Regression Analysis

❖ Regression Analysis

Regression analysis is a set of statistical processes for estimating the relationships between a dependent variable (often called the 'outcome variable') and one or more independent variables (often called 'predictors', 'covariates', or 'features').

In other words, Regression Analysis is a statistical tool used to determine the probable change in one variable for the given amount of change in another. This means, the value of the unknown variable can be estimated from the known value of another variable.

Example: Suppose that a researcher is investigating the factors that determine the rate of inflation. If the researcher believes that rate of inflation depends on the growth rate of money supply, he may estimate a regression model using the rate of inflation as dependent variable and the growth rate of money supply as independent variable

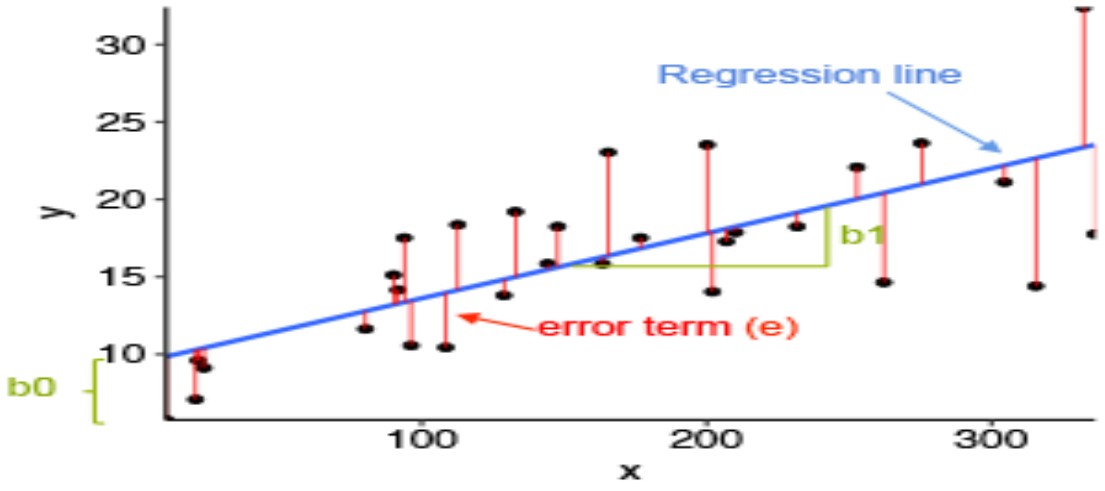
❖ Regression Line

The **Regression Line** is the line that best fits the data, such that the overall distance from the line to the points (variable values) plotted on a graph is the smallest.

The regression line of y on x is

$$\hat{y} = b_0 + b_1 x$$

Where, y is dependent variable, x is independent variable, b_0 is intercept term and b_1 is slope or coefficient of x.



❖ **Types of Regression**

❖ **Simple Linear Regression**

Simple liner regression is a model that assesses the relationship between a dependent variable and an independent variable. The simple linear regression model is

$$Y = \alpha + \beta X + \epsilon$$

Where, Y is dependent variable, X is independent variable, α is intercept term and β is slope or coefficient of X and ϵ is random error term.

❖ **Multiple Regression**

Multiple liner regression is a model that assesses the relationship between a dependent variable and two or more independent variables. The multiple linear regression model is
The three variable regression model is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

Where, Y is dependent variable, $X_i's$, $(i = 1,2)$ are independent variables, β_0 is intercept term and $\beta_i's$ are slope or coefficient of $X_i's$, $(i = 1,2)$ and ϵ is random error term.

In general, k-variable linear regression model is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots\cdots\cdots + \beta_{k-1} X_{k-1} + \epsilon$$

Where, Y is dependent variable, $X_i's$, $(i = 1,2, \dots\dots, k-1)$ are independent variables, β_0 is intercept term and $\beta_i's$ are slope or coefficient of $X_i's$, $(i = 1,2, \dots\dots, k-1)$ and ϵ is random error term.

