IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Optimizing Prediction Systems With Linear Regression

Dr.B.Azhagusundari
Associate Professor, Department of Computer Science NGM College Pollachi

Abstract

Linear regression is a fundamental statistical technique used to model the relationship between a dependent variable and one or more independent variables. By establishing a linear equation that best fits the observed data, it enables the prediction of the dependent variable based on known values of the independent variables. This method not only quantifies relationships but also provides critical insights for trend analysis and forecasting, making it an indispensable tool across various domains. As one of the most widely applied statistical techniques, linear regression plays a pivotal role in predictive modeling, data analysis, and informed decision-making. This paper explores the core principles of linear regression, examining its applications, significance in predictive systems, and its extensive use in fields such as finance, healthcare, and artificial intelligence..

Keywords: Continuous variable, linear regression, Prediction system, least squares method

I Introduction

The model of linear regression was first proposed by Sir Francis Galton in 1894. Linear regression is a statistical test applies to a data set to define and quantify the relation between the considered variables. Chan et al (2004) explain the Univariate statistical tests such as Chi-square, Fisher's exact test, t-test, and analysis of variance do not allow taking into account the effect of other covariates/confounders during analyses. Chan et al (2003) discuss the partial correlation and regressions are the tests that allow the researcher to control the effect of confounders in the understanding of the relation between two variables. Schneider et al (2010) discuss about statistical assessment of medical data is often to explain relationships between two variables or among several variables.

Regression analysis is a predictive modeling technique that estimate the relationship between two or more variables. Recall that a correlation analysis makes no assumption about the causal relationship between two variables. Regression analyses focus on the relationship between a dependent/target variable and independent variable/predictors. Here, the dependent variable is assumed to be the result of the independent variable(s). The value of predictors is used to estimate or expect the likely-value of the target variable.

2. Linear Regression Model

A linear regression model can be defined as the function approximation that represents a continuous response variable as a function of one or more predictor variables. While create a linear regression model, the purpose is to identify a linear equation that best predicts or models the relationship between the response or dependent variable and one or more predictor or independent variables.
Equation of linear regression with one dependent and one independent variable is defined by the formula

$$y = c + b * x$$

where y = estimated dependent score, c = constant , b = regression coefficient, x = independent variable.

2.1 Frame work for linear Regression predict model

Suppose you are an HR professional and want to determine:

- Whether age of an employee has a considerable effect on their maturity
- The significance of experience and capability on remuneration
- The significance of Intelligence and Emotional Quotient on problem handling capability
- How sedentary lifestyle at workplace affects employee output
- A Specific physical activity makes employees more energetic and lively at the workplace

All these are practice scenarios in an organization. But their impact is huge. How, as an HR professional, can you determine which variables have what impact on employee productivity. Regression analysis offers you the answer. It helps you explain the relationship between two or more variables.

The model uses ordinary Least Squares method, which determines the value of unknown parameters in a linear regression equation. Its aim is to minimize the difference between observed responses and the ones predicted using linear regression model. There are certain requirements that need to fulfill, in order to use this model. Otherwise the results can be confusing and ambiguous.

The phenomenon is, Work experience and remuneration are related variables. The linear regression model can help predict the remuneration slab of an employee given their work experience.

There are two lines of regression

- Y on X
- X on Y.

Y on X is when the value of Y is unknown. X on Y is when the value of X is unknown.

Suppose, the value of remuneration is = Y and the value of experience is = X

2.2 Selection of Line of Regression

The statistical representation above is an example to show how to develop econometric models when the value of one of the variables is known and another's unknown.

