Volume 9, Issue 09, September 2024 ISSN (online): 2456-0006

International Journal of Science Technology Management and Research

Available online at: www.ijstmr.com

HANDLING PROBLEM OF OVERFITTING FOR ACCURATE MEASURE OF REGRESSION MODEL

Jay Prajapati
M Tech CSE 4th semester
Jawaharlal Institute of Technology Borawan
Khargone (M.P.) 451228

Mr. Kamlesh Patidar
Assistant Professor CSE department
Jawaharlal Institute of Technology Borawan
Khargone (M.P.) 451228

Abstract: ***The Regression Analysis is a part of the linear regression technique. It examines an equation that reduces the distance between the fitted line and all of the data points. Determining how well the model fits the data is crucial in a linear model. A general idea is that if the deviations between the observed values and the predicted values of the linear model are small and unbiased, the model has a well-fit data. In technical terms, "Goodness-of-fit" is a mathematical model that describes the differences between the observed values and the expected values or how well the model fits a set of observations. This measure can be used in statistical hypothesis testing. According to statisticians, if the differences between the observations and the predicted values tend to be small and unbiased, we can say that the model fits the data well. The meaning of unbiasedness in this context is that the fitted values do not reach the extremes, i.e. too high or too low during observations. A linear regression model gives we the outlook of the equation which represents the minimal difference between the observed values and the predicted values. In simpler terms, we can say that r squared linear regression identifies the smallest sum of squared residuals probable for the dataset. Determining the residual plots represents a crucial part of a regression model and it should be performed before evaluating the numerical measures of goodness-of-fit, like R-squared. They help to recognize a biased model by identifying problematic patterns in the residual plots. If we have a biased model, we cannot depend on the results. If the residual plots look good, we can assess the value of R-squared and other numerical outputs.***

Keywords: Regression, R-squared , Observed, Expected ,Observations

I. INTRODUCTION

Regression Analysis is a set of statistical processes that are at the core of data science. In the field of numerical simulation, it represents the well-understood models and helps in interpreting machine learning algorithms. Their real-life applications can be seen in a wide range of domains, ranging from advertising and medical research to agricultural science and even different sports. In linear regression models, r squared interpretation is a goodness-fit-measure. It takes into account the strength of the relationship between the model and the dependent variable. Once we have a fit linear regression model, there are a few considerations that we need to address[8,9]:

- How well does the model fit the data?
- How well does it explain the changes in the dependent variable?

Regression Analysis is a well-known statistical learning technique that allows we to examine the relationship between the independent variables (or explanatory variables) and the dependent variables (or response variables). It requires to formulate a mathematical model that can be used to determine an estimated value that is nearly close to the actual value.

The two terms essential to understanding Regression Analysis:

- Dependent variables - The factors that we want to understand or predict.
- Independent variables - The factors that influence the dependent variable.

Consider a situation where we are given data about a group of students on certain factors: number of hours of study per day, attendance, and scores in a particular exam. The Regression technique allows we to identify the most essential factors, the factors that can be ignored and the dependence of one factor on others.

There are mainly two objectives of a Regression Analysis technique:

- Explanatory analysis - This analysis understands and identifies the influence of the explanatory variable on the response variable concerning a certain model.
- Predictive analysis - This analysis is used to predict the value assumed by the dependent variable.

MOTIVATION

In R-squared coefficient is the determination of a statistical tool, which measures the level of safety of any operation, which may be due to the performance of a particular benchmark indicator. R-squared (R2) is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model[10,11].

Whereas correlation explains the strength of the relationship, between an independent and dependent variable, R-squared explains to what extent the variance of one variable, explains the variance of the second variable. So, if the R2 of a model is 0.50, then approximately half of the observed variation can be explained by the model's inputs.
"Higher the "R-squared" the more variation is explained by our input variables so it is called the good model".

- However, the problem with R-squared is that it can remain the same or increase with the addition of many variants, even if they do not have a relationship with the output variables.
- This is where "Adjusted R square" helps. Adjusted R-square penalizes we for adding variables that do not enhance your existing model.
- Therefore, if we are creating a linear regression of many variables, it is always recommended that we use Adjusted R-squared to judge the beauty of the model. In case we have one input variable, the R-square and Adjusted R squared will be the same.
- Typically, the less important variables add to the model, the gap in R-squared and Adjusted R-squared increases.

