International Journal for Research in Engineering Application & Management (IJREAM)
ISSN : 2454-9150 Vol-06, Issue-05, Aug 2020

Linear Regression Model to Predict Number of Deaths for COVID-19 using OLS

[1]Archana Soni, [2]Dr. Sanjay Thakur, [3]Vijay Kumar Verma

[1]M. Tech. (C.S.E.) 4th Semester, [2]Professor, [3]Assistant Professor L.K.C.T Indore M.P. India.

[1]archi.soni123@gmail.com, [2]drsanjay2009@rediffmail.com, [3]vijayvermaonline@gmail.com

Abstract - Regression analysis is a form of predictive modeling technique which finds the relationship between a dependent and independent variable or predictor. This technique is used for forecasting, time series modeling and finding the causal effect relationship between the variables. Regression analysis is an important tool for modeling and analyzing data. Regression analysis is used to fit a curve or line to the data points, in such a manner that the differences between the distances of data points from the curve or line is minimized. There are various types of regression techniques available to for making predictions linear regression is one of them.. These techniques are mostly driven by three metrics number of independent variables, type of dependent variables and shape of regression line. In linear regression, the relationships are modeled using linear predictor functions whose unknown model parameters are estimated from the data. Such models are called linear models. In this paper we used linear regression model to predict number of death for COVID-19 infected patient. We used two models to find predicted error. We used real life data set for death prediction for three week data. We found that OLS method minimizes predicted error as compared to SSE method.

Keywords —Regression, COVID, Death, Prediction, Dependent, Independent, Variables

I. INTRODUCTION

Regression analysis is a form of predictive modeling technique which finds the relationship between a dependent and independent variable or predictor. This technique is used for forecasting, time series modeling and finding the causal effect relationship between the variables. Regression analysis is an important tool for modeling and analyzing data. Regression analysis is used to fit a curve or line to the data points, in such a manner that the differences between the distances of data points from the curve or line is minimized.

There are various types of regression techniques available to for making predictions linear regression is one of them.. These techniques are mostly driven by three metrics number of independent variables, type of dependent variables and shape of regression line. In linear regression, the relationships are modeled using linear predictor functions whose unknown model parameters are estimated from the data. Such models are called linear models. Linear Regression establishes a relationship between dependent variable (Y) and one or more independent variables (X) using a best fit straight line (also known as regression line).

It is represented by an equation $Y=a + b\times X + e$, where a is intercept, b is slope of the line and e is error term. This equation can be used to predict the value of target variable based on given predictor variable(s).

II. BUILDING A LINEAR REGRESSION MODEL

Two most common techniques through which a Linear Regression model is built are :

A. Ordinary Least Squares
B. Gradient Descent
C. Regularization

A. Ordinary Least Squares Method

Ordinary Least Squares method is used for multiple linear regressions. The OLS method corresponds to minimizing the sum of square differences between the observed and predicted values

B. Gradient Descent

When there are one or more inputs you can use a process of optimizing the values of the coefficients by iteratively minimizing the error of the model on your training data.

This operation is called Gradient Descent and works by starting with random values for each coefficient. The sum of the squared errors is calculated for each pair of input and output values. A learning rate is used as a scale factor and the coefficients are updated in the direction towards minimizing the error.

 DOI : 10.35291/2454-9150.2020.0556

C. *Regularization*

There are extensions of the training of the linear model called regularization methods. These seek to both minimize the sum of the squared error of the model on the training data (using ordinary least squares) but also to reduce the complexity of the model (like the number or absolute size of the sum of all coefficients in the model).

III. LINEAR REGRESSION MODEL

The regression bit is there, because what you're trying to predict is a numerical value.

There are a few concepts to unpack here:

- Dependent Variable
- Independent Variable(s)
- Intercept
- Coefficients

In a Simple Linear Regression Model with single x and y, the form of the model would be-

Properties of Regression Line

Here are a few features a regression line has:

- Regression passes through the mean of independent variable (x) as well as mean of the dependent variable (y).
- Regression line minimizes the sum of "Square of Residuals". That's why the method of Linear Regression is known as "Ordinary Least Square (OLS)". We will discuss more in detail about Ordinary Least Square later on.
- B_1 explains the change in Y with a change in x by one unit. In other words, if we increase the value of 'x' it will result in a change in value of Y.

