Regression Analysis

By Akshita and Komal Yadav

April 2022

1 ABSTRACT

Regression is a statistical method used in finance, investing, and other disciplines that attempts to determine the strength and character of the relationship between one dependent variable (usually denoted by Y) and a series of other variables (known as independent variables).Regression can be classified into 8 major categories namely,Linear Regression ,Logistic Regression ,Polynomial Regression Support Vector Regression ,Decision Tree Regression ,Random Forest Regression , Ridge Regression , Lasso Regression. In this paper , we have studied research papers published on linear regression, logistic regression , decision tree and random forest . The analysis is made and a summary is done at last.

2 INTRODUCTION

In statistical modeling, regression analysis is a set of statistical processes for estimating the relationships between a dependent variable (often called the 'outcome' or 'response' variable) and one or more independent variables (often called 'predictors', 'covariates', 'explanatory variables' or 'features'). The earliest form of regression was the method of least squares, which was published by Legendre in 1805, and by Gauss in 1809. Legendre and Gauss both applied the method to the problem of determining, from astronomical observations, the orbits of bodies about the Sun (mostly comets, but also later the then newly discovered minor planets). Gauss published a further development of the theory of least squares in 1821, including a version of the Gauss–Markov theorem.The term "regression" was coined by Francis Galton in the 19th century to describe a biological phenomenon. The phenomenon was that the heights of descendants of tall ancestors tend to regress down towards a normal average (a phenomenon also known as regression toward the mean). In the 1950s and 1960s, economists used electro mechanical desk "calculators" to calculate regressions. Before 1970, it sometimes took up to 24 hours to receive the result from one regression. Regression methods continue to be an area of active research. Regression analysis is primarily used for two conceptually distinct purposes.First, regression analysis is widely used for prediction and forecasting, where its use has substantial overlap with the field of machine learning. Second, in some situations regression analysis can be used to infer causal relationships between the independent and dependent variables. Importantly, regressions by themselves only reveal relationships between a dependent variable and a collection of independent variables in a fixed data set. The challenges faced during the analysis of this paper were selecting the right papers from a vast quantity of papers for the purpose and aggregating them into appropriate categories. Regression Analysis involves very lengthy and complicated

procedure of calculations and analysis. Assumption regarding cause and effect relationship in between the variables to be remain unchanged may not always stand true.

2.1 APPLICATIONS

1. Forecasting : The most common use of regression analysis in business is for forecasting future opportunities and threats. Demand analysis, for example, forecasts the amount of things a customer is likely to buy. When it comes to business, though, demand is not the only dependent variable. Regressive analysis can anticipate significantly more than just direct income. For example, we may predict the highest bid for an advertising by forecasting the number of consumers who would pass in front of a specific billboard.
2. Reliable source -Many businesses and their top executives are now adopting regression analysis (and other types of statistical analysis) to make better business decisions and reduce guesswork and gut instinct. Regression enables firms to take a scientific approach to management. Both small and large enterprises are frequently bombarded with an excessive amount of data. Managers may use regression analysis to filter through data and choose the relevant factors to make the best decisions possible.
3. CAPM: The Capital Asset Pricing Model (CAPM), which establishes the link between an asset's projected return and the related market risk premium, relies on the linear regression model. It is also frequently used in financial analysis by financial analysts to anticipate corporate returns and operational performance.The beta coefficient of a stock is calculated using regression analysis. Beta is a measure of return volatility in relation to total market risk. Because it reflects the slope of the CAPM regression, we can rapidly calculate it in Excel using the SLOPE tool.
4. Comparing with competition:It may be used to compare a company's financial performance to that of a certain counterpart. It may also be used to determine the relationship between two firms' stock prices (this can be extended to find correlation between 2 competing companies, 2 companies operating in an unrelated industry etc). It can assist the firm in determining which aspects are influencing their sales in contrast to the comparative firm. These techniques can assist small enterprises in achieving rapid success in a short amount of time.

3 REGRESSION

For regression , we have explored 4 papers which are being described as further.

3.1 Categorical Variables in Regression Analysis: A Comparison of Dummy and Effect Coding by Hussain Alkhurasi (2012) [?]

The paper begins with an overview of the coding methods for categorical variables. Then, a description of the dummy and effect coding along with an example that illustrates their use and interpretation in the Regression analysis is being provided. the paper addressed the use of these coding methods in designs with unequal sample sizes. Coding methods can be defined as the ways in which membership in a group can be represented in a mutually exclusive and exhaustive manner. So, any categorical variable with k categories can be represented by creating (k-1) dummy variables . This process involves assigning one numerical value, which is called a code, to all subjects of a particular group and a different numerical value to

others. Dummy Coding method -This method represents group membership with dummy variables that take on values 0and 1. Thus, membership in a particular group is coded one whereas non-membership in the group is coded zero. In most common applications, one group receives 0s on all dummy variables and functions as the reference group When dummy coding is used in the regression analysis, the overall results indicate whether there is a relationship between the dummy variables and the dependent variables. The values of the intercept and the regression coefficients of the resulted regression model is obtained using least squares estimation . The regression model from the dummy coding can be written as:

$$Y_{ij} = B_0 + \sum_{j=1}^{k-1} B_j D_{ij} + \epsilon_{ij} \tag{1}$$

[**?**]

The dummy coding is the preferred method when to compare several treatment group with a control group. In this case, the control group may serve as the reference group and the regression coefficients would then reflect the treatment-control mean differences . However, this method does not test the differences between specific treatment means as well as the effect of a particular treatment defined as the deviation between the treatment mean and the grand mean .

Effect Coding Method - In this method, the dummy variables take on the values 1, 0, and -1. the coding method used for effect coding is similar to that used for dummy coding except for the way in which the reference group is identified. Using dummy coding, the reference group is coded 0, but in the effect coding it is coded -1 When effect coding is used in the regression analysis, the overall results for the regression model (R^2 and F) are the same as in the dummy coding However, the interpretations of the intercept and the regression coefficients are different. The regression model from the effect coding can be written as follows:

$$Y_{ij} = B_0 + \sum_{j=1}^{k-1} B_j D_{ij} + \epsilon_{ij} \tag{2}$$

[**?**] Effect coding is appropriate when each group is compared with the entire set of groups rather than with a reference group. In other words, effect coding is useful in testing the effect of a treatment defined as the deviation between the treatment mean and the grand mean. However, to determine which means differ significantly from each other, one of the methods for multiple comparisons of means has to be applied (Cohen and Cohen, 1983).

Thus , we can summarize

Points of contrasts	Dummy coding	Effect Coding
Coding system	0 and 1	1,0,-1
Intercept	Mean of group coded all 0s	Grand mean of all groups
Regression Coefficient	$Y_j - Y_0$	$Y_j - Y_{\cdot\cdot}$
uses	Compare several experimental groups with a control group	Test treatment effect
Effect of unequal sample sizes	unaffected by sample sizes	intercept = unweighted average of the group means

3.2 Exploratory regression analysis: A tool for selecting models and determining predictor importance by Michael T. Braun and Frederick L. Oswald (2011) [?]

In this paper we reviewed the methods for establishing predictor importance and provides a program (in Excel) for implementing them.The program investigates all sub models and produces several indices of predictor importance Determining the most important variables in a model which to include in the model, and which of the included variables contribute the most to prediction is critical from both the practical and theoretical perspectives.Practically, it is often essential to select a subset of tests that is both cost- and time-efficient, and that has adequate criterion-related validity.Theoretically , good theories are parsimonious, containing only those constructs essential to understanding behavioral phenomena. Relative importance is defined as the proportionate contribution each predictor in a linear regression model makes to the model R2, considering both its unique contribution and its contribution when combined with other predictor variables.Although there are no unambiguous measures of relative importance when predictor variables are correlated, some approaches are well motivated and have been shown to provide meaningful results. After selecting a set of predictors and conducting linear regression analysis, evaluation is made on the predictors' importance in the regression model by examining the size of the standardized regression weight associated with each variable. Variables with larger weights are more important than those with smaller weights.

To overcome the problem of determining relative importance from regression weights, importance indices are computed instead.The three types of importance indices: incremental R2, general dominance weights, and relative importance weights.

Incremental R2 It reflects the unique criterion variance accounted for by a predictor after all other variance accounted for by the remaining predictors has been partial ed out of the criterion. More specifically, in a hierarchical regression where predictors are inserted into the model in a step-wise fashion, the incremental R2 for a predictor is the increase in R2 when that predictor is entered last, indicating the unique impact of that predictor in the model

General dominance weights Dominance analysis produces general dominance weights that are computed by averaging a given predictor's incremental validity across all possible sub-models that involve that predictor (i.e., given p predictors, there are 2p – 1 possible sub-models).

Relative importance weights Relative importance weights are a third type of importance index, computed by first transforming a set of p predictors into a new set of predictors that are not correlated with one another, yet are correlated as highly as possible to their counterpart.

3.3 How to use linear regression and correlation in quantitative method comparison studies By P. J. Twomey, M. H. Kroll [?]

