International Research Journal of Human Resource and Social Sciences
ISSN(O): (2349-4085) ISSN(P): (2394-4218)
Impact Factor 6.924 Volume 9, Issue 12, December 2022
Website- www.aarf.asia, Email : editoraarf@gmail.com

A STUDY OF FAMILY OF REGRESSION MODELS FOR LINEAR REGRESSION

HARIPAL SINGH

RESEARCH SCHOLAR, SUNRISE UNIVERSITY, ALWAR

DR. SATENDRA KUMAR

ASSOCIATE PROFESSOR, SUNRISE UNIVERSITY, ALWAR

ABSTRACT

Analysing the data and using it for the predictions of the future events has become a most important aspect in this era where data is so rapidly generated everywhere. The most popular and commonly used statistical model for predicting the response or outcome variable is Regression. Many linear and non-linear models were developed for analysis of the data and estimation. In statistical literature one may find a large number of research articles on various regression models and the reason for the large number of publications is the large variety of regression models. The aim of all these regression models is-It indicates significant relationships between the predictor variables and the response variableit indicates the strength of the impact that each predictor variable has on the response variable. Comparisons between the effects of predictor variables are also possible in Regression analysis even if they are measured on different scales. That helps to drop or estimate variables that are not really useful while identifying the best set of variables for building a predictive model.The three factors are considered important while determining the type of model. and they are i) the number of independent or predictor variables in the model, ii) the functional form of regression model or the shape of the regression line whether linear or non-linear and iii) the nature of the dependent or response variable in the model. Most of the classical

regression models are based on one or more of the above considerations in different proportions. It is still possible to develop a new type of regression model by using a new combination of the above considerations. Before attempting to develop a new model, it is necessary to understand the eight most commonly used regression models. The models are briefly described below.The most popular and commonly used technique of predictive modelling in Regression models is a Linear Regression. It is represented by a straight line (as the relation is linear). This regression line is optimal in the sense that it minimizes the total error sum of squares of prediction.

KEYWORDS: Family, Regression Models, Linear Regression, linear and non-linear models

INTRODUCTION

Statistics has always been considered to be a branch of science that consists of data collection and analysis of the collected data and interpreting the results obtained from such analysis related to the dataset in hand. The purpose of statistical activities is to develop and apply methodologies for extracting useful information from data. Statistical activities may include one or more of the following.

- Sample surveys and design of experiments for collection of the data for understanding the phenomenon under investigation.
- Visualization of the collected sample data is important because it helps us to understand the nature and structure of collected data.
- The relationships present between different variables are modelled with the objective of obtaining prediction of the response variable when values of other (predictor) variables are known.
- Drawing statistical inference about underlying parameters of statistical models and testing various hypotheses about these parameters.

In the pre-computer ages, data used to be processed manually. It used to be a very tedious and time consuming process. During the computer age, statistical packages were developed to process the data, which can vary from small to large magnitude. During recent times, when data is generated rapidly in big volumes and is required to be processed very quickly. The new

phenomenon of collecting data of huge magnitude, generated almost continuously over time and mostly stored automatically is known as the big data revolution. The big data revolution has resulted in having data everywhere and the need of the hour is a smart and accurate analysis of such data. For example, big data that are generated in a company (data related to its employees) are to be processed for formulating company policies on recruitment of new employees. Some malls and departmental stores are interested in analysing sales data and then predicting customer behaviour with a view to provide offers or good services to the customers. This big data is so large and complex that traditional methods cannot manage it. There is a need for fast and accurate methods of data analysis that can be implemented on a digital computer. The study reported in this thesis is related to one such method, known as regression. The overall goal of present study is to develop and present most generalized form of regression. This generalized regression model includes all variations and formulations of the regression model that are already developed as well as the variations that may be developed in the future can be shown to be special cases of this generalized model. Let us first understand the background of the statistical regression model as a prediction model or the model that provides a prediction formula.

