CHAPTER FIVE

REGRESSION ANALYSIS

Mr. Mohammad Manjur Alam (Manju)
Associate Professor
Department of Computer Science and Engineering
International Islamic University Chittagong.
Mobile: 01715966144hyst
FB and Email: manjuralam44@yahoo.com

Regression is the functional relationship between two variables and of the two variables one may represent cause and the other may represent effect. The variable representing cause is known as independent variable and is denoted by X. The variable X is also known as predictor variable or repressor. The variable representing effect is known as dependent variable and is denoted by Y. Y is also known as predicted variable. The term "regression" was used by a famous Biometrician Sir. F. Galton (1822-1911) in 1877.

Example: The productions of paddy of amount y is dependent on rainfall of amount x. Here x is independent variable and y is dependent variable.

Assumptions

1. The x's are non-random or fixed constants

2. At each fixed value of X the corresponding values of Y have a normal distribution about a mean.

3. For any given x, the variance of Y is same.

4. The values of y observed at different levels of x are completely independent.

Objectives of Regression Analysis:

i. To estimate the relationship that exits, on the average, between the dependent variable and independent variables.

ii. To determine the effect of each independent variable on the dependent variable, controlling the effects of the others independent variables.

iii. To predict the value of the dependent variable for a given value of the explanatory variables

Simple Linear Regression Model

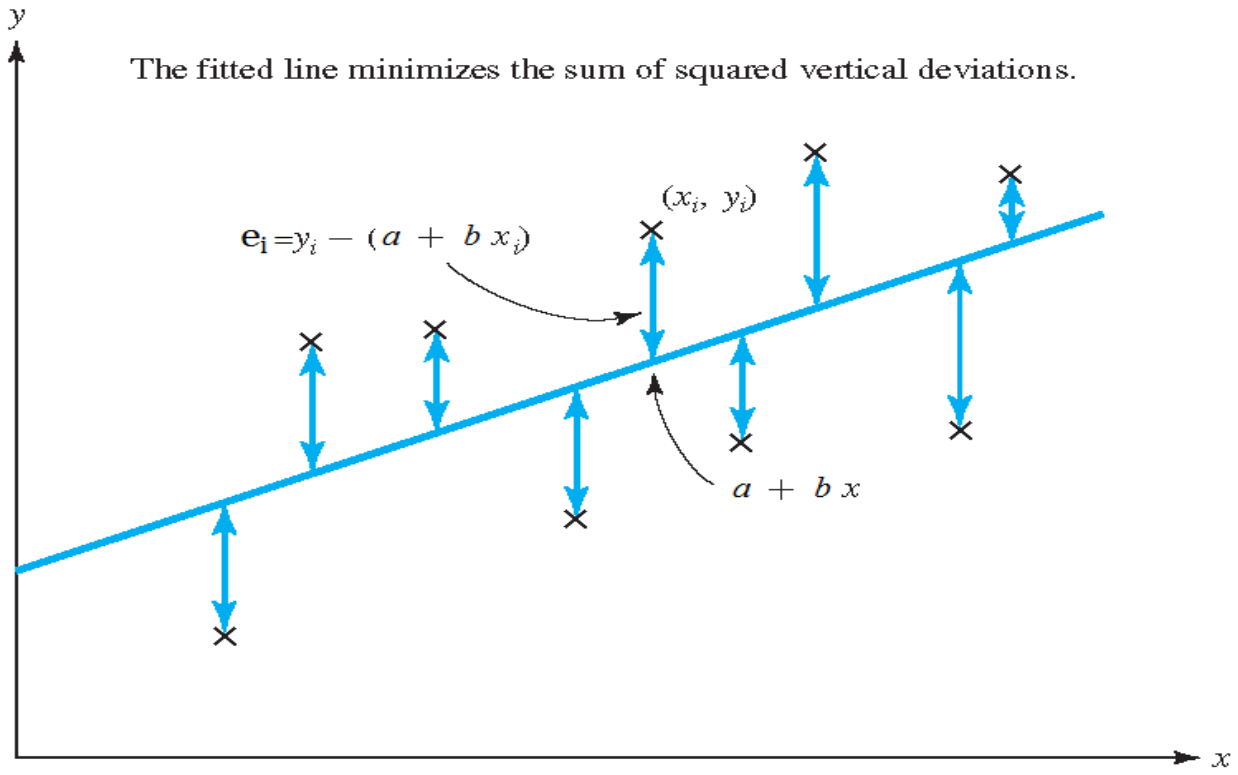
The simplest form of the regression model that displays the relation between X and Y is a straight line, which appears as follows: $\mathbf{Y}=a + bx$