Linear regression methods attempt to find the best linear relationship between data points, whereas correlation coefficients evaluate the relationship between the two. If the two methods have a relationship, the points are plotted on a scatter diagram, with the concentrations for one method on the x-axis and the matching concentrations from the second method on the y-axis of a two-dimensional figure

A straight line can often be drawn through the middle of the points, indicating a linear

relationship and this line is the data's regression line, and it has the mathematical form y = mx + c, where m is the estimated line's slope, and c is the estimated line's y intercept. The correlation coefficient, r, measures the degree of linear relationship between two variables, x and y, and measures how close the observations are to the regression line. r is a unitless variable that ranges from -1 to +1, with a negative sign indicating that one variable increases as the other decreases and a positive sign indicating that one variable increases as the other increases.

Types of linear regression

Parametric and non-parametric linear regression techniques are the two basic types of linear regression techniques

1. Parametric methods – the distribution of the residual values (and not the data points themselves) in the y axis has a Gaussian distribution.

2. Unweighted parametric approaches - the variability of the population's distribution of y values is the same for all x values.

In OLR, the regression line is calculated by minimising the squared residuals in the y direction, hence the name "least squares" regression (Figure 6). The x variable is assumed to be error-free. Any disparities caused by this effect, on the other hand, diminish as the number of data points increases.

Cook's distance is another concept that is dependent on the data point's residual — the bigger the residual, the larger the Cook's distance, and thus the more probable the data point will have an impact on the linear regression equation. As a result, the greater the Cook distance, the more likely it is that a point is an outlier.

In contrast to OLR, which does so in the y (vertical) axis, Deming's linear regression minimises the distances between data points that are orthogonal (at right angles) to the regression line. When the variances of x and y differ by a fixed ratio, the distance between the data points is minimised at an angle to the regression line, and this is independent of the ratio. Almost all examiners use OLR outside of the clinical laboratory. It has the advantage of knowing and understanding its limitations because it is parametric and derives directly from mathematical principles. When properly obtained data with a significant spread of distribution is used, the results are repeatable and easy to interpret.

Types of correlations

Parametric and non-parametric correlation techniques are the two basic types of correlation approaches. The parametric method is the Pearson r product moment correlation coefficient, while the nonparametric method is Spearman's rho (q) rank.

Visual data analysis

The first stage, when examining a patient, is a visual inspection. This helps you to visually evaluate if there is a correlation between the data. The next step is to examine the relationship between the data. Clearly, a linear regression model is only valid if the data has a linear relationship. The next step is visual examination of the scatter and residual plots for outliers and influential points. Finally, the residual plot can be used to determine whether or not the SD is constant.

3.4 REGRESSION ANALYSIS FOR CORRELATED DATA by Kung-Yee Liang and Scott L. Zeger(1986) [?]

This paper has reviewed approaches to regression analysis of correlated data organized in clusters.

For continuous data, the most commonly used measure of dependence between two responses, YI and Y2, is the correlation coefficient

$$p = Cov(Y_1, Y_2)/[Var(Y_1)Var(Y_2)]^{1/2} \tag{3}$$

[?]

The correlation coefficient is dimensionless, taking values in the range [- 1, 1]. The correlation p is close to 0 when there is little dependence. Strong dependence is indicated when p approaches either 1 or - 1 . A positive correlation indicates that Y I tends to be larger than expected if Y2 is and vice versa. For discrete, in particular dichotomous data, the correlation coefficient is a poor measure of association because it is constrained by the mean parameters ILJ, and IL2. Specifically, for dichotomous variables, YI and Y2, the correlation is given by

$$P = Pr(Y_1 = Y_1 = 1) - 2m_1m_2/[m_1(1-m_1)m_2(1-m_2)]^{1/2} \tag{4}$$

[?]

To analyze clustered data, we must model both the regression of Y on x and the within-cluster dependence. If the responses are independent of each other, GLMs, can be used for diverse types of responses. For correlated data, GLMs are not sufficient, and other approaches that address the dependence are needed. We reviewed three different modeling approaches: marginal, random effects, and observation driven. In marginal models, the regression of Y on x and the within-cluster dependence are modeled separately. whereas the other approaches attempt to address both issues simultaneously through a single model. In the random effects model its is assumed that the parameters vary from cluster to cluster, thus reflecting natural heterogeneity caused byvarious unmeasured factors. The random effects model is especially useful when the objective is to make inference about individuals. In observation-driven models, the regression inferences is depending on the choice of the within-cluster dependence model.Better methods for choosing the best model from a set of candidates are needed. Random effects models are very difficult to estimate except in the linear and log-linear case but are still attractive to use.

4 LINEAR REGRESSION

4.1 Modified One-Parameter Liu Estimator for the Linear Regression Model by Adewale F. Lukman , B. M. Golam Kibria , Kayode Ayinde,4 and Segun L. Jegede (2020) [?]

In this paper , a modified Liu estimator is proposed to solve the multicollinearity problem for the linear regression model. In general , The linear regression model (LRM) is

$$y = x\theta + \epsilon$$

[?]

The parameters in above equation of linear regression model are mostly estimated by the ordinary least square (OLS) estimator defined as below:

$$\theta = (x'x)^{-1}x'y$$

[?]

The performance of the estimator is conditional on the non-violation of the assumption of the LRM model that the predictor variables are independent. In reality applications, we observed that the predictor variables grow together, which results in the problem termed multicollinearity. The consequence of this on the OLS estimator is that it reduces its efficiency and it becomes unstable.And thus in this paper , a new one-parameter Liu-type estimator for the regression parameter is proposed when the predictor variables of the model are linearly related. Ridge Regression Estimator (RRE). is defines as

$$\theta_k = (x'x + kI_p)^{-1}x'y$$

[?] Liu Estimator. Is obtained by modifying the LRM

$$\theta_d = (x'x + I_p)^{-1}(x'x + dI_p)\theta$$

[?]

On simulation of the given method, on various samples and under various conditions the results obtained are as follows:- Both ridge regression and Liu estimators are widely accepted in the linear regression model as an alternative to the OLS estimator to circumvent the problem of multicollinearity. In this study, we saw a modified Liu estimator, which possesses a single parameter which places it in the class of the ridge and Liu estimators. Theoretical comparisons, simulation study, and real-life applications evidently show that the proposed estimator consistently dominates the existing Liu estimator and ridge regression estimator under some conditions. We recommend the use of this estimator for the linear regression model with multicollinearity problem. We noted that the proposed estimator can be extended to other regression models, for example, logistic regression, Poisson, ZIP,gamma, and related models, and these possibilities are under current investigation.

4.2 Interpreting Multiple Linear Regression By Laura L. Nathans , Frederick L. Oswald, Kim Nimon [?]

This paper presents a guidebook of variable importance measures that inform MR results.The focus is on two general families of techniques. One family provides different methods of rank ordering individual predictors‘ contributions to an overall regression effect or R2 and the other family involves partitioning R2 into the unique and shared variance contributions of the independent variables.

These two families are aligned with the framework of LeBreton, Ployhart, and Ladd (2004), who categorized methods of variable importance into those that assess (a) direct effects, which quantify the contribution of each independent variable to the regression equation when measured in isolation from other independent variables; (b) total effects, which quantify the each independent variable‘s contribution to the regression equation when the variance contributions of all other predictors in the regression model have been accounted for; or (c) partial effects, which quantify the each independent variable‘s contribution to the regression equation.

Beta Weight The regression weight for each given independent variable is interpreted as the expected difference in the dependent variable score between people who differ by one unit on that independent variable, with all other independent variable scores held constant (Hoyt, Leierer, and Millington, 2006; Johnson, 2004). When variables are not standardized (i.e., scaled in their original metric), regression weights are called B weights. The focus in this paper is on beta weights rather than B weights, because beta weights are more comparable

across independent variables due to being scaled in the same standardized metric. According to Pedhazur (1997), beta weights are computed to weight the independent variables so that when the weights are multiplied by variable scores, their sum is maximally correlated with the dependent variable. And the sole reliance on using beta weights to interpret MR is only justified in the case where predictors are perfectly uncorrelated. In the absence of shared variances between independent variables, each standardized beta weight is equal to the zero-order correlation between the independent and dependent variable. The major advantage of beta weights is that they provide a measure of variable importance that is easily computed and provides an initial rank ordering of independent variables‘ contributions to a MR equation that accounts for contributions of other independent variables.

Zero-Order Correlation Zero-order correlations reflect the bivariate relationships between independent and dependent variables. According to Hinkle, Wiersma, and Jurs (2003), the correlation coefficient is, an index that describes the extent to which two variables are related. The correlation coefficient reflects both the magnitude and direction of the relationship between two independent variables. If a correlation coefficient is positive, an increase (or decrease) in one variable is related to an increase (or decrease) in the other variable in the coefficient.

Product Measure Pratt (1987) proposed the product measure, which is calculated by multiplying the variable‘s zero order correlation (its relationship to the dependent variable in isolation from of other independent variables) by its beta weight (which accounts for contributions of all other predictors to the regression equation).

Structure Coefficients A structure coefficient is the bivariate correlation between an independent variable variable and the predicted value resulting from the MR model, where represents the predicted dependent variable scores. The major difference between a zero-order correlation and a structure coefficient is that the structure coefficient is scaled to remove the difference of the multiple R2. A special case that highlights the usefulness of structure coefficients in identifying how the variance assignment process for a particular regression equation occurs is the suppression case

Commonality Coefficient Developed in the 1960s as a method of partitioning variance (Mayeske et. al, 1969; Mood, 1969, 1971; Newton and Spurrell, 1967), commonality analysis partitions the R^2 that is explained by all independent variables in a MR into variance that is unique to each variable and variance that each possible subset of independent variables share.