REGRESSION AS A PREDICTIVE MODEL

One of the most popular statistical methods utilized in building statistical models that can address prediction problems is regression analysis. As a consequence, regression is known as the most commonly used technique of prediction modelling. The literature on regression models is full of several types of regression. However, most of the practicing analysts know a very few types of regression. Almost all the analysts know linear regression and logistic regression. Some of them may know the concept of regularization and hence may be aware of ridge regression and LASSO regression. The statistical literature on regression mentions more than 15 different types of regression. In addition, some methods are used for partitioning data so that a better regression model can be obtained by suitably partitioning given data. The reason for most of the analysts not being aware of many types of regression is that these are either too complex in nature or are not covered by standard text books on regression. The present study aims at generalizing most of these various types of regression so that any particular type of regression can be shown to be a special case of this generalized form. It is then easy to remember the generalized form of regression as the standard regression model, with the option of specifying conditions for each

particular case as a special case. Every special case is applicable under some restrictions, either on the parameter space, on the data space, or on the relationship that the predictor variable(s) has with the response variable. Some such special cases are known by names, whereas some of these are known as restricted or regularized regression models. A data scientist cannot spend too much time in understanding the intricacies of modelling in order to identify the most appropriate model to be used in the given situation. This study will help a data scientist to take a systematic approach to developing the most appropriate regression model so that the model is optimal in terms of the performance and, at the same time, is computationally least complex among competing candidate models. Every regression model is developed with the ultimate goal of predicting the response variable corresponding to known or given values of the predictor (explanatory) variable(s).

The Multiple Linear Regression Model

One of the typical problems in data analysis is problem of prediction. A prediction problem involves several variables, where one variable is supposedly affected by variations in the other variables. The affected variable shows the effect of variations in other variables and is therefore called the dependent or response variable. The remaining variables may or may not affect the response variable directly, but we cannot conclude one way or the other unless we carry out an appropriate statistical analysis of data on these variables. We therefore begin with a dataset that has the response variable, conventionally denoted by the letter Y and a number of potential predictor variables, conventionally denoted by the letter W with a subscript to distinguish between the different predictor variables. Suppose a dataset has „n‟ no. of observations on each of (P+1) number of variables. The dependent or the response variable is one of these (P+1) no. of variables. The remaining P no. of variables are predictor variables. Denote the response (target) variable by Y and P no. of predictor (or explanatory) variables by W1, W 2… W P. Further, suppose that there are n observations on each of the P+1 variables. It is a common practice to organize these n * (P+1) values in the form of a data matrix. This data matrix obviously has n rows, every row representing one observation, and P+1 columns, every column representing one variable, P predictor variables and one response variable. The objective of prediction analysis is to determine a mathematical relationship or association between predictor variables and the response variables. The purpose of doing so is to use this mathematical relation to predict the target variable for given values of the explanatory variables.

Multiple linear regression is most widely used and popular prediction model in statistical literature. It stipulates that the explanatory variables are a related linearly to the response variable. This model is defined by the expression written as,

Y = b0 + b1 W 1 + b2 W 2 + … + bP W P +ε, where b0, b1, b2, …,bP are called regression coefficients and *e* is called the residual. The best multiple linear regression model is obtained by least square principle, where the sum of squared errors (SSE) is minimized. Errors are measured by taking deviation of observed values of explanatory variable from its predictions. This model is called the least squares regression model and the method used for deriving this model is called Ordinary Least Squares (OLS) method.

It should be noted that the multiple linear regression model is optimal (that is, best) only under certain conditions or assumptions. These assumptions are as follows.

1. Linearity of the relationship. The relation or association of the explanatory and the target variable is assumed to be linear. Linearity is often verified visually with help of scatter plots of the predictor variables and the response variable. There is no formal statistical test for linearity of the relationship of a predictor variable with the response variable. It should be noted that the concept of linearity is one of the most misunderstood, misconceived, and misinterpreted in the literature. The linearity in the linear regression relates to the assumption that the regression formula is linear function of parameters. It is usually stated that the model is called linear regression because it‟s expression involves the power of predictor variable at most one. It is enough to mention that polynomial regression is an example of multiple linear regression, in which the regression formula is not linear in the predictor variables, but is linear in its parameters.

