Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 08, August - 2017

# Review of Regression Analysis Models

Aviral Gupta
B.Tech Student
Department of Information Technology
Maharaja Agrasen Institute of Technology
New Delhi, India

Akshay Sharma
B.Tech Student
Department of Information Technology
Maharaja Agrasen Institute of Technology
New Delhi, India

Dr. Amita Goel
Associate Professor
Department of Information Technology
Maharaja Agrasen Institute of Technology
New Delhi, India

*Abstract*— **In statistics and data analysis, we often need to establish a relationship between the various parameters in a data set. This relationship is important for prediction and analysis. Regression Analysis is such a technique. This work mainly focuses on the different Regression Analysis models used nowadays and how they are used in context of different data sets. Picking the right model for analysis is often the most difficult task and therefore, these models are looked upon closely in this research. While a Linear Regression Analysis model is used to fit linear data, a Polynomial Regression Analysis model focuses on a data set representing polynomial relationship between data parameters. Logistic Regression model is used in a scenario where we need a binary type of prediction. When the data set becomes complex, these models may suffer from issues like Underfitting and Overfitting. Ridge and Lasso Regression are considered the best models to deal with this type of situation. Ridge regression is used when data suffers from multicollinearity, that is independent variables are highly correlated. Lasso regression differs from ridge regression in a way that it uses absolute values in the penalty function, instead of squares. Using these models in the right way and with right data set, Data Analysis and Prediction can produce the most accurate results.**

***Keywords— Regression; Underfitting; Overfitting; Regularization***

## I. INTRODUCTION

In statistics, a statistical model is a part of mathematical model which incorporates a set of assumptions and is basically concerned with generation of data which is similar and synthesized from a larger sample. A statistical model is basically a data-generating process.

The suppositions used by a statistical model represents a combination of probability distributions, where some of which adequately represent the particular data set. It is the innate use of probability which is unique to statistical models.

In statistical modelling, Regression Analysis is a technique to find out the relationship between different variables. Regression looks closely into how a dependent variable is affected upon varying an independent variable while keeping the other independent variables constant [1].

Regression analysis is used by mathematicians and data scientists alike for prediction and forecasting. It involves fitting the right model with respect to the given data set and then using that model to make further predictions. The ideal model showcases all the relationships accurately. Naturally, a tool based on regression analysis can provide valuable insights to an economist or a manager. The various uses or advantages of Regression Analysis are as follows:

i. Can be used to predict the future: By using the relevant model to a data set, Regression Analysis can accurately predict a lot of useful information like Stock Prices, Medical Conditions and even Sentiments of the public.
ii. Can be used to back major decisions and policies: Results from regression analysis adds a scientific backing to a decision or policy and makes it even more reliable as it likelihood of success is then high.
iii. Can correct an error in thinking or disabuse: Sometimes, an anomaly between the prediction of regression analysis and a decision/thinking can help correct the fallacy of the decision.
iv. Provides a new perspective: Large data sets realise their potential to provide new dimensions to a study through the application of Regression Analysis.

Therefore, Regression analysis is a very important tool for a Data Scientist working with Data Sets. To yield the correct results from different types of Data Sets with different relationships, different types of Regression Analysis models are used.

## II. LINEAR REGRESSION

Linear regression is the most simple regression analysis technique. It is the most commonly regression analysis mechanism in predictive analysis.

At the core of linear regression analysis is to find a line that could satisfy the scatter plots as efficiently as possible [2].
We plot many lines in linear regression analysis and then find out which of the line among them satisfies the scatter plot most efficiently.
An example of linear regression in its simplest form is:

$$Y = ax + b + \varepsilon$$

Here $Y$ is a dependent variable, whose value depends proportionally to $x$. Linear regression analysis is used to find the coefficients $a$ and $b$ such that, the equation of the line formed satisfies the scatter points most efficiently. $\varepsilon$ is the error term.