There are two types of commonality coefficients: unique effects and common effects. Unique effects reflect how much variance an independent variable contributes to a regression equation that is not shared with other independent variables. A unique effect is a measure of total effect, as it is only calculated when all independent variables have been entered into the regression equation.

4.3 A New Regression Model: Modal Linear Regression By WEIXIN YAO and Longhai Li (2014) [?]

This article develops a new data analysis tool called modal linear regression in order to explore high-dimensional data. Modal linear regression models the conditional mode of a response Y given a set of predictors x as a linear function of x.

Suppose we have collected a random sample

$$(xi; yi), i = 1, ..., n$$

where x_i is a p-dimensional column vector and y_i is observation of a continuous response variable Y . A new regression model called modal linear regression (MODLR) is proposed that assumes that the mode of f (y | x) is a linear function of the predictor x. MODLR measures the centre using the 'most likely' conditional values rather than the conditional average.

Compared with other regression models, the proposed MODLR has the following features:
(i) MODLR attempts to capture the 'most probable' value—the mode (instead of the mean, median or quantile) of the conditional distribution of Y given x.
ii) MODLR may provide shorter prediction intervals than other linear regression approaches for a nominal confidence level, because an interval around a conditional mode can cover more samples than an interval of the same length around a conditional mean.
(iii) MODLR is robust to outliers that do not follow the same relationship exhibited by the majority of a sample and is also robust to heavy-tailed conditional error distributions.
(iv) MODLR is well justified in situations where conditional distributions are highly skewed.

Modal linear regression Suppose that a response variable Y given a set of predictor x is distributed with a probability density function f (y | x). Assume that the mode of f (y | x), denoted by Mode (Y | x) = argmaxy(f (y | x)), is unique. The proposed MODLR method assumes that Mode(Y| x) is a linear function of x, that is,

$$Mode(Y \mid x) = x^T \beta$$

[**?**] We assume that the first element of x is 1.

Simulation study and application The modal regression estimator requires a selection of the bandwidth. The asymptotically optimal bandwidth formula contains the unknown quantities, that is, the vth derivative of the conditional density of ϵ given x.

4.4 A comparison of random forest regression and multiple linear regression for prediction in neuroscience By Paul F. Smith (2013) [?]

Even though indicator variables can be used to contain nominal variables, the predictor variables should be numerical. The predictor variables can be numerical, ordinal, or nominal in nature.

The data is frequently divided into training and test sets (e.g., 90:10), and the mean square error (MSE) between the training and test data is calculated as a measure of the model's success.

'Random forests,' which were created by selecting a portion of data from the training set at random. To forecast the target variable with the minimum MSE, the multiple regression tree solutions are averaged.

Multiple linear and random forest regression All of the analyses were carried out with the help of the R programming language (2012). The data was split into training and test data sets in a 90:10 ratio. Multiple linear regressions (MLRs) were performed on the training data set, employing one neurochemical variable at a time as the response variable

and the other 8 as predictor variables, in addition to the categorical predictor variables, age, and housing, based on the concerns outlined above. In all cases, the response neurochemical variable was a continuous variable, expressed as a concentration. The remaining 8 neurochemical predictor variables were all continuous, but age and housing were nominal, thus they were transformed to binary indicator variables, with young equaling 0 and aged equaling 1, standard equaling 0, and enriched equaling 1. The magnitude of the adjusted R2, the residual standard error (RSE) for the regression, the t-test results for the various predictor variables, and the analysis of variance (ANOVA) for the regression can all be used to determine the MLRs' success.

RFR modelling necessitates deciding on m, the number of variables (a subset of the available p predictor variables) that will be used to determine the choice at each tree node. Because each target neurochemical variable had ten predictor variables, it was chosen to make m equal to the integer component of the square root of p, i.e. m = 3. The total number of trees that would need to be fitted was fixed at 1000.

5 SIMPLE LINEAR REGRESSION

5.1 Advanced Statistics: Linear Regression, Part I: Simple Linear Regression by Keith A. Marill (2004) [?]

Simple linear regression is a subtype of linear regression in which there is a single outcome or dependent variable and a single predictor or independent variable.

Fundamental Assumptions In simple linear regression, the equation for a line is used to model the relationship between two variables. The equation is z=kx+ c, where k is a coefficient that represents the slope of the linear relationship between the variables x and z and c is a constant. The constant c is known as the "z intercept" because it is the value of z when x=0 and the regression line crosses the z-axis.

1.There is some linear relationship between the predictor and outcome variable. As the points' values rise along the x-axis, their values rise along the y-axis as well. Rather than a curve or other shape, the cloud of points appears to revolve around a straight line. **2.The variation around the regression line is constant (homoscedasticity).** It's possible that some points are further away from the regression line than others. The average variance of the points from the regression line stays nearly the same as the eye wanders laterally down the x-axis, which is known as homoscedasticity.

3.The variation of the data around the regression line follows a normal distribution at all values of the predictor variable. The data points at any given value of x will form a bell-shaped or normal curve around the value of the regression line at that point if they are examined. The regression line will be near to the majority of points, with fewer points being far away.

4. The deviation of each data point from the regression line is independent of the deviation of the other data points. The relationship between the value of one point and the regression line has no influence on the value of another point in the dataset.

THE METHOD OF LEAST SQUARES The goal of formulating the regression line is to maximise the portion of the data points that can be assigned to the regression while minimising the residual. Three data points with result values of 1, 2, and 6 are shown in Figure 11. The sample's mean outcome value is 3, and the sample's mean outcome value for the three points is represented by a broken straight line drawn horizontally. In addition,

a regression line has been created over the graph. The sum of the vertical distance from the mean outcome line to the regression line (reg) and the distance from the regression line to the data point can be used to depict the variation of each data point from the mean outcome.

VARIABLE TRANSFORMATIONS After changing the independent predictor variable, a linear regression equation can be created. For example, in simple linear regression, let z = c + kx, and now let x=s2, where x is a transform of the variable s. To summarise, transforms enable the use of the well-developed mathematical framework in linear regression to model some data sets that would otherwise fail to meet the required assumptions.

5.2 Correlation and Simple Linear By Kelly H. Zou, Kemal Tuncali, Stuart G. Silverman (2003) [?]

Simple Linear Regression The purpose of simple regression analysis is to evaluate the relative impact of a predictor variable on a particular outcome. Typical steps for regression model analysis are the following: (a) determine if the assumptions underlying a normal relationship are met in the data, (b) obtain the equation that best fits the data, (c) evaluate the equation to determine the strength of the relationship for prediction and estimation, and (d) assess whether the data fit these criteria before the equation is applied for prediction and estimation.

Least Squares Method The main goal of linear regression is to fit a straight line through the data that predicts Y based on X. To estimate the intercept and slope regression parameters that determine this line, the least-squares method is commonly used. A set of regression parameters are found such that the sum of squared residuals (ie, the differences between the observed values of the outcome variable and the fitted values) are minimized (14). The fitted y value is then computed as a function of the given x value and the estimated intercept and slope regression parameter.

Limitations and Precautions The following understandings should be considered when regression analysis is performed. (a) To understand whether the assumptions have been met, determine the magnitude of the gap between the data and the assumptions of the model. (b) No matter how strong a relationship is demonstrated with regression analysis, it should not be interpreted as causation (as in the correlation analysis). (c) The regression should not be used to predict or estimate outside the range of values of the independent variable of the sample.

6 MULTIPLE LINEAR REGRESSION

6.1 Assumptions of Multiple Regression: Correcting Two Misconceptions by Matt N. Williams,Carlos Alberto Gomez Grajales ,Dason Kurkiewicz (2013) [?]

In this paper, we see that multiple regression models estimated using ordinary least squares require the assumption of normally distributed errors in order for trustworthy inferences, at least in small samples, but not the assumption of normally distributed response or predictor variables. Secondly, we saw that regression coefficients in simple regression models will be biased (toward zero) estimates of the relationships between variables of interest when measurement error is uncorrelated across those variables, but that when correlated measurement

error is present, regression coefficients may be either upwardly or downwardly biased.

Assumption about the model: Linearity in the parameters The model that relates the response Y to the predictors

$$X_1, X_2, X_3 ... X_p$$

is assumed to be linear in the regression parameters This means that the response variable is assumed to be a linear function of the parameters (1, 2, 3… p), but not necessarily a linear function of the predictor variables

$$X_1, X_2, X_3 ... X_p$$

unfortunately repeat a common misconception in claiming that "Standard multiple regression can only accurately estimate the relationship between dependent and independent variables if the relationships are linear in nature". In reality, some types of non-linear relationships can be modeled within a linear regression framework. For example, a quadratic (U or reverse-U shaped) relationship between X and Y can be accommodated by including both X and X_2 as predictors, as in the equation:

$$y = b_0 + b_1 x_1 + b_2 x_2 + ... \tag{5}$$

This regression equation is still a linear regression equation, because Y is modeled as a linear function of the parameters.

