# LINEAR REGRESSION ANALYSIS

**Nishant Kumbhar**

Undergraduate
Dr DY Patil institute of technology

***Abstract:*** **Linear Regression is a fundamental statistical technique widely used for modelling and predicting relationship between variables. This paper presents applicability and performance of linear regression. The study begins with the introduction to linear regression and understanding key concepts related to it. We then proceed towards the formula and equations of regression. To evaluate the effectiveness of linear regression we have included the data set of happiness vs income from Kaggle. Our analysis involve accuracy and predictive power. We have discussed the pros and cons of Linear Regression analysis. We have also assumed the data with minimal error.**

**In conclusion, this paper contributes to understanding regression analysis for beginners and understanding the key concepts related to it.**

**Introduction**:
Regression analysis is one of the important topic in statistics. Prediction is one of the key task in the field of statistics , regression provides the platform for user for prediction . Regression is widely used by data analyzers and researcher by comparing dependent and independent variable for daily market application .In recent, machine learning uses **regression algorithm** for prediction of continuous value

Having a good understanding in regression will help the researcher in multiple ways . For example - a data analyst at an e-commerce company can use the regression analysis to asses how festival(Independent variable) impact sales(dependent variable). By collecting historical data on sales and festival dates , they can build regression model .The model can used to predict the future sales based on upcoming festival dates. Such insights can help in building market strategies.

OBJECTIVES :

* **To understand the regression analysis at beginner level**
-Explain the fundamental concepts of regression , including dependent and independent variables , correlation coefficient and linear regression
-Provide examples to help readers to grasp the basic principle of regression
* Improve decision making in terms of prediction
- Explain how regression analysis helps quantifying relationships and making prediction thereby aiding in better decision making
*Offer practical insights based on regression analysis, Application in Real world content
-Offer guidelines for conducting regression analysis effectively -Discuss the challenges in regression analysis -

Understanding Regression:
*What actually regression is ?

-Regression is nothing but a term in statistics which is used to compare two or more sets or variable of data and extract meaningful output. The one set or variable of data should be dependent and other should be independent one .

*What are dependent and independent variable?
- Dependent variable are those which you want to predict or explain. As the example mentioned above, the sales are the dependent event or variable which depend on the festivals.

-Independent variables are those which are not affected by any external event or variable . As for the above example festival is an Independent variable which is not affected.

* How Regression work ?