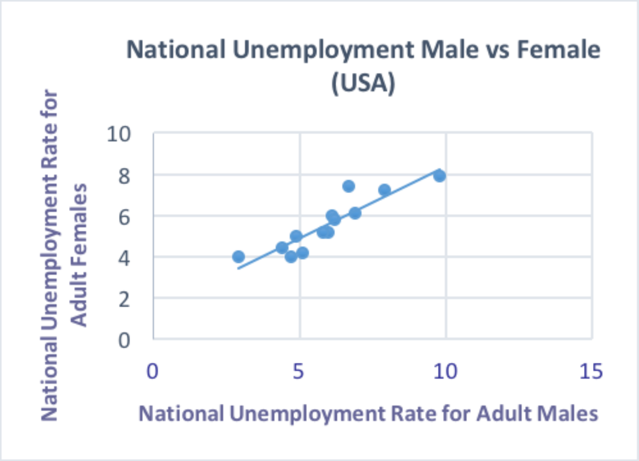

Figure 1: Linear Regression Example

## III. LOGISTIC REGRESSION

In Logistic Regression, the dependent variable is binary that is it has two values. It can have values like True/False or 0/1 or Yes/No. This model is used to determine the chance whether a dichotomous outcome depends on one or more free (independent) variables [3]. It uses logistic function to find out the association amongst dependent variable and or more free variables.

Logistic Regression has similarity with Linear Regression because it is an exceptional case of Generalized Linear Model. Yet, there are significant differences between the two Regressions. They are as follows:

i. In Logistic Regression, Conditional Distribution **y|x** is not a Gaussian distribution but a Bernoulli distribution.

ii. In Logistic Regression, the predicted outcomes are probabilities determined through logistic function and they are circumscribed between 0 and 1.

The value determined by Logistic Regression can be represented by the following equations:

$Odds = P(1 - P) =$ Probability that an event occurs/Probability that an event does not occur

$$\ln(Odds) = \ln\left(\frac{P}{1-P}\right)$$

$$logit(P) = \ln\left(\frac{P}{1-P}\right) = B0 + B1X1 + B2X2 + B3X3 + B4X4 \ldots\ldots\ldots. + BnXn$$

*Here P, is the probability of attribute under scrutiny.*

## IV. POLYNOMIAL REGRESSION

Polynomial Regression is a variant of Linear Regression in which the association between the free variable $X$ and the dependent variable $Y$ can be represented as any valid polynomial. This polynomial is of n$^{th}$ degree.

In normal Linear Regression,

$$Y = ax + b + \varepsilon$$

In many situations, such a linear relationship may not be sufficient. For example, in a game a team's score may increase by n$^{th}$ power to the amount of goals scored:

$$Y = ax^n + b + \varepsilon$$

Here, the relationship is not linear.

Hence, we can represent the Polynomial Regression as:

$$Y = a_o + a_1x + a_2x_2 + a_3x_3 \ldots\ldots\ldots\ldots\ldots\ldots + a_nx_n + \varepsilon$$

Where n is called the degree of the polynomial.

Even though Polynomial Regression model accepts a nonlinear association between $X$ and $Y$, still it is viewed as linear regression because its regression coefficients $a_o, a_1, a_2, \ldots.. a_n$ are linear.

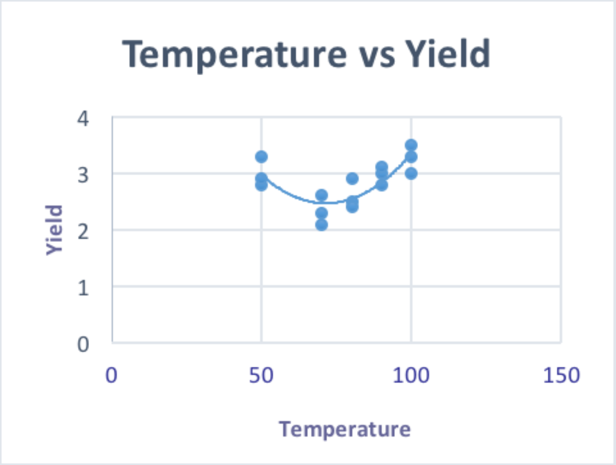

Figure 2: Polynomial Regression Example

## V. UNDERFITTING AND OVERFITTING

In machine learning and statistical analysis, the main cause of poor performance is either due to under fitting and overfitting.

