Unit II

Lesson 3

Regression – Introduction

Objective:

The study of regression is of immense importance to analyse and interpret the pattern and behavior of the given statistical information. With an understanding of regression, students can easily get to know the movement of the information thereby facilitate the future prediction of the data behavior.

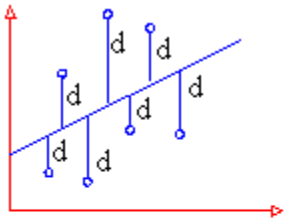
Introduction

Correlation gives us the idea of the measure of magnitude and direction between correlated variables. Now it is natural to think of a method that helps us in estimating the value of one variable when the other is known. Also correlation does not imply causation. The fact that the variables x and y are correlated does not necessarily mean that x causes y or vice versa. For example, you would find that the number of schools in a town is correlated to the number of accidents in the town. The reason for these accidents is not the school attendance; but these two increases what is known as **population**. A statistical procedure called **regression** is concerned with causation in a relationship among variables. It assesses the contribution of one or more variable called **causing variable** or **independent variable** or one which is being **caused** (**dependent variable**). When there is only one independent variable then the relationship is expressed by a straight line. This procedure is called **simple linear regression**.

Regression can be defined as a method that estimates the value of one variable when that of other variable is known, provided the variables are correlated. The dictionary meaning of regression is "to go backward." It was used for the first time by Sir Francis Galton in his research paper "Regression towards mediocrity in hereditary stature."

Lines of Regression: In scatter plot, we have seen that if the variables are highly correlated then the points (dots) lie in a narrow strip. if the strip is nearly straight, we can draw a straight line, such that all points are close to it from both sides. such a line can be taken as an ideal representation of variation. This line is called the line of best fit if it minimizes the distances of all data points from it.

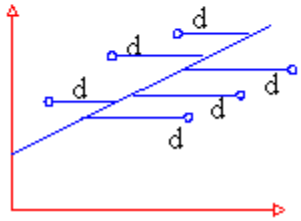
This line is called **the line of regression**. Now prediction is easy because now all we need to do is to extend the line and read the value. Thus to obtain a line of regression, we need to have a line of best fit. But statisticians don't measure the distances by dropping perpendiculars from points on to the line. They measure **deviations** (or **errors** or **residuals** as they are called) (i) vertically and (ii) horizontally. Thus we get two lines of regressions as shown in the figure (1) and (2).



(1) Line of regression of y on x

Its form is y = a + b x

It is used to estimate y when x is given



(2) Line of regression of x on y

Its form is x = a + b y

It is used to estimate x when y is given.

They are obtained by (1) graphically - by Scatter plot (ii) Mathematically - by the method of least squares.

ii. Let y = a + b y ..... (1) where a and b are given by the normal equations

$\sum y = n a + b \sum x$ ..... (2)

$\sum xy = a \sum x + b \sum x^2$ .... (3) where ' n ' be the number of pairs of values of x and y.

Equation (2) can be written as $\frac{\Sigma y}{n} = a + b\frac{\Sigma x}{n}$ but $\frac{\Sigma y}{n} = \bar{y}$ and $\frac{\Sigma x}{n} = \bar{x}$ .... means of y & x respectively.

Therefore, we have $\bar{y} = a + b\bar{x}$ ..... (4)

Thus the line passes through the point $(\bar{x}, \bar{y})$

Shifting the origin to $(\bar{x}, \bar{y})$ we get

$y - \bar{y} = a + b(x - \bar{x})$ ..... (5)

and (2) reduces to $\Sigma(y - \bar{y}) = n\,a + b\,\Sigma(x - \bar{x})$.

But $\Sigma(y - \bar{y}) = 0$ and $\Sigma(x - \bar{x}) = 0$

Therefore, $\boxed{a = 0}$

Also (2) becomes $\Sigma(x - \bar{x})(y - \bar{y}) = b\,\Sigma(x - \bar{x})^2$

$\therefore\ a = 0$

or $\Sigma xy = b\,\Sigma x^2$ where $x = x - \bar{x}$ and

$y = y - \bar{y}$ are deviations.

