Understanding Predictive Models Using R Software

Understanding Predictive Models Using R Software

By

Kirti Raskar

Cambridge Scholars Publishing

Understanding Predictive Models Using R Software

By Kirti Raskar

This book first published 2024

Cambridge Scholars Publishing

Lady Stephenson Library, Newcastle upon Tyne, NE6 2PA, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN: 978-1-0364-1529-7
ISBN (Ebook): 978-1-0364-1530-3

This book is dedicated to the memory of my mentor,
Dr. Sharad D. Gore
for all his support and faith in me.
My success would not have been possible
without his valuable guidance.

TABLE OF CONTENTS

ABBREVIATIONS

glm- generalized linear model
i.i.d.- independently identically distributed
lm- linear model
1Q and 3Q- first quartile and third quartile
t_value- value of t statistic
Pr(>|t|)- p value related to t statistic
Vs- Versus
S.E. – standard error
l.o.s.- level of significance
Df- degrees of freedom
rpart- recursive partitioning and regression tree
SSE- error sum of squares
MSE- mean of error sum of squares

Notations

$x_1, x_2 \ldots x_k$: k observations on variable X
μ and σ^2: means and variance
μ_x and μ_y: means of variables X and Y
σ_x^2 and σ_y^2: variances of variables X and Y
$\alpha, \beta_1, \beta_2 \ldots . \beta_k$: regression coefficients

CHAPTER 1

INTRODUCTION

Statistics has always been considered a branch of science that consists of data collection, data analysis, and interpretation of the results obtained from such analysis related to the dataset. Statistical activities aim 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 to understand the nature and structure of collected data.
- Modelling the relationships between variables to predict the response variable when values of other (predictor) variables are known.
- Drawing statistical inferences 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. Recently, data has been generated rapidly in significant volumes and is required to be processed very quickly. The new phenomenon of collecting data of enormous magnitude, generated almost continuously over time and primarily 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 to have a bright and accurate analysis of such data. For example, big data generated in a company (data related to its employees) will be processed to formulate company policies on recruiting new employees. Some malls and departmental stores are interested in analysing sales data and predicting customer behaviour to provide offers or good customer services. 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 relates to one such method, regression. The overall objective of the present study is to develop and present the most generalized form of the regression model so that all the 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 exceptional 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.

1.1 REGRESSION AS A PREDICTIVE MODEL

Regression analysis is one of the most popular statistical methods for building models that can address prediction problems. 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 practising analysts know very few types of regression. Almost all 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 to partition data to obtain a better regression model by suitably partitioning the data. Most analysts are not aware of many types of regression because they are either too complex or need to be covered by standard textbooks on regression. Every particular case applies under some restrictions, either on the parameter space, data space, or relationship between the predictor variable(s) and the response variable. Names know some of these exceptional cases, whereas some are known as restricted or regularized regression models. A data scientist can only spend a little bit of time understanding the intricacies of modelling to identify the most appropriate model to use in the given situation. This book will help a data scientist take a systematic approach to developing the most appropriate regression model so that the model is optimal in terms of performance and, simultaneously, is computationally least complex among competing candidate models. Every regression model is developed to predict the response variable corresponding to known or given values of the predictor variable(s).

THE MULTIPLE LINEAR REGRESSION MODEL

One of the most common problems in data analysis is the problem of prediction. A prediction problem involves several variables, where one variable is supposedly affected by variations in the other variables. The

affected variable is called the dependent or response variable. The other variables may or may not affect the response variable directly. However, we can only conclude one way or the other if we conduct an appropriate statistical data analysis of these variables. Therefore, we begin with a dataset with the response variable conventionally denoted by the letter Y and several potential predictor variables conventionally denoted by the letter X with a subscript to distinguish between the different predictor variables. Suppose a dataset has n observations on each of P+1 variables. One of these P+1 variables are the dependent or the response variable, and the remaining P variables are predictor variables. Let Y denote the response variable, and X_1, X_2, …, X_P denote the P predictor variables.

Further, there are n observations on each of the P+1 variables. Organizing these n * (P+1) values in the form of a data matrix is a common practice. This data matrix has n rows corresponding to the n observations and P+1 columns corresponding to the P predictor variables and one response variable. The objective of prediction analysis is to determine a mathematical relationship between the predictor variables and the response variables so that this mathematical relationship can be used to predict the value of the response variable when the values of the predictor variables are known.

Multiple linear regression is one of the most common prediction models in the statistical literature. This model stipulates a linear relationship between the predictor and response variables. The formula defines the multiple linear regression model

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_P X_P + \varepsilon,$$

where b_0, b_1, b_2, …, b_P are called regression coefficients, and ε is called the residual.

The best multiple linear regression model is obtained by minimizing the total of squared residuals, that is, errors measured by taking differences between observed values and predicted values of the response or dependent variable. The model obtained in this way is called the least squares regression model, and the method used to derive this model is known as the ordinary least squares (OLS) method.