When an algorithm in machine learning or in statistical analysis is used to predict future values by training itself from a training set, but in the end models itself too well. Then we say that there is an overfitting condition.

Overfitting condition causes poor performance, and therefore modelling needs to be done again to get a good fit condition [4].

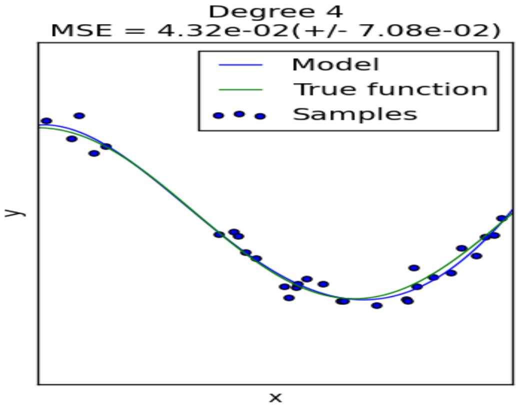

Figure 3: Overfitting Example

Underfitting refers to a condition when an algorithm is not able to model itself from training set, or cannot generalize itself to a new data. Since the modelling done is underfit, therefore it is obvious to say that the model will obviously affect the performance [5].

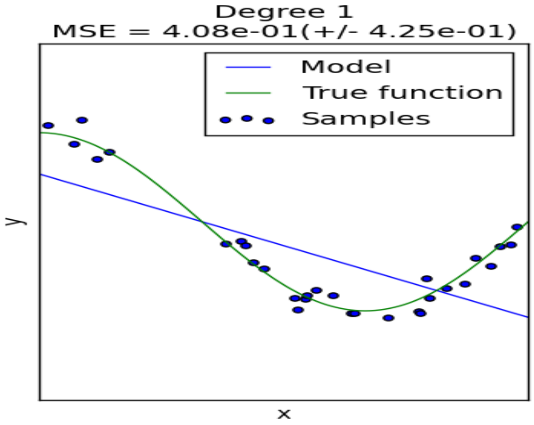

Figure 4: Underfitting Example

## VI. RIDGE REGRESSION

Ridge regression is used to perform L2 regularization. L2 regularization, adds a factor dependent on the sum of squares of the coefficients to our modelling algorithm [6].

Modelling done using Linear Regression or polynomial regression might lead to an overfitting condition, therefore Ridge Regression is used to minimize the overfitting condition by adding a new factor to the least square objective of Linear Regression. We define the Ridge Regression mathematically as:

$$\Delta + \alpha^*(sum\ of\ squares\ of\ coefficients)$$

Here $\Delta$ is the objective derived from Linear or Polynomial Regression techniques and $\boldsymbol{\alpha}$ is the parameter which is used to balance how much amount of emphasis needs to be given to $\Delta$ and magnitude of the coefficients.

There are various types of values that $\boldsymbol{\alpha}$ can take:

i. $\boldsymbol{\alpha = 0}$: When $\alpha$ becomes 0, the objective above becomes same as simple Linear or Polynomial Regression. Here the coefficients we will get for our plot will be same as of simple Linear or Polynomial Regression.

ii. $\boldsymbol{\alpha = \infty}$: Since now $\alpha$ is infinite, the coefficients of the plot will vanish.

iii. $\boldsymbol{0 < \alpha < \infty}$: Depending on the value of $\alpha$, weightage will be given to the different sections of the mathematical expression above.

## VII. LASSO REGRESSION

Lasso regression is used to perform **L1 regularization**. L1 regularization, adds a factor dependent on the sum of absolute value of the coefficients to our modelling algorithm [7].

