WWW.IJRPR.COM

**International Journal of Research Publication and Reviews**

Journal homepage: www.ijrpr.com ISSN 2582-7421

# Goodness of Fit for Linear Regression using R squared and Adjusted R-Squared

***Manisha Keer [a], Dr. Harsh Lohiya [b], Mr. Sudeesh Chouhan [c]***

***Research Scholar, Sri Satya Sai University of Technology & Medical Science, Sehore, India***
***Assistant Professor, Sri Satya Sai University of Technology & Medical Science, Sehore, India***
***Assistant Professor, Sri Satya Sai University of Technology & Medical Science, Sehore, India***

**ABSTRACT**

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 R squared in regression acts as an evaluation metric to evaluate the scatter of the data points around the fitted regression line. It recognizes the percentage of variation of the dependent variable. R-squared is the proportion of variance in the dependent variable that can be explained by the independent variable. 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. The Adjusted R-squared value is similar to the Multiple R-squared value, but it accounts for the number of variables. In proposed work we use both R-squared and Adjusted R-squared to build a machine learning model in which we can identify the effect of other variable on learn regression model. We used combination of feature and calculated R-squared and Adjusted R-squared. We found that not each variable can be used to create a model sometime a variable has no correlation and negative effect on the model.

**Keywords:** Regression, Adjusted, Dependent, Independent R-squared, Factors

## 1. INTRODUCTION:

Regression Analysis is a well-known statistical learning technique that allows us to examine the relationship between the independent variables and the dependent variables. It requires formulating 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.

The technique generates a regression equation where the relationship between the explanatory variable and the response variable is represented by the parameters of the technique.

We can use the Regression Analysis to perform the following:

- To model different independent variables.
- To add continuous and categorical variables having numerous distinct groups based on a characteristic.
- To model the curvature using polynomial terms.

- To determine the effect of a certain independent variable on another variable by assessing the interaction terms.

## 2. LITERATURE SURVEY

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 [9].

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 14 2020)" Introduction: and Aims: No valid treatment or preventative strategy has evolved till date to counter the SARS CoV 2 (Novel Corona virus) 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 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. Keeping these projected mortality data in mind, current measured for containment of COVID-19 must be strengthened or supplemented .

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. 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 regressions. 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.

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 graphical representations and tables are presented to illustrate the models. The findings from this study have 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.

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. 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 .

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. 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 Frequents 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 Frequents method using OLS.

In 2017 Radek Silhavy et al proposed " Analysis and selection of a regression model for the Use Case Points method using a stepwise approach". They investigate 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 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.

In 2016 Sandhya Jain et al , 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. 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. The aim of this 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 characterized and incompletely validated.

In 2015 Supichaya Sunthornjittanon et al proposed "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 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 the significant category is found deeper analysis is conducted of the fungi category. 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.

In 2014 Lin Yu et proposed "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.

## 3. RESIDUALS

Residuals identify the deviation of observed values from the expected values. They are also referred to as error or noise terms. A residual gives an insight into how good our model is against the actual value but there are no real-life representations of residual values. Consistent with the stochastic error (differences between the expected and observed values must be random and unpredictable). Residuals identify the deviation of observed values from the expected values. They are also referred to as error or noise terms.

A residual gives an insight into how good our model is against the actual value but there are no real-life representations of residual values.