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

1. **Linearity of the relationship**. The relationship between the predictor variables and the response variable is assumed to be linear. Linearity is often verified visually with the help of scatter plots of the predictor variables and the response variable. There is no formal statistical test for the linearity of the relationship between a predictor variable and the response variable. It is important to note that linearity is one of the most misunderstood, misconceived, and misinterpreted in the literature. The linearity in the linear regression refers to the fact that the regression formula is a linear function of the parameters. It is usually stated that the regression is called linear regression because the regression formula is a linear function of the predictor or independent variables. Polynomial regression is an example of multiple linear regression, where the regression formula is not linear in the predictor variables but is linear in its parameters.
2. **Independence of the predictor variables**. The P predictor variables are mutually independent. In other words, the predictor variables do not affect one another. Consequently, no predictor variable affects the relationship of any other predictor variable with the response variable. It is not possible to have predictor variables that are independent of one another. However, the relationships among predictor variables should be weak enough 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 need more utility.
3. **Independence of errors**. The error term is independent of the response variable and the predictor variables. This assumption is often verified by drawing a scatter plot of residuals against observed values of the response variable and predicted values of the response variable.
4. **Normality of errors**. The residuals (observations on the error variable) are distributed generally with zero mean and a constant variance sigma-square. After fitting the model and computing the predicted values and residuals, this assumption is verified through a standard Q.Q. plot of residuals.
5. **Homoscedasticity of errors**. The variance of the error terms does not change with either observed or predicted values of the response variable. This assumption is also verified graphically by drawing a scatter plot of residuals against the response variable (using either observed or predicted values).

As an illustrative example, consider the problem of analysing the following data 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
y	1	1	0	0	1	1	1	0	1	0
Hr.	5	6	3	2	2	4	5	3	5	3

No.	11	12	13	14	15	16	17	18	19	20
y	1	1	0	1	1	0	1	1	1	1
Hr.	5	4	2	3	4	3	5	4	6	5

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

Call: lm (formula = y ~ Hr.)

Coefficients: (Intercept): .0.2827, Hr.: 0.2488

We have the following result if we want information regarding the significance of these estimates.

Coefficients:	Estimate	St. Error	t-value	Pr(>|t|)	
(Intercept)	0.28271	0.26518	1.066	0.30047	
Hr.	0.24879	0.06403	3.885	0.00108	**

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘’ 1
Residual standard error: 0.3562 on 18 degrees of freedom
Multiple R Squared: 0.4561, Adjusted R Squared: 0.4259
F Statistic: 15.1 on 1 and 18 D.F., p-value: 0.001085

Since y is a binary variable, taking values 0 and 1, the predicted values of y are rounded in the following table to show observed values of y and rounded (to the nearest integer) values of predicted values of y.

	1	2	3	4	5	6	7	8	9	10
y	1	1	0	0	1	1	1	0	1	0
Hr.	1	1	0	0	0	1	1	0	1	0

	11	12	13	14	15	16	17	18	19	20
y	1	1	0	1	1	0	1	1	1	1
Hr.	1	1	0	0	1	0	1	1	1	1

Second, consider the logistic regression model since the response variable y is binary. The “glm” (generalized linear model) function in R is used with the option family = “Binomial” to obtain the following result.

Call: glm (formula = y ~ Hr., family = "binomial")

Coefficients:	Intercept	Hr.
	6.451	2.101

Degrees of Freedom: 19 Total (i.e. Null); 18 Residual
Null Deviance: 24.43
Residual Deviance: 12.89 AIC: 16.89
Call: glm (formula = y ~ Hr., family = "binomial")
Deviance Residuals: Min 1Q Median 3Q Max

1.1156 .0.4480 0.1854 0.5148 2.1674

Coefficients:	Estimate	Std. Errors	z- values	Pr(>|z|)	
(Intercept)	6.4509	3.1777	2.03	0.0424	*
Hr.	2.1012	0.9715	2.163	0.0305	*

Significant. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘’ 1
(Dispersion parameter for binomial family taken to be 1)
Null deviance: 24.435 on 19 degrees of freedom
Residual deviance: 12.891 on 18 degrees of freedom
AIC: 16.891
Number of Fisher Scoring iterations: 6

Again, the following table shows the observed and predicted values (after rounding to the nearest integer) of y

	1	2	3	4	5	6	7	8	9	10
y	1	1	0	0	1	1	1	0	1	0
$\hat{y}$	1	1	0	0	0	1	1	0	1	0

	11	12	13	14	15	16	17	18	19	20
y	1	1	0	1	1	0	1	1	1	1
$\hat{y}$	1	1	0	0	1	0	1	1	1	1

Notice that simple linear and logistic regression produce identical predictions and have an error of 10 % (2 predictions are wrong in a sample of size 20) in predicting the value of y.

Now, let us deviate from using a mathematical function for prediction and adopt a different approach. This approach splits the given data according to the predictor variable to minimize the impurity of the resulting subsets regarding the response variable. If the impurity of a subset is reduced to zero, the prediction is without any uncertainty and is free from error. Let us try partitioning the given data in different ways to identify the best partition for predicting the value of the response variable with a minimum classification error.

We wish to classify the output (y) according to the number of study hours (Hr.). We know that, for a categorical response variable with binary output, we use the "Gini index" (or Gini impurity measure) formula to assess the classification accuracy of the given dataset.

