# Comprehensive Analysis of Different Regression Techniques and its Applications used in Machine Learning

**Deepansh Sharma[1], Gulsimar Kaur[2], Himanshu Arora[3], Ishan sharma[4] , Jyoti Diwakar[5] , Muskan Jain[6]**

*[1-5]B. Tech Student, Chandigarh Engineering College, Landran, Punjab, India*
*[6]B. Tech Student, Krishna Engineering College, Ghaziabad, Uttar Pradesh, India*

**Abstract -** *Regression Analysis is an assortment of factual methods that fill in as an explanation behind drawing acceptances about associations among the various interrelated elements. This technique is pretty much applicable in almost everywhere in study field which includes biological sciences, social studies and relapse examination. The main purpose of this research paper is to evolve the vital hypothesis for the factual Regression Technique and also to demonstrate the hypothesis with a collection of wide variety of various models looked over economic aspects, demography, sketching and engineering concepts. To assemble the relevant content which is independent from various insights, mathematics based on variable and numerical investigation is included. Linear Regression computation is by R language is also discussed.*

***Key Words*: Regularizer, Penalty, Correlation, Multicollinearity, Autonomous**

## 1. INTRODUCTION

Regression examination gives information on connection between a (dependent) variable and at least one (indicator) autonomous factors to the degree that data is contained in the information. The goal of regression study is to demonstrate the reaction variable as a component of the indicator factors. This method is utilized for anticipating, time arrangement demonstrating and finding the causal impact relationship between the factors. For instance, connection between rash driving and number of street mishaps by a driver is best concentrated through regression. Regression investigation is a significant instrument for displaying and breaking down information. Here, we fit a bend/line to the information focuses, in such a way, that the contrasts between the separations of information focuses from the bend or line are limited. The duplicity of fit and the exactness of end rely upon the information utilized. Consequently, non-delegate or inappropriately incorporated information bring about poor fits and ends. In this way, for compelling utilization of relapse investigation one must investigate the information assortment process, discover any confinements in information gathered, and restrict ends in like manner. When a relapse examination relationship is acquired, it very well may be utilized to anticipate evaluation of the variable, distinguish agents that impact the reaction most, or confirm conjectured causal models of the reaction. The approximation of each marker variable can be evaluated through the tests which are measurable on the assessed coefficients (multipliers) of the indicator factors.

Regression Analysis is a nifty factual strategy which can be employed over a communication to elect how many particular free aspects are affecting ward factors. The potential circumstances to co-ordinate regression examination are to produce necessary, noteworthy marketing bits of knowledge are unending. Whenever somebody proposing a speculation in your business which exhibit that one component, whether you are capable to control that component easily or not, impacts business, recommend regression analysis to determine the exactness of surety in that theory. This helps in taking decision of accelerated educated business choices, assign assets with more proficiency, at last lift your primary concern Regression Analysis determines the connection between at least two factors. How about we grasp this with a basic model: Assume someone needs to assess expansion in offers which are provided by organization, depend on present monetary situations. The ongoing association data which exhibits that the advancement in bargains is near more than multiple times the development in the economy. Make use of this understanding, one can anticipate upcoming organization offers which are dependent on present and past data.

There are different points of interest of using regression examination. They are according to the accompanying:

- It signifies the huge connections between subordinate variable and free factor.
- It demonstrates the quality of effect of different free factors on a needy variable.

Regression investigation additionally enables one to decide about the various influences of factors estimated on various scales, such as, the value changing impact and the measurement of restricted time operations. These advantages help economic specialists/information investigators/information researchers to kill and assess the best arrangement of factors to be utilized for building prescient models.

## 2. VARIOUS REGRESSION TECHNIQUES

There are various sorts of regression system accessible to make expectations. These methods are determined by three measurements (count of autonomous factors, dependent variable types and regression line shape).

### 2.1 Linear Regression

This regression is generally known as demonstrating strategy. The procedure of linear regression includes the Dependent variable which is continuous in nature, variable(s) which are autonomous can be continual or discrete, and the straight line formed in this technique. Linear Regression forms a connection between subordinate variable (Y) and to a certain degree one autonomous factor(X) utilizes the line which fits in best manner considered by the term Regression line [1].