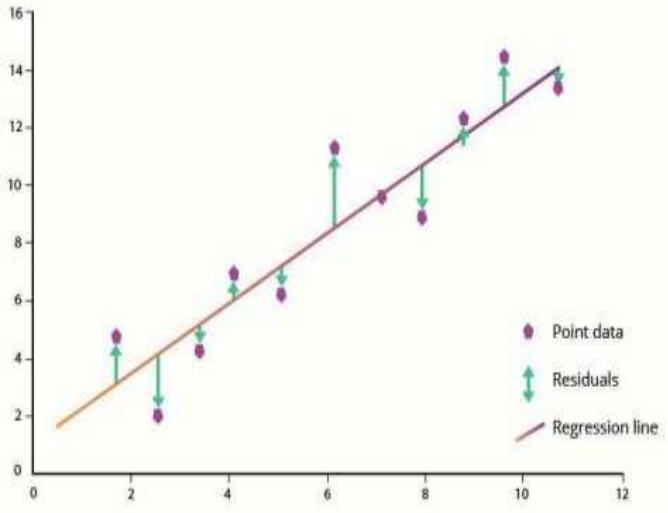


Figure 1 Observed values and expected values

The calculation of the real values of intercept, slope, and residual terms can be a complicated task. However, the Ordinary Least Square (OLS) regression technique can help us to speculate on an efficient model. The technique minimizes the sum of the squared residuals. With the help of the residual plots, we can check whether the observed error is.

## 4. R-SQUARED

R squared value in machine learning is referred to as the coefficient of determination or the coefficient of multiple 2 determination in case of multiple regression.

R squared in regression acts as an evaluation metric to evaluate the scatter of the data points around the fitted regression line. It recognizes the percentage of variation of the dependent variable. R-squared is the proportion of variance in the dependent variable that can be explained by the independent variable

$R^2$ = Variance explained by the model

Total Variance

The value of R-squared stays between 0 and 100%:

0% corresponds to a model that does not explain the variability of the response data around its mean. The mean of the dependent variable helps to predict the dependent variable and also the regression model

On the other hand, 100% corresponds to a model that explains the variability of the response variable around its mean. If value of R is large, have a better chance of your regression model fitting the observations. We can get essential insights about the 2 regression model in this statistical measure; we should not depend on it for the complete assessment of the model. It does not give information about the relationship between the dependent and the independent variables. It also does not inform about the quality of the regression model. Hence, as a user, should always analyze R along with other variables and then derive conclusions about the 2 regression model. We have a visual demonstration of the plots of fitted values by observed values in a graphical manner. It illustrates how R-squared values represent the scatter around the regression line.

## 5. HOW TO INTERPRET R SQUARED

The simplest r squared interpretation is how well the regression model fits the observed data values. Let us take an example to understand.

Consider a model where the R2 value is 70%. Here r squared meaning would be that the model explains 70% of the fitted data in the regression model. Usually, when the R2 value is high, it suggests a better fit for the model. The correctness of the statistical measure does not only depend on R2 but can depend on other several factors like the nature of the variables, the units on which the variables are measured, etc. So, a high R-squared value is not always likely for the regression model and can indicate problems too