The Gini index is given by 1- Σ (pi) ^{2,} where pi is the probability of an observation being classified as belonging to the distinct class i.

Let us check all the possible binary classifications and then think about finding the best classification.

Notice that the variable Hr. in the given example is an ordinal variable, even though it is a discrete (and hence categorical) variable. Due to its ordinal nature, partitioning of the given data according to the variable Hr. It will be restricted by the condition that the entire range of its values should be split at a single point so that one subset is to the left while the other subset is to the right of the splitting point.

Consider all the possible ways of partitioning the given data according to the variable Hr. Taking values 2, 3, 4, 5, and 6 and then calculating the Gini index for each partition to identify the best partitioning, we find the following four different ways of partitioning the given dataset according to the variable Hr.

1. 2 and (3, 4, 5, 6)
2. (2, 3) and (4, 5, 6)
3. (2, 3, 4) and (5, 6)
4. (2, 3, 4, 5) and 6

Table of Hr. and y values

Hr.↓ \ y→	0	1
2	2	1
3	4	1
4	0	4
5	0	6
6	0	2

	Possible partitions according to the values of Hr.				
Hr.	2	3	4	5	6
	2	3	4	5	6
	2	3	4	5	6
	2	3	4	5	6

The partition sets and values of the response variable in each partition set.

Hr.↓ \ y→	**0**	**1**
2	2	1
3,4,5,6	4	13
Hr.↓ \ y→	**0**	**1**
2,3	6	2
4,5,6	0	12

Hr.↓ \ y→	**0**	**1**
2,3,4	6	6
5,6	0	8
Hr.↓ \ y→	**0**	**1**
2,3,4,5	6	12
6	0	2

Let us measure the impurity of the response variable in each of the partitioned sets.

For the first possible partition

When Hr. = 2, we have observations (2, 0), (2, 0), and (2, 1)

So, from the above table, we find that, out of the 3 cases where Hr. = 2, the proportion of the cases where y = 0 is 2/3 and the proportion of the cases where y = 1 is 1/3

That is, $p_1 = 2/3 \quad p_2 = 1/3$

Gini index $= 1 - (2/3)^2 - (1/3)^2 = \underline{0.4444}$

When Hr. = 3, 4, 5, or 6 (that is, the last four rows of the above table), we have observations (3,0), (3,0), (3,0), (3,0), (3,1), (4,1), (4,1), (4,1), (4,1), (5,1), (5,1), (5,1), (5,1), (5,1), (5,1), (6,1), and (6,1). Out of the total of 17 observations, the proportion of cases where y = 0 is 4/17, and the proportion of cases where y = 1 is 13/17

That is, p_1= 4/17 p_2 = 13/17

Gini index = $1 - (4/17)^2 - (13/17)^2 = \underline{0.359862}$

For the second possible partition

When Hr. = 2 or 3, we have observations (2, 0), (2,0), (2,1), (3,0), (3,0), (3,0), (3,0) and (3,1). There is a total of 8 observations, and the proportion of cases where y = 0 is 6/8, and the proportion of cases where y = 1 is 2/8

That is, p_1= 6/8 p_2 = 2/8

Gini index = $1 - (6/8)^2 - (2/8)^2 = \underline{0.3750}$

When Hr. = 4, 5 or 6, we have observations (4,1), (4,1), (4,1), (4,1), (5,1), (5,1), (5,1), (5,1), (5,1), (5,1), (6,1) and (6,1). There is a total of 12 observations, and the proportion of cases where y = 0 is 0/12, and the proportion of cases where y = 1 is 12/12

That is, p_1=0/12 p_2 = 12/12

Gini index = $1 - (0)^2 - (1)^2 = \underline{0}$

For the third possible partition

When Hr. = 2, 3 or 4, we have observations (2,0), (2,0), (2,1), (3,0), (3,0), (3,0), (3,0), (3,1), (4,1), (4,1), (4,1), (4,1). There are a total of 12 observations; the proportion of cases where y = 0 is 6/12, and the proportion of cases where y = 1 is 6/12

That is, p_1= 1/2 p_2 = 1/2

Gini index = $1 - (6/12)^2 - (6/12)^2 = \underline{0.5}$

When Hr. = 5 or 6, we have observations (5,1), (5,1), (5,1), (5,1), (5,1), (5,1), (6,1), (6,1). There are a total of 8 observations, and the proportion of cases where y = 0 is 0/8, and the proportion of cases where y = 1 is 8/8

That is, p_1= 0 p_2 = 1

Gini index = $1 - (0)^2 - (1)^2 = \underline{0}$

For the fourth possible partition

When Hr. = 2, 3, 4, 5, we have observations (2,0), (2,0), (2,1), (3,0), (3,0), (3,0), (3,0), (3,1), (4,1), (4,1), (4,1), (4,1), (5,1), (5,1), (5,1), (5,1), (5,1) and (5,1). There is a total of 18 observations; the proportion of cases where y = 0 is 6/18, and the proportion of cases where y = 1 is 12/18

That is, p_1= 6/18 p_2 = 12/18

Gini index = $1 - (6/18)^2 - (12/18)^2 = \underline{0.4444}$

When Hr. = 6, we have observations (6, 1), (6, 1). There are a total of 8 observations; the proportion of cases where y = 0 is 0/2, and the proportion of cases where y=1 is 2/2

That is, p_1= 0 p_2 = 1

Gini index = $1 - (0)^2 - (1)^2 = \underline{0}$

The above calculations conclude that the second partition is the best as the Gini index is minimum for partition 2 (zero Gini index represents a pure node).

