A Review of Regression Models in Machine Learning

Sunil Kumar, Vaibhav Bhatnagar *

Department of Computer Applications, Manipal University Jaipur, Dehmi Kalan, Jaipur, Rajsthan 303007, India

Abstract

Machine learning is one of the active fields and technologies to realize artificial intelligence (AI). The complexity of machine learning algorithms creates problems to predict the best algorithm. There are many complex algorithms in machine learning (ML) to determine the appropriate method for finding regression trends, thereby establishing the correlation association in the middle of variables is very difficult, we are going to review different types of regressions used in Machine Learning. There are mainly six types of regression model Linear, Logistic, Polynomial, Ridge, Bayesian Linear and Lasso. This paper overview the above-mentioned regression model and will try to find the comparison and suitability for Machine Learning. A data analysis prerequisite to launch an association amongst the innumerable considerations in a data set, association is essential for forecast and exploration of data. Regression Analysis is such a procedure to establish association among the datasets. The effort on this paper predominantly emphases on the diverse regression analysis model, how they binning to custom in context of different data sets in machine learning. Selection the accurate model for exploration is the most challenging assignment and hence, these models considered thoroughly in this study. In machine learning by these models in the perfect way and thru accurate data set, data exploration and forecast can provide the maximum exact outcomes.

Keywords: Dependent variable (DV); Independent variable (IV); Regression, Multicollinearity; Least-squares guesses; Slope; Variance; Ordinary least squares (OLS); Multicollinearity.

1. Introduction

Regression is the procedure of typical association between two or greater than two variables of interest concerning original elements of the data-set. It also command to launch the behavior of the association in the middle of variables on interest that are describing the practical association between the variables and thus afford an instrument for prediction or forecasting. It is a method of analysis and recognizes the relationship between two or greater than two variables of interest. The method that is adapted to execute regression analysis helps to realize whichever aspects are significant, whichever aspects fail to notice and in what way an individual promoting one and all. Regression castoff for prediction and forecasting. Regression is a subset of supervised machine learning techniques to predict the pattern of data. “It is a task of predicting a continuous quantity usually one dependent variable (DV) and one or more than one independent variable (IV)” [1]. “This process regresses desirability of a predictor variable in consequence of the predicted variable. Another way testing to realize how the value of Y modifies with respect to changes in X” [2]. A group of machine learning procedures that help as a foundation for illustrates interpretations about associations among consistent variables. These procedures are appropriate in practically all studies, with the common practical estimation, the utmost routine of all data analysis procedures.

One of the machine learning procedures that provide clues to opt appropriate equations to find and identify trends into a dataset is regression. It is a statistical process for models in Machine Learning applying for prediction and forecasting. This has a substantial connection to the field of machine learning, apply diagonally to different businesses

* Corresponding author
E-mail address: vaibhav.bhatnagar15@gmail.com

ARK: https://n2t.net/ark:/47543/JISCOM2022.v3i1.a30

and productions, for example a farmer in reducing their losses and to get the best prices for their crops. In contest of model review and study of different regression methods matrices, techniques parameters to find predictions based on three parameters: independent variables count, regression shape line and dependent variable type in machine learning.

This study aims to study the relationship among various types of regressions with a suitable factor in machine learning. Based on factors categorize the regression model for machine learning which helps for the selection of the model. The main purpose is to select the suitable types of regression depends upon the user's needs. Here need means what users want to predict and forecast and how it can be achieved? This review will support selecting a suitable regression machine learning methodology. It is also helpful to decide high size or low-size datasets. i.e., when the data size is small select linear or as suitable, and when data size is big which is deep learning model advice suitable regression. It suggests the mythology to select predictor variables and emphasize the limitation of the hypothesis.

2. Literature Review

According to E C Alexopoulos [3] "Linear regression is the procedure that estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable which should be quantitative. Logistic regression is similar to a linear regression but is suited to models where the dependent variable is dichotomous. Regression analysis in order to model its relationship. There are various types of regression analysis. All these methods allow us to assess the impact of multiple variables on the response variable."

According to *Yuki Hiruta and Yasushi Asami* [4] "A simple way for evaluating the different types of regression models from two points of view: the 'data fitting' and the 'model stability'. To figure out whether the relative positions among models coordinated by two criteria are reasonable enough or not, used the models that have a smoothing or complexity parameter, and confirmed whether the two criteria are able to represent such tradeoffs."