Modelling done using Linear Regression or polynomial regression might lead to an overfitting condition, therefore as Ridge Regression is used to minimize the overfitting condition by adding a new factor to the least square objective of Linear Regression, similarly Lasso Regression is used to minimize the overfitting condition by adding a different factor to the least square objective of Linear or Polynomial Regression, the factor added is proportional to the absolute sum of the coefficients.

We define the Lasso Regression mathematically as:

$$\Delta + \alpha^*(absolute\ sum\ of\ coefficients)$$

Here $\Delta$ is the objective derived from Linear or Polynomial Regression techniques and $\boldsymbol{\alpha}$ is the parameter which is used to balance how much amount of emphasis needs to be given to $\Delta$ and magnitude of the coefficients.

There are various types of values that $\boldsymbol{\alpha}$ can take:

i. $\boldsymbol{\alpha = 0}$: When $\alpha$ becomes 0, the objective above becomes same as simple Linear or Polynomial Regression. Here the coefficients we will get for our plot will be same as of simple Linear or Polynomial Regression.

ii. $\boldsymbol{\alpha = \infty}$: Since now $\alpha$ is infinite, the coefficients of the plot will vanish.

iii. $\boldsymbol{0 < \alpha < \infty}$: Depending on the value of $\alpha$, weightage will be given to the different sections of the mathematical expression above.

## VIII. CONCLUSION

Regression is a really effective tool in statistical analysis of data. Different Regression models are used in different situations. Their usability depends on a case to case basis usually on the basis of type data and the relationship between data. While a Linear Regression Model is best suited to data representing a linear relationship between two variables, a Polynomial Regression Model is used in case of multiple variables having a polynomial relationship. Logistic Regression model is used in the situations where we need to make a binary type of prediction based on the data like Yes/No or True/False or 0/1. Ridge regression is used when data suffers from multicollinearity that is, independent variables are highly correlated. To compensate this problem, Ridge Regression, uses the shrinkage parameter, which uses the squares of coefficients to reduce the multicollinearity.

Lasso regression differs from ridge regression in a way that it uses absolute values in the penalty function, instead of squares. This leads to penalizing values which causes some of the parameter estimates to turn out exactly zero. Larger the penalty applied, further the estimates get shrunk towards absolute zero.

## ACKNOWLEDGMENT

It gives us immense pleasure to express our deepest sense of gratitude and sincere thanks to our highly respected and esteemed guide Dr. M.L. Sharma (Head of Department, Department of Information Technology), MAIT Delhi, for his valuable guidance, encouragement and help for completing this work.

## REFERENCES

[1] “Comprehensive Guide Regression”, https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/.

[2] Schneider A, Hommel G, Blettner M., “Linear Regression Analysis”, Part 14 of a Series on Evaluation of Scientific Publications. Deutsches Ärzteblatt International. 2010;107(44):776-782.

[3] Ngokkuen, Chuthaporn & Grote, Ulrike, "Geographical Indication for Jasmine Rice: Applying a Logit Model to Predict Adoption Behavior of Thai Farm Households," Quarterly Journal of International Agriculture, Humboldt-UniversitÃ¤t zu Berlin, vol. 51.

[4] Piotrowski, A.P., Napiorkowski, J.J., “A comparison of methods to avoid overfitting in neural networks training in the case of catchment runoff modeling”, J. Hydrol. 476, 97–111.

[5] van der Aalst, W. M., Rubin, V., Verbeek, H. M. W., van Dongen, B. F., Kindler, E., & Günther, “Process mining: a two-step approach to balance between underfitting and overfitting. Software & Systems Modeling”, 9(1), 87-111.

[6] Kidwell, Jeannie S., and Lynn Harrington Brown. “Ridge Regression as a Technique for Analyzing Models with Multicollinearity”, Journal of Marriage and Family, vol. 44, no. 2, 1982, pp. 287–299.

[7] Tibshirani, R., “Regression shrinkage and selection via the lasso”, J. Royal. Statist. Soc B., Vol. 58, No. 1, pages 267-288.M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.