2. Independence among predictor variables. The P independent variables are mutually independent. In other words, the predictor variables do not affect one another. As a consequence, no predictor variable affects the relationship of any other predictor variable with the response variable. It is not possible to have predictor variables to be totally independent of one another. However, the relationships among predictor variables should not be so strong as to make the variance-covariance matrix of predictor variables singular or near singular. There is no formal measure of multicollinearity among the predictor variables. Some indicators like the variance inflation factor (VIF) are available, but such indicators have a limited utility.

3. Independence of errors. The error term is independent of the target variable as well as of the explanatory variables. This assumption is most often verified by drawing a scatter plot of residuals against observations on the response variable against its predictions of the response variable.

4.Normality of errors. Residuals (that is, observations on the error variable) are distributed normally with mean (μ) equal to zero and variance (σ2) should be constant. This assumption is verified through a normal Quantile-Quantile plot of error terms when the model is fitted and computing the predicted values and residuals are computed.

5. Homoscedasticity of errors. The error variance does not change with response variable, either observed or predicted. This assumption can also be verified only graphically by drawing a scatter plot of residuals against the response variable (using either observed values of predicted values).

For illustration, consider the problem of analysing the data given below of 20 observations on the result of an examination (y) and number of study hours (Hr.). The purpose of the analysis is to develop a method for predicting the result of the examination (y) when the number of study hours (Hr.) is given (or specified).

No.	1	2	3	4	5	6	7	8	9	10	11	12
y	1	1	0	0	1	1	1	0	1	0	1	1
Hr.	5	6	3	2	2	4	5	3	5	3	5	4

No.	13	14	15	16	17	18	19	20
y	0	1	1	0	1	1	1	1
Hr.	2	3	4	3	5	4	6	5

First, consider the simple linear regression model for this purpose. It is easy to enter values of Hr. and y in R as a data frame and use the function lm to apply simple linear regression of y on Hr. This provides the following estimates of the regression coefficients.

TYPES OF THE REGRESSION MODEL

Many different types of regression model are found in statistics literature. We briefly review some of these models, so that it can be made clear how these can be shown to be special cases of the proposed generalization.

1. Linear Regression. Simple linear regression is the simplest regression model. It is interesting to note that origin of simple linear regression in order to understand why the normal (that is, Gaussian) distribution is so important in the linear regression model. Suppose, W and Y are the two random variables following bivariate normal distribution with respective means $\mu_w$and μ_y, respective standard deviations σ_w and σ_y, and correlation coefficient ρ (rho). Then, the marginal distribution of W is N(mean=μ_w, standard deviation= σ_w).

The conditional distribution of Y | W = w is a normal distribution having mean=[$\mu_y - \rho * \sigma_y / \sigma_w$ (w – μ_w)]and variance=[$\sigma_y^2 (1 - \rho^2)$] . This shows that the conditional expectation of Y given W = w is linear function of independent variable w. Since regression of Y on W is defined as the conditional expectation of Y | W, linear regression is linear when W and Y follow a bivariate normal distribution. Further, the residual is also distributed normally with zero mean and variance which is independent of W or Y. The multiple linear regression is similarly the natural choice when the dependent variable Y and the P independent variables W 1, W 2, …, W P jointly follow the multivariate normal distribution. Derivation of the regression is straightforward if we notice that conditional distribution of Y given W 1 = w1, W 2 = w2, …, W P = wP is a linear function of the variables w1, w2, …, wP.

2. Polynomial Regression. The scatterplot of W and Y sometimes indicates that the relationship of W with Y may not seem be linear. It is common in such cases to try the polynomial regression. Polynomial regression model involves higher order powers of the predictor variable W and is of the following form. γγ

$$Y = \gamma_0 + \gamma_1 W + \gamma_2 W^2 + \gamma_3 W^3 + \ldots + \gamma_k W^k + \varepsilon.$$

It is the expression of the kth degree polynomial regression. What is interesting to note that the polynomial regression is non-linear in predictor variable and not in regression coefficients If we define W 1 = W, W 2 = W^2, W 3 = W^3, … W k = W^k, then the polynomial regression model will be same as the multiple linear regression model. This is this reason why not much literature is

available on multinomial regression. The polynomial regression model sometimes contains terms that involve products of predictor variables or their powers, such as W 1 ⋅ W 2, W 2 ⋅ W 3,

$W_1^2 \cdot W_2$, and so on. Even then, new predictor variables are introduced for each of these, so that the new model is still a multiple linear model of regression. It can be very tempting to fit highly involved polynomial regression models to data because the more complex the polynomial expression the better is the fit of the model. What is important to keep in mind is the fact that introducing to many terms in expression of polynomial regression can lead to the overfitting problem.