According to R. Reka [5] "Regression is one of the techniques for prediction analysis and data mining task. Each one has its own sense. These techniques vary in terms of type of response variable, explanatory variable and distribution, focused on the dissimilar types of regression techniques premeditated for various types of analysis and which types of regression used in context of different data sets."

According to *Dastan Hussen Maulud* and *Adnan Mohsin Abdulazeez* [6] "Regression modeling is a statistical method commonly used in research, particularly for observational studies. The proper choice of regression model, the choosing and presence of model variables are the key actions which should be established and properly controlled in order to achieve valid statistical results because the unavailability or misapplication of an appropriate regression modeling may cause to inaccuracies results."

3. Tools Explored for Regression Analysis

There are many software tools for regression analysis. Selection of tools be contingent on data set. Such as for logistic regression, data set must not have continuous, it should binary dependent variable. But, the best one can be declared based on requirements, few software to perform regression analysis are: (a) Statistical Practice for the Social Science (SPSS) (b) National Council for the Social Studies (NCSS) (c) Stata (d) Minitab (e) SAS (f) R (g) Matlab and (h) KNIME. The favourable tool for regression analysis is R and SAS, R can use the in-built function of linear model (LM) where the structure of model can be easily define and with summary function the outcomes are detailed. Here I explored in excel with free add-in RegressIt, which provides a 2-way interface between Excel and R, perform analysis in R without writing any code with combination of TableCurve2D for fitting data.

4. Different types of Regression

4.1. Linear Regression Model

This is trivial and easy form of regression [9] it represents by the formula of a straight line

$$Y = \alpha + \beta x \quad (1)$$

This equation controls the suitable standards for alpha and beta to expect dependent variables upon a given value of independent variables. Here independent variable(s) may constant or distinct and finds a linking among the regressed and regressor with best fit straight line. There is no connotation between two variables if the grid line in a simple linear regression marks smooth. There are two types of linear regression association positive and negative. When the representation of line tends on upper side of Y-intercept and sharp-line tends on upper side, positive linear association occurs. When the representation of line tends on lower side on the Y-intercept and some part of the representation graph tends towards X-intercept a negative linear association occurs. This model accepts that the association between input and output is linear. It does not sustenance other complicated trends. Real life examples used in forecasting by environmental scientist. Machine can adopt this process to forecast the weather and other side.

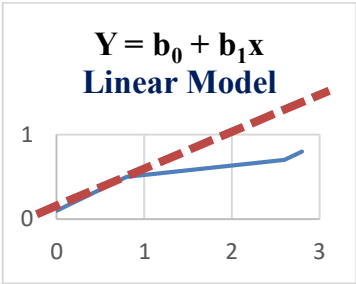

Fig.1. Linear Regression Model

4.2. Logistic Regression Model

Logistic regression built the possibility of occurrence of either success or failure. The dependent variables are binary in nature. The rate of dependent variable ranges from lower limit 0 to upper limit 1 denoted by the following calculation given below.

$$\text{Odds} = p/(1-p) \quad (2)$$

$$\beta_0+\beta_1X_1+\beta_2X_2+\ldots\ldots\ldots\ldots+\beta_nX_n \quad (3)$$

The nature of p means the possibility of survival characteristic of interest, one indicates the possibility of happening of incidence and zero indicates the possibility of not happening incidence.

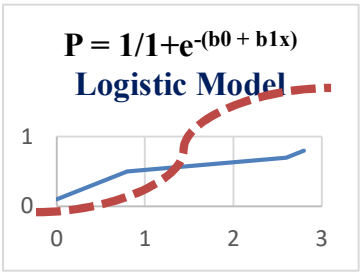

Fig.2. Logistic Regression Model

For example, Spam detection, credit card fraud, tumour prediction. This regression Model is useful in such a situation where there are only two conditions either yes or no which denoted with 1 and 0. The predicted values trends below the average limit between 0 and 1, generally = 0.5. It accepts all values below 0.5 and predicts the truth of reality in cyber security. Machine can use in cyber security applications to identify cyber fraud.

"This model is used to determine the chance whether a dichotomous outcome depends on one or more free (independent) variables" [35]. It is like linear regression but Conditional Distribution is Bernoulli instead of Gaussian and possibility circumscribed between zero and one.