$\therefore\ b = \frac{\Sigma xy}{x^2} = \frac{\Sigma xy}{n\sigma x^2}$ .... since

$\sigma x = \sqrt{\frac{\Sigma x^2}{n}} \Rightarrow n\sigma x^2 = \Sigma x^2$

But $r = \frac{\Sigma xy}{n\sigma x\,\sigma y}$

Therefore, $b = r.\frac{\Sigma xy}{\sigma x}$ Also denoted by b_{yx}

Putting values of 'a' and 'b' in (5) we get

$y - \bar{y} = r\frac{\sigma y}{\sigma x}(x - \bar{x})$

i.e. $y - \bar{y} = b_{yx}(x - \bar{x})$.....(6)

Equation (6) is the equation of the **line of regression of y on x**.

Lesson 4

Coefficient of Regression

called the coefficient of regression of y on x which is obviously the slope of this line. Interchanging x and y in equation (6), the equation of the line of regression of x and y is given by

$$x - \bar{x} = r.\frac{\sigma x}{\sigma y}(y - \bar{y})$$

i.e. $x - \bar{x} = b_{xy}(y - \bar{y})$.....(7)

Naturally b_{xy} is the slope of this line which is equal to

Note:

i) The point of intersection of these regression line is the point $(\bar{x}, \bar{y})$.

ii) $b_{xy} \times b_{yx} = r\frac{\sigma y}{\sigma x} \times r\frac{\sigma x}{\sigma y} = r^2$

$\therefore r = \pm\sqrt{b_{xy} \times b_{yx}}$

iii) $b_{yx} = \dfrac{\Sigma uv - \dfrac{(\Sigma u)(\Sigma v)}{n}}{\Sigma u^2 - \dfrac{(\Sigma u)^2}{n}}$

$= \dfrac{\Sigma xy}{\Sigma x^2}$ for ungrouped data

$b_{yx} = \dfrac{\Sigma fuv - \dfrac{(\Sigma fu)(\Sigma fv)}{n}}{\Sigma fu^2 - \dfrac{(\Sigma fu)^2}{n}}$

$= \dfrac{\Sigma xy}{\Sigma x^2}$ for grouped data

$$\text{and } b_{xy} = \frac{\Sigma uv - \frac{(\Sigma u)(\Sigma v)}{n}}{\Sigma fv^2 - \frac{(\Sigma u)^2}{n}}$$

$$= \frac{\Sigma xy}{\Sigma y^2} \text{ for ungrouped data}$$

$$b_{xy} = \frac{\Sigma fuv - \frac{(\Sigma fu)(\Sigma fv)}{n}}{\Sigma fv^2 - \frac{(\Sigma fv)^2}{n}}$$

$$\frac{1}{r} \cdot \frac{\sigma y}{\sigma x} = \frac{\Sigma xy}{\Sigma y^2} \text{ for grouped data}$$

Example A panel of two judges A and B graded dramatic performance by independently awarding marks as follows:

Performance No. :	1	2	3	4	5	6	7	8
Marks by A :	36	32	34	31	32	32	35	38
Marks by B :	35	33	31	30	34	32	36	?

Solution:

Sr.No.	x - x			y - y			xy
	x_i	x	x^2	y_i	y	y^2	6
1	36	+3	9	35	+2	4	0
2	32	-1	1	33	0	0	-2
3	34	1	1	31	-2	4	6
4	31	-2	4	30	-3	9	-1
5	32	-1	1	34	1	1	1
6	32	-1	1	32	-1	1	6
7	35	2	4	36	3	9	
$n = 7$	$\Sigma x_i = 135$		$\Sigma x^2 = 21$	$\Sigma y_i = 231$		$\Sigma y^2 = 28$	$\Sigma xy = 16$

$$\therefore \bar{x} = \frac{\Sigma x}{n} = \frac{232}{7} = 33 \text{ and } \bar{y} = \frac{\Sigma y}{n} = \frac{231}{7} = 33$$

$$\text{Now } r = \frac{\Sigma xy}{\sqrt{x^2} \times \sqrt{y^2}} = \frac{16}{\sqrt{21} \times \sqrt{18}}$$

$$= \frac{16}{\sqrt{588}} = \frac{16}{24.22} = 0.65$$

$$\text{Also } \sigma x = \sqrt{\frac{\Sigma x^2}{n}} = \sqrt{\frac{21}{7}} = \sqrt{3} = 1.73$$

$$\sigma y = \sqrt{\frac{\Sigma y^2}{n}} = \sqrt{\frac{28}{7}} = 2$$

The equation of the line of regression of y on x

$$y - \bar{y} = r.\frac{\sigma y}{\sigma x}(x - \bar{x})$$

Inserting x = 38, we get

y - 33 = 0.74 (38 - 33)

y - 33 = 0.74 □ 5

y - 33 = 3.7

y = 3.7 + 33

y = 36.7 = 37 (approximately)

Therefore, the Judge B would have given 37 marks to 8th performance.