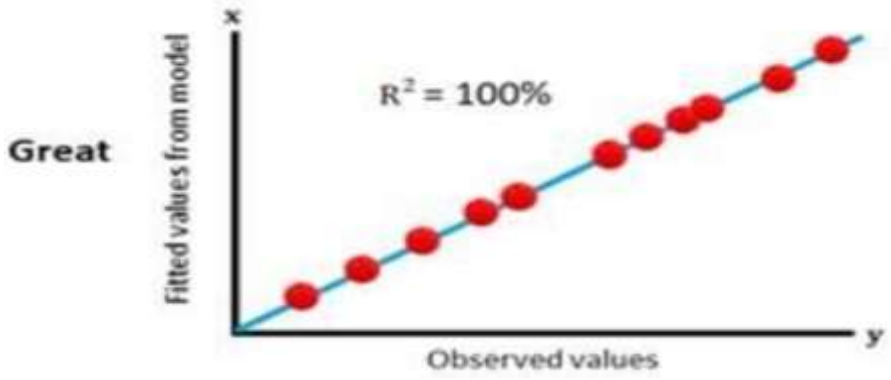


Figure 2 R squared with Value 100%

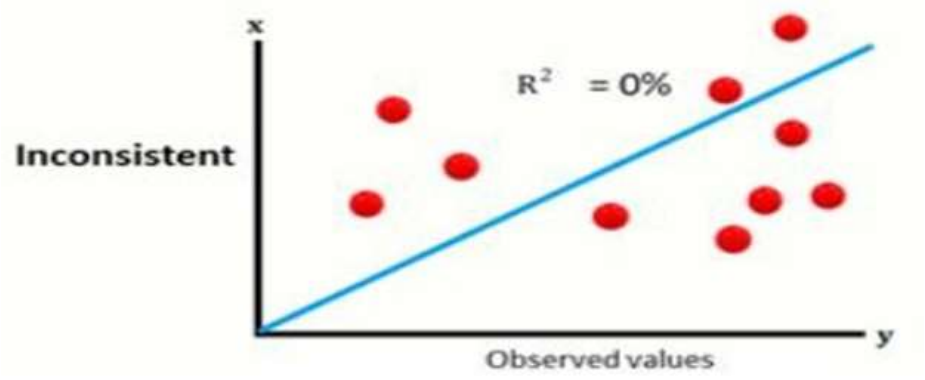


Figure 3 R squared with value 80%

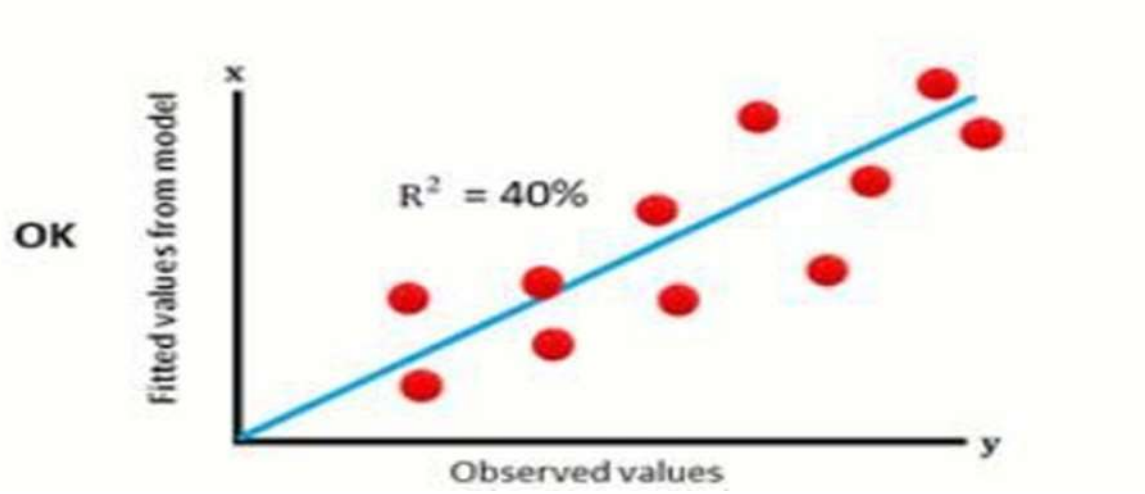


Figure 4 R squared with value 40%

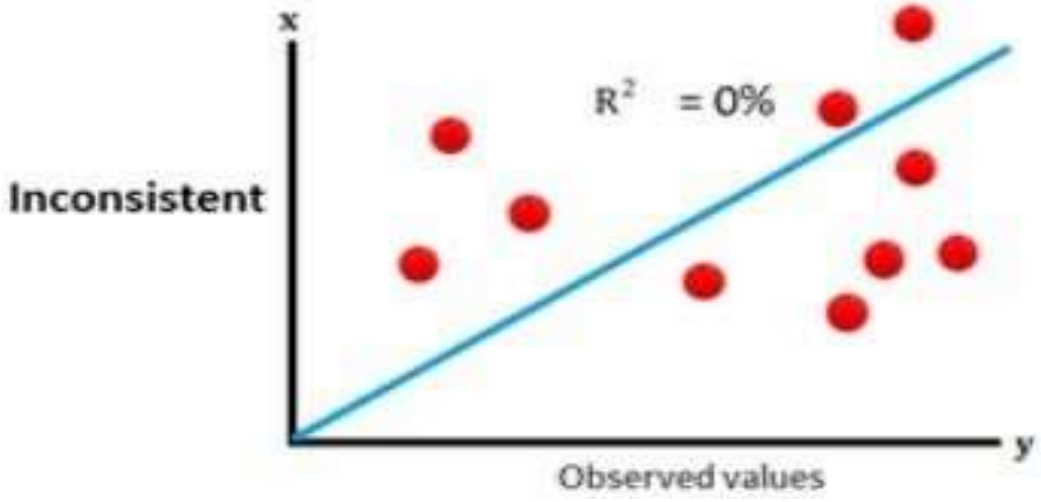


Figure 5 R squared with value 0%

## 6. PROPOSED APPROACH

***Explained variation***

Explained variation is the difference between the predicted value (y-hat) and the mean of already available 'y' values (y-bar) .