Where $\hat{Y}$ denotes the predicted value of Y, a is the intercept and b is the slope of the straight line. In regression terminology, b is the regression coefficient of Y on X. This straight line is called the fitted line of Y.

The least-Squares Method:

The least-squares method is a technique for minimizing the sum of the squares of the differences between the observed values and estimated values of the dependent variable. That is the least-squares line is the line that minimizes $\Sigma \mathbf{e_i^2} = \Sigma(\mathbf{Y_i} - \mathbf{bX_i} - \mathbf{a})^2$

Here Σe_i^2 is called sum of squares of errors (SSE).



To minimizes SSE with respect to a and b, from calculus we know that the partial derivatives of SSE with respect to a and b must be 0. Then $\frac{\delta}{\delta a}\sum e_i^{\,2}$ $=-2\Sigma\,(Y_i-bX_i-a)=0$

$\frac{\delta}{\delta b}\sum e_i^{\,2}$ $=-2\Sigma\,(Y_i-bX_i-a)\,Xi=0$

Which concludes $\Sigma Y_i=n\,a+b\,\Sigma\,X_i$

and $\Sigma X_i Y_i=a\Sigma X_i+b\Sigma X_i^2$

$$b_{yx}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$ and $\mathbf{a}=\bar{\mathbf{Y}}-\mathbf{b}\bar{\mathbf{X}}$

Regression coefficient

The mathematical measures of regression are called the coefficient of regression.

Let, (x_1,y_1), (x_2,y_2)…….. (x_n,y_n) be the pairs of n observations. Then the regression coefficient of y on x is denoted by byx and defined by

$$b_{yx}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

Again, the regression coefficient of x on y is denoted by bxy and defined by

$$b_{xy}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$$

Regression lines:

If we consider two variables X and Y, we shall have two regression lines as the regression line of Y on X and the regression line of X on Y. The regression line of Y on X gives the most probable values of Y for given values of X and The regression line of X on Y gives the most probable

values of X for given values of Y. Thus we have two regression lines. However, when there is either perfect positive or perfect negative correlation between the two variables, the two regression lines will coincide i.e, we will have one line.

Regression equation:

The regression equation of y on x is expressed as follows:

y = a + bx, where y is the dependent variable to be estimated and x is the independent variable, a is the intercept term (assume mean) and b is the slope of the line.

Here, $a = \bar{y} - b\bar{x} = \frac{\sum y}{n} - b\frac{\sum x}{n}$ and $b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$

$$= \frac{\sum x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sum_{i=1}^{n} x_i^{\,2} - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}$$

Similarly, the regression equation of x on y is expressed as follows:

x = a + by, where x is the dependent variable to be estimated and y is the independent variable, a is the intercept term (assume mean) and b is the slope of the line.

$$a = \bar{x} - b\bar{y} = \frac{\sum x}{n} - b\frac{\sum y}{n}$$

And $b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\sum x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sum_{i=1}^{n} y_i^{\,2} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}}$

Properties of regression coefficient.