Example The two regression equations of the variables x an y are

x = 19.13 - 0.87 y and y = 11.64 - 0.50 x

Find (1) Mean of x's

(2) Mean of y's

(3) Correlation coefficient between x and y

Solution:

1. Calculation of Mean

Since lines of regression pass through means
i,e. $(\bar{x}, \bar{y})$

$\bar{x} + 0.87\,\bar{y} = 19.13$ ------- (1)

$50\,\bar{x} + \bar{y} = 11.64$

i.e. $\bar{x} + 2\,\bar{y} = 23.28$ ------- (2)

Subtracting (1) from (2) we get

$1.13\,\bar{y} = 4.15$ or $\bar{y} = 3.67$

Inserting $\bar{y} = 3.67$ in (2) we get $\bar{x} = 15.94$

\ Mean of x's = 15.94 and Mean of y's = 3.67

2. Calculation of ' r '

x = 19.93 - 0.87 y

Therefore, $b_{xy} = r \cdot \frac{\sigma x}{\sigma y} = -0.87$ -----(3)

and y = 11.64 - 0.50 x

Therefore, $b_{yx} = r \cdot \frac{\sigma y}{\sigma x} = -0.50 \quad -----(4)$

From (3) and (4)

$$r^2 = \pm \sqrt{b_{xy} \times b_{yx}} = \pm \sqrt{-0.87 \times -0.50}$$

$$= \pm \sqrt{0.435}$$

$r = \pm 0.66$

But regression coefficient are negative

r = - 0.66

Example In a partially destroyed laboratory record of an analysis of correlation data, the following results are legible:

Variance of x = 9

Regression equations : 8 x - 10 y + 66 = 6

40 x - 18 y = 214

What are (1) Means of x's and y's (2) the coefficient of correlation between x and y (3) the standard deviation of y ?

Solution:

1. Means:

   8 x - 10 y = -66 ----- (1)

   40 x - 18 y = 214 ----- (2)

   Solving (1) and (2) as

40 x - 50 y = -330 ----- (1)

40 x - 18 y = 214 ----- (2)

-32 y = -544

y = 17

□□Mean of y's $(\bar{y})$17

Substituting y = 17 in (1) we get 8x - 10□□ 17 = -66

or 8x = 104 □ x = 13

□ Mean of x's $(\bar{x})$= 13

2. Coefficient of correlation between x and y

40 x = 18 y + 214

$$x = \frac{18}{40} y + \frac{214}{40}$$

$$\therefore\ b_{xy} = \frac{18}{40} = 0.45$$

Also -10 y = - 8 x - 66

$$y = \frac{8}{10} x + \frac{66}{10}$$

Therefore, $r = \sqrt{b_{xy} \times b_{yx}} = \sqrt{.45 \times .8} = 0.6$

3. Standard deviation of y

□ Variance of x i.e. □ x^2 = 9 □ □ x = 3

Now byx = $b_{yx} = r.\frac{\sigma y}{\sigma x}$ or $0.8 = 6 \times \frac{\sigma y}{3} = 2\sigma y$

□ □ y = 0.4

Example From 10 observations of price x and supply y of a commodity the results obtained □ x = 130, □□y = 220, □□x^2 = 2288, □ xy = 3467

Compute the regression of y on x and interpret the result. Estimate the supply when the price of 16 units.

Solution: The equation of the line of regression of y on x

y = a + b x

Also from normal equations

□ y = n a + b □ x and □ xy = a □ x + b □ x^2 we get

220 = 10 a + 130 b □ (1)

3467 = 130 a + 2288 □ (2)

Solving (1) and (2) as

2860 = 130 a + 1690 b
3467 = 130 a + 2288 b

On subtraction

□ 607 = 598 b □□b = 1.002

Putting b = 1.002 in 220 = 10 a + 130 b, we get a = 8.974.

Hence the 3 equation of the line of regression of y on x is y = 8.974 + 1.002 x

When x = 16, we get

y = 8.974 + 1.002 (16)

y = 25.006

Example If a is the acute angle between the two regression lines in the case of two variables x and y show that

$$\tan\theta = \frac{1-r^2}{r} \times \frac{\sigma x \sigma y}{\sigma x^2 + \sigma y^2}$$

with usual meanings. Explain the significance when r = 0 and r = □ 1.

