Volume 9, Issue 08, August 2024 ISSN (online): 2456-0006

International Journal of Science Technology Management and Research

Available online at: www.ijstmr.com

Reducing Problem of over Fitting Due to Several Factors in Linear Regression Model

Naval Singh Bhaghel
M Tech CSE 4th sem
LNCT (Bhopal) Indore Campus
Indore M.P. India

Dikshika Maliwad
Assistant Professor CSE department
LNCT (Bhopal) Indore Campus
Indore M.P. India

Abstract*: The two basic types of regression are simple linear regression and multiple linear regression, although there are nonlinear regression methods for more complicated data and analysis. Simple linear regression uses one independent variable to explain or predict the outcome of the dependent variable Y, while multiple linear regression uses two or more independent variables to predict the outcome. Analysts can use stepwise regression to examine each independent variable contained in the linear regression model. Regression can help finance and investment professionals. For instance, a company might use it to predict sales based on weather, previous sales, gross domestic product (GDP) growth, or other types of conditions. In statistical analysis, regression is used to identify the associations between variables occurring in some data. It can show the magnitude of such an association and determine its statistical significance. Regression is a powerful tool for statistical inference and has been used to try to predict future outcomes based on past observations. An economist may, for example, hypothesize that as a person increases their income, their spending will also increase. If the data show that such an association is present, a regression analysis can then be conducted to understand the strength of the relationship between income and consumption and whether or not that relationship is statistically significant.

Keywords: Linear, Regression , Predict, Dependent , Relationship , Statistical

I. INTRODUCTION

In statistical modeling, **regression analysis** is a set of statistical processes for estimating the relationships between a dependent variable (often called the 'outcome' or 'response' variable, or a 'label' in machine learning parlance) and one or more independent variables (often called 'predictors', 'covariates', 'explanatory variables' or 'features'). The most common form of regression analysis is linear regression, in which one finds the line (or a more complex linear combination) that most closely fits the data according to a specific mathematical criterion. For example, the method of ordinary least squares computes the unique line (or hyperplane) that minimizes the sum of squared differences between the true data and that line (or hyperplane). For specific mathematical reasons (see linear regression), this allows the researcher to estimate the conditional expectation (or population average value) of the dependent variable when the independent variables take on a given set of values. Less common forms of regression use slightly different procedures to estimate alternative location parameters (e.g., quantile regression or Necessary Condition Analysis[1]) or estimate the conditional expectation across a broader collection of non-linear models (e.g., nonparametric regression)[9].

Regression analysis is primarily used for two conceptually distinct purposes. First, regression analysis is widely used for prediction and forecasting, where its use has substantial overlap with the field of machine learning. Second, in some situations regression analysis can be used to infer causal relationships between the independent and dependent variables. Importantly, regressions by themselves only reveal relationships between a dependent variable and a collection of independent variables in a fixed dataset. To use regressions for prediction or to infer causal relationships, respectively, a researcher must carefully justify why existing relationships have predictive power for a new context or why a relationship between two variables has a causal interpretation. The latter is especially important when researchers hope to estimate causal relationships using observational data[10]

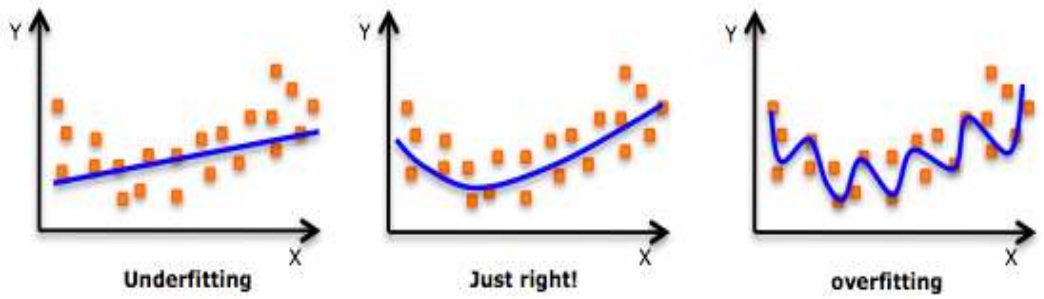

Fig 1 Different curve of regression

II. How Does it Regression analysis

Regression analysis is a way of mathematically sorting out which of those variables does indeed have an impact. It answers the questions: Which factors matter most? Which can we ignore? How do those factors interact with one another? And, perhaps most important, how certain are we about all these factors? In regression analysis, those factors are called "variables." You have your *dependent variable* — the main factor that we're trying to understand or predict. In Redman's example above, the dependent variable is monthly sales. And then we have we *independent variables*— the factors we suspect have an impact on we dependent variable. To conduct a regression analysis, we gather the data on the variables in question. (Reminder: We likely don't have to do this we self, but it's helpful for we to understand the process we data analyst colleague uses.) We take all we monthly sales numbers for, say, the past three years and any data on the independent variables we're interested in. So, in this case, let's say we find out the average monthly rainfall for the past three years as well. Then we plot all that information on a chart that looks like this[8,9]

