Linear Regression Model Minimizing Prediction Error Using SSE and OLSE a Comparison

Deepshikha Patidar[1] Mr. Ramiz Sheikh[2]
[1]P.G. Research Scholar [2]Assistant Professor
[1,2]Department of Computer Science and Engineering
[1,2]JIT Borawan Khargone, India

Abstract— Regression analysis is a form of predictive modeling technique which investigates the relationship between a dependent or target 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 / 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. Regression analysis estimates the relationship between two or more variables. There are multiple benefits of using regression analysis. It indicates the significant relationships between dependent variable and independent variable. It indicates the strength of impact of multiple independent variables on a dependent variable. In this paper we proposed apply Ordinary Least Square approach in linear regression model to minimize error in prediction. our object is to find a line so that all object are best fit around the predicted line.

Keywords: SSE, OLSE, Linear Regression

I. INTRODUCTION

In statistics and machine learning, linear regression is one of the most popular and well understood algorithms. Regression is a statistical method used in finance, investing, and other disciplines that attempts to determine the strength and character of the relationship between one dependent variable (usually denoted by Y) and a series of other variables (known as independent variables [6,8,9].

Linear Regression is one of the machine learning algorithms where the result is predicted by the use of known parameters which are correlated with the output. It is used to predict values within a continuous range rather than trying to classify them into categories. The known parameters are used to make a continuous and constant slope which is used to predict the unknown or the result.

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. Most commonly, the conditional mean of the response given the values of the explanatory variables (or predictors) is assumed to be an affine function of those values; less commonly, the conditional median or some other quantile is used. Like all forms of regression analysis, linear regression focuses on the conditional probability distribution of the response given the values of the predictors, rather than on the joint probability distribution of all of these variables, which is the domain of multivariate analysis.

Linear regression has many practical uses. Most applications fall into one of the following two broad categories:

- If the goal is prediction, forecasting, or error reduction,[clarification needed] linear regression can be used to fit a predictive model to an observed data set of values of the response and explanatory variables. After developing such a model, if additional values of the explanatory variables are collected without an accompanying response value, the fitted model can be used to make a prediction of the response.
- If the goal is to explain variation in the response variable that can be attributed to variation in the explanatory variables, linear regression analysis can be applied to quantify the strength of the relationship between the response and the explanatory variables, and in particular to determine whether some explanatory variables may have no linear relationship with the response at all, or to identify which subsets of explanatory variables may contain redundant information about the response.

II. ASSUMPTIONS OF LINEAR REGRESSION

There are four principal assumptions which justify the use of linear regression models for purposes of inference or prediction

1) linearity and additivity of the relationship between dependent and independent variables. The expected value of dependent variable is a straight-line function of each independent variable, holding the others fixed. The slope of that line does not depend on the values of the other variables. The effects of different independent variables on the expected value of the dependent variable are additive.
2) Statistical independence of the errors, in particular, no correlation between consecutive errors in the case of time series data.
3) Homoscedasticity constant variance of the errors, versus time in the case of time series, data, versus the predictions, versus any independent variable
4) Normality of the error distribution. If any of these assumptions is violated , if there are nonlinear relationships between dependent and independent variables or the errors exhibit correlation, heteroscedasticity, or non-normality, then the forecasts, confidence intervals, and scientific insights yielded by a regression model may be at best inefficient or at worst seriously biased or misleading.

III. TECHNIQUES TO BUILD A LINEAR REGRESSION MODEL

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

1) Ordinary Least Squares
2) Gradient Descent
3) Regularization

1) 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

2) 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.