Solution: The slopes of the two regression lines are

$$m_1 = b_{yx} = r.\frac{\sigma y}{\sigma x} \text{ and } b_{xy} = r.\frac{\sigma x}{\sigma y} \text{ or } \frac{1}{r}\cdot\frac{\sigma y}{\sigma x} = m_2$$

$$\therefore \tan\theta = \left|\frac{m_1 - m_2}{1 + m_1.m_2}\right|$$

$$\therefore \tan\theta = \left|\frac{r\frac{\sigma y}{\sigma x} - \frac{1}{r}\frac{\sigma y}{\sigma x}}{1 + \frac{\sigma y}{\sigma x}\cdot\frac{1}{r}\frac{\sigma y}{\sigma x}}\right|$$

$$\therefore \tan\theta = \left| \frac{\frac{\sigma y}{\sigma x}\left(r - \frac{1}{r}\right)}{1 + \left(\frac{\sigma y}{\sigma x}\right)^2} \right| = \left| \frac{\frac{\left(r^2 - 1\right)}{r} \cdot \frac{\sigma y}{\sigma x}}{\frac{\sigma x^2 + \sigma y^2}{\sigma x^2}} \right|$$

$$\therefore \tan\theta = \frac{\left(1 - r^2\right)}{r} \times \frac{\sigma x \sigma y}{\sigma x^2 + \sigma y^2}$$

Lesson 5

Assumptions & Non-Linear Regression

Classical assumptions for regression analysis include:

- The sample must be representative of the population for the inference prediction.
- The error is assumed to be a random variable with a mean of zero conditional on the explanatory variables.
- The variables are error-free. If this is not so, modeling may be done using errors-in-variables model techniques.
- The predictors must be linearly independent, i.e. it must not be possible to express any predictor as a linear combination of the others. See Multicollinearity.
- The errors are uncorrelated, that is, the variance-covariance matrix of the errors is diagonal and each non-zero element is the variance of the error.
- The variance of the error is constant across observations (homoscedasticity). If not, weighted least squares or other methods might be used.

These are sufficient (but not all necessary) conditions for the least-squares estimator to possess desirable properties, in particular, these assumptions imply that the parameter estimates will be unbiased, consistent, and efficient in the class of linear unbiased estimators. Many of these assumptions may be relaxed in more advanced treatments.

Assumptions include the geometrical support of the variables (Cressie, 1996). Independent and dependent variables often refer to values measured at point locations. There may be spatial trends and spatial autocorrelation in the variables that violates statistical assumptions of regression. Geographic weighted regression is one technique to deal with such data (Fotheringham et al., 2002). Also, variables may include values aggregated by areas. With aggregated data the Modifiable Areal Unit Problem can cause extreme variation in regression parameters (Fotheringham and Wong, 1991). When analyzing data aggregated by political boundaries, postal codes or census areas results may be very different with a different choice of units.

Non-Linear Regression – Introduction

Nonlinear regression is a form of regression analysis in which observational data are modeled by a function which is a nonlinear combination of the model parameters and depends on one or more independent variables. The data are fitted by a method of successive approximations

The data consist of error-free independent variables (explanatory variable), *x*, and their associated observed dependent variables (response variable), *y*. Each *y* is modeled as a random variable with a mean given by a nonlinear function *f*(*x*,β). Systematic error may be present but its treatment is outside the scope of regression analysis. If the independent variables are not error-free, this is an errors-in-variables model, also outside this scope.

Other examples of nonlinear functions include exponential functions, logarithmic functions, trigonometric functions, power functions, Gaussian function, and Lorentzian curves. Some functions, such as the exponential or logarithmic functions, can be transformed so that they are linear. When so transformed, standard linear regression can be performed but must be applied with caution. See Linearization, below, for more details.

In general, there is no closed-form expression for the best-fitting parameters, as there is in linear regression. Usually numerical optimization algorithms are applied to determine the best-fitting parameters. Again in contrast to linear regression, there may be many local minima of the function to be optimized and even the global minimum may produce a biased estimate. In practice, estimated values of the parameters are used, in conjunction with the optimization algorithm, to attempt to find the global minimum of a sum of squares.

References:

Kendall and Stuart, "The Advanced Theory of Statistics"