The y-axis is the amount of sales (the dependent variable, the thing we're interested in, is always on the y-axis), and the x-axis is the total rainfall. Each blue dot represents one month's data—how much it rained that month and how many sales we made that same month. Glancing at this data, we probably notice that sales are higher on days when it rains a lot. That's interesting to know, but by how much? If it rains three inches, do we know how much we'll sell? What about if it rains four inches? Now imagine drawing a line through the chart above, one that runs roughly through the middle of all the data points. This line will help we answer, with some degree of certainty, how much we typically sell when it rains a certain amount.This is called the "regression line," and it's drawn (using a statistics program like SPSS or STATA or even Excel) to show the line that best fits the data. In other words, explains Redman, "The red line is the best explanation of the relationship between the independent variable and dependent variable[7]

Challenges with regression analysis

Correlation does not equal causation. You can show a relationship between any two variables, but that does not prove that one of the variables causes the other. Some people think when they see a positive relationship in a regression analysis that it is a clear sign of cause and effect. However, as we discussed before, regression analysis only shows the relationship between variables, not the cause and effect. You must be careful that you are not making assumptions about relationships that do not actually exist in real life.
The independent variable may be something you can't control. For instance, you know that rain increases sales volumes, but you cannot control the weather. Does that variable even matter? You can control a lot of internal factors; your marketing, store layout, staff behaviour, features and promos. Waiting for it to rain is not a good sales strategy. In the analysis, the Y-axis always contains the dependent variable, or what you are trying to test. In this case, sales figures. The X-axis represents the independent variable, the number of inches of rain. Looking at this simple fictional chart, you can see that sales increase when it rains, a positive correlation. But it doesn't tell you exactly how much you can expect to sell with a certain amount of rainfall. This is when you add a regression line.
This is a line that shows the best fit for the data, and the relationship between the dependent and independent variable. In this example, you can see the regression line intersects the data, showing visually a prediction of what would happen with any amount of rainfall.

ERROR TERMS

Regression analyses do not predict causation, just the relationship between variables. While it is tempting to say that it is obvious that the rainfall level affects sales figures, there's no proof that this is the case. Independent variables will never be a perfect predictor of a dependent variable. The error term is the figure that shows you the certainty with which you can trust the formula. The larger the error term, the less certain that regression line is. The error term might be 50 percent, indicating that variable is no better than chance. Or, it could be 85 percent, showing that there is a significant likelihood the independent variable affects the dependent variable. Correlation does not equal causation – it might not be the rain causing that increase in sales, it could be another independent variable. While the variables seem to be linked, it is possible that there is something else altogether, and only by running multiple analysis will a business be able to gain a clearer understanding of the factors involved. It is almost impossible to predict a direct cause and effect in regression analysis. This is why regression analyses usually include a number of variables, so that it's more likely that you're finding the actual cause of the sales increase or decrease. Of course, including multiple independent variables can create a messy set of outcomes, however good data scientists and statisticians can sort through the data to get accurate results.

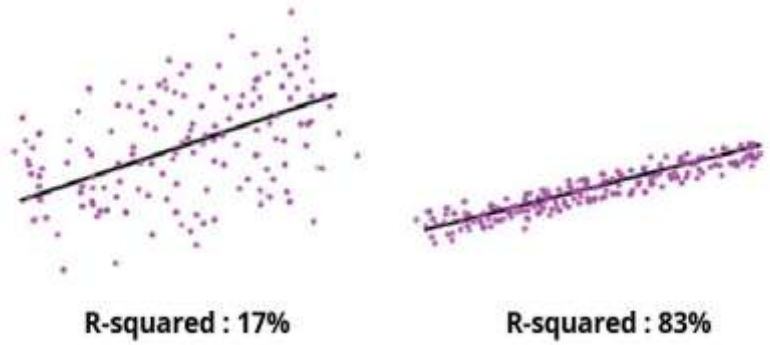

Fig 2 represent the scatter around the regression line.

The other thing that can help is knowledge of the business. The store might sell more products on days with heavier rainfall , but if the data scientists talk to the sales staff, they may find out that more people come in for the free coffee that is given away on rainy days. If that is the case, is the cause of increased sales the rain, or the free coffee? This means the business needs to do a bit of market research.

