Vol. 1, No. 2; Apr-Jun (2022)

Quing: International Journal of Innovative Research in Science and Engineering

Available at https://quingpublications.com/journals/ijirse

Linear Regression Analysis Theory and Computation

G. Asha[*]

Assistant Professor, Department of Computer Science and Applications, Don Bosco College (Arts & Science), Thamanangudy, Karaikal, PY, IND.

M. Sindhuja

Assistant Professor, Department of Computer Science and Applications, Don Bosco College (Arts & Science), Thamanangudy, Karaikal, PY, IND.

ARTICLE INFO

Received: 23-03-2022
Received in revised form: 26-04-2022
Accepted: 27-04-2022
Available online: 30-06-2022

Keywords:

Linear Regression;
Single Linear Regression;
Multiple Linear Regression;
Polynomial Regression;
Machine Learning.

ABSTRACT

In a statistical method, linear regression is used to estimate the relationship between a dependent variable based on the value of an independent variable. It is a forecasting model in which one or more independent variables are utilised to predict a variable. Of all statistical models, the linear regression model is the most commonly used. In this paper, there are three kinds of linear regression discussed (i) single linear regression, (ii) multiple linear regression, and (iii) polynomial regression. It also shows how we can perform manual linear regression analyses using model datasets. Every model is frequently subjected to hypothesis testing to ensure that accurate outcome is expected.

DOI: https://doi.org/10.54368/qijirse.1.2.0002

1.0 INTRODUCTION

In terms of machine learning algorithms, one of the best-known, most trustworthy, and oldest is linear regression. It makes use of a mathematical model to illustrate how two or more variables are related. Independent variables (input) and dependent variables (output) are the two types of variables used in regression (output). Independent variables: These values are plotted on the X-axis, and they are used to estimate the dependent variable because their values are unaffected by the effects of other variables. These values are plotted on the Y-axis as dependent variables. If there is some manipulation of independent variables, the value is modified. In the case of linear regression, two variables are taken into consideration. As a first step, the variables should be significant predictors of the independent variables, and as a second step, the regression line should be as reliable

[*] *Corresponding author's e-mail:* gnanaasha.asha@gmail.com (G. Asha)

as possible. The regression line (linear equation) is the line that best fits the data. The least square method will be used to calculate this line.

Regression analysis is a form of predictive modelling that looks at two or more variables to develop a model. We have one dependent variable and more than two independent variables in regression analysis, and the independent variables are responsible for the change in the dependent variable. We must estimate the new values of the dependent variable as a function of the current values of the independent variable to determine the relationship between the dependent and independent variables. A constant numerical value is needed for the dependent variable. The following are some of the applications of regression analysis: calculating the strength of predictors (used to assess the influence of an independent variable on dependent variables), predicting an effect (used to determine how often the dependent variable changes as one or more independent variables change), and trend forecasting (to predict future values).

1.1 Statistical Learning Perspective

1.1.1 Single Linear Regression

In this case, the algorithms are attempting to learn data in the context of a hypothetical function (f). That is, we will state the following relationship between input and output (Brownlee, 2016):

$$\text{Output} = f(\text{input})$$

The main outputs are the data it's used as input. The main outputs were data that are expected to occur in the future. Answer parameters apply to the new output results. The following is a reflection of that.

$$\text{Output variables} = f(\text{input variables})$$

1.1.2 Multiple Linear Regression

There are several input variables in this case. Let's call the set of input variables an input vector. The following diagram illustrates the relationship between input and output variables (Brownlee, 2016).

$$\text{Output variable} = f(\text{input vector})$$

$$\text{Dependent variable} = f(\text{independent variable})$$

The input variable is abbreviated as x or X, and the output variables are abbreviated as y or Y. As a result, the formation is

$$Y = f(X)$$

If we are using more than input variables, then we will be termed as X_1, X_2, X_3, ……, X_n.

1.2 Computer Science Perspective

There are several variations between the viewpoints of computer science and statistics. We'll be using agreements transfer function properties whenever building a model that makes projections. Characteristics may also be referred to as functions (Brownlee, 2016).

$$\text{Output attribute} = \text{program}(\text{input attributes})$$

$$\text{Output} = \text{program}(\text{input features})$$

$$\text{Prediction} = \text{program}(\text{instance})$$

1.3 Reasons Why We Want a Regression Model

There are plenty of reasons for using a regression model. Here we will discuss some reasons.

1) **Descriptive –** Regression model is evaluated to find the significant relationship between the independent variables and dependent variables.
2) **Correction –** Corrections are made for covariates and cofounders in the analysis of data.
3) **Predictors –** To measure the risk variables that have a substantial impact on a dependent variable.
4) **Estimation –** It is used to estimate the number of new events that will take place in the future.

1.4 Submissions

- **Economic Growth:** This term is used to estimate the state's financial growth in the coming quarters.
- **Product Price:** Product price can be used to forecast a product's future price.

  Estimate how many houses the builder will be able to construct in the coming months and what the price will be.
- **Score Prediction:** To use a defender's latest work, decide that most goals him and she will score in upcoming matches.
- **Automobiles:** Evaluate engine efficiency using test data.
- **Forecasting:** It involves forecasting future patterns and making fair forecasts.

The proposed research study is classified into the following groups. The first session deals with the introduction of the study, and the literature review is discussed in section two. The standard regression analysis method is presented in section three. The method of multiple linear regression is defined in the fourth section. Polynomial regression is covered in section five. At last, the conclusion is defined in section six, and the proposed projects are defined in section seven.

2.0 LITERATURE REVIEW

Regression is used to investigate dependency. In research ventures, regression analysis is essential. Regression analysis is the essential aspect of regression methodology. Regression aims to review a dataset in a straightforward, functional, and forthright manner. Drawing the correct graph for the database is the most crucial step in regression analysis. Scatterplot is a two-dimensional graphical method that can be used to display regression data. The multivariate regression matrix is an easy way to arrange many scatterplots at once (Weisberg, 2013). Regression analysis aims to create a mathematical model that explains how variables interact (Seber and Lee, 2003).