Remark 1. In the above example, we have only one predictor variable taking five discrete values to be partitioned into two subgroups, and therefore we have four possible partitions. With a predictor taking six values, we shall have five possible partitions. Thus, in general, for a single predictor variable taking n1 discrete values, we shall have (n_1-1) possible partitions.

With two predictor variables taking n_1 and n_2 values, respectively, we shall have (n_1-1)·(n_2-1) possible partitions. Suppose there are k predictor variables, taking n_1, n_2, …, and nk discrete values, respectively. We shall then have (n_1-1)·(n_2-1)···(nk-1) as the number of possible partitions that must be examined for determining the best partition for predicting the value of the response variable. Thus, the number of possible partitions increases geometrically as the number of predictor variables increases.

Remark 2. Suppose the predictor variable is nominal (categorical but not ordinal) in an example like the above. Further, suppose that the variable has five possible values. The number of possible partitions (of the dataset) induced by this variable will be given by $^5C_1 + {}^5C_2 + {}^5C_3 = 30$ ($= 2^5 - 2$). As the number of possible values of the predictor variable increases, the number of possible partitions will increase. Thus, the problem will

increasingly become more complex with increasing possible values of predictor variable.

Let us consider the prediction problem where the predictor variable is a real-valued continuous variable with k distinct values in the observed data. The number of distinct ways to partition data according to this particular predictor variable is k-1, the same as the number of gaps between observed successive values of the predictor variable when arranged according to their order of magnitude, ascending or descending.

1.2 TYPES OF THE REGRESSION MODEL

The statistics literature contains many different types of regression models. We briefly review some of these models to explain how they can be shown to be exceptional cases of the proposed generalization.

1. Linear Regression. The simplest regression model is the simple linear regression. It is interesting to note the origin of simple linear regression to understand why the normal (Gaussian) distribution is so important in the linear regression model. Suppose two random variables X and Y follow the bivariate normal distribution with means μ_x and μ_y, standard deviations σ_x and σ_y, and correlation coefficient ρ (rho). Then, the marginal distribution of X is normal with mean μ_x and standard deviation σ_x. The conditional distribution of Y given X = x is normal with mean $\mu_y - \rho * \sigma_y / \sigma_x (x - \mu_x)$ and variance $\sigma_y^2 (1 - \rho^2)$. This shows that the conditional expected value of Y given X = x is a linear function of x. Since the regression of Y on X is defined as the conditional expectation of Y given X, linear regression is linear when X and Y follow a bivariate normal distribution.

Further, the residual is usually distributed with zero mean and variance that does not depend on X or Y. The multiple linear regression model is similarly the natural choice when the response variable Y and the P predictor variables $X_1, X_2, \ldots, X_P$ jointly follow the multivariate normal distribution. Derivation of the regression is straightforward if we notice that the conditional distribution of Y given $X_1 = x_1, X_2 = x_2, \ldots, X_P = x_P$ is a linear function of $x_1, x_2, \ldots, x_P$.

2. Polynomial Regression. The scatterplot of X and Y sometimes indicates that the relationship between X and Y may not be linear. It is common in such cases to try the polynomial regression. The polynomial regression model involves higher-order powers of the predictor variable X and is in the following form. $Y = \beta_0 + \beta_1 X + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k + \varepsilon$. This is the formula of the k-th degree polynomial regression. What is interesting to note

is that the polynomial regression is nonlinear in the predictor variable, not in regression coefficients. If we define $X_1 = X$, $X_2 = X^2$, $X_3 = X^3$, … $Xk = X^k$, then the polynomial regression model will be the same as the multiple linear regression model. This is the reason why not much literature on multinomial regression is available. The polynomial regression model sometimes contains terms that involve products of predictor variables or their powers, such as $X_1 \cdot X_2$, $X_2 \cdot X_3$, $X_1^2 \cdot X_2$, and so on. Even then, new predictor variables are introduced for each of these, so the regression model is still a multiple linear regression model. It can be very tempting to fit highly involved polynomial regression models to data because the more complex the polynomial expression, the better the model's fit. What is essential to keep in mind is the fact that introducing too many terms in the polynomial regression model can lead to the problem of Overfitting.

3. Logistic Regression. The response variable in the logistic regression model is binary; it only has two possible values, 0 and 1, often described as failure and success, respectively. The linear regression model is inappropriate because a straight line is unbounded, while the binary response variable is bounded. The problem of predicting the value of the response variable is changed to the problem of predicting the probability p of success, that is, the probability of the response variable taking the value of 1. The probability p is bounded between 0 and 1 and still cannot be predicted with the 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. Log (p / (1-p)) covers the entire natural 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 predictor variables. So, the logistic regression model is given by $\log (p/(1-p)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_P + \varepsilon$. What is important to note in the case of logistic regression is that the effect of changes in any predictor variables does not affect the original response variable linearly. As a result, the interpretation of the regression coefficients in the case of logistic regression is not straightforward. The effects of predictor variables are multiplicative and not additive, as in the case of linear regression.

