Regression Analysis

(by Francis Galton,1886, Karl Pearson, 1903)

- Regression describes the relationship between two variables based on observed data
- Regression predicts the value of the value of the response variable (or outcome/dependent variable) from the independent variable (or predictor).
- Suppose y is a dependent variable (outcome/response variable)
- x is an independent variable (predictor/regressor)
- The two variables will be closely related.
- Regression analysis is the study of dependency
- Suppose age (x) and weight (y) are correlated, then we may be interested to find out the estimated value of weight (y- dependent variable) for a given value of age (x- independent variable).

Regression model

- The linear regression model is the single most useful model for prediction.
- The basic linear regression model of y on x can be written as

$$y = \alpha + \beta x + \varepsilon$$

Where,

- y is known as dependent variable or explained variable as it is dependent on x
- x is called independent or explanatory variable

$$y = \alpha + \beta x + \varepsilon$$

- α is intercept, value (or average value) of y when x is absent (zero)
- β is the slope coefficient, measures the average change (increase/decrease) in y for a unit change (increase/decrease) in x.
- the coefficients α and β are unknown parameters, known as regression coefficients
- ε is known as random error term. The disturbance term ε represents all those factors that affect the dependent variable but are not taken into account.

The parameters α and β are calculated by ordinary least squares (OLS) method (by Friedrich Gauss, German mathematician) as.

$$\widehat{\beta} = \frac{\sum xy - n\overline{x}\,\overline{y}}{\sum x^2 - n(\overline{x})^2}$$

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

The estimated/fitted regression equation is

$$\hat{y} = \hat{\alpha} + \hat{\beta}x$$

Example 1: A study is designed to check the relationship between smoking and longevity. A sample of 11 men of 50 years and older was taken and the average number of cigarettes smoked per day and the age at death was recorded, as summarized in the table.

Cigarettes	5	23	25	48	17	8	26	35	4	23	11
Longevity	80	78	60	53	85	84	79	72	92	65	81

i. Can you establish the relationship between number of cigarettes and longevity ?

ii. Fit a regression line of longevity on number of cigarettes

iii. What will be longevity if the number of cigarettes smoked per day is 5?

iv. Find coefficient of determination and find how much variation of longevity is explained by the average number of cigarettes smoked per day.

v. What is your interpretation for the model?

i. Pearson's correlation coefficient will be used to find the relationship as follows:

Cigarettes x	Longevity y	xy	x2	y2
5	80	400	25	6400
23	78	1794	529	6084
25	60	1500	625	3600
48	53	2544	2304	2809
17	85	1445	289	7225
8	84	672	64	7056
26	79	2054	676	6241
35	72	2520	1225	5184
4	92	368	16	8464
23	65	1495	529	4225
11	81	891	121	6561
∑x=225	∑y=829	∑xy =15683	∑x2 =6403	∑y2= 63849

$\bar{x} = \frac{\sum x}{n}$ =225/11 =20.45, and

$\bar{y} = \frac{\sum y}{n}$ =829/11 = 75.36

Then the Pearson's correlation coefficient, r is

$$r = \frac{\sum xy - n\bar{x}\bar{y}}{\sqrt{(\sum x^2 - n\bar{x}^2)(\sum y^2 - n\bar{y}^2)}}$$

$$r = \frac{15683 - 11 \times (20.45) \times (75.36)}{\sqrt{[6403 - 11 \times (20.45)^2] \times [63849 - 11 \times (75.36)^2]}}$$

$$= -0.805$$

Interpretation: There is negative and very strong relationship between number of cigarettes smoked per day and the longevity. That is when number of cigarettes smoked per day increase, the longevity decreases

ii. It is required to fit a regression model where

Dependent variable y=longevity and

Independent variable, x= Cigarettes

By the method of ordinary least square (OLS)

$$\hat{\beta} = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2} = \frac{15683 - 11 \times (20.45)(75.36)}{6403 - 11 \times (20.45)^2} = -0.704$$

and

$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$ = 75.36 – (-0.704)(20.45) = 89.76

Thus the fitted regression model is

$\hat{y} = \hat{\alpha} + \hat{\beta}x$ = 89.76 + (-0.704).(x)

Interpretation of the model

$\hat{\alpha}$=89.76 means (y) the longevity will be 89.76 years, when the number of cigarettes smoked per day (x) is 0

It implies that who did not smoke cigarette at all, on an will live up to 89.76 years

$\hat{\beta}$ =-0.704,

As it is negative, the longevity (y) will decrease

So $\hat{\beta}$ =-0.704means (y) the longevity will decreased 0.704 years, when the number of cigarettes smoked per day (x) is increased by 1

iv.

if the number of cigarettes smoked per day is 5 , that is x=5, then the longevity (y)is

$$\hat{y} = 89.76 + (-0.704)x = 89.76 + (-0.704) \times 5 = 86.27\ years$$

Similarly you can check if the number of cigarettes smoked per day is 10 , that is x=10, then the longevity (y) is

$$\hat{y} = 89.76 + (-0.704)x = 89.76 + (-0.704) \times 10 = 82.72\ years$$

v. The coefficient of determination is calculated as

$R^2 = r^2$

where r is correlation coefficient

The coefficient of determination R^2 is on of the important tools to verify the strength or fitness of the model

We got that r= -0.805, then $R^2 = 0.65$.

We can say that 65% of the variation in the longevity that is age at death is explained by taking into account the average number of cigarettes smoked per day.

Or, we can say that 65% of the variation in age at death is explained by the average number of cigarettes smoked per day

Note: It has to be mentioned that the independent or explained variable x can be more than one but the interpretation abbot model will be almost the same

- Example.2

Father's Height (inches)	Son's Height (inches)
58	60
60	62
62	63
64	64
65	64
66	65
67	66
70	67
73	69
75	70

i. Establish the relationship between Father's Height and Son's Height ?

ii. Fit a regression line of Son's Height on Father's Height

iii. What will be Son's Height if the Father's Height is 74 inches?

iv. Find coefficient of determination and find how much variation of Son's Height is explained by the Father's Height .

v. What is your interpretation for the model?