The Adjusted R-squared value is similar to the Multiple R-squared value, but it accounts for the number of variables. This means that the Multiple R-squared will always increase. When a new variable is added to the prediction model, but if the variable is a non-significant one, the Adjusted R-squared value will decrease.

II. LITERATURE REVIEW

In 2017 N J Gogtay et al 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. They need to taken into account an "error" or "deviation" that is likely to occur when this equation is used[1].

In 2018 Ira Sharma et al proposed "Linear Regression Model to Identify the Factors Associated with Carbon Stock in Chure Forest of Nepal". Their aims to assess the factors associated with carbon stock in Chure forest of Nepal. The data were obtained from Department of Forest Research and Survey (DFRS) of Nepal. A multiple linear regression model and then sum contrasts were used to observe the association between variables such as stem volume, diameter at breast height, altitude, districts, number of trees per plot, and ownership of the forest. 95% confidence interval (CI) plots were drawn for comparing the adjusted carbon stocks with each of the factors and with the overall carbon stock. The linear regression showed a good fit of the model (adjusted $R2$ = 83.75%) with the results that the stem volume (sv), diameter at breast height (dbh), and the number of trees per plot showed statistically significant (p value ≤ 0.05) positive association with carbon stock. 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 [2].

In 2018 Shen Rong et al 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. 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 [3].

In 2018 Syarifah Diana Permaia et al 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 perform 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 [4].

In 2018 Katarina Valaskova et al 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 [5].

In 2019 Anjali Pant et al 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 [6].

In 2019 Hazlina Darman et al 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 models 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. 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[7].

In 2019 Gaurav Pandeya et al proposed " SEIR and Regression Model based COVID-19 outbreak predictions in India". According to WHO reports, COVID-19 is a severe acute respiratory syndrome which is transmitted through respiratory droplets and contact routes. Analysis of this disease requires major attention by the Government to take necessary steps in reducing the effect of this global pandemic. They proposed a study to outbreak of this disease has been analyzed for India till 30th March 2020 and predictions have been made for the number of cases for the next 2 weeks. SEIR model and Regression model have been used for predictions based on the data collected from John Hopkins University repository in the time period of 30th January 2020 to 30th March 2020. The performance of the models was evaluated using RMSLE and achieved 1.52 for SEIR model and 1.75 for the regression model. The RMSLE error rate between SEIR model and Regression model was found to be 2.01. Also, the value of R0 which is the spread of the disease was calculated to be 2.02. Expected cases may rise between 5000-6000 in the next two weeks of time. This study will help the Government and doctors in preparing their plans for the next two weeks. Based on the predictions for short-ter In this study, two machine learning models SEIR and Regression were used to analyse and predict the change in spread of COVID-19 disease[8].

In 2019 K. K. Baseer et al proposed "Analysing various Regression Models for Data Processing". Regression Analysis (RA) is utilized for prediction and determination, where its utilization has generous cover with the field of Artificial Intelligence. RA is a measurable procedure's for assessing the relationship among variables (one dependent and one or more independent). Its helps us to predict and that is why it is also called as predictive analysis model. They used vehicle data like velocity with which traffic move's, gradient, actual velocity to predict the velocity profile of the vehicle. They analyzed various regression models like linear regression, multivariate linear regression and nonlinear regression. The outcome of this work is to write a function for every model that everyone can reuse that without using pre-defined functions in languages and plotting the given data to best fit for analyzing. They discussed various regression models and their uses. Regression analysis helps us to predict things accurate; we can develop this using exponential, logistic type of regression. They used matlab software and wrote a code for that without using pre-defined functions like fitln(), fitlm()[9].

In 2020 Khushbu Kumari et al proposed "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[10].

In 2020 Samit Ghosal et al 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. Validated database was used to procure global and Indian data related to coronavirus 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. Using auto-regression technique and using week 5 death count as input the linear regression model predicted week 6 death count in India to be 467, while keeping at the back of our mind the risk of over-estimation by most of the risk-based models. Projected death rate (n) is 211 and467 at the end of the 5th and 6th week from now, respectively. 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 [11].