1. Zero conditional mean of errors :The errors are assumed to have a mean of zero for any given value, or combination of values, on the predictor variables.
2. Independence of errors :The errors are assumed to be independent . Breach of this assumption leads to biased estimates of standard errors and significance, though the estimates of the regression coefficients remain unbiased, yet inefficient state that independence of observations is required for linear regression, which is not entirely correct.
3. Homoscedasticity (constant variance) of errors: The model errors are generally assumed to have an unknown but finite variance that is constant across all levels of the predictor variables. This assumption is also known as the homogeneity of variance assumption. If the errors have a variance that is finite but not constant across different levels of the predictor/s , ordinary least squares estimates will be unbiased and consistent as long as the errors are independent, but will not be efficient.
4. Normal distribution of errors: Normally distributed errors are not required for regression coefficients to be unbiased, consistent, and efficient but this assumption is required for trustworthy significance tests and confidence intervals in small samples.

The predictor variables are assumed to be measured without error.

6.2 Contrast Coding in Multiple Regression Analysis: Strengths, Weaknesses, and Utility of Popular Coding Structures by Matthew J. Davis (2010) [?]

The present paper provides a description of the most popular coding structures, with emphasis on their strengths, limitations, and uses. Dummy coding, described by Cohen and Cohen in 1983, is the simplest coding structure that allows the researcher to examine group mean differences. Dummy coding only uses 1s and 0s, and is completed by creating up to

(k − 1) contrasts; thus in a two group example dummy codes would be created by giving one group a 1 and the others a 0 (Fox, 1997). More complex dummy codes can be created though for variables with multiple categories. Completion is just as easy though, first group one is given a 1 in the first contrast with all other groups receiving a 0, in the second contrast group two receives a 1 with all other groups receiving a 0, and so forth (Wendorf, 2004).

One benefit is that dummy code structure works especially well with nominal and more specifically dichotomous data. A second benefit to such a coding structure is the ease of interpretation. Because for groups coded as 0, the intercept of the regression equation is the mean of these groups, thus if we know the mean of the 0 coded groups we already understand half of the regression equation.

Limitations of dummy coding

1. Contrast coding Contrast coding was created as an extension of dummy coding to examine mean differences between groups The basic premise of this coding structure is that it requires the researcher to assign contrasts that sum to 0 across all subjects.

Limitations:contrasts are created in an orthogonal, balanced (same n for all groups) design, then interpretation of main effects and interactions is fairly straightforward and accurate; however, when designs become unbalanced or non-orthogonal, such interpretations can become confounded.

2. Effects coding is very similar to contrast coding,h this coding structure the same process is completed as that for dummy coding except the last group receives a −1 on all contrasts, thus only (k −1) contrasts are used in this coding type Effects coding allows researchers to test mean differences between two groups by using simple contrasts, but does not allow the complex contrasts of contrast coding.

One of the benefits of this coding structure is the ease of interpretation. With effects coding, the slope is simply the difference between the mean of the group coded as 1 and the grand mean of all the group.However, one limitation of this coding structure is it only tests the difference between simple contrasts and does not allow the researcher to test hypotheses for both simple and complex contrasts.

6.3 Correct and Incorrect use of Multilinear Regression By Michelle Sergent (1995) [?]

In this paper an example of correct and incorrect use of Multilinear Regression is presented in detail; the quality of the coefficients and the goodness of the prediction depend on the experimental design, and the value of R^2 gives no information at all about them. Multi linear Regression is applied when experimenters wish to investigate the relationship between a block of predictor variables (X), whose values are fixed by the experimenter, and one or more responses (Y), measured at each experiment. let us consider the following linear model:

$$\eta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k \tag{6}$$

[?]

With variable to be explained, experimental response,representing one of the characteristics of the phenomenon studied; β_j = coefficient to be estimated (j = 0, 1 ..... k);
X_j =independent variable or predictor or explanatory variable (j =1 ..... k).

When an experiment is performed (assigning values to the variables X), we obtain, taking into account the experimental error e_j, not the value corresponding to η_j, but an estimate

called y_j :

$$y_j = \eta_j + e_j$$

There are statistical tests which can be applied in order to evaluate the significance level of a regression analysis., such as: Squared multiple regression coefficient R^2 . Sum of Squares of Residuals(SSR) Using the formula given below , we find SSR for the calculated results and infer that candidate 2 has best outcome because the best regression is usually the one for which R2 is close to 1 and SSR is close to zero.

$$SSR = (Y_{exp} - Y_{comp})^2 \tag{7}$$

[?]

The flowchart summarizes all the information that can be obtained during the different steps of regression.

6.4 Multiple linear regression analysis By Nikolaos Pandis (2016) [?]

In linear regression using a single independent continuous variable. In linear regression analysis, we can include more parameters in an effort to find factors that better predict an outcome. This analysis is known as multiple linear regression analysis. We can use a combination of continuous and categorical predictors in a multiple regression model, as well as interaction terms between categorical predictors or between categorical and continuous predictors, to analyze the combined effect of those parameters on the outcome.

Multiple Linear Regression Output			
Variable	Coefficient	Standard Error	P value
Age	0.049	0.005	< 0.001
Gender	-0.044	0.131	0.736
wbc	-0.004	0.021	0.845
hb	-0.009	0.002	< 0.001
lac	0.740	0.047	< 0.001

[?]

The Table shows that in the adjusted model, initial crowding remains a significant predictor (P 5 0.001), whereas sex shows a non significant association (P 5 0.62) with days to alignment.

The Figure shows the predicted values separately for boys and girls. The predicted number of days to alignment is slightly higher in boys than in girls, but this difference is assumed to be fixed at every value of irprtx (the interaction between sex and irprtx is very weak).

6.5 The Steps to Follow in a Multiple Regression Analysis By Theresa Hoang Diem Ngo, La Puente, CA 2012 [?]

Model building, model adequacy, model assumptions – residual tests and diagnostic plots, potential modelling difficulties and solutions, and model validation are the five processes to follow in multiple regression analysis. DATA SET The goal is to develop a multiple regression model to predict a vehicle's invoice using a data set called Cars from the SASHELP library. The bill is calculated based on the number of cylinders, engine horsepower, length, MPG city, MPG highway, weight, wheelbase, drivetrain, make, and kind of vehicle.

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon \tag{8}$$

[?] Response variable(y) = invoice Independent Variables(x_i) = cylinders, engine, horsepower, length, MPG city, MPG highway, weight, wheelbase, drive train, make, and type.
STEP 1. MODEL BUILDING
As independent variables, you can employ first-order or second-order terms, interaction terms, and dummy variables. Two variable screening approaches that can help analysts find the most important elements that contribute to the response variable are stepwise regression and the all-possible-regressions selection procedure.
1) Stepwise Regression determines the independent variable (s) added to the model at each step using t-test.
2) The All-Possible-Regressions Selection Procedure returns all possible models with the recommended independent variable(s) that are related with the following criteria at each phase. The analyst chooses the potential independent variables to include in the model based on these criteria.

STEP 2. MODEL ADEQUACY
The following criteria are important for checking the utility of the model:
1) Global F test: To see if the independent variables as a group are significant in predicting the response variable.
2) $100(1-\alpha)\%$ Confidence intervals and t-tests: Inferences about the β parameters.
3) R^2_{adj}: After adjusting for sample size and number of factors, the total sample variation of the response variable y that is explained by the model. Both R^2 and R^2_{adj} are indicators of how well the prediction equation fits the data.
4) Root MSE or s: The estimated standard deviation of the random error. The interval $\pm 2s$ is an approximation of the accuracy in predicting y based on a specific set of independent variables.
5) Coefficient of variation (CV): The ratio of the response variable's sample mean to the estimated standard deviation. Models with a CV of 10% or less typically produce accurate predictions (Mendenhall and Sincich 108).
STEP 3. MODEL ASSUMPTIONS
Random error: $\epsilon \sim N(0, \sigma)^2$ [?]
All pairs of random errors are independent.
Using the data to obtain the least squares estimates $\beta_0 + ... + \beta_k$ the error value can be estimated to detect the deviation between the observed and the predicted value of y.
STEP 4. POTENTIAL MODELING PROBLEMS AND SOLUTIONS
Analysts should be aware of potential concerns while building a multiple regression model, many of which are caused by assumptions that have been broken. Some of these flaws can only be minimised, while others can be fixed to improve the accuracy of the model.
STEP 5. MODEL VALIDATION
Models that fit the sample data well may not be statistically useful when applied to a fresh data set due to changes or unknown events that may occur in the future. In addition to ensuring that the model is adequate, it is critical to validate its performance in practise. The following model validation procedures have been presented.
1) Examine the predicted values: The model is either inaccurate or the parameter coefficients are poorly calculated if the predicted values appear to be unreasonable and much beyond the response variable's range. If the expected numbers appear to be correct, continue to evaluate the model validity.
2) Examine the model parameters: If coefficients have the opposite sign as predicted, have unusually large or tiny values, and/or are inconsistent when applied to new data, they have been inadequately estimated and/or multicollinearity exists.
3) Apply the model to the new data for prediction: Use $R^2_{prediction}$ and MSE to measure the model validity.
4) Perform data-splitting: The sample data is split into two sections: one for estimating model parameters and the other for validating predictions.

7 LOGISTIC REGRESSION

7.1 Alternatives to logistic regression models in experimental studies By Francis L. Huang [?]

When evaluating binary outcomes in psychology or education experiments, logistic regression models (LRMs) are frequently used. However, one disadvantage of LRMs is that the results are often difficult to comprehend. This paper discuss the linear probability model, the log-binomial model, and the modified Poisson regression model as alternatives to LRMs in experiment analysis.