3. Logistic Regression. The response variable in the logistic regression model is binary, that is, has only two possible values, 0 and 1, often described as failure and success, respectively. Obviously, one can say that linear regression model is not seem to be appropriate because a straight line is unbounded, while the binary response variable is bounded. The problem of predicting the value of the target variable (response) is changed to problem of predicting probability p of success, that is, probability of value 1 of the response variable. The probability p is bounded between 0 and 1 and still cannot be predicted with help of a linear function. The odds ratio p / (1-p) is non-negative and is not bounded above. The logistic regression model uses log (p / (1-p)) as the response variable. It may be noted that log (p / (1-p)) covers the entire real line as p moves over the unit interval from 0 to 1. As it is real-valued, this function can be predicted using a linear function of the repressors. So, expression for logistic regression is given as follows. log (p/(1-p)) = γ_0+ γ_1 W 1 +γ_2 W 2 + … + γ_p W P + ε.

What is important to note in case of the logistic regression is that no predicted variable affects original response linearly. As a result, coefficients in logistic regression cannot be easily interpreted. The effects of predictor variables are multiplicative and not additive, like in case of linear regression. Moreover, logistic model of regression violates the assumption of homoscedasticity. Further, since the response variable is binary, it does not follow the normal distribution. Even residuals also do not follow the normal distribution for the same reason.

4. Quantile regression. The model of quantile regression differs from the model of linear regression in the following sense. The linear regression model attempts to find the conditional expectation of the response variable corresponding to given values of explanatory variables, while

quantile regression model attempts to locate the specified quantile of conditional distribution of response corresponding to given values of the explanatory variable(s). In other words, quantile regression model extends of the concept of a quantile to the concept of conditional quantile. If F(y) = P(Y < y) is a distribution function of a random variable Y, then the qthquantile (Q (q)) of Y is defined in terms of the inverse distribution function Q(q) = inf { y : F(y) > q } = F-1(q) for 0 < q < 1. For example, the median is Q (1/2). Given a random sample y1, y2, …,yn, it is known that the sample median (Md) minimizes the total absolute deviation of sample values around it, namely minMd$\sum_{i=0}^{n}$| yi – Md |. It should then be obvious that the quantile regression model cannot use the squared error loss function, because the quantiles do not have the property of the sample mean that the squared deviations are minimum when taken from the mean. Instead, the objective function that is minimized in quantile regression is given by minz in R {Sum wq(yi – z), where wq(w) = w (q – I(w < 0)) and I() is the indicator function. The linear regression model uses the property of sample mean of minimizing the total of square of the deviations from the sample observations and applies it to obtain the conditional mean as the optimal solution to the problem of minimizing squared error loss function. Similarly, the quantile regression model extends the property of the sample quantile that minimizes the total weighted deviations from sample values when the weights are given by the function wq() defined above and obtains the conditional quantile function for any specified quantile q, 0 < q < 1. Quantile regression is preferred when data is heteroscedastic or distribution of the dependent variable is skewed. Quantile regression is also robust to outliers. Note that the coefficients of regression in quantile regression are very different from the regression coefficients in linear regression. If this does not happen, then the use of quantile regression is not justified.