III. PROPOSED APPROACH

Adjusted-R^2 measures the proportion of variations explained by only those independent variables that really help in explaining the dependent variable. Unlike R^2, the Adjusted-R^2 **punishes** for adding such independent variables that don't help in predicting the dependent variable (target). Let us mathematically understand how this feature is accommodated in Adjusted-R^2. Here is the formula for adjusted R^2

$$R_a^2 = 1 - \left(\frac{n-1}{n-k-1}\right) \times (1 - R^2)$$

Where

$n = number\ of\ obervations$
$k = number\ of\ independent\ variables$
$R_a^2 = adjusted\ R^2$

Let's take an example to understand the values changes of these metrics in a Regression model

For Example,

TABLE 1
R^2 and Adjusted-R^2

Independent Variable	R^2	Adjusted-R^2
X_1	67.8	67.1
X_2	88.3	85.6
X_3	92.5	82.7

In this example for a regression problem statement, we observed that the independent variable X_3 is **i**nsignificant or it doesn't contribute to explain the variation in the dependent variable. Hence, adjusted-R^2 is decreased because the involvement of in-significant variable harms the predicting power of other variables that are already included in the model and declared significant

IV. EXPERIMENTAL ANALYSIS

Now we used combination of different feature to find squared value. We take combination of LED and Vehicle 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 Flex and Vehicle paper have negative affect the sales.

TABLE 2

R-Squared Value and Adjusted R-Squared value feature selection

Feature	R-Squared Value	Adjusted R-Squared
LED Hoarding Board	0.61	0.61
Flex Hoarding Board	0.33	0.33
Vehicle wraps	0.33	0.32

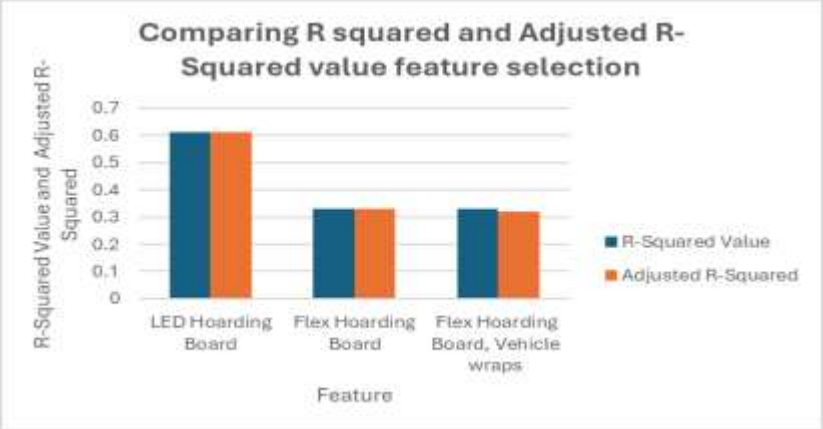

Figure 1 R-Squared Value and Adjusted R-Squared value feature selection

CONCLUSION

Regression Analysis is a well-known statistical learning technique that allows we to examine the relationship between the independent variables It requires we to formulate a mathematical model that can be used to determine an estimated value that is nearly close to the actual value. The two terms essential to understanding Regression Analysis, Dependent variables - The factors that we want to understand or predict Independent variables - The factors that influence the dependent variable. There are mainly two objectives of a Regression Analysis technique, Explanatory analysis - This analysis understands and identifies the influence of the explanatory variable on the response variable concerning a certain model, Predictive analysis, This analysis is used to predict the value assumed by the dependent variable. Determining how well the model fits the data is crucial in a linear model. A general idea is that if the deviations between the observed values and the predicted values of the linear model are small and unbiased, the model has a well-fit data.

REFERENCE

1. NJ Gogtay, SP Deshpande, UM Thatte "Principles of Regression Analysis" Journal of The Association of Physicians of India Vol. 65 April 2017
2. 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.
3. 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.
4. 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.
5. 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.
6. 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
7. 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.
8. 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}.
9. 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.
10. 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.
11. 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