Lets compare two sets of data with the help of bar charts

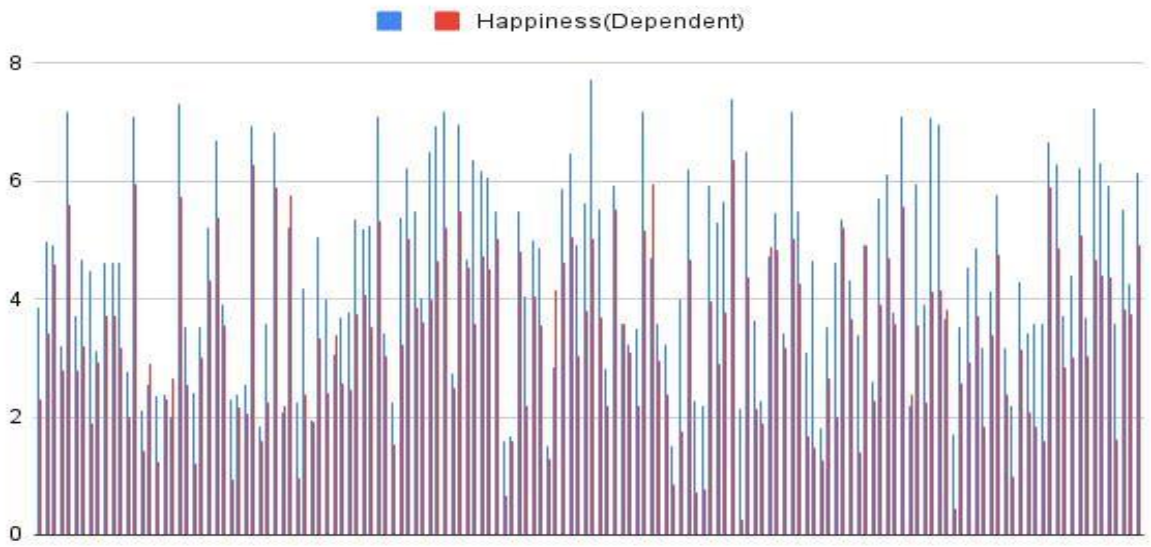


Fig 1

The following chart is of happiness vs income

-In the above chart happiness is considered as dependent variable and income is considered as independent variable

-Red is representing happiness rate and blue is representing the income

- By the bar chart representation it is difficult to predict the outcome or drive any relationship between them

-Hence, Regression uses scatterplot for the representation of data

Following is the scatterplot for the same value of above graph

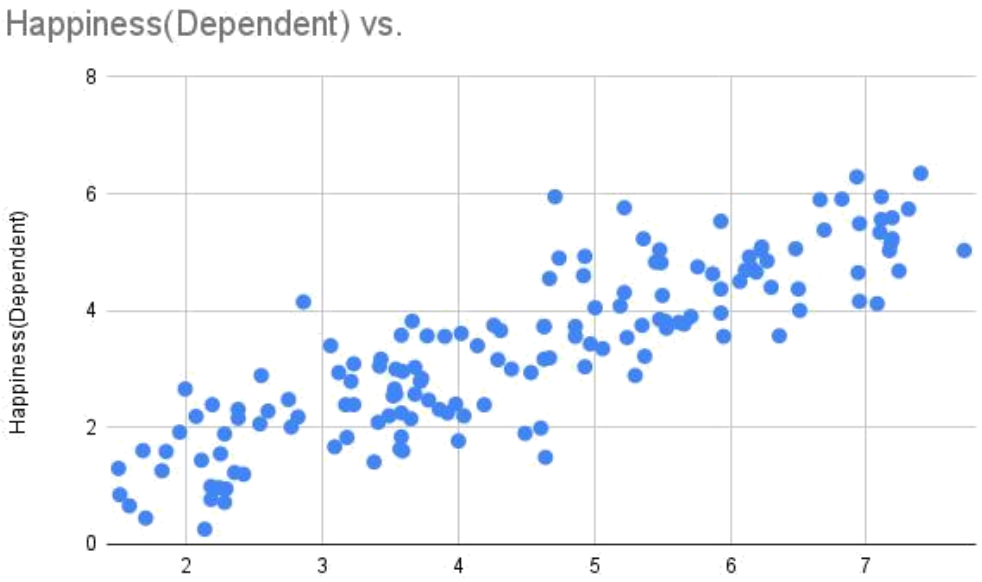


fig 2

-Generally in Regression dependent variable is marked on the Y-axis and Independent variable is marked on X-axis

-Upon observing we see that the tendency of point on the scatter plot is in a particular direction

-The general movement of the paired point is best explained by the straight line know as Regression line. This helps to discover whether is the relationship is linear or nonlinear

*Regression line-The best fit line which represent the relationship between two variables . The line is drawn in such a way that the all point is at minimum distance from line

There are two methods for drawing a regression line

1]Free hand-In free hand drawing , the movement and spread if the points is observed and line is drawn from maximum density

2]Method of least square-Method of least square for fitting of straight line requires minimizing the squares of vertical deviations

Following are the steps to find the best fit lines

- Calculate the mean of both dependent and independent variable (*X) and(*Y)
- Calculate standard deviations of both x and y(u can use the formula or STDEV key word in datasheet)
- Calculate $b = \frac{\Sigma(Xi - *X)(Yi - *Y)}{\Sigma(Xi - *X)2}$
- Or b can calculated as r multiplies ratio of standard deviation of y is to x
- Calculate $a = * Y - (b * X)$