Moreover, the logistic regression model violates the assumption of homoscedasticity. Further, since the response variable is binary, it does not follow the normal distribution. Even the residuals do not follow the normal distribution for the same reason.

4. Quantile regression. The quantile regression model differs from the linear regression model in the following sense. The linear regression model

attempts to find the conditional expectation of the response variable corresponding to the given values of the predictor variable(s). In contrast, the quantile regression model attempts to find the specified quantile of the conditional distribution of the response variable corresponding to the given values of the predictor variable(s). In other words, the quantile regression model extends from the concept of a quantile to the concept of a conditional quantile. If Y is a random variable with the distribution function F(y) = P (Y < y), then the q^{th} quantile (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, …, y n, it is known that the sample median (Md) minimizes the total absolute deviation of sample values, namely Ę Æ"ˬ Á" $\sum_{i=0}^{n}$ | yi − Md |. It should then be evident that the quantile regression model cannot use the squared error loss function because the quantiles do not have the property of the sample mean, and the squared deviations are minimum when taken from the mean. Instead, the objective function that is minimized in quantile regression is given by $\min_z$ in R {Sum w_q(y i – z), where w_q(x) = x (q – I (x < 0))" and I () is the indicator function. The linear regression model uses the property of the sample mean to minimize the total of squared deviations from sample observations and applies it to obtain the conditional mean as the optimal solution to the problem of minimizing the 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 w_q() 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 the distribution of the response variable is skewed. Quantile regression is also robust to outliers. It is important to note that the regression coefficients in quantile regression differ significantly from those in linear regression. If this does not happen, then quantile regression is not justified.

5. Ridge Regression. We shall quickly understand the concept of regularization in regression before defining ridge regression. Regularization is a way to handle the problem of Overfitting, where the model fits well with the training data but could perform better on test data. Regularization introduces a penalty function in the objective function and controls the complexity of the model through the penalty function. Regularization is proper when the number of predictor variables is too large compared to the sample size and when there is multicollinearity among the predictor variables. The two most common methods of regularization use the L1 norm and L2 norm as the penalty function. The L1 norm imposes a penalty to the

sum of absolute values of the regression coefficients. The L2 norm imposes a penalty to the sum of squares of the regression coefficients. Ridge regression uses the L2 norm as the regularization method. The objective function for the ridge regression is to minimize $\sum(y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - \ldots - \beta_p x_{pi})^2 + \lambda \sum \beta_i^2$. The standard equations for ridge regression have the solution given by $\hat{\beta} = (X' X + \lambda I)^{-1} X' y$. Ridge regression was initially proposed to resolve the problem of multicollinearity. A consequence of regularization is that it is no longer meaningful to assume the error terms to be normally distributed.

6. LASSO regression. Lasso regression was proposed as an alternative to ridge regression by using the L1 norm instead of the L2 norm that ridge regression uses. LASSO is the acronym for Least Absolute Shrinkage and Selection Operator. It minimizes the objective function $\sum(y_i - \beta_1 x_{1i} - \beta_2 x_{2i} - \ldots - \beta_p x_{pi})^2 + \lambda \sum |\beta_i|$. LASSO does not regularize the intercept because all variables are standardized before fitting the model.

Consequently, 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 the ith lp of statistical software. As the name suggests, LASSO automatically selects variables for use in the model and resolves the multicollinearity problem. In this sense, LASSO is better than ridge regression. However, the 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 based on 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 for use in multicollinearity, but it is still being determined whether ridge regression is better or lasso regression is better. Elastic net regression combines ridge and lasso regression using both the L1 and L2 norms. All the variables are standardized before fitting the regression model; therefore, the model has no intercept term. The objective function for elastic net regression is $\sum(y_i - \beta_0 - \beta_1 X_1 - \beta_2 X_2 - \ldots - \beta_p X_p)^2 + \lambda_1 \sum \beta_i^2 + \lambda_2 \sum |\beta_i|$. Elastic net regression does not assume errors to be normally distributed.

8. Principal Components Regression (PCR). Principal Components Regression (PCR) is used when there are too many predictor variables, and it is required to select the most appropriate of them or when there is multicollinearity among the predictor variables. PCR is implemented by

finding the principal components, selecting the first k principal components that explain the specified percentage of the total variation, and then fitting multiple linear regression of the response variable on scores of the selected principal components. Principal components regression is used when either there are too many predictor variables, or the predictor variables exhibit multicollinearity. Since principal components are orthogonal, they are stochastically independent of each other and hence cannot have multicollinearity. Even though PCR achieves a reduction in the dimensionality of data, it is not a genuine feature selection method. Instead of selecting variables that should be included in the model, PCR constructs fewer variables than the original ones. However, it is essential to note that none of the original variables can be ignored or removed from the analysis because all variables must compute principal components before selecting some. If the original variables are normally distributed, then the principal components are also normally distributed, being linear functions of the original variables. The normality and homoscedasticity of residuals are not affected by PCR.