The flat surface is now an efficient regression model. For several factors, everybody accepts the statistical model. One of the most critical reasons is finding triggers by looking at relationships between variables (Seber and Lee, 2003). Creating the regression analysis requires the separation of predictor variables (Raftery *et al.,* 1997). Correlation is also a subject of regression analysis. Correlation is a calculation of how strongly the values of another influence the singular values. The relationship between two variables is based on the following questions: Is there any correlation between two variables, and does the value of one variable determines the other (Zou *et al.,* 2003).

There are two essential aspects to a regression problem. One is to figure out which scientific formula is the most effective. The other question is how we complete the effective database schema (Tanaka *et al.,* 1982).

In a Univariate or Single Linear Regression, the purpose is to examine the correlation between a single independent variable and a dependent variable and develop a linear relationship equation between the two variables and interpret the regression results (Schneider *et al.,* 2010). In their study, Schneider *et al.,* (2010) opted for the Multiple Linear Regression (or) Multivariate approaches to examine the relationship between two or more independent variables with one dependent variable. "The equation y = mx + c, which defines the line of best fit for the relationship between y (dependent variable) and x (independent variable), is used in the linear regression analysis (Kumari and Yadav, 2018)".

3.0 SINGLE LINEAR REGRESSION PROCESS

The model data is presented in Table 1. It explains the positive correlation between the variables X and Y. We need to expect the assessment of Y from X, the assessment of X is high, and thereafter the assessment of Y is similarly high. For simplicity, the following five data have been taken for this study. The scatter plot for the above model data is showing up in Figure 1.

Table 1 – Model Data of Variable X and Y

VAR X	VAR Y
1	4
2	5
3	2
4	5
5	6

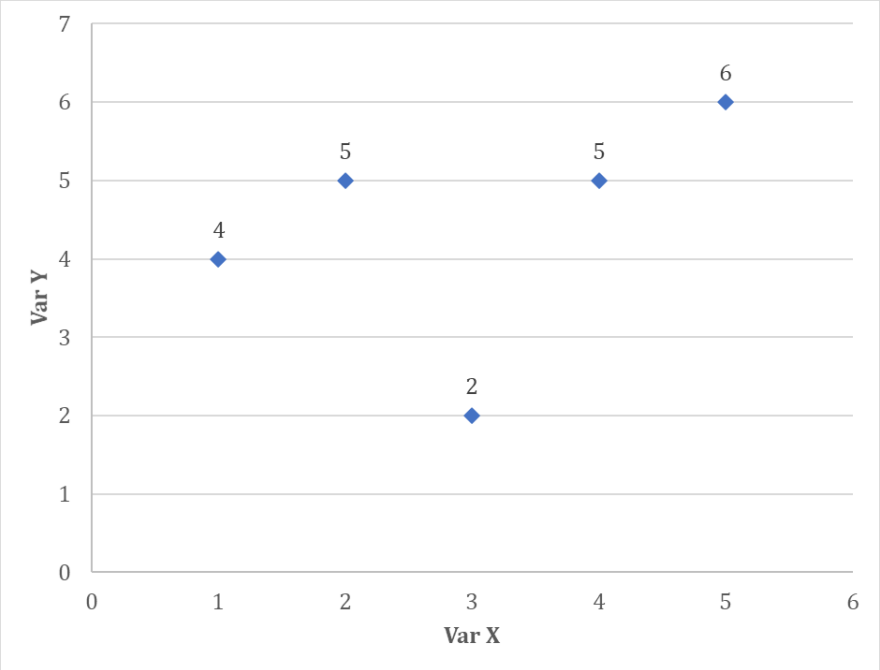

Figure 1 – Scatter Plot for the Model Data of Variable X and Y

Linear regression is used to determine the best fit line for the data based on the core interests. The most proper line is called regression or linear line. The line contains the expected score on Y for every impetus for X is showed up in Figure 2. The vertical line connecting the concentrations to the direct line illustrates the assumption bumbles. We know from the Figure 2, that a couple of values are incredibly near the line, so the prediction error is slighter. Unexpectedly, a couple of values are far away from the line, which has an enormous prediction error.

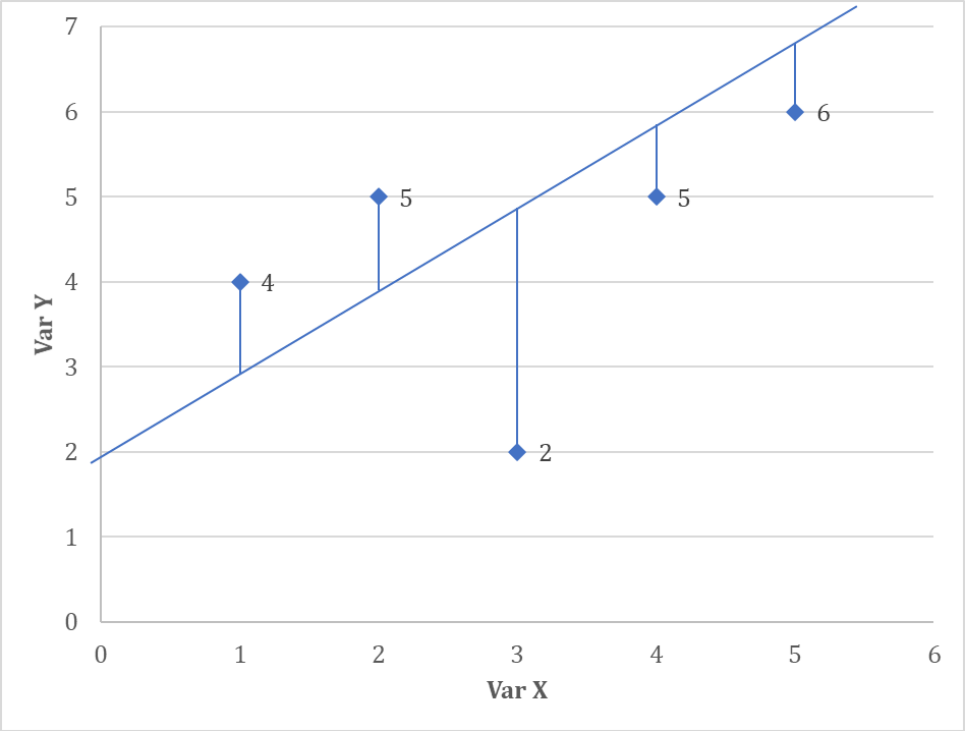

Figure 2 – Regression Line for the Model Data of Variable X and Y

Two sorts of relationships exist between factors. One is a positive (or) direct relationship, and another is a negative (or) reverse relationship. We said that our model shows a positive relationship. For example, let us examine the impact of spending time in social networks on students' grades in the examination. We consider the spending time in social networking as X-axis and the grade scored by the students as Y-axis. The study reveals that the students who spend more time in the social network scored a low grade. So, this is the model for negative relationships. In the positive relationship, the regression line is outlined by the following equation **y = mx + c**. Where m denotes the positive slope, c denotes the y-intercept of the line. For negative relationships, the equation is y = - mx + c, where -m denotes the negative slope of the line.