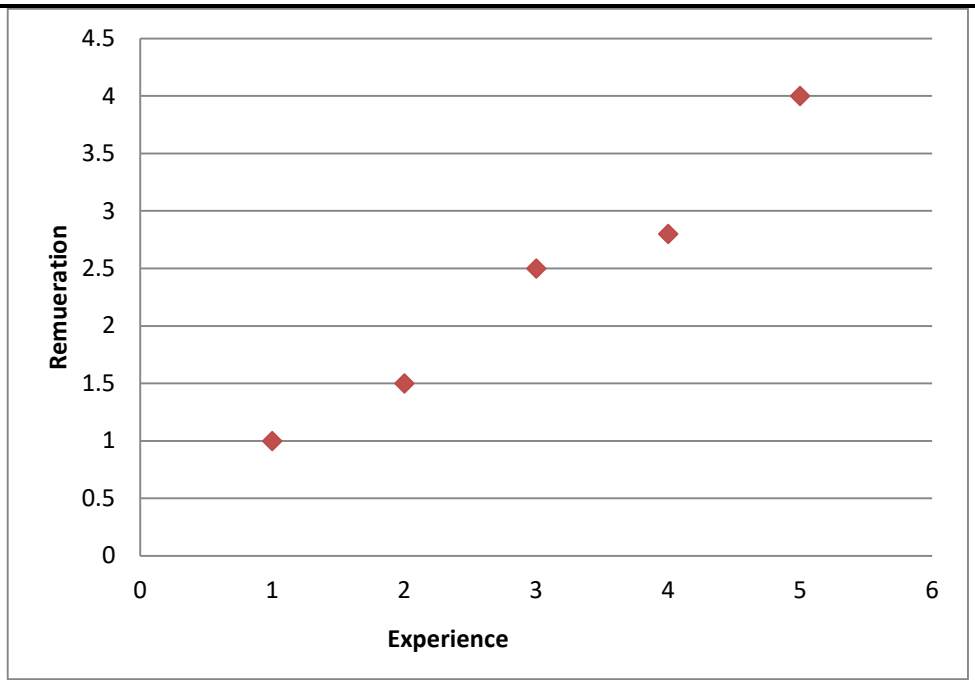
This is because remuneration may depend on the work experience of an individual but the vice versa is not true. Experience doesn't depend on remuneration. Therefore, to carefully choose the dependent variable and then the line of regression.

A linear regression equation shows the percentage increase or decrease in the value of dependent variable (Y) with the percentage increase or decrease in the value of independent variable (X).

Let us suppose the values of X and Y are known.

Experience : X	1	2	3	4	5
Remuneration : Y	1	1.5	2.5	2.8	4

Table 1



Graph 1

The white line connecting all the dots in the graph1 above represents the error or prediction. Now want to find the best-fitted line of regression to minimize the error of prediction. The aim is to help find the best-fitted line of regression.

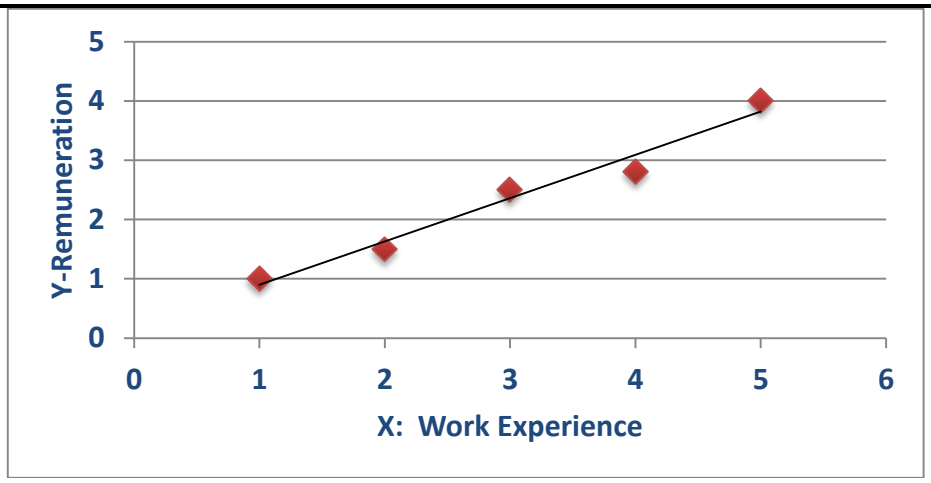
2.3 Best-Fitted Linear Regression

By using the ordinary least squares method

Let us continue with the above example:

	X	**Y**	**XY**	**X-X'**	**Y-Y'**	**(X-X')(Y-Y')**	**$(X-X')^2$**	**$(Y-Y')^2$**
	1	1	1	-2	-1.16	2.32	4	1.346
	2	1.5	3	-1	-0.66	0.66	1	0.436
	3	2.5	7.5	0	0.34	0.34	0	0.116
	4	2.8	11.2	1	0.64	0.64	1	0.410
	5	4	20	2	1.84	3.68	4	3.386
Sum	**15**	**10.8**	**42.7**			**7.64**	**10**	**5.69**
Mean	**X' = 3 (15/5)**	**Y' = 2.16 (10.8/5)**						

Table 2



Graph 2: Regression Line

The primary equation is:

Y = a + b(X) +e (error term)

In this case, e is zero because it is assumed that the independent variable (X) has negligible errors.