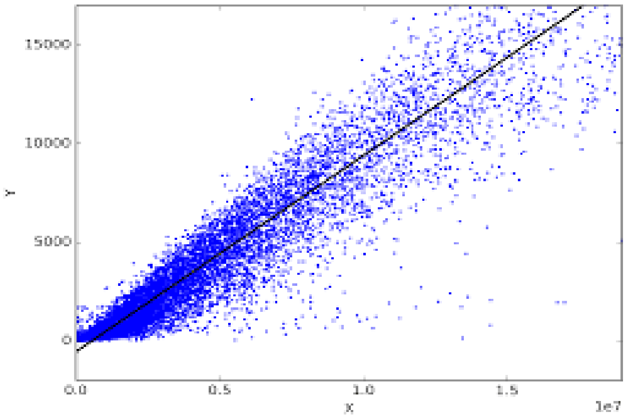


**Fig -1**: Straight Line Graph

Linear Regression Analysis is denoted with a condition Y=a+b*X + e, where intercept is denoted by a, b is the inclination of the line and e signifies the occurrence of inaccuracy. The utilization of this condition helps to foresee the estimation of chosen variable which is dependent in nature [2].

### 2.2 Logistic Regression

One of the most generally utilized regression procedures in the business which is broadly applied crosswise over misrepresentation location, Visa scoring and clinical preliminaries, any place the reaction is parallel has a significant preferred position. One of the significant upsides is of this well-known calculation is that one can incorporate more than one ward variable which can be consistent or dichotomous. The other significant preferred position of this managed AI calculation is that it gives an evaluated an incentive to gauge the quality of relationship as indicated by the remainder of factors. Notwithstanding its prominence, scientists have drawn out its impediments, referring to an absence of powerful system and furthermore an incredible model reliance [3] [4].

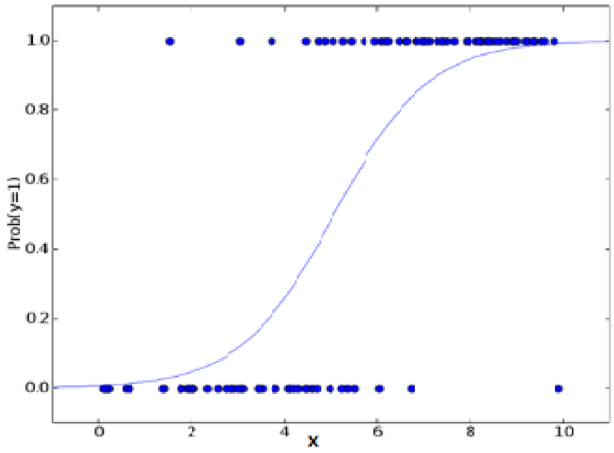


**Fig -2**: Probability Graph

### 2.3 Polynomial Regression

An equation is called to be a polynomial regression equation is it satisfies the condition which if the intensity of autonomous variable is multiple. The condition underneath speaks to a polynomial condition:

$$Y = a + b * x ^ 2$$

As a result, in polynomial regression technique, there is no formation of straight line for the best fit line. The curve shaped line is formed in which all the data points are fitted [5] [6].

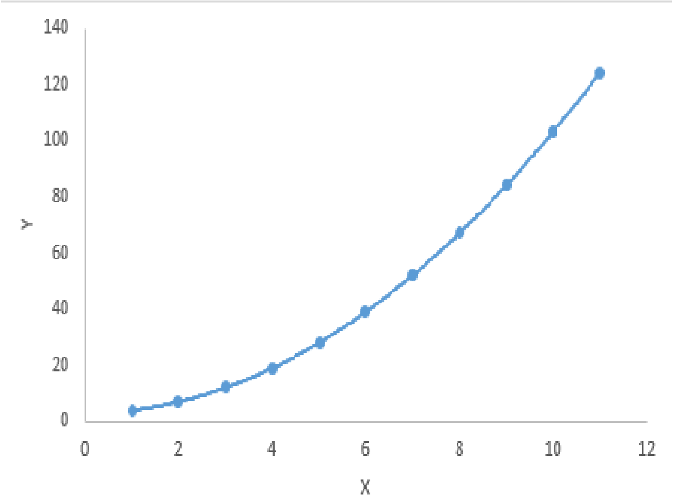


**Fig -3**: Parabolic Graph

## 2.4 Stepwise Regression

Stepwise regression is commonly utilized when there is need to manage different factors which are considered as free. Here, the alternatives of free factors are finished by involving well programmed procedures instead of including human interactions.