It is the variation in 'y' that is explained by a regression model.

$$\text{Explained variation} = \bar{y} - \hat{y}$$

***Unexplained variation***

Unexplained variation is the difference between true/actual value (y) and y-hat. It is the variation in 'y' that is not captured/explained by a regression model. It is also known as the residual of a regression model.

$$\text{Unexplaied variation} = y - \hat{y}$$

***Total variation***

It is the sum of unexplained variation and explained variation. It is also the difference between y and y-bar.

$$\text{Total variation} = (y - \hat{y}) + (\bar{y} - \hat{y}) = (y - \bar{y})$$

Here, we've calculated explained variation, unexplained variation and total variation of a single sample (row) of data. However, in the real world, we deal with multiple samples of data, so we need to calculate the squared variation of each sample and then compute the sum of those squared variations. This would give us a single number metric of variation. To achieve this, we need to slightly modify the formulae of the variations as shown below.

$$SS\tilde{O}explained = \sum n(y- \bar{y})2$$

$$i=1$$

$$SSresidua = \sum n (yi - y) 2$$

$$i=1$$

$$SSTotal = \sum^{n} (y - \bar{y}) ^{2}$$

$$i=1$$

**Where**

$$SS_{explained} = explained\ variation\ sum\ of\ squares$$
$$SS_{residual} = unexplained\ variation\ sum\ of\ squares$$
$$SS_{Total} = total\ variation\ sum\ of\ squares$$

R-squared (aka coefficient of determination) measures the variation that is explained by a regression model. R-squared of a regression model is positive if the model's prediction is better than a prediction which is just the mean of the already available 'y' values, otherwise it is negative. Below is the theoretical formula of R-squared.

$$R2 = \frac{SSexplained}{SSTotal}$$

The above formula is theoretically correct, but only when the R-squared is positive. The formula doesn't return a negati ve R-squared as we are computing the sum of squares in both the numerator and denominator, which makes them always positive, thereby returning a positive R-squared. We can derive the right formula (the one used in practice and also returns negative R-squared) from the above formula as shown below.

$$SSTotal = SSresidual + SSexplained$$

$$SSexplained = SSTotal - SSresidual$$

$$R^2 = \frac{SSTotal - SSresidual}{SSTotal}$$

$$R2 = \frac{1 - SSresidual}{SSTotal}$$

## 7. IMPLEMENTATION ENVIRONMENT

This chapter, we used python 3.9 for creating model. In the proposed approach we used different combination of attribute to create model. We calculate the value of R and adjusted R square. We want to found the effect of the attribute on regression model. The experiments were performed on Intel Core i3processor1GB main memory and RAM: 4GB Inbuilt HDD: 400GB OS: Windows7. The algorithms are .implemented in using Python 3.9 and data set is used in csv file

We used Advertising.csv data set. The Advertising.csv data set contain TV, Radio and Newspaper attribute. TV: advertising dollars spent on TV for a single product in a given market (in thousands of dollars )

Radio: advertising dollars spent on Radio Newspaper: advertising dollars spent on Newspaper What is the response?