9. Partial Least Squares (PLS) Regression. Partial Least Squares (PLS) regression is an alternative to PCR when predictor variables are highly correlated. It can also be used when the number of predictor variables is enormous. PLS and PCR construct new variables from the original predictor variables using orthogonal (linear) transformations, but the two methods have a significant difference. PCR constructs principal components to capture the maximum variability among predictor variables without considering the response variable.

Conversely, PLS sequentially finds linear combinations of predictor variables that have maximized the correlation coefficient with the response variable, subject to the condition of independence with earlier linear combinations. Due to the condition of independence with earlier linear combinations, successive correlations are partial correlations conditional on earlier linear combinations. This is the reason why the name of this method is partial least squares. Compared to PCR, PLS has been found to work with fewer predictor dimensions (linear combinations of predictor variables) than the number of original predictor variables.

10. Support Vector Regression (SVR). Support vector regression is based on the support vector machines algorithm and, as such, can handle cases of linear and nonlinear regression models. Support vector regression uses kernel functions to identify the optimal regression model without any consideration for the model's functional form in the sense of linearity or

non-linearity of the function. Support vector regression minimizes the sum of squares of the coefficients (that is, L2-norm) rather than minimizing the sum of squared errors, as in the case of linear regression. The error term involves only constraints but not the objective function. The constraint on the error term is that the absolute error is less than a specified value epsilon, known as the maximum error. It is then easy to note that the support vector regression will be within epsilon distance from observations in the given data set. This is made possible by relaxing the requirement of specifying the algebraic form of the regression model that fits best with the given data.

Consequently, support vector regression is not expressed in an algebraic form. Instead, the model is stored in the computer and is used to predict values of the response or dependent variable corresponding to values of predictor or independent variables in test data. Support vector regression is a non-parametric method and, therefore, does not depend on the assumption of normality of errors that the Gauss-Markov setup requires. It is also interesting to note that support vector regression focuses not on predicting the response but on keeping the error term within the specified limit.

11. Ordinal Regression. Linear regression and most of its variants are applicable when the response or dependent variable is numerical (interval scale) and continuous. Logistic regression is appropriate when the response variable is nominal. In the simple form, logistic regression assumes that the response variable is binary; that is, it takes one of two possible values, often labelled as success and failure, and coded as 1 and 0. The logistic regression model converts this binary (nominal) variable into a continuous variable by treating log (p/ (1-p)) as the response variable, where p is the probability of success in a single trial. The new response variable log (p/ (1-p)), a log of the odds ratio, is called the logit function. There are situations where the response variable (Y) is ordinal, and the interest is in predicting the category of the response variable. Since Y is ordinal, its values (or categories) can be ordered or ranked, but the differences between successive categories may not be comparable. Ordinal regression is employed in such cases. Suppose the response variable Y has k different possible values, such as y_1, y_2, y_3, …, y_k, so $y_1 < y_2 < y_3 < \ldots < y_k$. Without loss of generality, these k distinct values can be coded as 1, 2, 3, …, k respectively and called scores. Let p_1, p_2, p_3, …, p_k be probabilities of the k possible values of Y. Corresponding sample proportions can estimate p_1, p_2, p_3, …, p_k, but the interest is in prediction, not estimation. The prediction problem is addressed by converting probabilities to cumulative logits (log-odds) of the k possible values in an ascending order. In other words, the response variable is transformed to take (k-1) distinct values given by z_j = (prob. of score < j) /

(prob. of score > j) for j = 1, 2, 3, …, (k-1). The last (highest) score does not have an odds ratio because the probability of a score being less than or equal to it is 1. This can equivalently be written as

zj = (prob. of score < j) / (1 – prob. of score < j), j=1, 2, …, k-1.

The ordinal regression model is then given by log (z_j) = α_j – β * X, where X is the predictor variable. This model can be modified to accommodate two or more predictor variables, such as simple linear regression, which is extended to multiple linear regression to accommodate two or more predictor variables. An obvious question is why linear regression is not used when the response variable is ordinal. The answer lies in the fact that differences between successive levels of an ordinal variable are not equal. Ordinal regression models the changes in log odds or logits as proportional to changes in predictor variables. The ordinal regression model is then stated as follows.

$$Z j = \alpha_j – \beta * X + \varepsilon,$$

Where epsilon (ε) is the error term, ordinal regression is similar to logistic regression. However, it is more general in the sense that it can handle situations where the response variable is categorical but not nominal (including binary). The response variable is ordinal.

12. Poisson regression. Poisson regression is used when the response variable represents count data that the Poisson distribution can model. An example of count data is the number of occurrences of a rare event during a specified period.

It is important to note that Poisson regression is not suitable when the response variable is discrete but takes non-integer values. The Poisson regression model expresses the log of the expected value of the response variable Y as a linear function of the predictor variables. That is, the Poisson regression model is defined as follows.

$$\log [E (Y \mid X = x)] = \alpha + \beta * x + \varepsilon,$$

Where epsilon is the error term, this model is also known as the log-linear model due to the log term on the left-hand side of the equation. The model is alternatively written as $E (Y \mid X = x) = e^{(\alpha + \beta * x + \varepsilon)}$.