3.1 Mathematical Implementation

Let's calculate the mean of X. Mean of X is 3, and the mean of Y is 4.4. The goal is to find out the best fit line. To draw the line, we must calculate the value of m and c.

$$m = \frac{\Sigma(x - x')(y - y')}{\Sigma(x - x')^2}$$

To find out m, we must find X-X'. It is the distance of all the points through the line y = 3. X' is the mean value of the X. So, we have X' as 3. For example, the first value has the value 1, and the X' is 3. So, X-X' is 1 – 3 = -2. Repeat the same process to find X-X' for all values. A step-by-step calculation of the best fit line is represented in Table 2.

Table 2 – Step-by-Step Calculation of the Best Fit Line

X	Y	X-X’
1	4	-2
2	5	-1
3	2	0
4	5	1
5	6	2

Next, we have to know Y-Y’. It is the distance of all the points through line X = 4.4. Y’ is the mean value of the Y. So, we have Y’ as 4.4. From this value, we can easily calculate the distance of all the points through line X. A step-by-step calculation of the best fit line is represented in Table 3.

Table 3 – Step-by-Step Calculation of the Best Fit Line

X	Y	X-X’	Y-Y’
1	4	-2	-0.4
2	5	-1	0.6
3	2	0	-2.4
4	5	1	0.6
5	6	2	1.6

Next, we must calculate $(X-X')^2$. A step-by-step calculation of the best fit line is represented in Table 4.

Table 4 – Step-by-Step Calculation of the Best Fit Line

X	Y	X-X’	Y-Y’	$(X-X')^2$
1	4	-2	-0.4	4
2	5	-1	0.6	1
3	2	0	-2.4	0
4	5	1	0.6	1
5	6	2	1.6	4

Next, we must calculate the product of (X-X’) and (Y-Y’). A step-by-step calculation of the best fit line is represented in Table 5.

Table 5 – Step-by-Step Calculation of the Best Fit Line

X	Y	X-X’	Y-Y’	$(X-X')^2$	(X-X’)(Y-Y’)
1	4	-2	-0.4	4	0.8
2	5	-1	0.6	1	-0.6
3	2	0	-2.4	0	0
4	5	1	0.6	1	0.6
5	6	2	1.6	4	3.2

Next, we can easily find out the value of m, the total sum of (X-X’)(Y-Y’) is 4, and the total sum of $(X-X')^2$ is 10. So, m = 4 / 10 that is m = 0.4. Substitute the value of m in the equation to find out c.

4.4 = 0.4 * 3 + c. So, c = 3.2. We find out the values as m = 0.4 and c = 3.2, from this we can find the regression line y = mx + c. that is Y = 0.4 X + 3.2.

For the given value m = 0.4 and c = 3.2 let us find out the predicted value for Y for the dataset X = {1, 2, 3, 4, 5}. For example, the first value is 0.4 * 1 + 3.2. So, Y_P=3.6. Repeat the same process to find out all the values. It is shown in the following Table 6.

Table 6 – Predicted Value for Y

X	Y_P
1	3.6
2	4
3	4.4
4	4.8
5	5.2

We must calculate the distance between the original value and our predicted value to find out the error. The goal of our algorithm is to reduce the distance.

3.2 Goodness Test

Next, we endeavour to test the respectability of our new model. There is plenty of methods open to testing the tolerability. Permit us to discuss the R-Square procedure.

3.3 R-Square Method

R-Square value is the quantitative evaluation of how close the data fits in backslide line. It is also known as the coefficient of confirmation or the coefficient of various ends. This square system's legitimacy makes it a strong prediction that this same system would provide extraordinary results from the lower square strategy. The distance of actual mean against distance expected mean. That is just the assessment of R^2. We need to find the R^2 by going with the condition.

$$R^2 = \frac{\Sigma(y_p - y')^2}{\Sigma(y - y')^2}$$

From the above definitions, we are evidently understood that Y_P is the expected worth and the Y is the principal worth. To find the R^2 regard, we recently decided (Y-Y'). Permit us to find out $(Y-Y')^2$. The calculation is presented in Table 7.

Table 7 – Step-by-Step Calculation of R^2

X	Y	Y-Y’	$(Y-Y')^2$
1	4	-0.4	0.16
2	5	0.6	0.36
3	2	-2.4	5.76
4	5	0.6	0.36
5	6	1.6	2.56

Next, we will find out Y_P - Y’. We already calculated Y_P, and the Y_P - Y’ is calculated, and values are furnished in Table 8.

Table 8 – Step-by-Step Calculation of R^2

X	Y	Y-Y'	$(Y-Y')^2$	Y_P	$Y_P - Y'$
1	4	-0.4	0.16	3.6	-0.8
2	5	0.6	0.36	4	-0.4
3	2	-2.4	5.76	4.4	0
4	5	0.6	0.36	4.8	0.4
5	6	1.6	2.56	5.2	0.8

Next, for finding R^2 we have to find out the last value $(Y_P - Y')^2$. This is tabulated in the following Table 9.