At final the line can be represented as Y = a + bx + e

e represent error which usually occur in every sample

- can calculate the regression line by above formula or by using software available

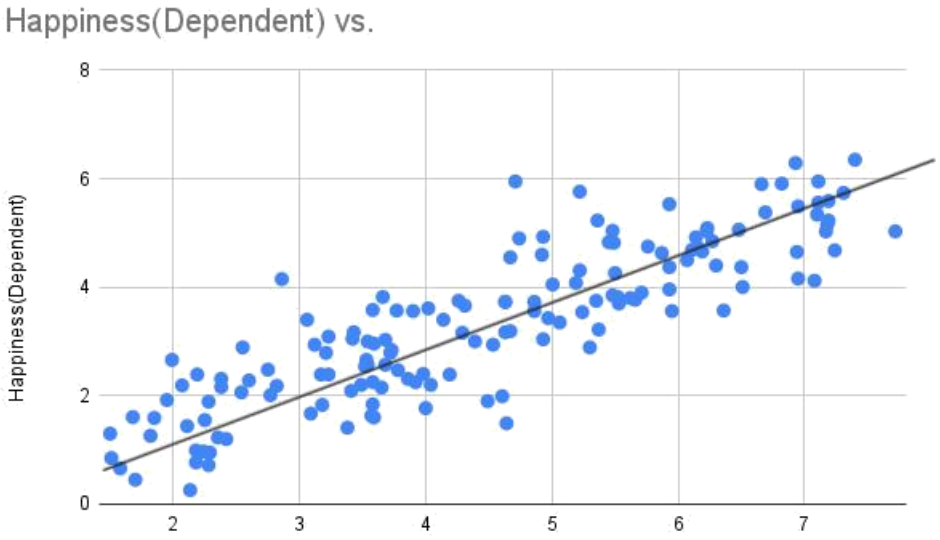


Fig 3

Before preceding towards how to predict dependent term from independent we need to know about correlation coefficient

*Correlation coefficient- Correlation coefficient is a measure of relationship or association between two variables or dataset.

- the value of correlation coefficient ranges from –1 to 1

- -1 depicts that when one variable is changing other on is decreasing which is in statistics also know as negative correlation

- 1 depicts the positive correlation where both the values are increasing simultaneously

- 0 depicts there is no linear relationship between two variables or dataset

* Calculation of correlation coefficient

Formula: $r = 1 \times \Sigma(Xi - *X)(Yi - *Y) \times 1$ —

n Sx

- Sx and Sy are the standard deviation of x and y

- U can just use spss software or use CORREL keyword in datasheet for calculating correlation coefficient

*PREDICTION

The value of r we calculated for happiness vs income dataset

Is 0.866

Lets learn how to predict from one problem

- The mean of happiness is 3.34 units

- The mean of income is 4.460k rupees

- Sd of happiness is 1.4

- Sd of income is 1.64

*Lets calculate the happiness of person whose income is 2.6k rupees:

1] the income 2.6k is below average by 1.86 which is 0.58 Sd from mean

2]next we just apply the formula

Mean of dependent variable - r× Sd from mean × Sd of dependent variable

- Here we minus from the mean as the income is below average

3.34 - 0.86×0.58×1.4

= 2.65 units

Therefore we can predict that a person who has income 2.6k can have happiness rate of approximately 2.65 units

As Regression is one of the powerful tool in predicting but it faces some challenges . Lets discuss:

* Challenges faced in regression

- Effects of outlier – In fig 3 the points are generally near the regression line thereby easy to predict. Outliers are the points which significantly far from the regression line . These points doesn't come under the influence of regression area. Sometime removing this element can fail the regression analysis

- Regression analysis is the model which conduct prediction on data provided. Ethics cant be predicted

- Regression analysis model cannot be used in the field of crime investigation which directly depend on human

- Regression takes limited data . Taking in account for more data cant give effective output. We have taken 2 sets of data in this analysis

- We have to consider the data provided is with minimum errors .Error can effect the efficiency

- Regression analysis become way more complex when it has deal with interval type values