*Sales: sales of a single product in a given market (in thousands of items)*

*What else do we know?*

*Because the response variable is continuous, this is a regression problem*

There are 200 observations (represented by the rows), and each observation is a single market

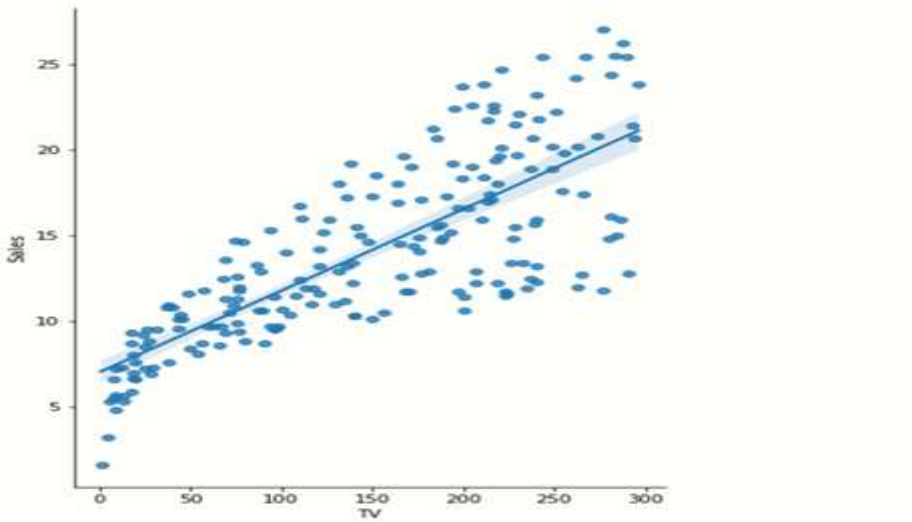


Figure 6 Relationships between advertising on TV and sales

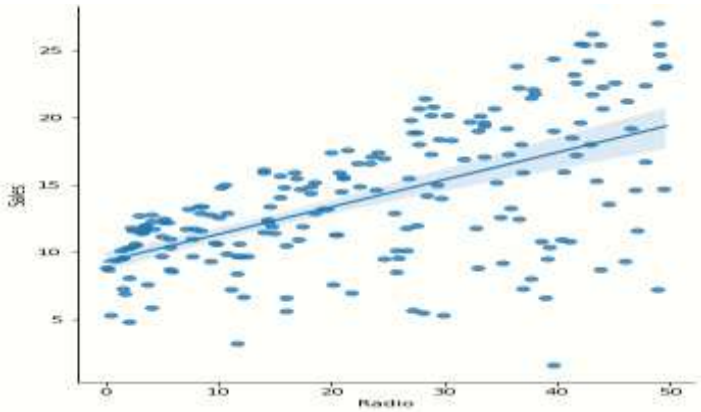


Figure 7 Relationships between advertising on Radio and sales

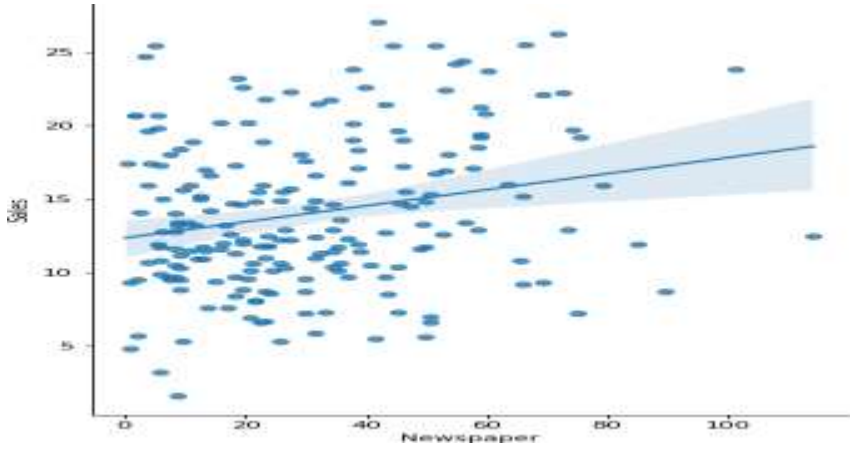


Figure 8 Relationships between advertising on Newspaper and sales

## 8. COMPARATIVE 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 Radio and Newspaper 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 R-Squared Value and Adjusted R-Squared value**

| Feature | R-Squared | Adjusted R- Value Squared |
|---|---|---|
| TV | 0.61 | 0.61 |
| Radio | 0.33 | 0.33 |
| Radio, Newspaper | 0.33 | 0.32 |