By observing different statistical factors such as A/C metrics, T Stats and R Squares the stepwise regression can be achieved and huge factors are also recognised. Also, by including and excluding co-variant the specific time helps to fulfil best fit stepwise regression blueprints. Here, are some majorly used stepwise regression techniques procedures are entitled below:

- Standard stepwise regression completes two things. It includes and expels indicators varying for each progression.
- The selection of forward strategy begins with the help of the most prominent forecasting factor for the available model and some variables are included for each progression.
- The backward elimination approach is stated by considering all the predictors which are involved in the given model and then expels those variables which are less noteworthy in each specific progression. So, the main purpose of demonstrating the stepwise regression is to ensure maximum forecasting capability by involving least number of variables which are useful in prediction. This technique is also helpful in controlling those data sets which have high dimensionality in nature [7] [8].

## 2.5 Ridge Regression

Edge Regression is helpful in those conditions where data comes across the multicollinearity means there is higher bond between the variables which are generally called as free factors. In multicollinearity method, in spite of unbiases between the least square estimates (OLS), but some variances are usually observed which means that there are some fluctuations between the observed value and the original value. In Ridge Regression method, the standard errors are eliminated by including bias degree to the regression predictors. The shrinkage parameter lambda (λ) is used to rectify the multicollinearity issue. Above, we saw the condition for direct relapse.

y=a+ b*x

This condition likewise has an error term. The total condition becomes: Ridge regression solves the multicollinearity problem through shrinkage parameter λ (lambda). Look at the equation below.

$$= \underset{\beta\in\mathbb{R}^p}{\operatorname{argmin}} \underbrace{\|y - X\beta\|_2^2}_{\text{Loss}} + \lambda \underbrace{\|\beta\|_2^2}_{\text{Penalty}}$$

In the above equation, two factors are involved. The first factor defines the last square and the second one explains about summation lambda of β2 (beta- square) where β is considered as the coefficient. The low variance is targeted by adding β coefficient to least square term [9] [10].

## 2.6 Lasso Regression

Like Ridge Regression, Lasso (Least Absolute Shrinkage and Selection Operator) regression coefficients perfect size is penalised. Besides this, it helps in reduction of variability and helps in the improvement of the models related to linear regression. The difference between lasso regression & ridge regression is that in ridge regression rather than using squares, accurate values are used for penalising the function [11] [12].

## 2.7 ElasticNet Regression

It is defined as the combination of two techniques namely Lasso and Ridge Regression on prior basis, it is prepared with the usage of L1 and L2 which act as a regularizar. ElasticNet regression is highly recommended in those cases where there is high correlation between many features. Any number of variables can be selected in this regression technique**.**

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}(\|y - X\beta\|^2 + \lambda_2\|\beta\|^2 + \lambda_1\|\beta\|_1).$$

A practical advantage of trading-off between Lasso and Ridge is that, it allows Elastic-Net to inherit some of Ridge's stability under rotation [13] [14].

## 3. IMPORTANCE OF REGRESSION ANALYSIS IN BUSINESS

Regression investigation is about information. It assists organizations with understanding the information focuses they have and use them – explicitly the connections between information focuses – to settle on better choices, including anything from anticipating deals to understanding stock levels and market interest. Of all the business investigation systems, relapse examination is regularly alluded to as one of the most noteworthy [16] [17].

### 3.1 Predictive Analytics

Predictive investigation for example determining future chances and analysing various risks involved in business is considered as an eminent usage of regression technique in marketing and business at global level. The study of demand analysis,such as prediction of items which will be presumably purchased by the customer.