Asking their customers why they purchased something on a specific day. It may be that the coffee drew them in, the rain made them stay, and then they saw a product they have been intending to buy. Therefore, the cause of increased sales is the rain, but you need to factor in the free coffee too. One without the other will not result in the same outcome.

II. LITERATURE SURVEY

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. Stepwise multiple linear regression models and residual analysis were used to analyze the impact of model complexity. The impact of each variable was studied using correlation analysis.. 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[1].

In 2017 N J Gogtay, S P Deshpande and UM Thatte proposed "Principles of Regression Analysis". Regression analysis is a statistical tool that helps evaluate relationships between a dependent variable and one or more independent or predictor variables. More specifically, it helps us understand how the dependent variable changes with changes in the independent variable and thus finds its application in forecasting and predicting. The technique must however be used with clear understanding of the assumptions in each type of regression analysis, their limitations and the potential error that can occur when models are applied to a larger population. They apply this equation to the population for making a prediction, and able to predict either the systolic blood pressure perfectly. Hence, they need to taken into account an "error" or "deviation" that is likely to occur when this equation is used[2].

In 2017 Erasmus University Rotterdam proposed "Using linear regression to establish empirical relationships". The linear regression model using OLS provides a powerful tool for investigating the relationship between an outcome variable and multiple explanatory variables that are potentially correlated with each other. The impact of one variable can be investigated, controlling for other variables or confounding factors (as long as these are observed). Under relatively weak assumptions, the linear regression model can be interpreted as describing a conditional expectation. By construction, the linear regression model provides the best linear approximation (or the best linear predictor) of the dependent variable. This makes linear regression useful in empirical work, even if there is no behavioral content in the model. A regression can be used to predict the outcome variable in cases where it is not observed and can thus provide a useful tool to answer "what if" questions for policymakers. The specification of a regression model should be chosen carefully and should involve some statistical testing. Carefully specifying the model is particularly crucial if estimates for the coefficients of interest appear very sensitive to the specification used or to the set of explanatory variables included in the regression. Policy makers can use linear regression models to test the impact of a proposed policy intervention. The model can be used to predict an outcome variable after changing one or more of the explanatory variables to reflect the proposed policy intervention[3].

In 2018 Shen Rong and Zhang Bao-wen proposed " The research of regression model in machine learning field". The paper herein will 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. Linear regression model that corresponds to the practical situation is proposed , which is to set up simple linear regression model based on practical problem and then to implement the following with the help of the latest and most popular Python3.6. Python3.6 boasts the features of pure object-oriented, platform independence and concise and elegant language. 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[4].

In 2018 Syarifah Diana Permaia and Heruna Tantyb proposed "Linear regression model using bayesian approach for energy performance of residential building" . Bayesian views a parameter as a random variable, it means the value is not a single value. The modeling method that most commonly used by researchers is linear regression model. The Frequentist methods that are often used in linear regression are Ordinary Least Square (OLS) and Maximum Likelihood Estimation (MLE). Along with the Bayesian development, several studies have shown better modeling results than the Frequentist method. On the other hand, Bayesian approach is also used when assumptions in linear regression model using OLS are not met. They performs linear regression modeling with Bayesian approach. The analysis showed that linear regression model using OLS does not met all assumptions. It means the model is not good enough. Then, Bayesian approach can be used as an alternative for the model. The comparison of Bayesian and Frequentist modeling results using several criteria such as RMSE, MAPE and MAD. The results showed that the linear regression method using Bayesian approach is better than Frequentist method using OLS[5].

In 2018 Katarina Valaskova and Tomas Kliestik, Lucia Svabova proposed "Financial Risk Measurement and Prediction Modeling for Sustainable Development of Business Using Regression Analysis". The issue of the debt, bankruptcy or non-bankruptcy of a company is presented in this article as one of the ways of conceiving risk management. They use the Amadeus database to obtain the financial and accounting data of Slovak enterprises from 2015 and 2016 to calculate the most important financial ratios that may affect the financial health of the company. The main aim of the article is to reveal financial risks of Slovak entities and to form a prediction model, which is done by the identification of significant predictors having an impact on the health of Slovak companies and their future prosperity. Realizing the multiple regression analysis, they identified the significant predictors in conditions of the specific economic environment to

estimate the corporate prosperity and profitability. The results gained in the research are extra important for companies themselves, but also for their business partners, suppliers and creditors to eliminate financial and other corporate risks related to the unhealthy or unfavorable financial situation of the company [6].

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[7].