## 9. CONCLUSION

R-squared is the proportion of variance in the dependent variable that can be explained by the independent variable. 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. The Adjusted R-squared value is similar to the Multiple R-squared value, but it accounts for the number of variables. In The proposed work we used both R-squared and Adjusted R-squared to build a machine learning model in which we can identify the effect of other variable on learn regression model. We used combination of feature and calculated R-squared and Adjusted R-squared. We found that not each variable can be used to create a model sometime a variable has no correlation and negative effect on the model .WE used real life data set to develop machine learning model using R-squared and Adjusted R-squared. We implement the model in python. By the experimental analysis we found that only some of the variable are use for creating model not all variable are useful.

**REFERENCES**


1. 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):1 870 -1877.
2. Supichaya Sun thornjittanon "Linear Regression Analysis on Net Income of an Agrochemical Company in Thailand" Portland State University PDX Scholar University Honors T heses University Honors College.
3. Sandhya Jain , Sunny Chourse, "Regression Analysis – Its Formulation and Execution In Dentistry", Journal of Applied Dental and Medical Sciences NLM ID: 10 167 1413, ISSN:2454-2 288, Volume 2, Issue 1, January- March 2016.
4. 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
5. 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
6. Syarifah Diana Permaia, Heru na Tanty b 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 International Conferen ce on Computer Science and Computational Intelligence 2018.
7. Anjali Pant, R.S. Rajput "Linear Regression Analysis Using R for Research and Development" See discussions, stats, and auth or 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, January20 19.
9. Khushbu Kumari, Suniti Yadav, "Linear Regression Analysis Study" http://www.j-pcs.org on Friday, July 17, 2020, IP: 1 57.34.76.130] , Journal of the Practice of rdiovascular 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 202 0) Contents lists available at Science Direct Diabetes & Metabolic Syndrome: Clinical Research & Reviews journal .
11. KosukeImai "Using the Predicted Responses from List Experiments as Explanatory Variables in Regression Models Advance Access" publication November 1 1, 11. 2014 Political Analysis (2015 ) 23:18 0– 196doi:10.1093/pan/mpu017.
12. Gaurav Pandeya, Poonam Chaud harya "SEIR and Regression Model based COVID -19 outbreak predictions in India" Department of CSE & IT, The NorthCap University, India DeenDayal Upadhyaya College, University of Delhi, India Defence Research & Development Organization, India a{Email: gaurav16csu120@ncuindia.edu}.\
13. K. K. Baseer, Vikram Neerugatti, Analysing various Regressio n Models for Data Processing International Journal of Innovative Technology an d Exploring Engineering (IJITEE) ISSN: 2278 -3075, Volume-8 Issue-8 June, 2019.

14. Rebecca Bevans. Revised on July 17, 2020.An introduction to simple linear regression Published on February 19, 2020 .

15. Variations Based on Linear Regression Julien I.E. Hoffman, in Biostatistics for Medical and Biomedical Practitioners , 2015.

16. A Brief Overview of Linear Models Jean-François Dupuy, in Statistical Methods for Overdispersed Count Data, 2018.

17. M. Wegmuller, J. P. v onder Weid, P. Oberso n, and N. Gisin, "High resolution fib er distributed measurements with coherent OFDR," in Proc. ECOC'00, 200 0, paper 11.3.4, p. 109.

18. R. E. Sorace, V. S. Rein hardt, and S. A. Vaughn, "High -speed digital-to-RF con verter," U.S. Patent 5 668 8 42, Sept. 16, 1997.

19. S. M. Metev and V. P. Veik o, Laser Assisted Microtechnolog y, 2nd ed., R. M. Osgoo d, Jr., Ed. Berlin, Germany: Springer-Verlag, 1998. C. Toh, "Maximum Battery Life Rou ting to Support Ubiquitous Mobile Computing in Wireless Ad Hoc Networks," IEEE Co mmunications Magazine, p p. 2 -11,