Table 9 – Step-by-Step Calculation of R^2

X	Y	Y-Y'	$(Y-Y')^2$	Y_P	$Y_P - Y'$	$(Y_P - Y')^2$
1	4	-0.4	0.16	3.6	-0.8	0.64
2	5	0.6	0.36	4	-0.4	0.16
3	2	-2.4	5.76	4.4	0	0
4	5	0.6	0.36	4.8	0.4	0.16
5	6	1.6	2.56	5.2	0.8	0.64
		$\Sigma(Y-Y')^2$ =	**9.2**		$\Sigma(Y_P - Y')^2$ =	**1.6**

Next, we can easily find the R^2, and the $\Sigma(Y_P - Y')^2$ is 1.6 and $\Sigma(Y-Y')^2$ is 9.2. So, $R^2 = 1.6 / 9.2 = 0.2$. Its value is approximately 0.2. The result of R^2 indicated that the data points of the regression line are long away. In our example, if we may increase the value of R^2 to 0.8, it will predict the accurate result. If we may take the value of R^2 as very low as 0.03, the data points are very far away.

4.0 MULTIPLE LINEAR REGRESSION PROCESS

Various direct backslide measure is an increase of precise direct backslide measures. Alternately with a single direction, it uses more than one free factor. We have one pointer (free), and one response (subordinate) variable in the above fundamental direct backslide measure. Anyway, in various backslide, we have more than one marker variable and one response variable.

Various straight backslides used to examine the relationship between two or more independent variables and one dependent variable. Use various straight backslides if necessary to determine:

- The strength of the correlation between at least two independent variables and one dependent variable (for example, temperature, how rainfall as well as the proportion of fertiliser added have an impact on crop improvement).
- To determine the value of the dependent variable at one specific value of the independent variable (for example, the yield of a crop is anticipated to produce with a certain amount of rainfall, addition of fertiliser, and temperature changes).

In the direct backslide measure, we used an "X" to address the free factor. We'll have more than one independent element in various direct backslide, so we'll have multiple "X" in the given model dataset as exhibited in the Table 10. It tells the association between X_1, X_2, and Y. We need to expect Y from X_1, X_2. The assessment of X_1, X_2 is high, and a short time later, the assessment of Y is similarly high. To explain this, totally five data is using as two self-governing elements.

Table 10 – Sample Data of Independent Elements

Y	X_1	X_2
-3.9	4	8
2.5	5	5
11.5	6	3
3.5	3	7
5.7	2	1

The simplicity of using 2D scatterplots for model displays is greatly aided by the availability of one independent variable in Simple Linear Regression. It is possible to visualise the model in three dimensions if there are two or more independent variables: the x-axis represents the first explanation, while the y-axis represents the second explanation, and the z-axis represents the response. Here is a picture of the 3D scatterplot of the primary data (*vide* Figure 3). A model would be a line with a set of coordinates. It is challenging to visualise models with three or more independent variables. In other words, we can't describe what the equation looks like.

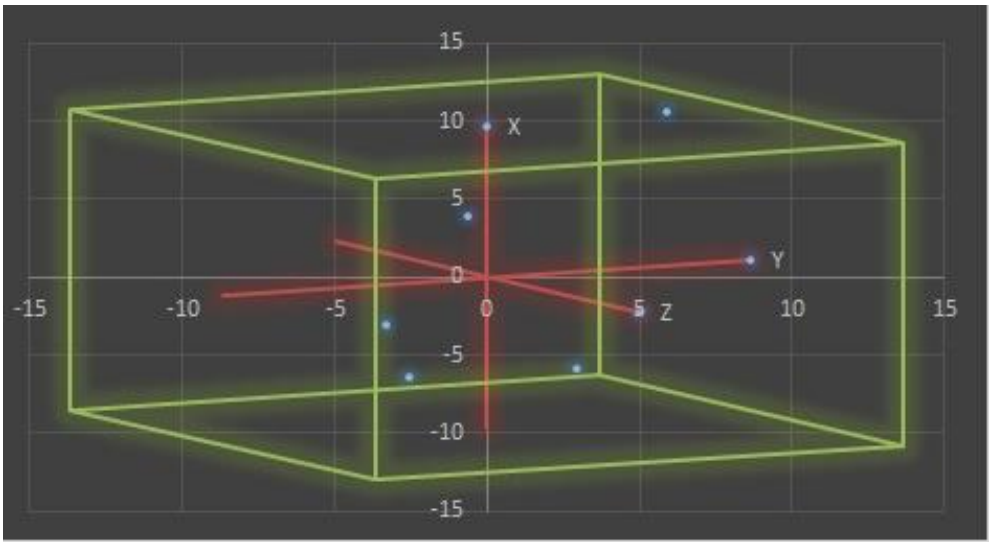

Figure 3 – 3D Scatter Plot for the Sample Data of Independent Elements

The formula for Multiple Linear Regression

Multiple linear regression has the following formula:

$$\mathbf{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots \beta_n X_n + \varepsilon}$$

Where,

Y = the dependent variable's expected value.

$\boldsymbol{\beta_0}$ = the slope of the Y-intercept (while all other variables are set to zero, the total value of Y is equal to zero).

$\boldsymbol{\beta_1 X_1}$ = β_1 indicates the coefficient of regression of the first independent variable (X_1) (i.e., the variation in the estimated Y value resulting from a change in the independent variable's value).

... = apply this procedure to whatever number of independent variables you want to examine.

$\boldsymbol{\beta_n X_n}$ = the final independent variable's regression coefficient.

$\boldsymbol{\varepsilon}$ = model error (i.e., there is a large degree of variance in our estimation of Y).

There are two kinds of direct backslide, (i) Ordinary Least Squares (OLS) and (ii) Generalized Least Squares (GLS). The standard differentiation between the two is that OLS acknowledges there is unquestionably not a strong connection between any two free factors. GLS oversees associated independent components by changing the data and a while later using OLS to build the model with changed data.

These strategies use the system for OLS. Along these lines, to manufacture a successful model, you ought to at first think about the associations between your components. For a more quantitative examination, pick free factors so that each pair has a Pearson relationship coefficient near nothing.

4.1 Best Fitting Line

The best fit line is called backslide direct line. The line contains the expected score on Y for every impetus for X_1, X_2 is showed up in the above picture (3D Scatter Plot). Diverse straight backslide learns three things to obtain the most robust line for each independent variable:

- The coefficients of backslide that lead to the most diminutive as a rule model bungle.
- The t-estimation of the overall model.
- The relevant p-regard (how probable it is that the t-estimation would have happened by chance if the faulty hypothesis of no relationship between the independent and dependent variables was significant).