In 2019 Hazlina Darman, Sarah Musa, and Rajasegeran Ramasamy, proposed "Predicting Students' Final Grade in Mathematics Module using Multiple Linear Regression" Learning analytics is the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs. They proposed, multiple linear regression model is developed to predict the students' score in Final Exam using their assessments' score. The response variable in this model is the students' score in Final Exam and the predictor variables are the assessment components (Test 1 and Test 2). The data were collected from a group of students in School of Actuarial Science, Mathematics, and Qualitative Study (SOMAQS), Asia Pacific University of Technology and Innovation (APU), Malaysia. In this research, a regression model has been developed with the aid of Statistical Package for Social Sciences (SPSS) analysis tool. The graphical representations and tables are presented to illustrate the models The findings from this study has achieved the objective of developing a model that can predict the students' performance in final exam. The analysis has shown that the students who perform well in Test 1 and Test 2 have better chances of getting good scores in final exam, and vice versa[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)" Introduction: and Aims: 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 coronavirus and related outcomes.. According to our analysis, if situation continue in present state; projected death rate (n) is 211 and 467 at the end of the 5th and 6th week from now, respectively. Keeping these projected mortality data in mind, current measured for containment of COVID-19 must be strengthened or supplemented [10].

III. PROPOSED APPROACH

Proposed algorithm has following steps

1. Draw the scatterplot.
   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.

IV. EXPERIMENTAL ANALYSIS

Now we used combination of different feature to find squared value. We take combination of Radio and Newspaper paper to see the effect on sales. We found that the value of R-Squared Value for Flex Hoarding Board and Vehicle wraps paper is 0.33 and the values of Adjusted R-Squared is 0.32. We found that combination of Radio and Newspaper paper have negative affect the sales.

TABLE 1

COD and Adjusted

Features	COD	Adjusted
Model 1	0.9	0.9
Model 2	0.33	0.32
Model 3	0.65	0.65

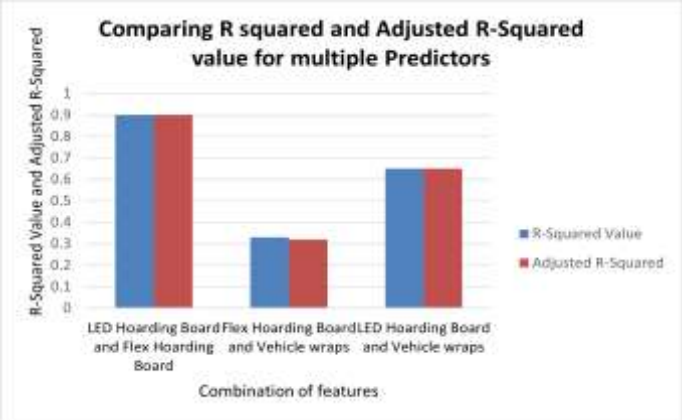

Fig 3 Coefficient of determination

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

REFERENCE

1. 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.
2. NJ Gogtay, SP Deshpande, UM Thatte "Principles of Regression Analysis" Journal of The Association of Physicians of India Vol. 65 April 2017
3. Erasmus University 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.
4. Shen Rong, Zhang Bao-wen " The research of regression model in machine learning field MATEC Web of Conferences 176, 01033 (2018) https://doi.org/10.1051/matecconf/201817601033IFID 2018.
5. Syarifah Diana Permaia, Heruna Tantyb Linear regression model using bayesian approach for energy performance of residential building 2018 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0/) 3rd International Conference on Computer Science and Computational Intelligence 2018.
6. Katarina Valaskova , Tomas Kliestik, Lucia Svabova and Peter Adamko Financial Risk Measurement and Prediction Modelling for Sustainable Development of Business Entities Using Regression Analysis Faculty of Operation and Economics of Transport and Communications, Received: 20 April 2018; Accepted: 20 June 2018; Published: 23 June 2018.
7. 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
8. Hazlina Darman, Sarah Musa Predicting Students' Final Grade in Mathematics Module using Multiple Linear Regression International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-7 Issue-5S, January 2019.
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 | Published by Wolters Kluwer – Medknow.
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) Contents lists available at Science Direct Diabetes & Metabolic Syndrome: Clinical Research & Reviews journal homepage: www.elsevier.com/locate/dsx

11. Gaurav Pandeya, Poonam Chaudharya "SEIR and Regression Model based COVID-19 outbreak predictions in India" Department of CSE & IT, The NorthCap University, India DeenDayalUpadhyaya College, University of Delhi, India Defence Research & Development Organization, India a{Email: gaurav16csu120@ncuindia.edu}.
12. K. K. Baseer, Vikram Neerugatti, Analysing various Regression Models for Data Processing International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-8 June, 2019.
13. Marno Verbeek "Using linear regression to establish empirical relationships Using linear regression to establish empirical relationships". IZA World of Labor 2017: 336 doi: 10.15185/izawol.336 | Marno Verbeek © | February 2017 | wol.iza.org