1. Regression coefficient is independent of change of origin but not of scale.

2. Regression coefficient lies between - $\propto$ to + $\propto$. i.e, -< $\propto$ byx < $\propto$.

3. Regression coefficient is not symmetric. i.e, $b_{xy} = b_{yx}$

4. The geometric mean of regression coefficients is equal to correlation coefficient

i.e, $r_{xy} = \sqrt{b_{yx} \times b_{xy}}$

5. The arithmetic mean of two regression coefficient is greater than correlation

Coefficient. i.e, $\left(\frac{b_{yx} + b_{xy}}{2}\right) \geq r_{xy}$

6. If one of regression coefficient is greater than unity the other must be less than

unity. i.e, $b_{xy} \geq 1$ and $b_{yx} < 1$

7. Regression coefficient is not pure number.

Coefficient of Determination R^2 :

- The coefficient of determination, r^2, is useful because it gives the proportion of the variance (fluctuation) of one variable that is predictable from the other variable. It is a measure that allows us to determine how certain one can be in making predictions from a certain model/graph. The coefficient of determination is the ratio of the explained variation to the total variation.
- The coefficient of determination is such that $0 < r^2 < 1$, and denotes the strength of the linear association between x and y.
- The coefficient of determination represents the percent of the data that is the closest to the line of best fit. For example, if r = 0.922, then $r^2 = 0.850$, which means that 85% of the total variation in y can be explained by the linear relationship between x and y (as described by the regression equation). The other 15% of the total variation in y remains unexplained.
- The coefficient of determination is a measure of how well the regression line represents the data. If the regression line passes exactly through every point on the scatter plot, it would be able

to explain all of the variation. The further the line is away from the points, the less it is able to explain.

Theorem: Show that correlation coefficient is the geometric mean of regression coefficients. i.e, rxy= $\sqrt{b_{yx} \times b_{xy}}$

Proof: Let, Let, (x_1,y_1), (x_2,y_2)…….. (x_n,y_n) be the pairs of n observations. Then the correlation coefficient between x and y is denoted by rxy and defined as,

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad \text{……..(1)}$$

$$b_{yx} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Again, the regression coefficient of y on x is,

$$b_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Again, the regression coefficient of x on y is,

$$b_{yx} \times b_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \times \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$\sqrt{b_{yx} \times b_{xy}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} = r_{xy} \text{ (proved)}$$

Theorem: The arithmetic mean of two regression coefficient is greater than correlation coefficient.

i.e, $\left(\frac{b_{yx} + b_{xy}}{2}\right) \geq r_{xy}$

Proof: Let, (x_1,y_1), (x_2,y_2)…….. (x_n,y_n) be the pairs of n observations. Then the regression coefficient of y on x is denoted by byx and the regression coefficient of x on y is denoted by bxy.

The arithmetic mean of byx and bxy is $\left(\frac{b_{yx}+b_{xy}}{2}\right)$ and the geometric mean is $\sqrt{b_{yx}\times b_{xy}}$

We know, Correlation coefficient is the geometric mean of regression coefficients.

i.e, $r_{xy}=\sqrt{b_{yx}\times b_{xy}}$

Since, A.M $\geq$ G.M

or, $\left(\frac{b_{yx}+b_{xy}}{2}\right)\geq\sqrt{b_{yx}\times b_{xy}}$

or, $\left(\frac{b_{yx}+b_{xy}}{2}\right)\geq$ r (proved

Uses of regression.

(i) Whether a relationship exists or not.

(ii) To find the strength of relationship.

(iii) Determination of mathematical equation.

(iv) Prediction the values of the dependent variables.

Distinguish between correlation coefficient and regression coefficient.

Correlation coefficient	Regression coefficient.
1. The numerical value by which we measure the strength of linear relationship between two or more variables is called correlation coefficient.	1. The mathematical measures of regression are called the coefficient of regression.
2. Correlation coefficient is independent of change of origin and scale of measurement.	2. Regression coefficient is independent of change of origin but not of scale.
3. Correlation coefficient lies between -1 to +1. i.e, -1< rxy < 1.	3. Regression coefficient lies between $-\infty$ to $+\infty$. i.e, $-<\infty$ byx < ∞.
4. Correlation coefficient is symmetric. i.e, rxy= ryx	4. Regression coefficient is not symmetric. i.e, bxy $\neq$ byx
5. It is always unit free.	5. Regression coefficient is not pure number.