4.3. Polynomial Regression Model

Polynomial regression functional only when the independent variable is greater than one. Polynomial equation denoted as:

$$Y = \alpha + \beta * x2 \qquad (4)$$

Linear datasets apply linear regression model but in case of non-linear datasets extremes output come out, in such situations Polynomial Regression adopted. Polynomial Regression is an alternative of Linear Regression the association between the independent variables and the dependent variable can be symbolized by polynomial.

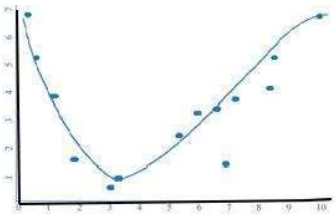

Fig.3. Polynomial Regression Model

It is more accurate to actual point of truth as usages polynomial in place of straight line, helpful in detection of truth. The regression line is a curve not a straight line for turn data circulation. Machine Learning can apply for this model where place of reaching human is difficult for example sea surface, dense forest, valley, hill areas.

4.4. Ridge Regression Model

When independent variables data is multicollinearity ridge regression [13] model applied. "In this multicollinearity data in smallest squares guesses are neutral, even though the least-squares guesses are neutral their variances are greater than the deviated observed value which is far from the accurate value. By collecting a degree of favor to the regression computes nearly, it decreases the ordinary faults" [14].

The equation can be denoted as:

$$Y = a + b*x + e \quad (5)$$

It resolves the multicollinearity badly-behaved through shrinkage constraint λ (lambda). Equation is:

$$= \underset{\beta \in Rp}{\operatorname{argmin}} \| y - x\beta \|_2^2 + \lambda \| \beta \|_2^2$$
$$Loss \qquad Penalty \quad (6)$$

This equation involves two mechanisms least square with the multiplication of β2; with the association of lambda (λ) where β represents coefficient. The regression line for the ridge model can be further to the least-square term to shrink the constraint and to have very little variance. This model machine can adopt for study water resources.

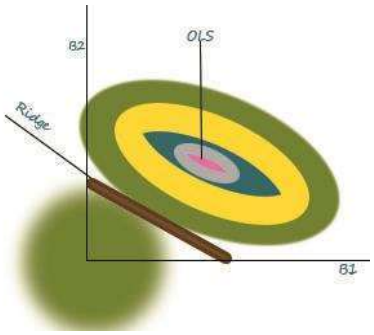

Fig.4. Ridge Regression Model

4.5. Bayesian Linear Regression Model

This model is based on the Bayes theorem to find out the value of regression constants. In the Bayesian model of regression, the subsequent circulation of the structures is resolute in place of outcome the least-squares. It is like both Linear Regression Model and Ridge Regression Model but more secure than the simple Linear Regression. Since it is a combination of linear and ridge, predicts the more accurate result. It draws the ranges and follows overlapping principle. So, it follows artificial intelligence. Machine can adopt this regression for audio/text recognition.

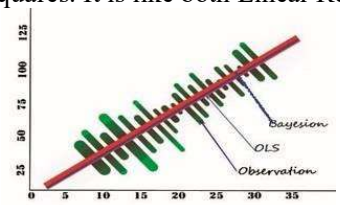

Fig.5. Bayesian Linear Regression Model

4.6. Lasso Regression Model

As per the name "Least Absolute Shrinkage and Selection Operator" is analogous to Ridge Regression Model but furthermore deals with strict the absolute size of the regression constants, accomplished of sinking the changeability and refining the correctness of linear regression models. "It disagrees ridge regression model focused to reliably pre-emptive the addition of the absolute principles of the guess's standards that ancestries a number of the constraint valuations to pick up to exactly zero" [15]. Machines are widely uses in agricultural crop, yield production and suitability for environment.

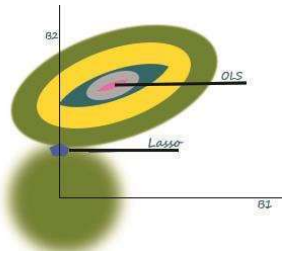

Fig.6.Lasso Regression Model

4.7. Comparative table

Table 1. Regression Model comparative Table.