When performing a regression with a binary dependent variable, logistic regression model (LRM) is often used rather than a standard linear model using ordinary least squares (OLS) regression. A linear probability model (LPM) is a linear model that uses OLS regression to predict a binary outcome. The dependent variable is simply the dependent variable coded as a 1 or 0. LPMs may be easier to estimate, and their results may be easier to communicate to a wider audience (Cleary and Angel, 1984; Dey and Astin, 1993).

The challenges associated with LRMs

1. The difficulty of interpreting model coefficients, successfully presenting results, and comprehending the magnitude of the effect size are all downsides of employing an LRM.

2. Logit units (i.e., log odds units), odds ratios (ORs), probabilities, and risk ratios (RRs) are all used to interpret logistic regression results, and mixing up the unit of interpretation can lead to highly misleading results (Liberman, 2005).

Should LPMs really not be used?

Using a continuous predictor to forecast LPM values can result in nonsensical projected probabilities above 1 or below 0. (DeMaris, 1995; Huang and Moon, 2013). LPM model residuals frequently show heteroscedasticity or uneven variances across different levels of X, which is a violation of the linear regression model assumption (Cohen et al., 2003; Fox, 1991).

The functional form of the relationship specified using OLS may be incorrect.

Other generalized linear model specifications

The LRM is expressed as

$$log(p/1-p) = \beta_0 + \beta_1 X_1 \tag{9}$$

[?] where p is the probability of an event occurring (Y= 1), the error term follows a binomial distribution, and the link between the outcome and the dependent variable uses a logit link (i.e., the log of the odds) function.

1.The log-binomial model instead of the logit link, the log link function is used instead where

$$log(p) = \beta_0 + \beta_1 X_1 \tag{10}$$

2. The Poisson regression model Binomial data, which consists of ones and zeroes, can be thought of as following a Poisson distribution, in which getting a value of two or larger is unlikely. A Poisson regression equation with one predictor can be written as

$$log(\mu') = \beta_0 + \beta_1 X_1, \tag{11}$$

[?] where (μ') is the expected number of events given X_1. The robust Poisson regression model is preferred over the log-binomial regression model because the results are similar, it is easier to specify, and it has less convergence concerns.

Current study

Results are studied using simulated datasets analysed using an LRM, an LPM, and a modified Poisson regression model in the context of an experimental setting, where individuals are randomly allocated to a treatment or a control condition.

Method

Data generating process

A LRM with one dichotomous (e.g., a treatment assignment variable) and one continuous predictor is used to simulate a binary result. The simulated model looked like this:

$$log(p/[1-p]) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \tag{12}$$

[**?**] where p was the probability of the event occurring, X1 was a dichotomous predictor drawn from a binomial distribution with a probability of 50% and X2 was a continuous normal predictor (M = 0, SD = 1). Both X1 and X2 were generated independently and Cor(X1, X2) =0.

HOW DOES MULTIPLE LOGISTIC REGRESSION WORK?

Without utilizing any predictors, the statistical program determines the baseline chances of having the outcome versus not having the outcome. This provides us with a consistent source of information (also known as the intercept). The chosen independent (input/predictor) variables are then inserted into the model, and for each of them, a regression coefficient (also known as "beta") and "P" values are calculated.

CAUTIONS AND PITFALLS

1. Choosing the right predictor variables

The correct variables to enter into the model are the key to a successful logistic regression model. While it may be tempting to incorporate as many input variables as possible, doing so can dilute actual correlations, resulting in huge standard errors and imprecise confidence intervals, or, on the other hand, identify fake associations.

2. Avoiding the use of highly correlated variables.

When input variables are highly correlated (known as multicollinearity), the impact of each on the regression model becomes less precise.

3. Restricting the number of variables entered into a multivariate logistic regression model

For each variable entered into a logistic regression model, it has been advised that the data should contain at least ten occurrences. As a result, if we want to uncover predictors of mortality using a sample of sixty deaths, we can only look at six (=60/10) predictor variables. The validity of this thumb rule, however, has been questioned.

4.Odds versus risk

It's important to keep in mind that logistic regression generates aORs for each predictor. The odds and the risk are two different things, and while the odds may appear to be large, the absolute risk may be minimal.

5. Handling continuous input variables

It's tempting to separate subjects into groups (e.g., age >50 years vs. age 50 years) when dealing with continuous data (such as age, height, or weight). This is a bad practise since the cutoffs are often arbitrary, resulting in the loss of some information.

7.2 Comparison between SVM and Logistic Regression: Which One is Better to Discriminate? By Diego Alejandro Salazar 2012 [?]

In this paper we saw the comparison, by statistical simulation, the mis-classification rate (MCRs) for support vector machines (SVM) and logistic regression (LR). When the data comes from a population in which individuals can be classified in one of two mutually exclusive categories. We consider different scenarios in which the training data set and other functional parameters are controlled. This control allows us to generate data sets with the specific characteristics and further decide whether SVM or LR should be used in that given particular situation.

Support Vector Machine (SVM) is a classification and regression method that combines computational algorithms with theoretical results; these two characteristics gave it a good reputation and have promoted its use in different areas. Moguerza and Muñoz (2006) and Tibshirani and Friedman (2008) consider a classification problem in which the discriminant function is nonlinear (Figure 1a), and there exists a kernel function$\boldsymbol{\phi}$ to a characteristic space on which the data is linearly separable as shown in (Figure 1b). On this new space, each data point corresponds to an abstract point on a p-dimensional space, where being p the number of variables in the data set.

Once the optimal margin hyperplane has been found, it is projected on the data's original space to obtain a discriminant function. For example, Figure 2(a) shows a data set in which two groups, linearly separable, are characterized by white and black dots that are not linearly separable. In Figure 2(b), the data is transformed to R where it is separable by a plane and, when it is projected back to the original space, a circular discriminant function is obtained.

In the Normal case, the MCR for the polynomial SVM model is higher (poor performance). On the other hand, the performances of LR and linear, radial and tangential SVM models are equivalent. When the sample sizes differ, the MCR of the tangential and polynomial kernel is lower than when the groups have the same number of individuals. However, the former presents lower MCRs. SVM models are a feasible alternative to RL. However, as shown for the Poisson, Exponential and Normal distributions, the polynomial SVM model is not recommended since its MCR is higher.In the case of multivariate and mixture of distributions, SVM performs better than LR when high correlation structures .Furthermore, SVM methods required less variables than LR to achieve a better (or equivalent) MCR.

7.3 Estimating predicted probabilities from logistic regression: different methods correspond to different target populations By Clemma J Muller and Richard F MacLehos [?]

Marginal standardization (predicted probabilities summed to a weighted average reflecting the confounder distribution in the target population); prediction at the modes (conditional predicted probabilities calculated by setting each confounder to its modal value); and prediction at the means(predicted probabilities calculated by setting each confounder to its mean value) are all discussed in this paper.

Logistic regression and predicted probabilities

The logit link is used in logistic regression to predict the log-odds of an event occurring. The model and results reported here are based on a simple logistic regression with a di-

chotomous exposure (E) and a single dichotomous confounder (Z), but they may easily be expanded to include numerous categorical or continuous confounders. Following maximum likelihood estimation of a logistic regression model, it is simple to estimate the probability of the outcome for any E = e and Z =z as follows:

$$p_{ez} = exp[\alpha + \beta_1 * e + \beta_2 * z]/(1 + exp[\alpha + beta_1 * e + beta_2 * z] \quad (13)$$

[?]

Method 1: marginal standardization The estimate of interest (e.g., rate, prevalence, or chances) is proportionally modified based on weight for each level of the confounding factor in this method (s). The exposure E is set to the (possibly counterfactual) level e for everyone in the dataset after performing logistic regression, and the logistic regression coefficients are used to calculate predicted probabilities for everyone at their observed confounder pattern and newly assigned exposure value.

Method 2: prediction of the modes Method 2 determines the outcome's projected probability for each exposure level, assuming that everyone in the population has the most common confounder values:

$$Pr(Y = 1|Set[E = e], Z = z^m \quad (14)$$

[?] where z^m reflects the modal value(s) of confounder vector Z.

Method 3: prediction at the means Method 3 determines the anticipated probability of the result based on exposure status, assuming that each individual in the dataset has the mean value of each confounder. It's written like this:

$$Pr(Y = 1|Set[E = e], Z = z^{'} \quad (15)$$

[?] **Marginal standardization vs prediction at the means** Consider the following scenario: the confounder of interest (Z) is gender, and half of the research sample is male (Z = 1) and half is female (Z = 0). We solely assess the probability of the result among the unexposed for the sake of simplicity. Assume that 50 percent of the study population's unexposed women and 99 percent of the research population's unexposed men experience the outcome, with $\alpha^{'} = 0, \beta_2 = 4.60$ from a logistic model fit to these data.