6. When r=0 then the variables are correlated.	6. When r=0 then two lines of regression are perpendicular to each other.

Application problem-1: A researcher wants to find out if there is any relationship between the ages of husbands and the ages of wives. In other words, do old husbands have old wives and young husbands have young wives? He took a random sample of 7 couples whose respective ages are given below:

Age of Husband(in years):x	39	25	29	35	32	27	37
Age of wife(in years):y	37	18	20	25	25	20	30

i. Compute the regression line of y on x.
ii. Predict the age of wife whose husband's age in 45 years.
iii. Find the regression line of x on y and estimate the age of husband if the age of his wife is 28 years.
iv. Compute the value of correlation coefficient with the help of regression coefficients.

Solution: The equation of the best –fitted regression line of y on x is $\hat{y} = a + bx$

Where, $b = \dfrac{\sum x_i y_i - \dfrac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sum_{i=1}^{n} x_i^{\,2} - \dfrac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}$ and $a = \bar{y} - b\bar{x}$

Computation table

x	y	x2	y2	xy
39	37	1521	1369	1443
25	18	625	324	450
29	20	841	400	580
35	25	1225	625	875
32	25	1024	625	800
27	20	729	400	540
37	30	1369	900	1110
$\sum x = 224$	$\sum y = 175$	$\sum x^2 = 7334$	$\sum y^2 = 4643$	$\sum xy = 5798$

Here, $b = \dfrac{\sum x_i y_i - \dfrac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sum_{i=1}^{n} x_i^{\,2} - \dfrac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} = \dfrac{5798 - \dfrac{(224)(175)}{7}}{7334 - \dfrac{(224)^2}{7}} = 1.193$

And $a = \bar{y} - b\bar{x} = \dfrac{\sum y}{n} - b\dfrac{\sum x}{n}$

$= \dfrac{175}{7} - (1.193)\dfrac{(224)}{7} = 25 - 38.176 = -13.176$

Hence the fitted regression line is $\hat{y} = a + bx = -13.176 + 1.193x$

Hence, if the age of husband is 45, the probable age of wife would be

$\hat{y} = -13.176 + 1.193x = -13.176 + 1.193 \times 45 = 40.51$ years.

The equation of the best –fitted regression line of y on x is $\hat{x} = a + by$

Where, $b = \dfrac{\sum x_i y_i - \dfrac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sum_{i=1}^{n} y_i^{\,2} - \dfrac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}}$

$= \dfrac{5798 - \dfrac{(224)(175)}{7}}{4643 - \dfrac{(175)^2}{7}} = 0.739$

And $a = \bar{x} - b\bar{y} = \dfrac{\sum x}{n} - b\dfrac{\sum y}{n}$

$= \dfrac{224}{7} - 0.739\dfrac{175}{7} = 13.525$

Hence the fitted regression line is $\hat{x}$ = a + by = 13.525+ 0.739y

Hence, if the age of wife is 28 years, the estimate age of husband is

$\hat{x}$ = a + by

= 13.525+ (0.739)(28) = 34.22 years.

Application problem-2: A research physician recorded the pulse rates and the temperatures of water submerging the faces of ten small children in cold water to control the abnormally rapid heartbeats. The results are presented in the following table. Calculate the correlation coefficient and regression coefficients between temperature of water and reduction in pulse rate.

Temperature of water	68	65	70	62	60	55	58	65	69	63
Reduction in pulse rate.	2	5	1	10	9	13	10	3	4	6

Also show that (i) $\left(\frac{b_{yx}+b_{xy}}{2}\right) \geq r_{xy}$

Solution: Calculating table of correlation coefficient and regression coefficients.

xi	yi	xi2	yi2	xiyi
68	2	4624	4	136
65	5	4225	25	325
70	1	4900	1	70
62	10	3844	100	620
60	9	3600	81	540
55	13	3025	169	715
58	10	3364	100	580
65	3	4225	9	195
69	4	4761	16	276
63	6	3969	36	378
$\sum x_i = 635$	$\sum y_i = 63$	$\sum x_i^2 = 40537$	$\sum y_i^2 = 541$	$\sum x_i y_i = 3835$