Parameters	Linear	Logistics	Ridges	Lasso	Polynomials	Bayesians' Regression
Variable Types	Continuous dependent variable	Categorical dependent variables	Multicollinearity variables	Inbuilt variable selection as well as parameter shrinkage	Fit a nonlinear equation by taking polynomial functions of the independent variable	Directing and forecasting DV on the basis of IV
Estimation Method	Least square estimation	Maximum like-hood estimation	Higher coefficient, more significant	Least absolute shrinkage and selection operator	Estimate the unknown parameters in ordinary differential equation models	Parameter estimation equal to subsequent distribution multiply with likelihood distribution
Graphical overfitting/ Overfitting	Straight line boundaries decide underfitting /Overfitting	Curve decide underfitting /Overfitting	Bias to important features for underfitting/ Overfitting	Shrink towards mean data with high multicollinearity[15] for underfitting/Overfitting	Frequently fitted using least squares values for underfitting/Overfitting	Draws a range of lines with different estimate model parameters for underfitting/Overfitting
Relationship between DV and IV	The linear relationship between the dependent and independent variable	The linear relationship is not mandatory	Depends on the shrinkage parameter	If there are too many features, some of them are eliminated done by setting the coefficients to zero	Linear or curvilinear/polynomial, "Independent bias spread through unpleasant null value and continuous inconsistency OLS" [16]	When the number of data points increases, the lines begin to overlap because there is less uncertainty in the model parameters.
Output	Predicted integer value	Predicted binary value (0 or1)	Factor rate recoil nonetheless not = 0,	Regular machine learning trends	"Forecast result of polynomial model behaves regression factors with multiplication of dependent variables " [17]	Obtained from probability distribution. The output is produced from a normal distribution (where mean and variance are normalized)

Applications	Business domain, forecasting sales	Classification problems, cybersecurity, image processing	Water resources study, Non-orthogonal problem	Agricultural predictions and forecasting	Predict Sea surface temperature	"Automation of speech, text, and AI, Recognition and classification of images" [17]

5. Conclusion

A real-time operational implementation can be achieved throw regression. Regression Models utilized based on user necessities and requirements, according to the circumstances. Their advantage turn on individual-base depends upon data and its relationship. Linear useful for two variables, polynomial for compound variables whereas Logistic applies whenever dualistic decision either one or zero. Ridge follows multicollinearity, shrinkage squares of coefficients to reduce multicollinearity while Lasso behaves absolute standards in the penalty, rather than squares this tends to penalty values to zero which eliminated from the function. Models compared as per characteristics of models' parameters adopted with the help of different aspects. The variables' relationship forms a path of the data flow for the most frequent datasets. The estimation method has defined on models studies and produces some trends in the form of lines and curves. It provides the right path of data trends in the machine Learning. This review analysis of the regression module emphasis on the physical and virtual structure of the machine Learning. However, the best model for a certain purpose and future usage of applications is still unclear, it may consider for future research. The review adds a value to identify the nature and trends of data with the regression analysis and establish a function to determine the future selection of the regression in machine learning.

References

[1] *What is Regression? Definition of Regression? By Great Learning Team Contributed* by: Prashanth Ashok Sep 26, 2020 https://www.mygreatlearning.com/blog/what-is-regression.

[2] Stephen E. Brown Previous version 1 *regression and regression analysis* published online: 15 Feb. 2007(Google Scholar).

[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] E C Alexopoulos Hippokratia, *Introduction to Multivariate Regression Analysis* (2010) Dec; 14(Suppl 1): 23–28.

[5] Yuki Hiruta and Yasushi Asami, *A Method for Comparing Multiple Regression Models* (2016) CSIS Discussion Paper No. 141 - January

[6] R. Rekha, *Computing dissimilar types of regression analysis models* (2020) AIP Conference Proceedings 2261, 030070

[7] Maulud et al. *A Review on Linear Regression Comprehensive in Machine Learning* (2020) Journal of Applied Science and Technology Trends Vol. 01, No. 04, pp. 140 –147

[8] Jiawei Han and Micheline Kamber (2006), *Data Mining Concepts and Techniques*, published by Morgan Kauffman, 2nd ed.

[9] Breese, J. S., Heckerman, D., and Kadie, C. (1998). *Empirical Analysis of Predictive Algorithms for Collaborative Filtering.* In Proceedings of UAI-1998: The Fourteenth Conference on Uncertainty in Artificial Intelligence.

[10] CiteSeer (2002). CiteSeer *Scientific Digital Library*. http://www.citeseer.com.