Therefore, it remains Y = a + b(X).

Let us now find the value of b.

$$b = \frac{\sum XY - (\sum Y)(\sum X)/n}{\sum (X - X')2}$$

Substitute the values in the above formula:

b = [42.7 – (15*10.8)/5] / 10 = 1.02

Therefore,

a = Y – b(X)

a = Y – 1.02(X) or a = ∑Y/n – 1.02 (∑X/n)

a = 2.16 – 1.02*3 = 2.16 – 3.06 **= -1.06**

By substituting the value of a=-1.06, b=1.02 and X, we can find the corresponding value of Y.

Y = a + b(X) = -1.06 + 1.02X

When X = 1 Y = -1.06 + 1.02*1 = -0.04

When X = 2 Y = - 1.06 + 1.02*2 = 0.98

When X = 3, Y will be 2

When X = 4, Y will be 3.02

When X = 5, Y will be 4.04

x	1	2	3	4	5
y	-0.04	0.98	2	3.02	4.04

Table 3

The best-fitted regression line will be displayed in Graph 3.

Graph 3: best-fitted regression line

Properties of the Best-Fitted Regression Line/Estimators

- The regression line passes through the X', which is 3 in this case. (Refer to Graph 3)
- b, the regression coefficient of X, is an average change in Y. In this case, b = 1.02 which is the average change in the values of y: -0.04, 0.98, 2, 3.02 and 4.04. (Refer Table 3)
- The regression line passes through Y', which is 2.16 in this case. (Refer Graph 3)

2.4 Coefficient of Determination – Assessing Goodness-of-Fit

Whenever regression model is used, the first thing should consider is – how well an econometric model fits the data or how well a regression equation fits the data. This is where the concept of coefficient of determination comes in. The regression models are generally fitted using this approach.

Denoted by R^2, its value lies between 0 and 1. When:

- $R^2 = 0$: the value of the dependent variable cannot be predicted from the independent variable.
- $R^2 = 1$, the value of dependent variable can be easily predicted from the independent variable. There are no errors in the data.

Higher the value of R^2, better fit the model is to the data.

Let's now understand how to calculate R^2. The formula for finding R^2 is:

$R^2 = \{ (1 / n) * \sum (X-X') * (Y-Y') \} / (\sigma x * \sigma y)^2$

Where n = number of observations = 5

$\sum (X-X') * (Y-Y') = 7.64$ (Ref Table 2)

σx is the standard deviation of X and σy is the standard deviation of Y

σx = square root of $\sum (X-X')^2 / n = \sqrt{10/5} = 1.414$

σy = square root of $\sum (Y-Y')^2 / n = \sqrt{5.69/5} = 1.067$

Now let's determine the value of R^2

$R^2 = \{ 1/5 (7.64) \} / (1.414 * 1.067)^2 = = 0.67$

Hence $\mathbf{R^2 = 0.67}$

Higher the value of coefficient of determination, lower the standard error. The result indicates that about 67% of the variation in remuneration can be explained by the work experience. It shows that the work experience plays a major role in determining the remuneration.

3. Conclusion

Correlation quantifies the strength of the linear relationship between a pair of variables, whereas regression expresses the relationship in the form of an equation. In this article, simple examples and illustrate linear regression analysis and encourage the readers to analyze their data by these techniques.

4.References

1. Schneider, Astrid, Gerhard Hommel, and Maria Blettner. "Linear regression analysis: part 14 of a series on evaluation of scientific publications." Deutsches Ärzteblatt International 107, no. 44 (2010): 776.
2. Freedman, David A. Statistical models: theory and practice. cambridge university press, 2009.
3. Chan, Y. H. "Biostatistics 201: linear regression analysis." Age (years) 80 (2004): 140.
4. Chan, Y. H. "Biostatistics 202: logistic regression analysis." Singapore medical journal 45, no. 4 (2004): 149-153..
5. Mendenhall, William M., and Terry L. Sincich. Statistics for Engineering and the Sciences. CRC Press, 2016.
6. Panchenko, D. "18.443 Statistics for Applications, Section 14, Simple Linear Regression." Massachusetts Institute of Technology: MIT OpenCourseWare (2006).