IV. LITERATURE SURVEY

In 2011 Turóczy Zsuzsannaa and Liviu Mariana proposed "Multiple regression analysis of performance indicators in the ceramic industry". They proposed a study in PhD thesis, which has the aim of enhancing the performances of industrial enterprises with mathematical models. The main goal is to increase the competitiveness, flexibility, adaptability and reactivity of enterprises in the ceramic industry. Since the ceramic sector represents an important part in the manufacturing industry, we focused on this sector, with the aim of evaluating the development of enterprises activating in this domain. The importance of this research lies in its uniqueness and effectiveness, as the performance indicators were analyzed with multiple regression analysis, in the case of an enterprise that produces technical ceramic products[1].

In 2013 Gianie Abdu, Purwanto "Analysis of Consumer Behavior Affecting Consumer Willingness to Buy in 7-Eleven Convenience Store". They proposed a method to find the relationship between the consumer behavior variables (cultural factor, social factor, personal factor and psychological factor) to the consumer willingness to buy a product in 7-Eleven convenience stores. Data was analyzed by using quantitative analysis. The interpretation of this research shows that the variables and dimensions of consumer behavior has a relationship with the consumer's willingness to buy a product in 7-Eleven Jatiwaringin, Jakarta, even so, there are some variables that has a relationship but not affecting the willingness to buy significantly. The variables that is mostly affecting the willingness to buy in this research shows that social factors giving more affect more that any variables among the consumer behavior variables[2].

In 2014 Kosuke Imai and Bethany Park proposed "Using the Predicted Responses from List Experiments as Explanatory Variables in Regression Mode". They address this gap by first improving the performance of a naive two-step estimator. Despite its simplicity, this improved two-step estimator can only be applied to linear models and is statistically inefficient. They develop a maximum likelihood estimator that is fully efficient and applicable to a wide range of models. They use a simulation study to evaluate the empirical performance of the proposed methods. They also apply them to the Mexico 2012 Panel Study and examine whether vote-buying is associated with increased turnout and candidate approval. The proposed methods are implemented in open-source software[3].

In 2015 Supichaya Sunthornjittanon "Linear Regression Analysis on Net Income of an Agrochemical Company in Thailand." . They analyze the ABC Company's data and verify whether the regression analysis methods and models would work effectively in the ABC Company based in Bangkok, Thailand. After the data are collected, models are created to examine the contribution of each of the company's financial factors to the net income of the company. The final model is selected using Stepwise Regression Methods. A linear regression line and equation for the model are generated to help observe and predict future trends. The model also shows which variables play the most important roles in the. After model selection method is processed, the consensus has shown that only the Income from Fungicide plays a statistically significant role in the net income of the company [4].

In 2016 Sandhya Jain , Sunny Chourse proposed "Regression Analysis – Its Formulation and Execution In Dentistry". Regression analysis is one such concept which explores the relationship between two or more quantifiable variables so that one variable can be predicted from other. Their aim article is to provide a simple yet holistic approach to the understanding of the concepts of Regression Analysis along with its use and misuse, advantages and disadvantages pertaining to the art and science of dentistry In the formulation and execution of an dental treatment plan, the variables involved in the decision making are often poorly

 DOI : 10.35291/2454-9150.2020.0556

characterized and incompletely validated. For these reasons we have to rely on the mean values or go for a wild guess[5].

In 2017 Radek Silhavy and Petr Silhavy " Analysis and selection of a regression model for the Use Case Points method using a stepwise approach". They investigates the significance of use case points (UCP) variables and the influence of the complexity of multiple linear regression models on software size estimation and accuracy. The best performing model (Model D) contains an intercept, linear terms, and squared terms. The results of several evaluation measures show that this model's estimation ability is better than that of the other models tested. Model D also performs better when compared to the UCP model, whose Sum of Squared Error was 268,620 points on Dataset 1 and 87,055 on Dataset 2. Model D achieved a greater than 90% reduction in the Sum of Squared Errors compared to the Use Case Points method on Dataset 1 and a greater than 91% reduction on Dataset 2 [6].