❖ **Regression Coefficient**

Regression coefficient measures the average linear relationship between a dependent variable and one or more independent variables in terms of original units of data.
Consider the simple linear regression model of Y on X is

$$Y = \alpha + \beta X + \epsilon$$

The regression coefficient of Y on X is

$$\beta_{YX} = \frac{Cov(X,Y)}{V(X)} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

$$= \frac{\sum X_i Y_i - \frac{\sum X_i \sum Y_i}{n}}{\sum {X_i}^2 - \frac{(\sum X_I)^2}{n}}$$

The regression coefficient of X on Y is

$$\beta_{XY} = \frac{Cov(X,Y)}{V(Y)} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(Y_i - \bar{Y})^2}$$

$$= \frac{\sum X_i Y_i - \frac{\sum X_i \sum Y_i}{n}}{\sum Y_i^2 - \frac{(\sum Y_I)^2}{n}}$$

❖ **Properties of Regression Coefficient**

Some important properties of regression coefficient is

- Regression coefficient is not symmetric measure i.e. $\beta_{YX} = \beta_{XY}$.
- Regression coefficient is not pure number.
- Regression coefficient is lies between $-\infty \; to \; +\infty$.
- Correlation coefficient is the geometric mean of regression coefficients i.e. $r_{XY} = \sqrt{\beta_{YX} \times \beta_{XY}}$.
- If $\beta_{YX} < 1$ then $\frac{1}{\beta_{XY}} < 1$.
- Regression coefficient is independent of change of origin but dependent on scale of measurement.

❖ **Ordinary Least Square Method**

In statistics, ordinary least squares (OLS) is a type of linear least squares method for estimating the unknown parameters in a linear regression model. OLS chooses the parameters of a linear function of a set of explanatory variables by the principle of least squares: minimizing the sum of the squares of the differences between the observed dependent variable (values of the variable being observed) in the given dataset and those predicted by the linear function.

❖ Estimate the parameters of simple linear regression model or two variable regression model by using ordinary least square method.

Solution:

Let us consider the simple linear regression model of Y on X is

$$Y = \alpha + \beta X + u$$

Where, Y is dependent variable, X is independent variable, α is intercept term and β is slope or coefficient of X and u is random error term.

If $\hat{\alpha}$, $\hat{\beta}$ and $\hat{u}$ are the least square estimator α, β and u respectively. Then the estimated model for SRF is

$$Y = \hat{\alpha} + \hat{\beta}X + \hat{u}$$

$$=> Y = \hat{Y} + \hat{u}$$

$$\hat{u} = Y - \hat{\alpha} - \hat{\beta}X \dots\dots\dots (i)$$

Using ordinary least square method, differentiate the equation (i) with respect to $\hat{\alpha}$ and $\hat{\beta}$ and setting it equal to 0. Then we have that

$$\frac{d}{d\hat{\alpha}}\sum \hat{u}^2 = 0$$

$$=> \frac{d}{d\hat{\alpha}}\sum (Y - \hat{\alpha} - \hat{\beta}X)^2 = 0$$

$$=> 2\sum (Y - \hat{\alpha} - \hat{\beta}X)(-1) = 0$$

$$=> \sum Y - n\hat{\alpha} - \hat{\beta}\sum X = 0$$

$$=> \hat{\alpha} = \frac{\sum Y}{n} - \hat{\beta}\frac{\sum X}{n}$$

$$\therefore \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$

And

$$\frac{d}{d\hat{\beta}}\sum \hat{u}^2 = 0$$

$$=> \frac{d}{d\hat{\beta}}\sum (Y - \hat{\alpha} - \hat{\beta}X)^2 = 0$$

$$=> 2\sum (Y - \hat{\alpha} - \hat{\beta}X)(-X) = 0$$

$$=> \sum XY - \hat{\alpha}\sum X - \hat{\beta}\sum X^2 = 0$$

$$=> \sum XY - (\bar{Y} - \hat{\beta}\bar{X}) \sum X - \hat{\beta} \sum X^2 = 0$$

$$=> \sum XY - \bar{Y} \sum X + \hat{\beta}\bar{X} \sum X - \hat{\beta} \sum X^2 = 0$$

$$=> \sum XY - \frac{\sum X \sum Y}{n} + \hat{\beta} \frac{(\sum X)^2}{n} - \hat{\beta} \sum X^2 = 0$$

$$=> \hat{\beta} \left(\sum X^2 - \frac{(\sum X)^2}{n} \right) = \sum XY - \frac{\sum X \sum Y}{n}$$

$$\therefore \hat{\beta} = \frac{\sum XY - \frac{\sum X \sum Y}{n}}{\left(\sum X^2 - \frac{(\sum X)^2}{n} \right)}$$

Therefore the fitted regression model is $\hat{Y} = \hat{\alpha} + \hat{\beta}X$

- **Residual**

Residual is the difference between observed value and predicted value. Mathematically

$$\hat{u} = Y - \hat{Y}$$

Where, $\hat{u}$ is the residual term, Y is the observed value and $\hat{Y}$ is the predicted or estimated value.