### 3.2 Operation Efficiency

The business and marketing procedures can also be optimised using various models of regression. There are numerous examples of optimization using regression such as in bakery, statistical models are created to estimate the temperature of the oven to bake cookies in it. Another perfect example of regression usage is in call centre, where one can examine the coordination between the call waiting time and various complaints registered by the various customers. The guesswork is eliminated by data driven decision approach. It helps in the advancement of business production by encouraging those areas having the maximum influence on the operational efficiency and income.

### 3.3 Optimizing processes

Having more information, and a comprehension of that information, can amplify effectiveness and refine forms with the goal that organizations can capitalize on them. Procedures that are enhanced by measurable information can assist organizations with working more astute.

### 3.4 Correcting Errors

Regression technique is not only limited for predictive decision but also plays a vital role in identification of judgement errors. It can be explained with an instance related to sales. If a store manager believes that by extending shopping duration timings sales will also increase. However, it can't be sufficient to provide a desired result because with the rise of shopping hours extra labor timings charges will also increase. Therefore, regression analysis techniques help in providing a measurable support to take better decisions to avoid any kind of mistakes occurs due to store assumptions.

## 4. LINEAR REGRESSION USING R LANGUAGE

Linear Regression model is generally used for the prediction of quantitative outcome variable (Y) based on single or multiple forecasting variables (X).

To compute the achievement of the predictive regression model two foremost metrics are usually considered:

- **Root Mean Square Error**: It is generally used to calculate the error of the predictive model. It is basically defined the average difference of the observed value and the predictive value. So, Root Mean Square (RMSQ)is defined as:

  RMSQ= Mean (observed value- Predictive Value)^2

  The model is usually considered better if it has lower RMSQ.

- **R- Square:** R Square id defined by the squared correlation between the observe value and predictive value by the model. The model will be better if it has higher R2. Here are simple guidelines to develop a predictive regression model:

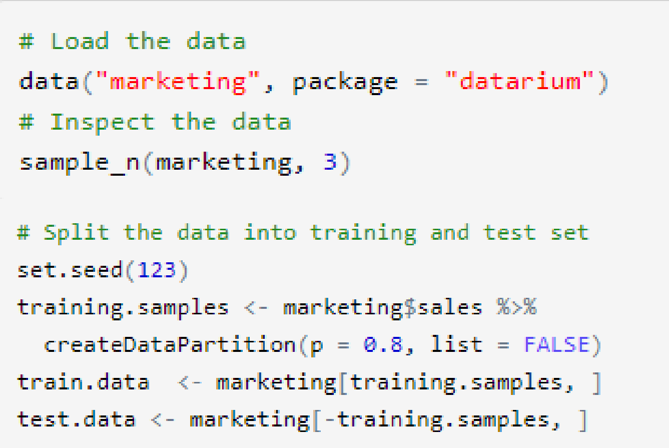
- Firstly, split the data into two categories namely training set and test set into 80% and 20 % respectively.
- Secondly, by using the training set develop the regression model.
- Finally, predictions are made by analyzing the test set and then assess the accuracy metrics of the model.

### 4.1 PACKAGES REQUIRED IN R FOR LINEAR REGRESSION

- **tidyverse** :It is used for easy data manipulation and visualization.
- **caret** : for easy machine learning workflow

### 4.2 DATA PREPARATION

To prepare the data, marketing data set is used to predict the sales units based on the money factor which is spent in three different media for advertisements namely youtube, facebook and newspaper.



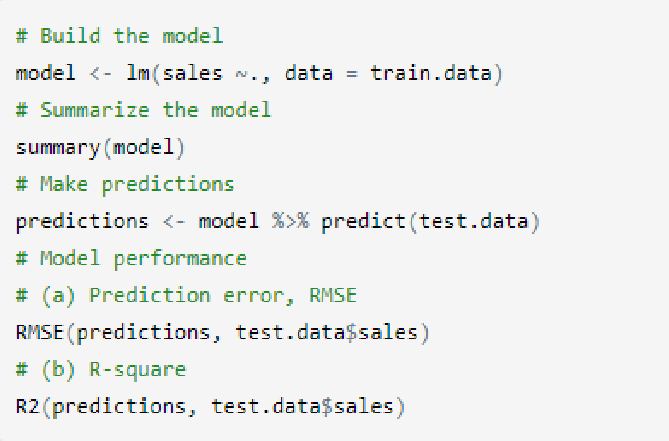
```
# Load the data
data("marketing", package = "datarium")
# Inspect the data
sample_n(marketing, 3)

# Split the data into training and test set
set.seed(123)
training.samples <- marketing$sales %>%
  createDataPartition(p = 0.8, list = FALSE)
train.data  <- marketing[training.samples, ]
test.data <- marketing[-training.samples, ]
```