In 2018 Shen Rong and Zhang Bao-wen proposed " The research of regression model in machine learning field". They analyze the sale of iced products affected by variation of temperature. They collected the data of the forecast temperature last year and the sale of iced products and then conduct data compilation and cleansing. They set up the mathematical regression analysis model based on the cleansed data by means of data mining theory. Regression analysis refers to the method of studying the relationship between independent variable and dependent variable. They call the corresponding library function to predict the sale of iced products according to the variation of temperature, which will provide the foundation for the company to adjust its production each month, or even each week and each day.. Moreover, the other situation as the profit will be affected by the lack of production since the rise of temperature will also be avoided[7].

In 2019 Anjali Pant and R.S. Rajput proposed "Linear Regression Analysis Using R for Research and Development" .The future forecasting opportunities and risks estimation are the most prominent prerequisite for a successful business. Regression analysis can go far beyond forecasting. The linear regression analysis technique is a statistical method that allows examining the linear relationship between two or more quantitative variables of interest. The rationale of the linear regression analysis technique is to predict an outcome based on historical data and finding a linear relationship. They discussed the implementation of linear regression using a statistical computing language R and consider that the suggested approach provides an adequate interpretation of research and business data. Introduction Software. They discussed simple linear regression and multiple linear regression. The chapter covers the fundamentals of linear regression, regression model equation, the test of significance, coefficient of determination, and residual with residual analysis. R is a potent statistical computation tool, all the computation of chapter conducted by using R. They explain R computations for the regression model with the help of two examples. Regression model also visualized with the help of some plots that are created with the help of R[8].

In 2020 Khushbu Kumari, Suniti Yadav "Linear Regression Analysis Study". Linear regression is a statistical procedure for calculating the value of a dependent variable from an independent variable. Linear regression measures the association between two variables. It is a modeling technique where a dependent variable is predicted based on one or more independent variables. Linear regression analysis is the most widely used of all statistical techniques. They explain the basic concepts and explain how we can do linear regression calculations in SPSS and excel. The techniques for testing the relationship between two variables are correlation and linear regression. Correlation quantifies the strength of the linear relationship between a pair of variables, whereas regression expresses the relationship in the form of an equation. They used simple examples and SPSS and excel to illustrate linear regression analysis and encourage the readers to analyze their data by these techniques [9].

In 2020 Samit Ghosal , Sumit Sengupta and Milan Majumder proposed "Linear Regression Analysis to predict the number of deaths in India due to SARS-CoV-2 at 6 weeks from day 0 (100 cases - March 14th 2020)" .No valid treatment or preventative strategy has evolved till date to counter the SARS CoV 2 (Novel Coronavirus) epidemic that originated in China in late 2019 and have since wrought havoc on millions across the world with illness, socioeconomic recession and death. They analysis tracing a trend related to death counts expected at the 5th and 6th week of the COVID-19 in India. Material and methods: Validated database was used to procure global and Indian data related to corona virus and related outcomes. Multiple regression and linear regression analyses were used interchangeably. Since the week 6 death count data was not correlated significantly with any of the chosen inputs, an auto-regression technique was employed to improve the predictive ability of the regression model[10].

V. PROBLEM STATEMENT

Regression analysis is the study of two variables in an attempt to find a relationship, or correlation A regression line is a straight line that attempts to predict the relationship between two points, also known as a trend line or line of best fit. Linear regression is a prediction when a variable (y) is dependent on a second variable (x) based on the regression equation of a given set of data.

In this paper we proposed we have taken data of corona virus patient. We have taken two variable age and death based on survey. Based on various survey we found that

 DOI : 10.35291/2454-9150.2020.0556

most of the person‘s death which have infected with COVID-19 are aged person (age greater than 55 year). To predict future death of COVID infected person we have following problems.