- **Coefficient of Determination**

The coefficient of determination (R^2 or r-squared) is a statistical measure in a regression model that determines the proportion of variance in the dependent variable that can be explained by the independent variable. In other words, the coefficient of determination tells one how well the data fits the model (the goodness of fit).

The coefficient of determination (R^2) is defined as

$$R^2 = \frac{\sum (Y_i - \hat{Y})^2}{\sum (Y_i - \bar{Y})^2}$$

The coefficient of determination is lies between 0 and 1.

- An R^2 of 0 means that the dependent variable cannot be predicted from the independent variable.
- An R^2 of 1 means the dependent variable can be predicted without error from the independent variable.

- An R^2 between 0 and 1 indicates the extent to which the dependent variable is predictable. An R^2 of 0.10 means that 10 percent of the variance in Y is predictable from X; an R^2 of 0.20 means that 20 percent is predictable; and so on.

❖ **Problem:** Hypothetical data on weekly family consumption expenditure (Y) and weekly family income (X) are given below:

Y	70	65	90	95	110	115	120	140	155	150
X	80	100	120	140	160	180	200	220	240	260

i. Fit the regression line of Y on X.
ii. How do you interpret the intercept and the slope of the regression line?
iii. If the weekly family income is 175 compute the weekly consumption expenditure.

Solution:

(i)

Y	**X**	Y^2	X^2	**XY**
70	80	4900	6400	5600
65	100	4225	10000	6500
90	120	8100	14400	10800
95	140	9025	19600	13300
110	160	12100	25600	17600
115	180	13225	32400	20700
120	200	14400	40000	24000
140	220	19600	48400	30800
155	240	24025	57600	37200
150	260	22500	67600	39000
Total **1110**	**1700**	**132100**	**322000**	**205500**

The regression line of Y on X is

$$Y = \alpha + \beta X + u$$

We know that the regression coefficient is

$$\beta = \frac{\sum X_i Y_i - \frac{\sum X_i \sum Y_i}{n}}{\sum {X_i}^2 - \frac{(\sum X_I)^2}{n}}$$

$$= \frac{205500 - \frac{1700 \times 1110}{10}}{322000 - \frac{(1700)^2}{10}}$$

$$= \frac{16800}{33000} = 0.509$$

And the intercept term is

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$

$$= \frac{\sum Y}{n} - \hat{\beta}\frac{\sum X}{n}$$

$$= \frac{1110}{10} - 0.509 \times \frac{1700}{10} = 24.470$$

Therefore the fitted regression line of Y on X is

$$\hat{Y} = 24.470 + 0.509X$$

(ii) **Interpretation:**

$\hat{\alpha} = 24.470$ indicates that if the values of weekly family income (X) 0 then average weekly family expenditure (Y) is 24.47.

$\hat{\beta} = 0.509$ indicates that if the values of weekly family income (X) goes up 1 unit then on average the values of weekly family expenditure (Y) goes up 0.509 unit.

(iii) If the weekly family income is 175 then the weekly consumption expenditure is

$$\hat{Y} = 24.470 + 0.509 \times 175 = 113.545$$

- **Importance of Regression Analysis**

The major importance of regression analysis are as follows:

- The regression equation provides a coincise and meaningful summary of the relationship between dependent variable and independent variable.
- The relationship can be used for predictive purpose i.e. It helps in predicting and estimating the value of dependent variable as price, production, sales etc.

- It helps to measure the variability or spread of values of a dependent variable with respect to the regression line.

❖ Comparison of Correlation and Regression

Some comparison of correlation and regression are as follows:

Correlation	Regression
1. Correlation analysis is used to measure strength of the association or linear relationship between two variables.	1. Regression measures the probable movement of one variable in term of another variable.
2. Correlation coefficient is a symmetric measure i.e. $r_{xy} = r_{yx}$.	2. Regression coefficient is not symmetric measure i.e. $\beta_{YX} = \beta_{XY}$.
3. Correlation indicate only linear relationship between the variables.	3. Regression indicate any type of relationship.
4. Correlation coefficient lies between -1 and $+1$.	4. Regression coefficient is lies between $-\infty \; to \; +\infty$.
5. Correlation coefficient is a pure number.	5. Regression coefficient is not pure number.
6. Correlation coefficient is independent on change of origin and scale of measurement.	6. Regression coefficient is independent of change of origin but dependent on scale of measurement.
7. Correlation is not very useful for further mathematical treatment.	7. Regression is widely used for further mathematical treatment.
8. There may be nonsense relationship between two variables in correlation.	8. There is no such nonsense relationship in regression.