5. Ridge Regression. Before defining ridge regression, we shall quickly introduce the concept of regularization used in regression. Regularization is a way to handle the problem of overfitting, where it can be seen that the model fits well to the training data but does not perform equally well on test data. Regularization controls the objective function with the help of penalty function. Regularization is useful in the case where sample size is too small as compared to no. of independent variables and also in presence of multicollinearity between the independent variables. Two most common methods of regularization use the L1 norm and L2 norm as the penalty function. The L1 norm, also known as absolute norm restricts the regression coefficients by the

condition that the absolute values of regression coefficients add to unity. The L2 norm , also known as quadratic norm restricts the regression coefficients by the condition that squares of regression coefficients adds up to unity. Ridge regression uses the L2 norm as the regularization method. The objective function for the ridge regression is to minimize $\Sigma(yi - \gamma_0 - \gamma_1 w1i - \gamma_2 w_{2i} - \ldots - \gamma_p w_{pi})^2 + \lambda \Sigma \gamma_i^2$. The normal equations for ridge regression have the solution $\hat{\gamma} = (W' W + \lambda I)^{-1} W' y$. Ridge regression was originally proposed to resolve the problem of multicollinearity. A consequence of regularization is that it is no more meaningful to assume the error terms to be normally distributed.

6. LASSO regression. Lasso regression was proposed as an alternative to the ridge regression by using L1 norm in place of L2 norm used in ridge regression. LASSO stands for acronym of Least Absolute Shrinkage and Selection Operator. It minimizes the objective function $\Sigma(yi - \gamma_1 w1i - \gamma_2 w2i - \ldots - \gamma_p wpi)^2 + \lambda \Sigma |\gamma_i|$. Note that LASSO does not regularize the intercept because all variables are standardized before fitting the model. As a consequence, the regression passes through the origin and hence has no intercept. There is no explicit mathematical solution for LASSO regression and hence the regression coefficients are obtained using an iterative process with help of a statistical software. As the name suggests, LASSO results in automatically selecting variables for use in the model and resolves the problem of multicollinearity. In this sense, LASSO is better than ridge regression. However, ridge has the advantage of being computationally more efficient. There is no straightforward comparison between ridge and LASSO. Both should be fitted to training data and selection should be on the basis of their performance on test data. The model that performs better for test data should be selected.

7. Elastic Net Regression. Elastic net regression was also proposed in presence of multicollinearity, but when it is not clear as to whether ridge regression is better or lasso regression is better. Elastic net regression is a mixture of ridge and lasso regression in the sense that it uses both L1 norm and L2 norm. All the variables are standardized before fitting the regression model, and therefore the model has no intercept term. The objective function for elastic net regression is shown below.

$\Sigma(yi - \gamma_0 - \gamma_1 W 1 - \gamma_2 W 2 - \ldots - \gamma_p W p)^2 + \lambda_1 \Sigma \gamma_i^2 + \lambda_2 \Sigma |\gamma_i|$. It is then obvious that elastic net regression does not assume errors to be normally distributed.

CONCLUSION

The research reported in the current thesis is an attempt to develop a generalization of all regression models so that different types of regressions become special cases of the generalized form. The most general form of regression is proposed to be linear in all variables, response as well as predictor. Special cases are obtained by defining the response variable or predictor variables suitably in consideration of problem under consideration. For example, logistic regression is obtained by defining the response variable as –log [p/(1-p)], where p= P[Y=1], and Y is binary variable representing the original response. Similarly, polynomial regression is obtained by defining additional predictor variables as powers or products of (possibly powers of) observed predictor variables. These are among more familiar cases and do not require any justification. What is not obvious is partition-based non-parametric regression models that are difficult to represent in a mathematical form. The regression tree model as defined in the literature assumes the response variable to have a single value for all observations at a leaf node. The research reported in this thesis proposes to fit a local linear regression at the leaf level. This leads to one linear regression model per leaf node. On one hand, it appears that the proposed modelling technique demands more work in fitting one linear regression model per leaf node. On the other hand, however, this technique not only improves the accuracy of the predictions obtained at the leaf node level, it also improves the precision of prediction by reducing the prediction error by making sure that observations in the leaf nodes are most suitable for a linear regression model. A common misunderstanding about the importance of assumptions found in the literature on regression analysis is that verifying assumptions is the justification for the model. The fact of the matter is that a model should be justified by the nature of the relationship between involved variables, rather than their statistical properties like normality, homogeneity, or homoscedasticity. The assumptions of the regression model are of two types. One of the assumptions is related to the behaviour of every random variable involved in the model, including the residual or error. What is often overlooked is consideration that error is normally distributed with zero mean and some finite variance equal to $\sigma 2$. The dependent variable Y is not assumed to follow any particular distribution. Similarly, predictor variables are also not assumed to follow any particular distribution. The only distributional assumption is that the conditional expectation of Y for specified values of predictor variables follows the normal distribution.