1. How to select following variables
   - Dependent Variable
   - Independent Variable(s)
   - Intercept
   - Coefficients

2. How the model Select data from the past to learn what's the relationship.

3. How minimize the predicted error for linear regression.

4. Find the line (hyper plane) that minimizes the vertical offsets

VI. OBJECTIVES

Our objects are

1. To best fit line is considered to be the line for which the error between the predicted values and the observed values is minimum. It is also called the regression line and the errors are also known as residuals.
2. Use OLS method and chooses the parameters for a linear function, the goal of which is to minimize the sum of the squares of the difference of the observed variables and the dependent variables.
3. By using Real life data set apply the proposed approach and compared it with predicted values.

VII. PROPOSED APPROACH

The underlying assumptions for linear regression are:

a. The values of independent variable “x” are set by the researcher
b. The independent variable “x” should be measured without any experimental error
c. For each value of “x,” there is a subpopulation of “y” variables that are normally distributed up and down the Y axis.
d. The variances of the subpopulations of “y” are homogeneous
e. The mean values of the subpopulations of “y” lie on a straight line, thus implying the assumption that there exists

Proposed algorithm has following steps

1. Draw the scatter plot.
   1) Linear or non-linear pattern of the data
   2) Deviations from the pattern (outliers).
2. Fit the least-squares regression line to the data and check the assumptions of the model by looking at the Residual Plot and normal probability plot (for normality assumption). If the assumptions of the model appear not to be met, a transformation may be necessary.
3. Use OLS techniques to transform the data and re-fit the least-squares regression line using the transformed data.
4. If a transformation was done check with minimum error.
5. Once a “good-fitting” model is determined, write the equation of the least-squares regression line. Include the standard errors of the estimates.

VIII. ILLUSTRATE WITH EXAMPLE

Linear regression is a statistical method of finding the relationship between independent and dependent variables. Let us take a simple data for death prediction for COVID-19 virus patients to explain the linear regression model.

Table 1 Sample COVID-19 patient data

S.NO	No of Deaths (Independent variable)	Age groups (Dependent variable)
1	2	15
2	3	28
3	5	42
4	13	64
5	8	50
6	16	90
7	11	58
8	1	8
9	9	54

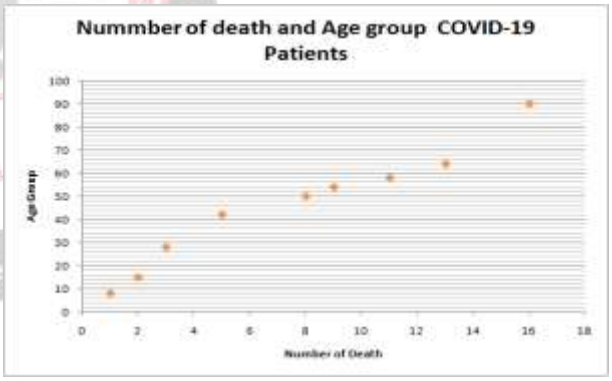

Figure 1 Scatter Plot Diagram

IX. IMPLEMENTATION RESULT AND ANALYSIS

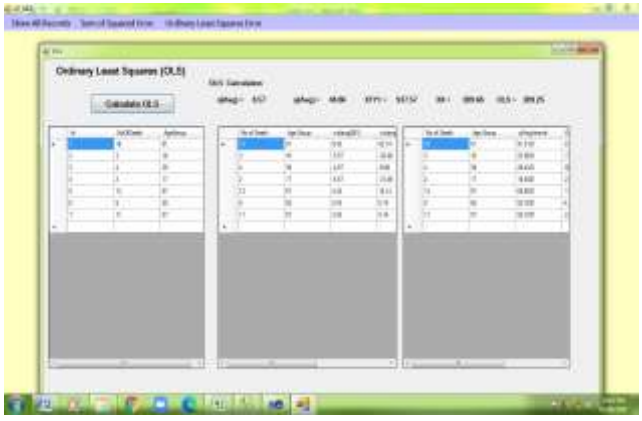

Figure 2 Third week death prediction and OLS error

 DOI : 10.35291/2454-9150.2020.0556

Number of COVID-19 patients and predicted death for first week for 1086 patients

We analyze COVID -19 infected patients and number of death for different age group for three week. Now we have taken 3252 COVID-19 infected patient for three week and prediction death for different age group using linear regression model. First week total number of infected patient are 1086 and predicted death are 58.