**Fig -4:** Data preparation Using Test set and Train set

### 4.3 COMPUTINNG LINEAR REGRESSION

To calculate the liner regression, lm()function in R is used . The code is explained in below:



```
# Build the model
model <- lm(sales ~., data = train.data)
# Summarize the model
summary(model)
# Make predictions
predictions <- model %>% predict(test.data)
# Model performance
# (a) Prediction error, RMSE
RMSE(predictions, test.data$sales)
# (b) R-square
R2(predictions, test.data$sales)
```

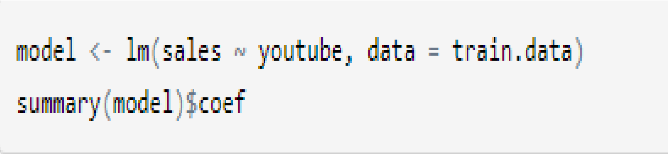
**Fig -5:** Usage of lm() function in R

### 4.4 LINEAER REGRESSION MODEL

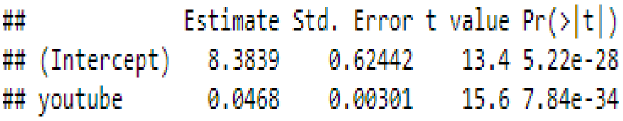
The basic purpose of simple regression is to foresee the continuous variable outcome (Y) on the basis of single predictive variable defined by (X). To predict the sales in accordance to youtube , the simple regression equation can be defined as:

Sales= b0+b1*youtube

The R function lm() can be used to determine the beta coefficients of the linear model, as follow:



```
model <- lm(sales ~ youtube, data = train.data)
summary(model)$coef
```

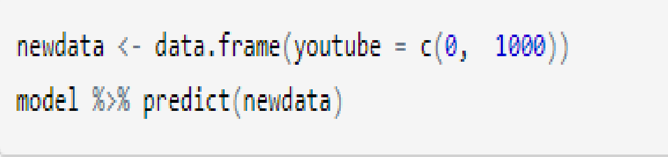


```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.3839    0.62442    13.4 5.22e-28
## youtube       0.0468    0.00301    15.6 7.84e-34
```

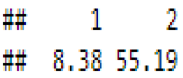
**Fig -6:** Linear Regression

In the above example, output estimation of the beta coefficients for regression and levels of significance defined by Pr(>|t|) is shown.

The estimated regression equation can be defined as: sales = 8.38 + 0.046*youtube. With the help of this formula, one can easily predict the unit sales for any advertisements of youtube. Also, one can predict the unit sales simply by using predict() function available in R, such as to predict sales unit for two youtube advertising budget: 0 and 1000.



```
newdata <- data.frame(youtube = c(0,  1000))
model %>% predict(newdata)
```



```
##     1     2
##  8.38 55.19
```

**Fig -6**: Prediction of Unit Sales

## 5. CONCLUSION

In this work, various regression analyses have been studied. It has been concluded that in regression analysis various statistical procedures are used to estimate the associations between the variables which are dependent or independent in nature. Regression analysis is popularly used to predict results in various applications of machine learning. Secondly, the wider scope of regression analysis can be seen in finding casual relationships among variables. Furthermore, few applications such as correction of errors in loan, operation efficiency in business have also been analysed. Various Results of Linear Regression using R language are also discussed.

## REFERENCES

[1] Navid Aghdaei, Georgios Kokogiannakis, Daniel Daly, Timothy McCarthy, "Linear Regression models for prediction of annual heating and cooling demand in representative Australian residential dwellings", International Conference on Improving Residential Energy Efficiency, IREE 2017, ppno: 79-86, 2017.