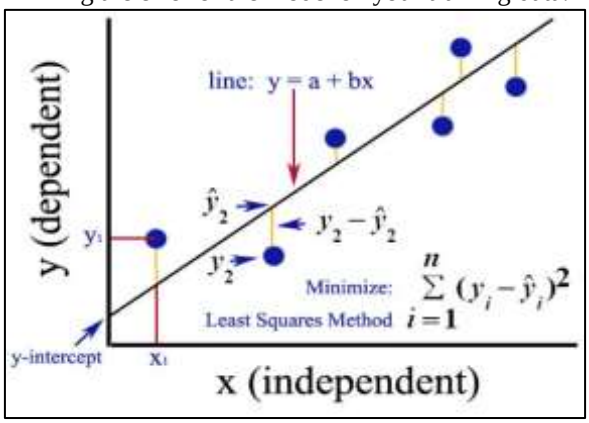

Fig. 1: Least square method

3) 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).

Two popular examples of regularization procedures for linear regression are:

- Lasso Regression: where Ordinary Least Squares is modified to also minimize the absolute sum of the coefficients (called L1 regularization).
- Ridge Regression: where Ordinary Least Squares is modified to also minimize the squared absolute sum of the coefficients (called L2 regularization).

IV. APPLICATIONS OF LINEAR REGRESSION

Linear Regression is a very powerful statistical technique and can be used to generate insights on consumer behavior, understanding business and factors influencing profitability. Linear regressions can be used in business to evaluate trends and make estimates or forecasts. For example, if a company's sales have increased steadily every month for the past few years, by conducting a linear analysis on the sales data with monthly sales, the company could forecast sales in future months[26,27].

Linear regression can also be used to analyze the marketing effectiveness, pricing and promotions on sales of a product. The beauty of linear regression is that it enables us to capture the isolated impacts of each of the marketing campaigns along with controlling the factors that could influence the sales.

Linear Regression can be also used to assess risk in financial services or insurance domain. For example, a car insurance company might conduct a linear regression to come up with a suggested premium table using predicted claims to Insured Declared Value ratio. In the credit card industry, a financial company maybe interested in minimizing the risk portfolio and wants to understand the top five factors that cause a customer to default. Based on the results the company could implement specific EMI options so as to minimize default among risky customers.

V. LITERATURE SURVEY

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 store. Vriables among the consumer behavior variables. The coefficient of determination (R2) is l 0.405 40.5% of the variation that occurs in the willingness to buy is described by all of the independent variables, cultural factor, social factor, personal factor, psychological factor, while the remaining 100% - 40.5% = 59.5% are explained by other variables which is not described in this research[1].

In 2014 Lin Yu 'A study of English reading ability based on multiple linear regression analysis". They contribute the exploration of ways of improving English reading ability by analyzing the influencing factors upon English reading ability. They started with questionnaires on the influencing factors upon English reading ability and analyses of exam questions evaluating reading ability, and discover that the subject factor, detail factor, inference factor, attitude factor and semantic factor are the major influencing factors upon English reading ability. Finally, based on the multiple linear regression models, a way of improving English reading ability by opportunely using intensive and extensive reading skills is presented [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. [3].

In 2015 Supichaya Sunthornjittan "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. 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. After running these additional models, it was discovered that the Net Income from Fungicide remained the most significant model. Time is also taken into account to see if it plays some role in the net

income, but after the analysis, it was found that time is not significant in this case[4].

In 2016 Sandhya Jain, Sunny Chourse proposed "Regression Analysis – Its Formulation and Execution in Dentistry". Prediction and estimation is the mainstay in the treatment planning in dentistry. With variations being common is many events of the oral cavity, it becomes important to have a methodology which can help us predict the happenings of the region in relation to each other. The regression analysis is a statistical technique that deals with the analysis of relationship between such variables. They help us to predict the duration, course and outcome of the treatment parameters. It is a complex process of predicting/estimating the magnitude of some unknown characteristics which might be involved in the growth and treatment of a given patient[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. 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 Ira Sharma and Sampurna Kakchapati proposed "Linear Regression Model to Identify the Factors Associated with Carbon Stock in Chure Forest of Nepal". The highest carbon stock was associated with sv more than 199m3/ha, average dbh more than 43.3 cm/plot, and number of trees more than 20/plot, whereas the altitude, geographical location, and ownership had no statistical associations at all. The results can be of use to the government for enhancing carbon stock in Chure that supports both natural resource conservation and United Nations-Reducing Emission from Deforestation and Forest Degradation program to mitigate carbon emission issues[7].