7.4 Logistic Regression Model Optimization and Case Analysis By Xiaonan Zou and Yong Hu [?]

In this paper, we look at the logistic mathematical model, define the error function, and use the gradient descent method to get the regression coefficient and the sigmoid function should be improved. And the number of iterations is lowered, the classification effect is improved, and the accuracy has remained essentially unaltered. Linear regression is a statistical analysis approach for determining the quantitative relationship between two or more variables that employs regression analysis in mathematical statistics. Consider the following two variables: (Y1, Y2,..., Y_i) is a dependent variable, and (X1, X2,..., X_i) is an independent variable. When the dependent and independent variables have a linear relationship, it's called a linear relationship. Its goal is to determine the best appropriate parameters and fit the hashed data points with a straight line. The major use of logistic regression is classification. The most significant distinction between it and linear regression is that its data points are not structured in line rows. The goal of logistic regression is to locate the classification boundary line, which is represented by the regression formula. For example,

When classifying, the function output 0 or 1 represents two classes to facilitate processing. The range of the dependent variable is either 0 or 1. Many functions fit the above criteria, but the Sigmoid function is now widely used. $\sigma(z) = \frac{1}{1+e^{-z}}$ [?] As shown in the diagram, $\sigma(0) = 0.5$. When z<0, the function approaches 0 and becomes 0 class, the function value satisfies the above classification function criteria. Logistic regression's classification procedure is as follows: It is expected that the input data features can be described as $(x_0, x_1, x_2, ..., x_n)$, and that each feature is multiplied by a regression coefficient $(w_0, w_1, w_2, ..., w_n)$, and that the input z is then summed as the sigmoid function:

$$z = w_0x0 + w_1x1 + ... + w_nxn \tag{16}$$

[?] which is,

$$z = w^T x \tag{17}$$

[?] Here, w is a row vector , the regression coefficient, x is the column vector,the input data of the classifier.

ANALYSIS OF REGRESSION COEFFICIENTS The classification algorithm's predicted value (y'=(z)) and the actual category label (y=0 or 1) have the minimum error, as a result, the error function Y can be written as:

$$Y = y^{'} - y \tag{18}$$

[?] Because the error is positive, it can be added to the absolute value.

$$Y = |y' - y| \tag{19}$$

[?] the error function can be defined as:

$$Y = \frac{1}{2}(y' - y)^2 \tag{20}$$

[?] To minimise the function $\frac{1}{2}(y' - y)^2$, the value of w is required. A gradient can be employed based on calculus understanding. The descent technique is iterated, and it has the following expression:

$$w = w - \alpha\frac{\delta Y}{\delta w} \tag{21}$$

[?] In equation(16) - (19), w is the regression coefficient row vector; Y is the error function; α is the iteration step size; y is the prediction category, $y = \sigma(z) = \sigma(w^T x)$ [?] ; y is the category label, y= 0,1. The gradient descent method requires the error Y to be as little as possible. The actual Y rises with the number of repetitions, according to the function specification, and the infinity approaches 0, which is not equal to 0. As a result, setting a specified threshold (such as e^7) and seeing if Y is smaller than this threshold or meets the predetermined number of repeats is the best method. In the gradient descent method, the step size is an empirical estimate based on the error function. In theory, the lower the real error function, the fewer iterations are required, and the algorithm can programme training samples faster. In combination with the first section, the 0 1 step function is unguided and the Sigmiod function is selected. When the Sigmiod function argument is close to 0, however, the dependent variable is far from the actual label (0, 1), resulting in a bigger

error function value. Although the gradient descent approach can find the minimal error, the number of iterations is considerable, which reduces the algorithm's efficiency. As a result, this study presents a radical Sigmiod function that transforms the Sigmiod function's bottom e into (where n>=1).

SOLVING MATHEMATICAL MODELS When the error approaches the smallest value, the gradient descent algorithm finds the regression coefficient. The gradient descent formula produced using the matrix approach is as follows:

$$w = w - \alpha n x^T (y' - y)(1 - y')^T y$$

[?]

Iteration can be used to get the regression coefficient at the minimum of the error function. The following is the first a SOLVING MATHEMATICAL MODELS algorithm:

Step 1: The regression coefficient, step size, number of iterations, and other parameters are all initialised.

Step 2: Continue steps 3-5 until the termination condition is satisfied (the termination condition is the number of iterations).

Step 3: Substitute the parameters and calculate the A= wx matrix.

Step 4: Substitute the result of step 3 calculation into a function, finds the value of $\sigma(A)$, and calculates y' - y.

Step 5: Using an iterative formula, update w.

7.5 Relating Patient Characteristics to Outcomes By Juliana Tolles [?]

Seymour et al1 published a new approach for assessing the probability of a patient dying of sepsis using the information on the patient's breathing rate, systolic blood pressure, and heart rate in a recent issue of JAMA as well as changed mentation. These clinical factors, referred to as "predictor" or "explanatory" or "independent variables" in the technique, were used to estimate the risk of a patient having an outcome of interest, which is referred to as the dependent variable.

Use of the Method

1. Why Is Logistic Regression Used? Using information or qualities that are assumed to be related to or impact such events, logistic regression can be used to estimate the chance that an event will occur or that a patient will have a specific outcome. Logistic regression can reveal which of the different factors under consideration has the strongest link to a certain result, as well as the degree of the potential influence.

2. Description of the Method Binary or dichotomous outcomes are those that can only have two values (for example, survived vs. died). The fraction of patients who experience the outcome of interest, or the probability that any single patient would experience that event, can be used to summarise the outcomes for groups of patients. The probability that an event will occur is divided by the probability that it will not occur is called the odds. The change in the odds of an outcome is measured as a ratio called the odds ratio (OR).

What Are the Limitations of Logistic Regression?

1. The number and suitability of the measured independent predictor variables determine the validity of a regression model. All biologically relevant elements should, ideally, be covered.

2. The variables must have a constant magnitude of association across the range of values

for that variable.
3. Many logistic regression analyses assume that the effect of one predictor is not influenced by the value of another predictor.

How Should the Results of Logistic Regression Be Interpreted in This Particular Study? The quick Sequential [Sepsis-related] Organ Failure Assessment was developed by Seymour et al using logistic regression to develop a novel clinical tool for determining the probability of mortality in patients with sepsis (qSOFA). Using respiration rate, systolic blood pressure, and Glasgow Coma Scale score, the qSOFA model is used to estimate the chance of in-hospital death in patients with suspected infection. The authors built a simplified model that could be applied to individual patients by counting the number of positive clinical predictors by giving all coefficients the same value.

7.6 Model building strategy for logistic regression: purposeful selection By Zhongheng Zhang [?]

Working example of Logistic Regression Here five variables are created age, gender, lac, hb and wbc for the prediction of mortality outcome. The outcome variable is binomial that takes values of "die" and "alive".

$> set.seed(888)$
$> age <- -abs(round(rnorm(n = 1000, mean = 67, sd = 14)))$
$> lac <- -abs(round(rnorm(n = 1000, mean = 5, sd = 3), 1))$
$> gender <- -factor(rbinom(n = 1000, size = 1, prob = 0.6), labels = c("male", "female"))$
$> wbc <- -abs(round(rnorm(n = 1000, mean = 10, sd = 3), 1),))$
$> hb <- -abs(round(rnorm(n = 1000, mean = 120, sd = 40)))$
$> z <- -0.1 * age - 0.02 * hb + lac - 10$
$> pr = 1/(1 + exp(-z))$
$> y = rbinom(1000, 1, pr)$
$> mort <- -factor(rbinom(1000, 1, pr), labels = c("alive", "die"))$
$> data <- -data.frame(age, gender, lac, wbc, hb, mort)$

Step one: univariable analysis

The first step is to use univariable analysis to explore the unadjusted association between variables and outcome.

Multiple Linear Regression Output			
Variable	Coefficient	Standard Error	P value
Age	0.049	0.005	< 0.001
Gender	-0.044	0.131	0.736
wbc	-0.004	0.021	0.845
hb	-0.009	0.002	< 0.001
lac	0.740	0.047	< 0.001

[?]

The error distribution and link function to be utilised in the model are described in the family argument. The family argument is given a Gaussian distribution with an identity link function. The univariable regression results can be displayed using the summary() method. For multivariable analysis, a P value of less than 0.25 and extra clinically relevant factors might be included. Literature supports a cutoff value of 0.25.

Step two: multivariable model comparisons

This stage fits the multivariable model that includes all of the variables identified in the

previous step. Variables that don't contribute to the model (e.g., those with a P value higher than the usual significance level) should be removed, and a new, smaller mode should be fitted. The parsimonious model is then compared to the original model using the partial likelihood ratio test to ensure that it fits as well as the original model. The coefficients of variables in the parsimonious model should be compared to the coefficients in the original model. If the change in coefficients () is greater than 20%, the removed variables have made a significant contribution to the adjustment of the effect of the remaining variables. These variables should be reintroduced into the model. This process of eliminating, adding variables, model fitting, and refitting continues until all variables are clinically and statistically inconsequential, while variables that remain in the model are crucial. Assume that the variable wbc is likewise included in our example because it is clinically relevant. The result shows that P value for variable wbc is 0.408, which is statistically insignificant. Therefore, we exclude it.

$> model2 <- glm(mort\ lac + hb + age, family = binomial)$

All variables in model2 are statistically significant. Then we will compare the changes in coefficients for each variable remaining in model2. The function coef() extracts estimated coefficients from fitted model.

Step three: linearity assumption In the step, continuous variables are checked for their linearity in relation to the logit of the outcome.