[2] Benat Arregi, Roberto Garay, "Regression analysis of the energy consumption of tertiary buildings", CISBAT 2017 International Conference – Future Buildings & Districts –Energy Efficiency from Nano to Urban Scale, CISBAT 2017, ppno: 9-14, 2017.

[3] Wei Chen, Xiaoshen Xie, Jiale Wang, Biswajeet Pradhan, Haoyuan Hong , Dieu Tien Bui, Zhao Duan, Jianquan Ma, "A comparative study of logistic model tree, random forest, and classification and regression tree models for spatial prediction of landslide susceptibility" CATENA, ppno: 147-160, 2017.

[4] Stephan Dreiseitl and Lucila Ohno-Machado, "Logistic regression and artificial neural network classification models: a methodology review", Journal of Biomedical Informatics 35, Volume 35, Issues 5–6 ppno: 352-359, 2003.

[5] Eve Bélisle , Zi Huang , Sébastien Le Digabel , Aïmen E. Gheribi, "Evaluation of machine learning interpolation techniques for prediction of physical properties", Computational Materials Science 98,ppno: 170-177, 2015.

[6] Prasanna Muralidhara, Jacob Hinkle and P. Thomas Fletcher, "A MAP ESTIMATION ALGORITHM FOR BAYESIAN POLYNOMIAL REGRESSION ON RIEMANNIAN MANIFOLDS", 2017 IEEE International Conference on Image Processing (ICIP), ppno: 215-219, 2017.

[7] Mahmoud Hosseinpour, Hojjat Sharifi & Yasser Sharifi, "Stepwise regression modeling for compressive strength assessment of mortar containing metakaolin", International Journal of Modelling and Simulation, volume38 Issue 4, ppno: 1-9,2018.

[8] Zeng Shuo , Liu Yaozong, Li Jun, "Metamodel for 2D magnet-rail relationship based on stepwise regression", International Journal of Applied Electromagnetics and Mechanics 56, volume 56 Issue 1, ppno: 75–89, 2018.

[9] Yang You, James Demmel, Cho-Jui Hsieh, Richard Vuduc, "Accurate, Fast and Scalable Kernel Ridge Regression on Parallel and Distributed Systems", International Conference on Supercomputing, ppno: 307-317, 2018.

[10] Piyush Kant Rai , Sarla Pareek , Hemlata Joshi and Shiwani Tiwari, "INDIRECT METHOD OF ESTIMATION OF TOTAL FERTILITY RATE AND STUDY ABOUT BIRTHS AVERTED DUE TO FAMILY PLANNING PRACTICES IN INDIA: A RIDGE REGRESSION APPROACH", Journal of Data Science, ppno- 647-676, 2018.

[11] BruceSpencer, OmarAlfandi, FerasAl-Obeidat, "A Refinement of Lasso Regression Applied to Temperature Forecasting", 8th International Conference on Sustainable Energy Information Technology (SEIT 2018),ppno: 728-735,2018.

[12] Zdravko Botev , Yi-Lung Chen , Pierre L'Ecuyer ,Shev MacNamara and Dirk P. Kroese, "EXACT POSTERIOR SIMULATION FROM THE LINEAR LASSO REGRESSION",2018 Winter Simulation Conference (WSC), ppno: 1706- 1717, 2018.

[13] Rymarczyk Tomasz , Kozlowski Edward ,Kłosowski Grzegorz , Rymarczyk pawel , Adamkiewicz Przemyslaw and Sikora Jan, "Elastic net method in the image reconstruction infiltration of water in the embankment", Conference on 2018 Applications of Electromagnetics in Modern Techniques and Medicine (PTZE), ppno:212-215, 2018

[14] Alexis Comber, Paul Harris, "Geographically weighted elastic net logistic regression", Journal of Geographical Systems, Volume 20 Issue 4, ppno: 317–341, 2018.

**Web References:**

[15]https://www.analyticsvidhya.com/blog/2015/08/ compreh ensive- guide-regression/

[16] https://smallbusiness.chron.com/application-regression-analysis-business-77200.html

[17] https://www.newgenapps.com/blog/business-applications-uses-regression-analysis-advantages/

[18] http://www.sthda.com/english/articles/40-regression-analysis/165-linear-regression-essentials-in-r/#computing-linear-regression