Regression Analysis

A population consist of 28 families. We are interested to predict the average height of sons knowing the heights of their father.

Let x be the height of father & y be the height of son

x	y	x	y	x	y	x	y
60	55	70	68	70	69	70	71
65	60	60	58	75	72	75	74
70	65	65	63	60	65	70	72
75	65	70	68	65	65	75	75
60	56	75	70	70	70	75	76
65	62	60	61	75	73	75	77
70	67	65	64	65	66	75	78

When we organize the son's height by their father's height we obtain a summary table

Father's height	Corresponding son's height	Total	mean
60	55, 56, 58, 61, 65	295	59
65	58, 62, 63, 64, 65, 66	378	63
70	64, 63, 67, 68, 69, 71, 72	476	68
75	66, 69, 70, 72, 73, 74, 75, 76, 77, 78	730	73

For a given x, there is a frequency distribution which is known as conditional distribution. The mean of this

From the table it is seen that when father's height is 60 the mean of son's height is 59, this is a conditional mean. It can be denoted by $\mu_{y/60} = 59$. In general conditional mean can be written as

$$\mu_{y|x}.$$

It is clear from the above discussion that $\mu_{y|x}$ is a function of x which is known as regression function.

i.e. $\mu_{y|x} = f(x)$

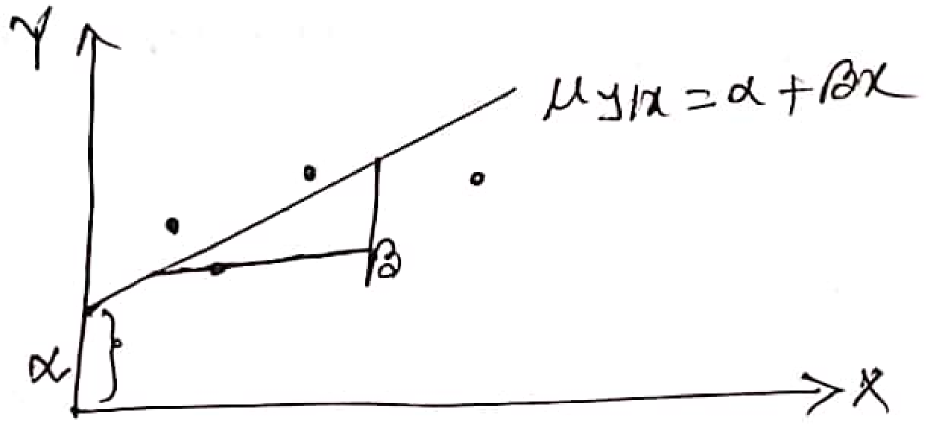
If we assume that the above function is a linear function it can be written as

$$\mu_{y|x} = \alpha + \beta x \quad \text{———} \quad (1)$$

where α and β are unknown constants of the regression function, mathematically they are the intercept and slope respectively

In regression analysis slope β is called the regression coefficient.

Equation (1) supposed that for the given values of α and β, the mean values of $\mu_{y|x}$ when plotted, an exact straight line will be found.



But in practice, this may not be the situation always. The observed values will tend to deviate from the $\mu_{y|x}$ values and then equation (1) is subject to some random error ε. Let the

---

Note: we may find from 100 values that $\mu_{y/60} = 59$ but when we take 1000 values it may be greater than or less than 59.

resulting value is y

i.e. $y = \mu_{y|x} + \epsilon_0$

where ϵ_0 is the random error and $-\alpha < \epsilon_0 < \alpha$.

$$\therefore \quad y = \alpha + \beta x + \epsilon_0 \quad \text{———} \quad (1)$$

Equation (I) is a mathematical model and equation (II) is a statistical model.

Q. Is father's height dependent on son's height or son's height dependent on father's height?

Dependent variable and Independent variable

If two variables are involved, the variable that is the basis of estimation is called independent variable and the variable whose value is to be estimated is called the dependent variable.

Regression Analysis:- Regression analysis is a statistical technique that serves as a basis for studying the dependence of one variable on one or more other variables.

Interpretation of parameters α & β:-

α : the parameter α is the average value of y for $x=0$. It is the y intercept

β: The parameter β represents the amount of change, on an average, in y for the one unit change in x.

β is also known as the regression coefficient.