The t-statistic and significant value are then computed for each coefficient of regression in the model.

4.2 Mathematical Implementation

The formula for multiple linear regression of two independent variables is

$$\mathbf{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \ (EQ - 1)}$$

In this equation X_1, X_2, and Y values are known values (*i.e.,* Example Dataset). β_0, β_1, β_2 are the unknown values. Let us calculate the unknown Values using the Normal Equation. The normal equation is given below:

$$\boldsymbol{\beta_0} = \overline{Y} - \beta_1 \overline{X_1} - \beta_2 \overline{X_2} \ (EQ - 2)$$

$$\boldsymbol{\beta_1} = \frac{(\Sigma x_2^2)(\Sigma x_1 y) - (\Sigma x_1 x_2)(\Sigma x_2 y)}{(\Sigma x_1^2)(\Sigma x_2^2) - (\Sigma x_1 x_2^2)} \ (EQ - 3)$$

$$\boldsymbol{\beta_2} = \frac{(\Sigma x_1^2)(\Sigma x_2 y) - (\Sigma x_1 x_2)(\Sigma x_1 y)}{(\Sigma x_1^2)(\Sigma x_2^2) - (\Sigma x_1 x_2^2)} \ (EQ - 4)$$

Next, we calculate the correlation coefficient and find the summation of values, as shown in the Table 11.

Table 11 – Calculation of Correlation Coefficient

Y	X_1	X_2	X_1^2	X_2^2	X_1*Y	X_2 *Y	X_1X_2
-3.9	4	8	16	64	-15.6	-31.2	32
2.5	5	5	25	25	12.5	12.5	25

11.5	6	3	36	9	69	34.5	18
3.5	3	7	9	49	10.5	24.5	21
5.7	2	1	4	1	11.4	5.7	2
27.1	**20**	**24**	**90**	**148**	**119**	**108.4**	**98**

Now let us calculate the prediction variable using the below prediction formula; N is the number of the dataset used in the given example, therefore N = 5.

The General Prediction Formula is:

$$\Sigma x_i^2 = \Sigma x_i^2 - \frac{(\Sigma x_i^2)}{N} \, (EQ - a)$$

$$\Sigma x_i \, y = \Sigma x_i \, y - \frac{(\Sigma x_i)(\Sigma y)}{N} (EQ - b)$$

In this above example dataset, we have taken two independent variables. The 'i' values are 1, 2.

For i = 1, the above formula **a** is:

$$\Sigma x_1^2 = \Sigma x_1^2 - \frac{(\Sigma x_1^2)}{N} \quad (EQ - 5)$$

Substitute the summation values $\Sigma x_1^2 = 90$ $(\Sigma x_1)^2 = 20$. Table 11 in equation number 5 and get the Σx_1^2 value.

$$\Sigma x_1^2 = 90 - \frac{(20^2)}{5} = 10 \, ; \, \Sigma x_1^2 = 10$$

For i=2, the above formula **a** is:

$$\Sigma x_2^2 = \Sigma x_2^2 - \frac{(\Sigma x_2^2)}{N} \, (EQ - 6)$$

Substitute the summation values $\Sigma x_2^2 = 148$; $(\Sigma x_2)^2 = 24$. Table 11 in equation number 6 and get the Σx_2^2 value.

$$\Sigma x_2^2 = 148 - \frac{(24^2)}{5} = 32.8 \, ; \, \Sigma x_2^2 = 32.8$$

For i=1, the above formula **b** is

$$\Sigma x_1 \, y = \Sigma x_1 \, y - \frac{(\Sigma x_1)(\Sigma y)}{N} \, (EQ - 7)$$

Substitute the summation values $\Sigma x_1 y = 119$; $(\Sigma x_1) = 20$; $(\Sigma y) = 27.1$. Table 11 in equation number 7; and get the $\Sigma x_1 y$ value .

$$\Sigma x_1 \, y = 119 - \frac{(20)(27.1)}{5} = 10.6; \, \Sigma x_1 \, y = 10.6$$

For i=2, the above formula **b** is

$$\Sigma x_2 \, y = \Sigma x_2 \, y - \frac{(\Sigma x_2)(\Sigma y)}{N} (EQ - 8)$$

Substitute the summation values $\Sigma x_2 y = 108.4$; $(\Sigma x_2) = 24$; $(\Sigma y) = 27.1$. Table 11 in equation number 8 and get the $\Sigma x_2 y$ value.

$$\Sigma x_2 \, y = 108.4 - \frac{(24)(27.1)}{5} = 21.68; \, \Sigma x_2 \, y = 21.68$$

$$\Sigma x_1 \, x_2 = \Sigma x_1 \, x_2 - \frac{(\Sigma x_1)(\Sigma x_2)}{N} (EQ - 9)$$

Substitute the summation values Σx_1x_2 =98; (Σx_1) =20; (Σx_2) =24. Table 11 in equation number 6 and get the Σx_2y value

$$\Sigma x_1\, x_2 = 98 - \frac{(20)(24)}{5} = 2\ ; \Sigma x_1\, x_2 = 2$$

Using the prediction formula, we can calculate the below values, as shown in the following table.

Σx_1^2	10
Σx_2^2	32.8
Σx_1y	10.6
Σx_1y	21.68
$\Sigma x_1\, x_2$	2

Now calculate the unknown values β_0, β_1, β_2 using the above table values, now the equation β_1 is;

$$\beta_1 = \frac{(\Sigma x_2^2)(\Sigma x_1 y) - (\Sigma x_1 x_2)(\Sigma x_2 y)}{(\Sigma x_1^2)(\Sigma x_2^2) - (\Sigma x_1 x_2^2)}$$

$$\beta_1 = \frac{(32.8)(\sum 10.8) - (2)(21.68)}{(10)(32.8) - (2^2)}$$

$$\boldsymbol{\beta_1 = 0.93925}$$

β_2 value is;

$$\beta_2 = \frac{(\Sigma x_1^2)(\Sigma x_2 y) - (\Sigma x_1 x_2)(\Sigma x_1 y)}{(\Sigma x_1^2)(\Sigma x_2^2) - (\Sigma x_1 x_2^2)}$$

$$\beta_2 = \frac{(10)(21.68) - (2)(10.6)}{(10)(32.8) - (2^2)}$$

$$\boldsymbol{\beta_2 = 0.60370}$$

The mean value of $\boldsymbol{\overline{Y}, \overline{X_1}, \overline{X_2}}$. The formula is given below

$$\overline{Y} = \frac{\Sigma Y}{N}; \overline{X_1} = \frac{\Sigma X_1}{N}; \overline{X_2} = \frac{\Sigma X_2}{N};$$