In 2019 Anjali Pant and R.S. Rajput proposed "Linear Regression Analysis Using R for Research and Development" . 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.. 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 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)" Introduction: and Aims: No valid treatment or preventative strategy has evolved till date to counter 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. [9].

VI. PROPOSED ALGORITHM

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.

VII. IMPLEMENTATION RESULT AND ANALYSIS

We have taken 3252 infected patient with COVID-19 for three week and we indentify number of deaths for COVID-19 patient. We have analysis number of death for different age group for prediction of number of deaths for future. We used SQL server 2010 to store data set. We have taken data from Dainik Bhaskar paper and from MY hospital Indore.
We have taken three week data and predict number of death for different age group and find the error using SSE and OLS. Our comparison is based on minimizing prediction by SSE and OLS.

Number of Patient weekly	Predicted Error using SSE	Predicted Error Using OLS
First week (1086)	3631.68	168.88
Seconds week (1032)	1952.86	151.25
Third week(1134)	4400.86	209.25

Table 5.4 Weekly infected COVID -19 patients' and predicted death error

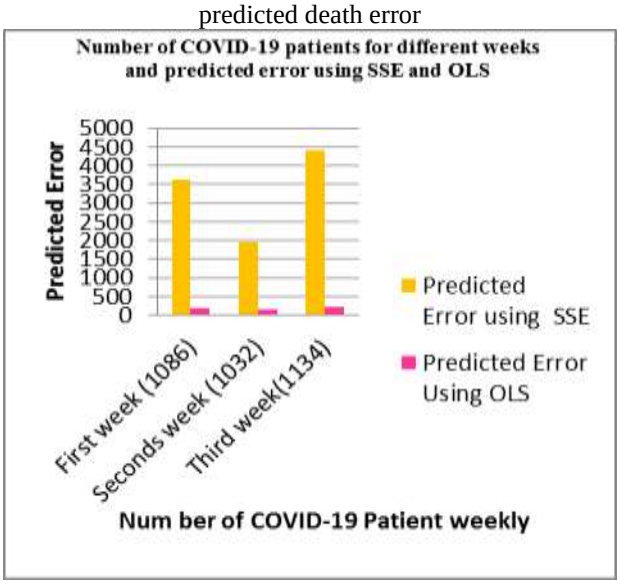

Fig. 5.4: Weekly infected COVID -19 patients' and predicted death error

VIII. CONCLUSION

Regression analysis is a form of predictive modeling technique which investigates the relationship between a dependent or target 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 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.

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.

REFERENCES

[1] 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

[2] Lin Yu "A study of English reading ability based on multiple linear regression analysis" Available online www.jocpr.com Journal of Chemical and Pharmaceutical Research, 2014, 6(6):1870-1877 Zbigniew Suraj Zdzisław Pawlak Polish mathematician and computer scientist, Polish Academy of Sciences and Warsaw University of Technology. Cairo University in December 2004.

[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) 23:180–196doi:10.1093/pan/mpu017.

[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) 1–14 Contents lists available at Science Direct The Journal of Systems and Software journal homepage: www.elsevier.com/locate/jss.

[7] Ira Sharma and Sampurna Kakchapati "Linear Regression Model to Identify the Factors Associated with Carbon Stock in Chure Forest of Nepal" Hindawi Scientific a Volume 2018, Article ID 1383482, 8 pages https://doi.org/10.1155/2018/1383482.

[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/publication/336981868

[9] 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) Contents lists available at Science Direct Diabetes & Metabolic Syndrome: Clinical Research & Reviews journal homepage: www.elsevier.com/locate/dsx