[11] Duda, R. O. and Hart, P. E., (1973). *Pattern Classification and Scene Analysis*. John Wiley & Sons.

[12] Hoerl AE, Kennard RW, (1970) *Ridge regression: biased estimation for nonorthogonal problems. Technomterics,* 12:55-67.

[13] Tibshirani R, (1996) *Regression shrinkage and selection via the lasso*. Journal of the Royal Statistical Society, Series B, 58:267-288.

[14] Developer Corner Hands (2020) *On-Implementation of Lasso and Ridge Regression* on 11/08/2020

[15] Quality Thought, (2019) *Machine Learning Materials Artificial Intelligence vs Machine Learning vs Deep Learning* Uploaded by Manish on Jun 15, 2019

[16] Takashi Isobe, Eric D. Feigelson and Michael G. Akritas and Gutti Jogesh Babu (1990*) Linear Regression in Astronomy.* I Received 1989 June 29; accepted 1990 May 22 (Google Scholar).

[17] Nadheesh Jihan, (2018) *Bayesian Learning for Machine Learning: Part II* - Linear Regression October 23, 2018

[18] Odd Aalen, University of TromsøNorway, *A Model for Nonparametric Regression Analysis of Counting Processes* (Conference Paper)

[19] C. Y. Wan, Naisyin Wang, Suojin Wang, *Regression Analysis When Covariates Are Regression Parameters of a Random Effects Model for Observed Longitudinal Measurements* published: 24 May 2004

[20] Meyer, P., *Un cours sur les integrales stochastiques. Séminaire de Probabilit6s X*. Lecture Notes in Mathematics.Springer-Verlag, Berlin 511, 246 - 400 (1975) (Google Scholar)

[21] Meyer, P., *Un cours sur les integrales stochastiques. Séminaire de Probabilit6s* X. Lecture Notes in Mathematics.Springer-Verlag, Berlin 511, 246 - 400 (1975) (Google Scholar)

[22] Prentice, R. L., *Exponential survivals with censoring and explanatory variables*. Biometrika, 60, 279 - 288 (1973).MathSciNetzbMATHCrossRef (Google Scholar)

[23] Prentice, R. L., *Models for survival analysis,Paper presented at the regional meeting of WNAR, IMS, and ASA,Corvallis*, Oregon, June 16, 1975 (Google Scholar)

[24] Rao, C. R., *Linear statistical inference and its applications*. Wiley, New York 1973.zbMATHCrossRef (Google Scholar)

[25] P. Kovac, D. Rodic, V. Pucovsky, B. Savkovic & M. Gostimirovic *Application of fuzzy logic and regression analysis for modeling surface roughness in face milling,* Published: 19 January 2012(Google Scholar)

[26] Chao-Ying Joanne Peng, Kuk Lida Lee&Gary M. Ingersoll *An Introduction to Logistic Regression Analysis and Reporting*, Published: 2 Apr. 20102

[27] Tunaru, R., Clark, E., & Viney, H. (2005). *An option pricing framework for valuation of football players*. Review of financial economics, 14(3-4), 281-295

[28] Carmichael, F., Forrest, D., & Simmons, R. (1999). *The labour market in association football: who gets transferred and for how much?*. Bulletin of Economic Research, 51(2), 125-150.

[29]Fullard, J. (2012). *Investigating Player Salaries and Performance in the National Hockey League*

[30] Peck, K. (2012). *Salary Determination in the National Hockey League: Restricted, Unrestricted, Forwards, and Defensemen*

[31] Louivion, S., & Pettersson, F. (2017). *Analysis of Performance Measures That Affect NBA Salaries*

[32] Castillo Ramirez, M., Viloria, A., Parody Muñoz, A., & Posso, H. (2017). *Application of Multiple Linear Regression Models in the Identification of Factors Affecting the Results of the Chelsea Football Team, International Journal of Control Theory and Applications*, 10(18), 7-13

[33] Newman, D. (2016) *Predicting Transfer Values in the English Premier League*

[34] Farrar, D. E., & Glauber, R. R. (1967). *Multicollinearity in regression analysis: the problem revisited. The Review of Economic and Statistics*, 92-107

[35] McCallum, B. T. (1970). *Artificial orthogonalization in regression analysis. The Review of Economics and Statistics,* 110-113