Now the value of β_0 is;

$$\beta_0 = \overline{Y} - \beta_1 * \overline{X_1} - \beta_2 * \overline{X_2}$$

$$\beta_0 = \frac{27.1}{5} - (0.9392) * \frac{20}{5} - (0.60370) * \frac{24}{5}$$

$$\boldsymbol{\beta_0 = -1.234}$$

We find out the values of β_0 = - 1.234, β_1= 0.93925, and β_2 = 0.60370. From this we can find the regression line using the formula $\mathbf{Y = \beta_0 + \beta_1X_1 + \beta_2X_2}$, the equation is $\mathbf{Y = -1.234 + 0.93925*X_1 + 0.60370*X_2}$

For the given value β_0 = 1.234, β_1= 0.93925, and β_2= 0.60370. Let us find out the predicted value for y for the dataset X_1= {4,5,6,3,2}, X_2 = {8,5,3,7,1} For example, the first value is Y_p = - 1.234 + 0.93925* (4) + 0.60370*(8). So, Y_p = 7.35 Repeat the same process to find out all the values. It is shown in the following table.

X_1	X_2	Y_P
4	8	3.5
5	5	6.4
6	3	6.2
3	7	5.8
2	1	1.24

4.3 Hypothesis Testing

This fragment inspects about hypothesis tests on the backslide coefficients in different direct backslide. Three sorts of theory tests are as often as possible coordinated for various rectilinear backslide models:

1) To determine the significance of a backslide: To assess the importance of the complete backslide model.
2) T-test: This test determines the significance of specific backslide coefficients.
3) F-test: This test may be used to evaluate the significance of several backslides coefficients at the same time. It may also be used to test specific coefficients.

4.4 T-tests in Multiple Linear Regression

H_0: The coefficient for a specific independent variable is 0.

OR

H_0: $\beta = 0 * i$, where $i = 1, 2, \ldots, v$

When all other explanatory factors in the model have been taken into account, this means that the specific explanatory variable being evaluated does not assist explain the outcome.

H_a: Coefficient for a particular independent variable is NOT 0

OR

H_a: $\beta \neq 0 * i$, where $i = 1, 2, \ldots, v$

This construes that the particular free factor being attempted helps explain the response variable in the wake of addressing the effects of the other self-governing components in the model. In different backslide, the coefficients and standard botches of the coefficients for all of the variables are settled reliant on the other intelligent elements being in the model.

We can standardize this exceptional test estimation of β_i into T bits of knowledge that follow a T scattering with levels of chance identical to df = n − k where k is the number of limits in the model. In this model, we have two variables we used, so k = 2.

$$T = \frac{\beta_i - 0}{SE_i} \sim t(df = n - k)$$

where SE_i represents the standard deviation of the distribution of the sample coefficients.

The p-a motivating force for every preliminary variable tests the invalid hypothesis that the variable has no association. If there is no association, there is no connection between the changes in the self-sufficient variable and the developments in the destitute variable. There is a deficiency regarding confirmation to gather that there is an effect at the model level. Actually hand, a p-regard that is more conspicuous than the importance level shows that there's lacking verification in your guide to derive that a non-zero relationship exists.

5.0 POLYNOMIAL REGRESSION

In specific events, it is over the top to anticipate the characteristics with a straight line. We need a best-fit twist, fractal, or spline instead of a direct line along these lines. Test disseminate plot that can't fit in the straight line is showed up in the Figure 4. On the off chance that there ought to be an event of uproar or goofs present in the dataset, we use the condition to demonstrate the assessment of limit as, $f = \sin 2\pi x + \varepsilon$. This is unquestionably not another improvement of the condition; it is comparable to unmistakable condition $y = mx + c$.

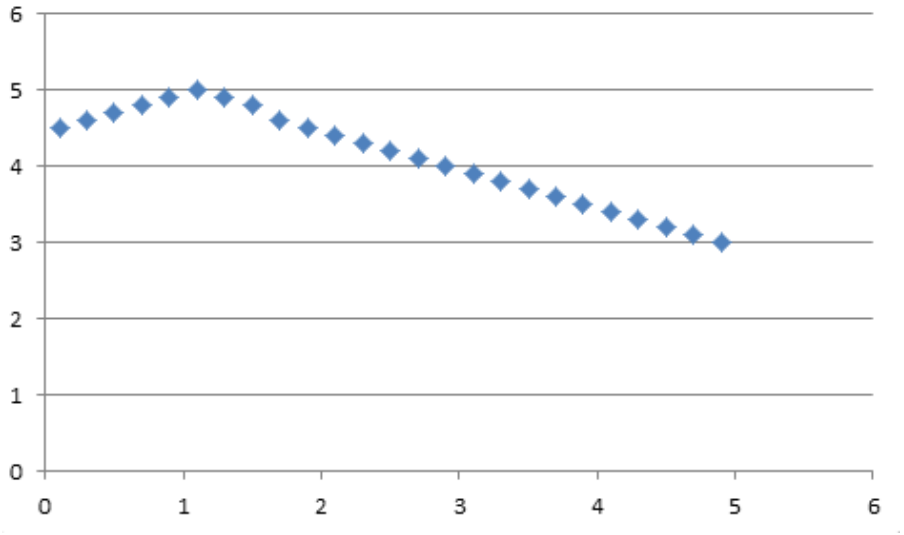

Figure 4 – Example for Polynomial Regression

Where ε denotes error present in data, and $\sin 2\pi$ is the modulated value of the data. Our goal is to minimize the error.