We know, $r_{xy} = \dfrac{\sum x_i y_i - \dfrac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sqrt{\left\{\sum_{i=1}^{n} x_i^{\,2} - \dfrac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right\}\left\{\sum_{i=1}^{n} y_i^{\,2} - \dfrac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right\}}}$

$= \dfrac{3835 - \dfrac{635 \times 63}{10}}{\sqrt{\left\{40537 - \dfrac{(635)^2}{10}\right\}\left\{541 - \dfrac{(63)^2}{10}\right\}}} = -0.94$

We know, the regression coefficient of y on x is, $b_{yx} = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$

$= \dfrac{\sum x_i y_i - \dfrac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sum_{i=1}^{n} x_i^{\,2} - \dfrac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} = \dfrac{3835 - \dfrac{635 \times 63}{10}}{40537 - \dfrac{(635)^2}{10}} = \dfrac{-1655}{2145} = -0.77$

Regression coefficient of x on y is, $b_{xy} = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$

$= \dfrac{\sum x_i y_i - \dfrac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sum_{i=1}^{n} y_i^{\,2} - \dfrac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}} = \dfrac{3835 - \dfrac{635 \times 63}{10}}{541 - \dfrac{(63)^2}{10}} = \dfrac{-1655}{1441} = -1.1$

(i) $\left(\frac{b_{yx}+b_{xy}}{2}\right) \geq r_{xy}$

Here, $\left(\frac{b_{yx}+b_{xy}}{2}\right) = \frac{(-0.77)+(-1.1)}{2} = -0.94 = r_{xy}$

Assignment Problem-1: The following data give the test scores and sales made by nine salesmen during the last year of a big departmental store:

Test Scores: y	14	19	24	21	26	22	15	20	19
Sales(in lakh Taka)	31	36	48	37	50	45	33	41	39

i. Find the regression equation of test scores on sales. Ans: $\hat{y}$ = -2.4 + 0.56x
ii. Find the test scores when the sale is Tk. 40 lakh. Ans: 20 lakh
iii. Find the regression equation of sales on test scores. Ans: $\hat{x}$ = 7.8 + 1.61y
iv. Predict the value of sale if the test score is 30. Ans: 56.1 lakh
v. Compute the value of correlation coefficient with the help of regression coefficients.

Assignment Problem-2: The following table gives the ages and blood pressure of 10 women:

Age in years x	56	42	36	47	49	42	72	63	55	60
Blood pressure y	147	125	118	128	125	140	155	160	149	150

Obtain the regression line of y on x. Ans: $\hat{y}$ = 83.76+ 1.11x

Estimate the blood pressure of a women whose age is 50 years. Ans: 139.26

Obtain the regression line of x on y.

Find correlation coefficient between x and y and comment.

Assignment Problem-3: Consider the following data set on two variables x and y:

x : 1 2 3 4 5 6

y : 6 4 3 5 4 2

Find the equation of the regression line y on x. Ans: $\hat{y}$ = 5.799-0.541x

Graph the line on a scatter diagram.

Estimate the value of y when x = 4.5 Ans: $\hat{y}$ = 3.486

Predict the value of y when x = 8. Ans: $\hat{y}$ =1.687

Assignment Problem-4: Cost accountants often estimate overhead based on production. At the standard knitting company, they have collected information on overhead expenses and units produced at different plants and what to estimate a regression equation to predict future overhead.

Units	56	40	48	30	41	42	55	35
Overhead	282	173	233	116	191	171	274	152

(i)Draw a scatter diagram and comment

(ii)Fit a regression equation.

(iii)Estimate overhead when 65 units are produced.

Assignment Problem-5: The following data refer to information about annual sales

(Tk.'000) and year of experience of a super store of 8 salesmen:

Salesmen	1	2	3	4	5	6	7	8
Annual sales (Tk.'000)	90	75	78	86	95	110	130	145
Year of experience	7	4	5	6	11	12	13	17

(i)Fit two regression lines.

(ii)Estimate sales for year of experience is 10

(iii)Estimate year of experience for sales 100000