Table2 Number of death predicted for first week

S. NO	No of Deaths	Age group
1	2	15
2	3	28
3	5	42
4	13	64
5	8	50
6	16	90
7	11	58

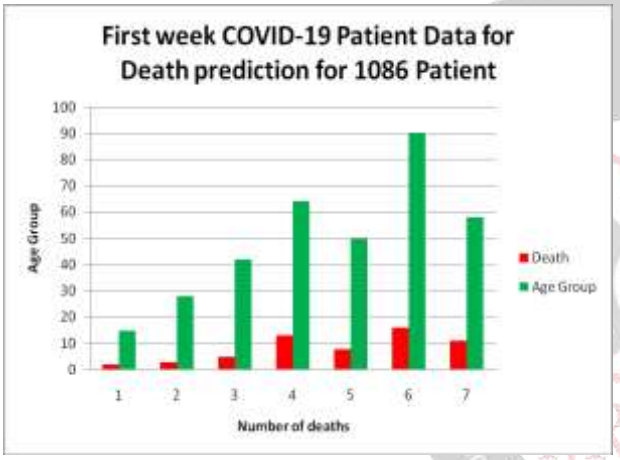

Figure 3 Number of predicted death for first week for different age group

X. CONCLUSION

In the proposed work we analyze COVID -19 infected patients and number of death for different age group for three week. Now we have taken 3252 COVID-19 infected patient and prediction death for different age group using linear regression model. We have taken three week data and predict number of death for different age group and find the error using SSE and OLS. By the experimental analysis we found that OLS method give more correct prediction a compared to SSE method. In future we can apply Gradient Descent Algorithm for optimization of proposed approach

REFERENCES

[1] Turóczy Zsuzsannaa,, Liviu *Mariana "Multiple regression analysis of performance indicators in theceramic industry"*. The Authors. Published by Elsevier Ltd. Selection and peer review under responsibility of Emerging Markets Queries in Finance and Business local organization. doi: 10.1016/S2212-5671(12)00188-8

[2] Gianie Abdu, Purwanto *Analysis of Consumer Behavior Affecting Consumer Willingness to Buy in 7-Eleven Convenience Store* Universal Journal of Management 1(2): 69-75, 2013 http://www.hrpub.org DOI: 10.13189/ujm.2013.010205

[3] Kosuke Imai *"Using the Predicted Responses from List Experiments as Explanatory Variables in Regression Models Advance Access"* publication November 11, 2014 Political Analysis (2015).

[4] Supichaya Sunthornjittanon *"Linear Regression Analysis on Net Income of an Agrochemical Company in Thailand"* Portland State University PDXScholar University Honors Theses University Honors College.

[5] Sandhya Jain , Sunny *Chourse "Regression Analysis – Its Formulation and Execution In Dentistry"*. Journal of Applied Dental and Medical Sciences NLM ID: 101671413 ISSN:2454-2288 Volume 2 Issue 1 January - March 2016

[6] Radek Silhavy , Petr Silhavy, Zdenka Prokopova *" Analysis and selection of a regression model for the Use Case Points method using a stepwise approach"* The Journal of Systems and Software 125 (2017 Science Direct The Journal of Systems and Software journal homepage: www.elsevier.com/locate/jss.

[7] Shen Rong, Zhang Bao-*wen " The research of regression model in machine learning field* "MATEC Web of Conferences 176, 01033 (2018) doi.org/10.1051/matecconf/2018.

[8] Anjali Pant R.S. Rajput *"Linear Regression Analysis Using R for Research and Development"* See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/

[9] Khushbu Kumari, Suniti Yadav "Linear Regression Analysis Study" [Downloaded free from http://www.j-pcs.org on Friday, July 17, 2020, IP: 157.34.76.130] Journal of the Practice of Cardiovascular Sciences .

[10] Samit Ghosal , Sumit Sengupta "Linear Regression Analysis to predict the number of deaths in India" due to SARS-CoV-2 at 6 weeks from day 0 (100 cases - March 14th 2020) Clinical Research & Reviews journal .

 DOI : 10.35291/2454-9150.2020.0556