5.1 Overfitting

It is the condition that the model gives the error in data. It happens when the model is more tedious. It uses many terms. The regression coefficients represent the noise rather than relationships (Gao, 2014). To minimize the error, we need polynomial expression such as:

$$Y = m_1x + m_2x^2 + m_3x^3 + \ldots\ldots + \ldots\ldots + m_nx^n$$

$$Y = \sum_{i=1}^{i=d} m_i x^i + c$$

Where $m_1, m_2, \ldots, m_n$ are the model parameters power of $x = 1,2,3, \ldots, n$ are the hyperparameters.

5.2 Mathematical Implementation

The example data is present in the Table 12, and the scatter plot for the sample data is shown in the Figure 5.

Table 12 – Sample Data of Xi and Yi

X_i	Y_i
1	1
3	6
4	1
7	8
9	20

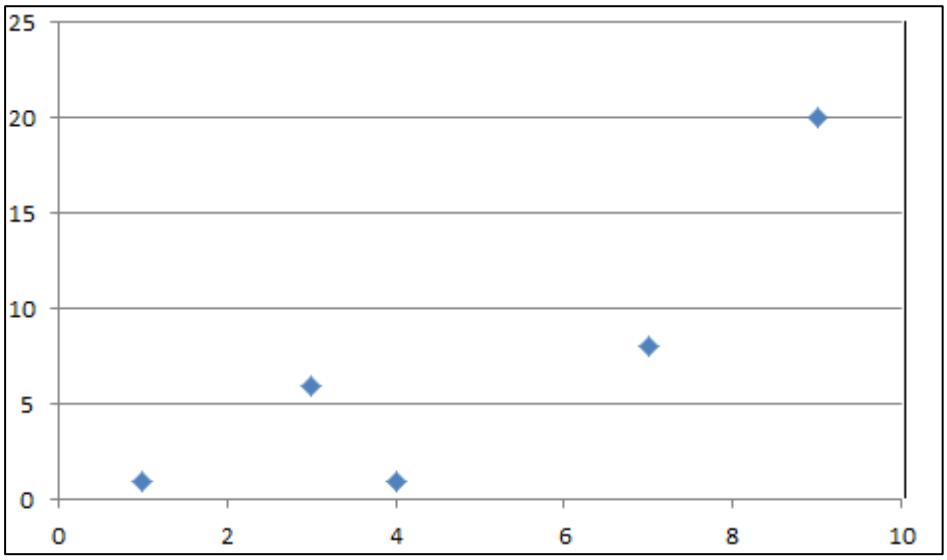

Figure 5 – Scatter Plot for the Sample Data of Xi and Yi

From the above scatter plot, we know that putting the linear line is difficult. So, predicting the value must have higher error rates. To know the best fit line for the above sample data, we must use polynomial regression.

First, we will calculate the linear best fit line. We take the polynomial equation as $y = b_0 + b_1x$. Where b_0 is the y-intercept and b_1 is the slope.

To find this, we must use the following matrix form

$$\begin{bmatrix} n & \Sigma x_i \\ \Sigma x_i & \Sigma x_i^{\,2} \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \begin{bmatrix} \Sigma y_i \\ \Sigma x_i y_i \end{bmatrix}$$

Using the above matrix, we have to calculate b_0 and b_1. Before going to do, we must know the value of Σx_i, Σy_i, $\Sigma x_i^{\,2}$, Σy_i *and* $\Sigma x_i y_i$. We will see all the values in the following table.

X_i	Y_i	X_i^2	Y_i^2	$X_i * Y_i$
1	1	1	1	1
3	6	9	36	18
4	1	16	1	4
7	8	49	64	56
9	20	81	400	180
$\Sigma X_i = 24$	**$\Sigma Y_i = 36$**	**$\Sigma X_i^2 = 156$**	**$\Sigma Y_i^2 = 502$**	**$\Sigma X_i * Y_i = 259$**

Let us fill it in the matrix form from the above table values, here n = 5 because we take 5 values only.

$$\begin{bmatrix} 5 & 24 \\ 24 & 156 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \begin{bmatrix} 36 \\ 259 \end{bmatrix}$$

We can simplify the matrix to find out the value of b_0 and b_1. The value of b_0 and b_1 are as follows.

$$\mathbf{b_0 = (-\ 2.94)}\ \textit{and}\ \mathbf{b_1 = 2.11}$$

Apply the values in the equation $y = b_0 + b_1x$

$$Y = (-2.94) + 2.11X \text{ (This is linear least square fit)}$$

It tells the slope is negative, and the intercept is negative.

Let us find out the predicted value for y for the dataset X= {1, 3, 4, 7, 9}. It is shown in the following table.

X	Y_P
1	-0.828
3	3.397
4	5.509
7	11.848
9	16.073

Our goal is to find the errors. To find the error value we must find $(Y- Y_P)^2$. It is displayed in the following table.

X	Y	Y_P	$(Y- Y_P)^2$
1	1	-0.828	3.343
3	6	3.397	6.775
4	1	5.509	20.338
7	8	11.848	14.807
9	20	16.073	15.417
$\Sigma(Y- Y_P)^2$			**60.68**

To find out the mean square error, use the following formula

$$\text{MSE} = \frac{\Sigma(y-yp)2}{n}$$

Where, MSE = Mean Square Error

Let us substitute the values

$$\text{MSE} = \frac{60.68}{5}$$

The answer is **12.136** (First Order)

Let us continue the following example to solve it in quadratic square fit. Then we will see the value is increased or decreased in the second order. We will predict the value with quadratic least fit. In this, we must include second-order polynomial X^2. Then the equation will be $y=b_o+b_1x+b_2x^2$

To predict the value, we must use the following matrix

$$\begin{bmatrix} n & \Sigma x_i & \Sigma {x_i}^2 \\ \Sigma x_i & \Sigma {x_i}^2 & \Sigma {x_i}^3 \\ \Sigma {x_i}^2 & \Sigma {x_i}^3 & \Sigma {x_i}^4 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} \Sigma y_i \\ \Sigma x_i y_i \\ \Sigma {x_i}^2 y_i \end{bmatrix}$$

Using the above matrix, we have to calculate b_0, b_1, and b_2. Before going to do, we must know the value of Σx_i, Σy_i, $\Sigma {x_i}^2$, $\Sigma {x_i}^3$, Σy_i, $\Sigma {x_i}^4$, $\Sigma x_i y_i$ and $\Sigma {x_i}^2 y_i$. We will see all the values in the following table.

