X} = \frac{\sum X}{N} = \frac{24}{5} = 4.8$

$\sum Y = (11+12+9+3+5) = 40$

$N = 5$

$\bar{Y} = \frac{\sum Y}{N} = \frac{40}{5} = 8$

Now let us calculate b_{yx} and b_{xy}

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

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

X	Y	XY	X^2	Y^2
3	11	33	9	121
4	12	48	16	144
8	9	72	64	81
7	3	21	49	9
2	5	10	4	25
24	**40**	**184**	**142**	**380**

- $b_{yx} = \dfrac{5 \times 184 - 24 \times 40}{5 \times 142 - (24)^2}$

  $= \dfrac{920 - 960}{710 - 576}$

  $= \dfrac{-40.}{134.} = -0.298$

- $b_{xy} = \dfrac{5 \times 184 - 24 \times 40}{5 \times 380 - (40)^2}$

  $= \dfrac{920 - 960}{1900 - 1600}$

  $= \dfrac{-40}{300} = -0.133$

- Regression Equation:
- Y on X
- $Y - \overline{Y} = b_{yx} (X - \overline{X})$
- $Y - 8 = -0.298(X - 4.8)$
- $Y - 8 = -0.298X + 1.4304$
- **$Y = -0.298X + 9.4304$**

Solve in terms of Y.

- X on Y
- $X - \overline{X} = b_{xy} (Y - \overline{Y})$
- $X - 4.8 = -0.133(Y - 8)$
- $X - 4.8 = -0.133Y + 1.064$
- **$X = -0.133Y + 5.864$**

Solve in terms of X.

a) X= -0.133Y + 5.864

If Y= 10

X= -0.133(10) + 5.864

X = 4.534

b) Y= -0.298X + 9.4304

If X = 10

Y= -0.298(10) + 9.4304

Y= 6.4504

- The following table gives the age of cars of a certain make and annual maintenance costs. Obtain the regression equation for costs related to age:

Age of Cars (in years)	2	4	6	8
Maintenance cost (in Rs. hundred)	10	20	25	30

- Also, estimate the annual cost for a ten-year-old car

- Let X = Age of cars,
  Y = Annual maintenance cost of cars

- We need Equation of Cost related to age i.e, Y on X
- Regression Equation of Y on X:
- $Y - \overline{Y} = b_{yx} (X - \overline{X})$

- So let us calculate mean first:
- $\sum X = (2 + 4 + 6 + 8) = 20$
- $N = 4$
- $\overline{X} = \frac{\sum X}{N} = \frac{20}{4} = 5.$

$\sum Y = (10 + 20 + 25 + 30) = 85$

$N = 4$

$\overline{Y} = \frac{\sum Y}{N} = \frac{85}{4} = 21.25$

Now let us calculate b_{yx}

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

X (Age)	Y(Cost)	X^2	Y^2	XY
2	10	4	100	20
4	20	16	400	80
6	25	36	625	150
8	30	64	900	240
$\sum X = 20$	$\sum Y = 85$	$\sum X^2 = 120$	$\sum Y^2 = 2025$	$\sum XY = 490$

- $b_{yx} = \frac{4 \times 490 - 20 \times 85}{4 \times 120 - (20)^2}$

$= \frac{1960 - 1700}{480 - 400}$

$= \frac{260}{80} = 3.25$

Regression equation Y on X:

$Y - \bar{Y} = b_{YX} (X - \bar{X})$

$Y - 21.25 = 3.25 (X - 5)$

$Y - 21.25 = 3.25X - 16.25$

$Y = 5 + 3.25X$

- (PART B)
- It means what will be the value of Y, if value of X = 10 (X = 10, Y =?) That is if the age of car(X) is 10, what is the cost (Y) for such a car.
  To find out value of Y for the given value of X, we will use regression
- equation Y on X.

  $Y = 5 + 3.25X$, as obtained in part (a)
- Substituting X = 10 in above equation, the estimated annual maintenance cost for a ten-year-old car is:
- $Y = 5 + 3.25 \times 10$
  $Y = 5 + 32.5$
  $Y = 37.5$ (in Rs. hundred)