Step four: interactions among covariates We examine for potential interactions between covariates in this stage. An interaction between two variables means that one variable's effect on the response variable is influenced by another variable.

Step five: Assessing fit of the model The final stage is to inspect the model's fit. Checking for model fit consists of two components: (I) summary measures of goodness of fit (GOF) and (II) regression diagnostics. The former employs a single summary statistic, such as the Pearson Chi-square statistic, deviation, sum-of-squares, and the HosmerLemeshow tests, to assess model fit. The difference between observed and fitted values is measured using these statistics.

7.7 Random forest versus logistic regression: a large-scale benchmark experiment By Raphael Couronné , Philipp Probst and Anne-Laure Boulesteix [?]

Logistic regression (LR) Let Y be the binary response variable of interest, and $X_1, ..., X_p$ denote the random variables used as explanatory variables in this work, which are referred to as features. The conditional probability $P(Y = 1|X_1, ..., X_p)$ is linked to $X_1, ..., X_p$ by the logistic regression model.

$$P(Y = 1|X_1...X_p) = \frac{exp(\beta_0 + \beta_1 X_1 + ... + \beta_p X_p}{1 + exp(\beta_0 + \beta_1 X_1 + ... + \beta_p X_p} \tag{22}$$

[?] where $_{0,1,...,p}$ are regression coefficients, which are estimated by maximum-likelihood from the considered dataset. The probability that Y = 1 for a new instance is then estimated by replacing the's with their estimated counterparts and the X's with their realizations in Eq (21).

@bookID,
author = author,
title = title,

date = date,
OPTeditor = editor,
OPTeditora = editora,
OPTeditorb = editorb,
OPTeditorc = editorc,
OPTtranslator = translator,
OPTannotator = annotator,
OPTcommentator = commentator,
OPTintroduction = introduction,
OPTforeword = foreword,
OPTafterword = afterword,
OPTsubtitle = subtitle,
OPTtitleaddon = titleaddon,
OPTmaintitle = maintitle,
OPTmainsubtitle = mainsubtitle,
OPTmaintitleaddon = maintitleaddon,
OPTlanguage = language,
OPToriglanguage = origlanguage,
OPTvolume = volume,
OPTpart = part,
OPTedition = edition,
OPTvolumes = volumes,
OPTseries = series,
OPTnumber = number,
OPTnote = note,
OPTpublisher = publisher,
OPTlocation = location,
OPTisbn = isbn,
OPTchapter = chapter,
OPTpages = pages,
OPTpagetotal = pagetotal,
OPTaddendum = addendum,
OPTpubstate = pubstate,
OPTdoi = doi,
OPTeprint = eprint,
OPTeprintclass = eprintclass,
OPTeprinttype = eprinttype,
OPTurl = url,
OPTurldate = urldate

Random forest (RF) The random forest (RF) is a "ensemble learning" technique that involves combining a large number of decision trees to reduce variance as compared to single decision trees. We will look at Leo Breiman's original version of RF in this paper. Each tree in the original version of RF is generated using the CART technique and the Decrease Gini Impurity (DGI) as the splitting criterion, with a bootstrap sample taken randomly from the original dataset. Only a certain amount of randomly selected features are considered as candidates for splitting when building each tree at each split. Due of the high number of trees, RF is typically thought of as a black-box method.

Hyperparameters This section presents the most important parameters for RF and their

common default values as implemented in the R package randomForest. The parameter ntree denotes the number of trees in the forest. The default value is ntree=500 in the package randomForest. The parameter mtry denotes the number of features randomly selected as candidate features at each split. A high value of mtry reduces the risk of having only non-informative candidate features. In the package randomForest, the default value is $\sqrt{p}$ for classification with p the number of features of the dataset. The parameter nodesize represents the minimum size of terminal nodes. The default is replace=TRUE, yielding bootstrap samples, as opposed to replace=FALSE yielding subsamples— whose size is determined by the parameter sampsize.

Performance assessment

Cross-validation The original dataset is randomly partitioned into k subsets of approximately similar sizes in a k-fold cross-validation (CV). One of the folds is chosen as the test set for each of the k CV iterations, while the other k 1 are utilised for training. The performances are finally averaged over the iterations. In our study, we perform 10 repetitions of stratified 5-fold CV. The folds are chosen such that the class frequencies are approximately the same in all folds. For each performance measure, the results are stored in form of an M × 2 matrix.

Performance measures Let $p_i^{'}$, I = 1,..., n indicate the estimated probability of the ith observation I = 1,..., ntest) belonging to class Y = 1, while the real class membership of observation I is simply denoted as yi, given a classifier and a test dataset of size n_{test}. The accuracy, or proportion of correct predictions is estimated as

$$acc = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} I(y_i = y_i^{'}) \tag{23}$$

[?]

The Area Under Curve (AUC), or probability that the classifier ranks a randomly chosen observation with Y = 1 higher than a randomly chosen observation with Y = 0 is estimated as

$$acc = \frac{1}{n_{0,test} n_{1,test}} \sum_{i:y_i=1} \sum_{j:y_i=0} I(p_i^{'} > p'_j) \tag{24}$$

[?]

Brier score measures the deviation between true class and predicted probability and is estimated as

$$brier = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} (p_i^{'} - y'_i)^2 \tag{25}$$

[?]

7.8 Common pitfalls in statistical analysis: Logistic regression By Priya Ranganathan, C. S. Pramesh1 , Rakesh Aggarwal [?]

HOW DOES MULTIPLE LOGISTIC REGRESSION WORK? Without utilizing any predictors, the statistical program determines the baseline chances of having the outcome versus not having the outcome. This provides us with a consistent source of information (also known as the intercept). The chosen independent (input/predictor) variables are then inserted into the model, and for each of them, a regression coefficient (also known as ”beta”) and ”P” values are calculated.

CAUTIONS AND PITFALLS
1. Choosing the right predictor variables The correct variables to enter into the model are the key to a successful logistic regression model. While it may be tempting to incorporate as many input variables as possible, doing so can dilute actual correlations, resulting in huge standard errors and imprecise confidence intervals, or, on the other hand, identify fake associations.
2. Avoiding the use of highly correlated variables When input variables are highly correlated (known as multicollinearity), the impact of each on the regression model becomes less precise.
3. Restricting the number of variables entered into a multivariate logistic regression model For each variable entered into a logistic regression model, it has been advised that the data should contain at least ten occurrences. As a result, if we want to uncover predictors of mortality using a sample of sixty deaths, we can only look at six (=60/10) predictor variables. The validity of this thumb rule, however, has been questioned.
4. Odds versus risk It's important to keep in mind that logistic regression generates aORs for each predictor. The odds and the risk are two different things, and while the odds may appear to be large, the absolute risk may be minimal.
5. Handling continuous input variables It's tempting to separate subjects into groups (e.g., age >50 years vs. age 50 years) when dealing with continuous data (such as age, height, or weight). This is a bad practise since the cutoffs are often arbitrary, resulting in the loss of some information.

7.9 Sparse Multinomial Logistic Regression: Fast Algorithms and Generalization Bounds Balaji Krishnapuram, Lawrence Carin, Fellow, IEEE, Mario A.T. Figueiredo, Senior Member, IEEE, and Alexander J. Hartemink 2005 [?]

Sparse classification algorithms aim to learn as sparse a classifier as possible, the likelihood of the weights in the presence of training data is typically regularized by some prior belief about the weights that promotes their sparsity Some of the sparse classification algorithms includes the relevance vector machine (RVM) ,the sparse probit regression (SPR) algorithm ,sparse online Gaussian processes , the informative vector machine (IVM) , and the joint classifier and feature optimization (JCFO) algorithm , These algorithms learn classifiers that are constructed as weighted linear combinations of basis functions; the weights are estimated in the presence of training data. In this paper, we addressed three issues related to sparse classifiers. First, we adopt a genuinely multiclass formulation based on multinomial logistic regression Second, by combining a bound optimization approach with a component-wise update procedure, a series of new fast algorithms for learning a sparse multiclass classifier . And at last, generalization bounds are derived . Given a dataset , we use sparsity-promoting Laplacian prior,

$$p(w) \propto exp(-\lambda||w||_1) \tag{26}$$

[?] where w is the weight vector corresponding to class i Next , by bound optimization is performed

$$w^{t+1} = w^t - B^{-1}g(w^t) \tag{27}$$

[?] The above equation is applied to maximum likelihood (ML) multinomial logistic regression as an alternative to IRLS (iteratively reweighted least squares) and how simple

variations can handle maximum a posteriori (MAP) multinomial logistic regression with either a Gaussian (l2-penalty) or Laplacian (l1-penalty) prior on the weights. Finally, the bound optimization iterations for ML multinomial logistic regression with g(w) and B is given below as in closed form

$$g(w) = \sum_{j=1}^{n}(y'_j - p_j(w))\otimes \tag{28}$$

[?] Multinomial Logistic Regression with Gaussian Prior

$$w^{t+1} = (B - \lambda I)^{-1}(Bw^t - g(w^t)) \tag{29}$$

[?] Here ,each iteration of this bound optimization method for multinomial logistic regression under a Gaussian prior is cheaper than each IRLS iteration for ML multinomial logistic regression. Multinomial Logistic Regression with Laplacian Prior , we get

$$w^{t+1} = \gamma^t(\gamma^t B\gamma^t - \lambda I)^{-1}\gamma^t(Bw^t - w^t)) \tag{30}$$

[?] multinomial logistic regression under a Laplacian prior for the same cost as the original IRLS algorithm for ML estimation . to optimise for the same The key idea will be to take the surrogate function and maximize it only with respect to one of the components of w, while holding the remaining components at their current values.. The resulted equation is

$$soft(a;\delta) = sign(a)max0, a - \delta \tag{31}$$

[?] Thus , we see that the objective function we optimize while learning an SMLR classifier is concave. This is not the case for the RVM or methods using Jeffreys priors This concavity has significant benefits for identification of unique maxima, for efficient computational implementation, and for derivation of useful generalization bounds. Second, many other sparse classification algorithms (including the RVM and Gaussian processes) can also be formulated for multiclass problems, but we expect our computational cost to scale more favorably. Third, the component-wise update procedure we described provides a natural mechanism for determining the inclusion and exclusion of basis functions; in the context of SMLR, the intuition behind traditional forward-backward feature selection heuristics is placed on a rigorous theoretical footing.