The mean of the Poisson distribution is equal to the variance. The situation where the sample variance is substantially more significant than the sample

mean cannot be represented by the Poisson distribution. This situation is called a case of over-dispersion, and it is recommended that the distribution of the response variable Y be assumed to be negative binomial. Another problem that may arise in Poisson regression is that there are too many occurrences of zero counts in data. This situation can be explained by the existence of two processes: one determining whether there is or is not any event (accounting for excess occurrences of zeros) and then the Poisson distribution of the number of occurrences of the event conditional on occurrence.

13. Negative Binomial Regression. The Poisson distribution is often generalized to the negative binomial distribution as a Poisson-gamma mixture by including the noise variable that follows the gamma distribution with unit mean and scale parameter v. The probability mass function (p.m.f) of the Poisson-gamma mixture distribution is given by

$f(y_i \mid \mu_i, \alpha) = \gamma(y_i + 1/\alpha) / [\gamma(y_i + 1)\,\gamma(1/\alpha)]\ \{(1/\alpha) / (1/\alpha + \mu_i)\}\ ^\wedge\ (1/\alpha)\ (\mu_i / (1/\alpha + \mu_i)^\wedge y_i,$

Where $\mu_i = t_i\,\mu$ and $\alpha = 1/v$.

The parameter mu is the mean exposure rate per unit of exposure (t), and heit$\mu_i = \mu * t_i$ is the mean exposure corresponding to exposure time ti. The mean of the negative binomial regression model is modelled as μ_i= exp $\{\ln(ti) + \beta_1\ x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k\ x_{ki}\}$. The negative binomial regression model for the observation yi is then written as

$P\ [\ Y = y_i \mid \mu_i, \alpha\ ] = \gamma\ (y_i + 1/\alpha) / [\ \gamma\ (1/\alpha)\ \gamma\ (y_i + 1)]\ \{1/\ (1 + \alpha\ \mu_i\ )\}^\wedge(1/\alpha)\ [\ \alpha\ \mu_i / (1 + \alpha\ \mu_i)]\ ^\wedge y_i.$

14. Quasi Poisson Regression. The variance of the Poisson distribution is equal to its mean. The variance of the negative binomial distribution is more significant than its mean. It is, therefore, natural to fit a negative binomial regression when there is over-dispersion in the response variable. However, it should be noted that the variance of the negative binomial distribution is a quadratic function of its mean. If the variance of the response variable is greater than the mean, but the variance is a linear function of the mean, then quasi-Poisson regression is used. Quasi-Poisson regression can also handle under-dispersion cases as effectively as over-dispersion cases.

15. Cox Regression (or Cox Proportional Hazards Model). Cox regression is used when data on an event's occurrences and time to event is available. Logistic regression analyses only occurrence data, ignoring the

time of occurrence, whereas Cox regression analyses data on occurrences of an event, including time to occurrence. Therefore, Cox regression is instrumental in survival analysis because the interest in survival analysis is not restricted to the number of survivals but also to time to failure for failed items. Cox regression thus sets the dual target: a continuous variable representing the time to the event and a discrete variable representing the event's status (i.e., whether or not the event occurs). The hazard rate is defined as $\lambda (t \mid Z) = \lambda_0 (t) * \exp (\beta'*Z)$, where Z is the vector of covariates (i.e., predictor or independent variables). The hazard ratio is then defined as $\lambda (t \mid Z) / \lambda_0(t) = \exp (\beta' * Z)$ or, equivalently, $\log [\lambda (t \mid Z) / \lambda_0 (t)] = \beta' * Z$. This shows that in Cox regression models the hazard ratio as a linear function of the covariates. Note that the hazard ratio is a function of the time t. When the hazard ratio is assumed to be independent of time t, it is a constant. This is why the model is called the proportional hazard model, as the hazard ratio maintains the same proportion over variation in the time t.

16. Tobit Regression. This regression model is used when the response variable is subjected to censoring. As a result of censoring, actual values of all the predictor variables are available, but only partially accurate values of the response variable are available (in uncensored cases). Values of the response variable in a particular range are all equal to a single value. Censoring can be one of the two types: right censoring and left censoring. Right censoring is when the study (or observation) period ends before the event of interest occurs in all the cases under study. For example, in a study regarding the attrition of customers of a company, the study period may be fixed to two years. In this case, the time of attrition is known for every customer that has attrition before the end of two years. All customers who have not attrition within the study period of two years are right-censored. Left censoring is the case when some events occur before the beginning of the study. Left censoring is rare and usually not considered in a general description of Tobit regression. Tobit regression makes the same assumptions as linear regression but is more sensitive to violations of assumptions due to its restrictive nature.

1.3 THE BIG DATA REVOLUTION AND ITS IMPACT ON PREDICTIVE MODELING

The traditional methods cannot handle such extensive data that involves complex calculations. For this purpose, many software types have been developed and are available for use. Nowadays, data is generated in large volumes and with a high velocity and needs to be processed efficiently in

the shortest possible time. Such datasets are called 'Big data'. Big data are massive (volume) and are still growing exponentially with time.

Some Examples of Big data are as follows.