X_i	Y_i	X_i^2	X_i^3	X_i^4	X_i*Y_i	$X_i^2*Y_i$
1	1	1	1	1	1	1
3	6	9	27	81	18	54
4	1	16	64	256	4	16
7	8	49	343	2401	56	392
9	20	81	729	6561	180	1620
24	**36**	**156**	**1164**	**9300**	**259**	**2083**

From the above table values, let us fill it in the matrix form.

$$\begin{bmatrix} 5 & 24 & 156 \\ 24 & 156 & 1164 \\ 156 & 1164 & 9300 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} 36 \\ 259 \\ 2083 \end{bmatrix}$$

We can simplify the matrix to find out the value of b_0, b_1, and b_2. The value of b_0, b_1 and b_2 are as follows.

$$\mathbf{b_0 = 4.08}, \mathbf{b_1 = -1.93} \textit{ and } \mathbf{b_2 = 0.39}$$

Apply the values in the equation $y=b_0+b_1x+b_2x^2$

$$Y = (4.08) + (-1.93)X + (0.39)X^2 \text{ (This is linear least square fit)}$$

Let us find out the predicted value for y for the dataset x= {1,3,4,7,9}. It is shown in the following table.

X	Y_P
1	2.542
3	1.852
4	2.700
7	10.022
9	18.882

Our goal is to find the errors. To find the error value, we must find $(Y- Y_P)^2$. It is displayed in the following table.

X	Y	Y_P	$(Y- Y_P)^2$
1	1	2.542	2.379
3	6	1.852	17.205
4	1	2.700	2.892
7	8	10.022	4.088
9	20	18.882	1.248
$\Sigma(Y- Y_P)^2$			27.812

Let us put the values

$$\text{MSE} = \frac{27.812}{5}$$

The answer is **5.5629**(Second Order)

So, we conclude that the error is reduced much. From this, we know as to increase the degree of a polynomial, the error is decreased. It shows it will try to fit as a curve, fractal, or spline. But if the higher-order is increased so much, it leads to overfitting. Overfitting leads to error-prone.

6.0 CONCLUSION

This paper analysed various kinds of immediate backslide examination and explained each type with some model dataset. Subsequently, we explained the little-by-little strategy about how the conjecture variable is resolved, the mathematical condition systems to calculate the assumption variable, and the best way to draw the best fit backslide line using theory testing. Finally, we assume that what are contrasts between straight backslide and polynomial backslide we saw is explained under.

- Polynomial backslide is connected to improving our model's closeness to the data by extending the solicitation for the associations between the segments and the response factors.
- In straight backslide, the condition that portrays the factor-response associations is Y=mx+c, where Y and x are vectors that depict the response variable and the factor variable, respectively. m and c are insinuated as the inclination and they catch of this straight condition.
- For polynomial cases, we would use higher powers of x to explain Y, as portray in $Y=m_1x+m_2x_2+c$ where m_1, m_2 are coefficients of the first and second powers of the factor. Thusly inside the polynomial backslide case, we mean to find if there are higher-demand associations among X and Y, past the immediate associations. We research that the higher solicitation associations are to improve model fit when we've inconvenience in building direct models to explain the case. Note that up to this point, we have pretty recently examined one factor (X) and its relationship with the response Y. In an alternate relationship case, we're intrigued about the impact of one yet various component on the response variable. This is now and again ordinarily illustrative of veritable issues an extraordinary stock single-factor versus response model as depicted beforehand.

7.0 FUTURE WORK

Linear regression is a crucial tool for statistical analysis. It includes the relationship between the description, estimation, and prognostication. This system has many applications, but it also has prerequisites and limitations that have always been considered to interpret finding the dependent variable with the independent variables. In our next article, we will discuss implementing these three processes in anyone of the real-time applications using python programming language and getting an accurate statistical value.

REFERENCES

Brownlee, J., (2016, Mar) *"Master Machine Learning Algorithms: Discover How They Work and Implement them from Scratch"*, United States: Jason Brownlee.

Gao, J., (2014) *"Machine Learning Applications for Data Center Optimization"*. Available from https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/42542.pdf.

Kumari, K., & Yadav, S., (2018) "Linear Regression Analysis Study". *Journal of the Practice of Cardiovascular Sciences,* 4(1), pp. 33-36. Available from https://www.j-pcs.org/text.asp?2018/4/1/33/231939.

Raftery, A. E., Madigan, D., & Hoeting, J. A., (1997) "Bayesian Model Averaging for Linear Regression Models", *Journal of the American Statistical Association,* 92(437), pp. 179-191. https://doi.org/10.1080/01621459.1997.10473615.

Schneider, A., Hommel, G., & Blettner, M., (2010, Nov) "Linear Regression Analysis: Part 14 of a Series on Evaluation of Scientific Publications". *Deutsches Ärzteblatt International,* 107(44), pp. 776-782. https://doi.org/10.3238/arztebl.2010.0776.

Seber, G. A. F., & Lee, A. J., (2003, Jan) *"Linear Regression Analysis"*, 2nd Edition, in: Book Series of Wiley Series in Probability and Statistics, A John Wiley & Sons Publication, New Jersey. https://doi.org/10.1002/9780471722199.

Tanaka, H., Uejima, S., & Asai, K., (1982, Nov) "Linear Regression Analysis with Fuzzy Model", *IEEE Transactions on Systems, Man, and Cybernetics,* 12(6), pp. 903-907. https://doi.org/10.1109/TSMC.1982.4308925.

Weisberg, S., (2013, Dec) *"Applied Linear Regression"*, 4th Edition, in: Book Series of Wiley Series in Probability and Statistics, A John Wiley & Sons Publication, New Jersey. Available from https://www.wiley.com/en-in/Applied+Linear+Regression%2C+4th+Edition-p-9781118386088.

Zou, K. H., Tuncali, K., & Silverman, S. G., (2003, Jun) "Correlation and Simple Linear Regression", *Radiology,* 227(3), pp. 617-622. https://doi.org/10.1148/radiol.2273011499.