Estimation of α and β

<u>Method of Least square :-</u>

consider a set of n pairs of values $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$

The population regression line

$$y_i = \alpha + \beta x_i + \varepsilon_i.$$

and estimated regression line

$$\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$$

$y_i - \hat{y}_i$ is called error term or residual

i.e. $e_i = y_i - \hat{y}_i$

$$= y_i - \hat{\alpha} - \hat{\beta} x_i$$

we will obtain $\hat{\alpha}$ and $\hat{\beta}$ by minimizing

$$\sum_{i=1}^{n} \varepsilon_i^2$$

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta} x_i)^2 \quad \text{—— (1)}$$

Differentiating (1) w.r.t $\hat{\alpha}$ and $\hat{\beta}$ and set them to zero

$$\frac{\delta}{\delta \hat{\alpha}}\left(\sum \varepsilon_i^2\right) = -2\sum\left(y_i - \hat{\alpha} - \hat{\beta}x_i\right) = 0$$

$$\Rightarrow \sum y_i - \sum \hat{\alpha} - \hat{\beta}\sum x_i = 0.$$

$$\Rightarrow \sum y_i = n\hat{\alpha} + \hat{\beta}\sum x_i \quad \text{—— (II)}$$

$$\frac{\delta}{\delta \hat{\beta}}\left(\sum \varepsilon_i^2\right) = -2\sum x_i\left(y_i - \hat{\alpha} - \hat{\beta}x_i\right) = 0.$$

$$\Rightarrow \sum x_i y_i - \hat{\alpha}\sum x_i - \hat{\beta}\sum x_i^2 = 0.$$

$$\Rightarrow \sum x_i y_i = \hat{\alpha}\sum x_i + \hat{\beta}\sum x_i^2 \quad \text{—— (III)}$$

multiplying (1) by $\sum x_i$ and (II) by n and then taking deduction we get

$$\sum x_i \sum y_i = n\hat{\alpha}\sum x_i + \hat{\beta}\left(\sum x_i\right)\left(\sum x_i\right)$$

$$\underset{-}{n\sum x_i y_i} \underset{-}{=} n\hat{\alpha}\sum x_i \underset{-}{+} n\hat{\beta}\sum x_i^2$$

$$\sum x_i \sum y_i - n\sum x_i y_i = \hat{\beta}\left(\sum x_i\right)^2 - n\hat{\beta}\sum x_i^2$$

$$\Rightarrow \boxed{\hat{\beta} = \frac{\sum x_i y_i - \dfrac{\left(\sum x_i\right)\left(\sum y_i\right)}{n}}{\sum x_i^2 - \dfrac{\left(\sum x_i\right)^2}{n}}}$$

we had $\hat{y}_i = \hat{\alpha} + \hat{\beta} x$.

$$\Rightarrow \sum \hat{y}_i = n\hat{\alpha} + \hat{\beta} \sum x_i.$$

$$\Rightarrow \frac{\sum \hat{y}_i}{n} = \hat{\alpha} + \hat{\beta} \frac{\sum x_i}{n}.$$

$$\Rightarrow \bar{y} = \hat{\alpha} + \hat{\beta}\bar{x}.$$

$$\Rightarrow \boxed{\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}}$$

So, $$\hat{\beta} = \frac{\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}$$

and $$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

Exam question may be as follows

*. Find least square estimate of α and β.

*. Importance of regression will be discussed in class

H.W Show that regression coefficient is independent on both origin and scale of measurement.

Example:- A departmental store has the following statistics on sales (y) for a period of last one year of 10 salesmen, who have varying years of sales experience (x).

(i) Find regression line of y on x.

(ii) Predict the annual sales volume of persons who have 12 and 15 years of sales experience.

Sales person i	years of experience (x)	Annual sales in (1000) taka
1	1	80
2	3	97
3	4	92
4	4	102
5	6	103
6	8	111
7	10	119
8	10	123
9	11	117
10	13	136

Solution :- we have to find the regression line $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$

For this purpose we have to find $\hat{\alpha}$ and $\hat{\beta}$ first.

The required computations are shown in the accompanying table

Sales person i	x_i	y_i	x_i^2	$x_i y_i$
1	1	80	1	80
2	3	97	9	291
3	4	92	16	368
4	4	102	16	408
5	6	103	36	618
6	8	111	64	888
7	10	119	100	1190
8	10	123	100	1230
9	11	117	121	1287
10	13	136	169	1768
	$\sum x_i = 70$	$\sum y_i = 1080$	$\sum x_i^2 = 632$	$\sum x_i y_i = 8128$

$$\bar{x} = \frac{70}{10} = 7 \qquad \bar{y} = \frac{1080}{10} = 108$$

$$\hat{\beta} = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}$$

$$= \frac{8128 - \frac{70 \times 1080}{10}}{632 - \frac{(70)^2}{10}} = 4$$

$\therefore \hat{\beta} = 4$

$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 108 - 4 \times 7 = 80$

i)

So, the regression line is

$\hat{y}_i = 80 + 4x_i$

ii)

we will now use the values of $\hat{\alpha}$ and $\hat{\beta}$ to estimate the sales for $x = 12$ and $x = 15$ years of experience.

Estimated sales for $x = 12$ is

$\hat{y}(12) = 80 + 4 \times 12 = 128$ (Thousand taka)

$\hat{y}(15) = 80 + 4 \times 15 = 140$ (Thousand taka).