1. Records of employees in big companies.
2. Data and records of transactions of account holders in banks.
3. Data collected on sales in big marts.
4. Data collected from social media sites.
5. Data related to healthcare
6. Production Data of Manufacturing Industries
7. Students' data and teachers' data in Academics
8. Data related to transportation systems like railways, airlines, road transport, and goods transportation.

The list is never-ending.

Characteristics of big data (The 4 V 's of big data)

1. Volume: The enormous volume (size) of big data makes it impossible for traditional data analysis methods to handle big data. This has created the need to develop methods to cope with the massive volume of big data. The volume of big data prompted IBM to develop significant data architecture, such as Hadoop, to handle big data. Due to its size, big data cannot be processed on any single processor. The advent of big data has provided an excellent motivation for parallel or concurrent computing, rather than the traditional sequential computing on a single processor. Prediction models for big data will also be developed using a multiprocessor system. As the first step in preparing data for such an analysis, data must be partitioned most suitably to derive the best possible results. This point will be relevant when we apply regression trees, where data gets partitioned to satisfy some optimality criterion.

2. Velocity: The velocity of significant data results from speedy data collection processes. In addition, the fact that several data collection mechanisms work simultaneously to collect data results in a high velocity of incoming or recorded data. The high velocity of data collection requires high speed in processing the data. The current situation with big data is such that a lot of big data has been generated and stored but has not been analysed till now. The unequal velocity of data generation from different sources has resulted in the development of distributed computing, where data generated from different sources is distributed over a network. Distributed computing

requires that the data be divided so that different parts of data are processed on different processors. Again, this point will be relevant when we discuss the partitioning of data as part of developing a regression tree.

3. Variety: Big data is generated by a large variety of sources, and hence, the resulting data has a lot of variation, not only in the content but also in the form. Big data is no longer restricted to numerical or text data. Big data also includes image data obtained from surveillance cameras, MRI and other medical imaging devices, remote sensing data acquired from satellites, audio and video data collected from auditions and competitions, etc. The variety of big data has resulted in three types of data: structured, unstructured, and semi-structured. Specific rules and formats regulate structured data and can be processed relatively quickly. Semi-structured data is partly structured and requires some pre-processing before subjecting it to any analysis. Unstructured data is still a challenge to the big data community due to the lack of any regulatory authority that can formalize the process of collecting, storing, accessing, and processing such data.

4. Veracity: The enormous volume and high speed of data generation in significant data results from multiple data sources. Such data cannot be regulated or controlled and cannot be expected to be homogeneous or homoscedastic. Consequently, traditional statistical data analysis methods heavily depend on the assumption that observations are independent and identically distributed (I.I. d.) are no longer appropriate for big data. The significant consequence of the first three V 's of big data results in an increased level of uncertainty. This is why veracity is relevant in the context of big data. As will be explained later in the context of the proposed generalization of regression models, the veracity of big data can be handled effectively by dividing data into subsets so that these subsets of data are more homogeneous and homoscedastic than the complete (original) data.

Big data are analysed using predictive techniques with the help of statistical software. Predictive modelling techniques are used to analyse the data and determine future response variable(s) values. Predictive modelling is used as a decision-making tool. Predictive modelling includes regression models (linear and logistic), decision trees, and neural networks. This technique uses statistical algorithms and machine learning techniques for analysis. Predictive analysis finds data trends and uses them to predict future values. Every predictive model must satisfy the specific needs of the researcher.

Multiple models can provide a good solution for any dataset; choosing the suitable model is the real challenge. For this purpose, it is necessary to

evaluate the performance of all models. The best model to achieve the goal is the best model with the best performance among the available alternatives. This explains why evaluating the model's performance fitted to the data is necessary. If the entire data set is used to build the model, then no data is left to test the model before it is used for future prediction.

For this reason, the original data set is divided into two parts. One part is used to fit the model and is called the 'Training' data set. The second part of the data is used to evaluate the model's performance developed on the training data, called the 'test' or 'validation' data set. The observations in the data are assigned to these two sets randomly. The observations in the data set are usually assigned to these two sets in the ratio of 80:20 or 70:30. Cross-validation can also be used for testing.

All the analyses in this study were performed using packages and R software facilities. R is a free and open-source software environment for statistical data analysis and data visualization. It is available for UNIX, Windows, and MacOS platforms. It provides many graphical and statistical techniques, such as regression modelling, probability distributions, statistical tests, clustering, classification, and time series analysis.

Performance measures in the classification problem

- Confusion matrix. It measures classification accuracy by counting false negatives, false positives, and correct predictions.
- Area under the curve (AUC). The operating characteristic (O.C.) function graph is drawn from the protocol of the particular model. The area under this curve is used as a measure. The higher the AUC, the better the performance of the model.

Components of the output of linear regression analysis

- Residuals. The output consists of the minimum, maximum, and three quartiles (including the median) for residuals. Note that the median value is always near zero.
- Coefficients. The output consists of estimated values of regression coefficients, including the intercept.
- Summary. The output consists of estimated values of the intercept and the regression coefficients, along with the following information for each.
  - **Std. Error**. Standard error of estimates of the regression coefficients, including the intercept.