REFERENCES

1. AaramKarallic (1992). Linear Regression in regression Tree Leaves. Proceedings of ECAI-92, pp 440-441.
2. Aaron K. Han (1986). Non-parametric Analysis of a Generalized Regression Model: The maximum Rank correlation coefficient. Journal of econometrics, Vol. 35, pp 303-316.
3. AchimZeileis, Christian Kleiber and SiemonJackman (2008). Regression Models for Count Data in R. Journal of Statistical Software, Vol.27, Issue 8.
4. Akhil Kumar, Vithala R. Rao and Harsh Soni (1995). An empirical comparison of Neural Network and Logistic Regression models. Marketing Letters, Vol.6, No. 4, pp 251-263.
5. Alex J. Smola, Bernhard Scholkopf and Klaus-Robert Muller. (1998). General Cost Functions for Support Vector Regression. In proceeding of 8th international conference on neural network, pp 79-83.
6. Aluisio JD Barrrows and Vania N Hirakata (2003). Alternatives for logistic regression in cross-sectional studies. BMC Medical Research Methodology, pp.3-21
7. Anantha M. Prasad, Louis R. Iverson and Andy Liaw (2006). Newer CART Technique. Ecosystems, Vol. 9, pp. 181-199.
8. Andreas Bulling, Jamie A Ward, Hans Gellersen and Gerhard Troster (2010). Eye movement Analysis for activity recognition using electro oculography. IEEE transactions on pattern analysis and machine intelligence, Vol. 33, Issue 4, pp 741-753.
9. Andrew C. Comrie (1997). Comparing Neural Network and Regression Models for Ozone Forecasting. Journal of Air & Waste management association, Vol. 47(6), 653-663.
10. Andrew C. Comrie (2012). Comparing Neural Networks and Regression Models for Ozone Forecasting. Journal of the Air & Waste Management Association, Vol. 47(6), 653 - 663.
11. Andrew S. Lan, MungChaiang and ChristophStuder (2018). Linearized Binary Regression. arXiv: 1802.00430vl [stat.ML]

12. Angel M. Felicisimo, Aurora Caurtero, Juan Remondo and EliaQuiros Rosado (2012). Mapping Landslide susceptibility. Landslides, Vol. 10, pp 175-189.

13. Arthur E. Hoerl and Robert W. Kennard (1994). Ridge Regression: Biased Estimation for Nonorthogonal Problems. American Statistical Association, Vol.42, No.1, pp 80-86.

14. Communications in Statistics, Vol. 32, No.2, pp 419-435.

15. B.M.E. Moret (2009). A New Implementation and Detailed Study of Breakpoint Analysis.Electrical and Computer Engineering Technical Reports.

16. Bin Gu et.al (2015). Incremental learning for υ- Support Vector Regression. Neural Networks, Vol. 67, pp 140-150.

17. Brigit Strikholam (2006). Determining the number of breaks in a piecewise linear regression model. SSE/EFI Working Paper Series in Economics and Finance 648, (2006)

18. Bruce E. Hansen (1992) Testing for Parameter Instability in Linear Models. Journal of Policy Modelling, Vol. 14, No. 4, pp 517-533.

19. BulentTiryaki (2009). Estimating Rock Cuttability using Regression Trees and Artificial Neural Networks. Rock Mech Rock Eng, Vol. 42, pp. 939-946.

20. Byron P. Roe and Hai- Jun Yang (2004).Boosted Decision Trees as an Alternative to ANNs for Particle Identification. Nuclear Instruments and Methods in Physical Research Section A Accelerators Spectrometers Detectors and Associated Equipment 543(2-3).

21. Cande V. Ananth and David G. Kleinbaum (1997). Regression Models for ordinal responses. International Journal of epidemiology (1997), Vol.26, No. 6, 1323-1333.