7.10 Ordinal Regression Analysis: Using Generalized Ordinal Logistic Regression Models to Estimate Educational Data [?]

In this paper,we show the use of generalized ordinal logistic regression models to predict mathematics proficiency levels using Stata and compare the results from fitting PO(proportional odds) models and generalized ordinal logistic regression models. The PO model is used to estimate the cumulative probability of being at or below a particular level of a response variable, or being beyond a particular level, which is the complementary direction. In this model, the effect of each predictor is assumed to be the same across the categories of the ordinal dependent variable. This means that for each predictor, the effect on the odds of being at or below any category remains the same within the model. This restriction is referred to as the proportional odds, or the parallel lines, assumption. The assumption of proportional odds is often violated, however, because it is strongly affected

by sample size and the number of covariate patterns The PPO model allows for interactions between a predictor variable that violates the PO assumption and different categories of the ordinal outcome variable. The generalized ordinal logistic regression model extends the PO model by relaxing the PO assumption. In this model, if the assumption is violated by a certain predictor, then its effect can be estimated freely across different categories of the dependent variable. The model is expressed as:

$$log(y_j) = ln(\pi_j(x)) \div [1 - \pi_j(x)] \tag{32}$$

[?] Compared to the PO model, the generalized ordinal logistic model provides a better solution when the proportional odds assumption is violated. The effects of the predictors which meet the PO assumption can be interpreted in the same way as that in the PO model. The effects of explanatory variables that violate the PO assumption must be interpreted separately at each comparison (i.e., being beyond a particular category versus at or below that category), and need more attention

8 DECISION TREE REGRESSION

8.1 A Decision Tree Regression based Approach for the Number of Software Faults Prediction By Santosh Singh Rathore and Sandeep Kumar [?]

We study the capability of decision tree regression (DTR) for predicting the amount of defects in two scenarios: intra-release prediction and inter-releases prediction in this study. The predictive accuracy of DTR is evaluated using absolute error and relative error, prediction at level l, and goodness-of-fit measure. Using the negative binomial regression (NBR) technique, Ostrand established a method for predicting the number of faults and fault density. Using varying ages and historical metrics of various files, the study was conducted throughout numerous releases of an inventory system. The findings revealed that fault prediction models based on NBR achieved significant results in terms of predicting the number of faults and fault density. Afzal investigated the use of genetic programming (GP) for predicting failure counts in a software system. Weekly defect counts were utilised as independent variables in the study, which was conducted over three industrial projects. The results showed that the GP-based fault prediction model has a high level of accuracy in predicting fault counts. Gao published a comprehensive review of various count models for predicting the number of defects. Five alternative count models were used in the investigation, which was conducted on some industrial software systems. The results showed that the zero-inflated negative binomial regression and the hurdle negative binomial regression achieved greater prediction accuracy among the various count models used.

Decision Tree Regression Decision tree regression is a type of tree-based structure used to predict the numeric outcomes of the dependent variable. To begin, a traditional decision-tree algorithm is employed to construct a tree. This decision tree employs a splitting criterion to reduce intra-subset volatility in the class-values of instances as they progress down each branch. As the root node, the characteristic that maximizes the expected error reduction is picked. Equation 32 contains the formula for calculating the

standard deviation reduction.

$$SDR = sd(T) - \sum_i \frac{|T_i|}{|T|} * sd(T_i) \tag{33}$$

[**?**] The tree is then pruned back to each leaf. Finally, to compensate for the severe discontinuities that would unavoidably exist between neighboring linear models at the pruned tree's leaves, a smoothing process is applied. The decision tree regression (DTR) method was chosen for the study because, unlike standard decision trees, it can predict the dependent variable's numeric outcomes. DTR can also handle datasets with high dimensionality, and the tree formed by DTR is substantially smaller than that generated by CART.

DTR based Fault Prediction Model In two different scenarios, intra-release prediction, and inter-releases prediction, we design and assess fault prediction models. We employ a 10-fold cross-validation approach for intra-release prediction. The cross-validation method is based on the leave-one-out principle, in which one part of each iteration is utilized as a testing dataset, while the remaining nine parts are used to train the fault prediction model. The procedure is repeated up to ten times, with the results averaged between iterations. We used the fault dataset from past releases as a training dataset for inter-releases prediction, and the fault prediction model was validated on the current release of the software system (testing dataset).

8.2 Algorithm of Building Regression Decision Tree Using Complementary Features By Sergey Saltykov [?]

It's possible that there are two features, each of which is a rather weak predictor, but together they are a strong predictor. At the same time, in the same dataset, there is a third feature, which is a medium strength predictor. If the third feature is removed from the dataset, a tree may be formed in which there is both the first and second feature in the chain. And the accuracy of such a tree may be higher than the tree formed on the dataset of three features. Removing some of the dataset's features may improve the accuracy of the model being built. The Random Forest method, for example, employs this heuristic approach. It is assumed that in order for the regression tree to be the most accurate, the average values of the most homogeneous subdatasets should be placed in various leaf components of the tree in some way. The addition of an informational structure to the dataset, namely, the complementary binary relation on the set of features, would improve the accuracy of the regression decision tree's prediction of the target variable for a specific class of cases. It is predicted that adding an informative structure to the dataset, such as a complementary binary relation on a collection of features, would increase the accuracy of the regression decision tree's prediction of the target variable for a given class of cases. A synthetic dataset is given in table 1 for testing complementary features. There are 19 samples, three characteristics, and a target value in this dataset. Using the CART procedure and a variation of the CART procedure, we attempt to construct a two-level regression decision tree. Because the f3 feature is a reasonably strong predictor of the target variable, while the f1 and f2 features are both poor predictors, if the f3 feature is not removed from the dataset, it will "overshadow" the f1 and f2 features, preventing them from appearing in the decision tree. When f3 is removed from the dataset, however, it becomes out that f1 and f2 together can dramatically raise the fraction of the explained variation - 52.63 times. Consider the overall dataset's Spearman's correlation of features with the target variable. Simple math shows that f1 and f2 have no statistically meaningful correlation with the target variable. The f3 characteristic, on the other hand, exhibits a negative correlation of =0.49. We may conclude that f3 is a considerably stronger predictor than f1 and f2 throughout the entire dataset. The CART approach does not use correlation when finding an ideal split, but this does not change the conclusions concerning predictor strength. As a result, the CART technique only uses f3 features to build a decision tree on a full dataset with all three features. That is, only the feature f3 split wins at each stage of the optimal split selection process.

8.3 Study and Analysis of Decision Tree Based Classification Algorithms By Harsh H. Patel , Purvi Prajapati [?]

Decision Tree Root nodes, branches, and leaf nodes make up a Decision Tree. Every internal node tests an attribute, the result of the test is on the branch, and the class label is on a leaf node as a result. A root node is the parent of all nodes and, as the name implies, is the highest node in a Tree. A decision tree is a tree in which each node (attribute) represents a feature, each link (branch) represents a decision (rule), and each leaf represents a result.

Decision Tree Algorithms

Algorithm Name	Classification	Description
CART (Classification and Regression Trees)	Uses Gini Index as a metric.	By applying numeric splitting, we can construct the tree based on CART
ID3 (Iterative Dichotomiser 3)	Uses Entropy function and Information gain as metrics.	The only concern with the discrete values. Therefore, continuous dataset must be classified within the discrete data set
C4.5	The improved version on ID 3	Deals with both discrete as well as a continuous dataset. Also, it can handle the incomplete datasets. The technique called "PRUNNING", solves the problem of over filtering
C5.	0 Improved version of the C4.5	C5.0 allows to whether estimate missing values as a function of other attributes or apportions the case statistically among the results
CHAID (CHi square Automatic Interaction Detector)	Predates the original ID3 implementation.	For a nominal scaled variable, this type of decision tree is used. The technique detects the dependent variable from the categorized variables of a dataset.

8.4 The role of decision tree representation in regression problems – An evolutionary perspective By Marcin Czajkowski , Marek Kretowski [?]

The focus of this study is on regression trees, which can be thought of as decision tree variants. Each leaf of the simplest regression tree has a constant value, which is usually the target attribute's average value.A linear (or nonlinear) regression function replaces the constant value in each leaf of the regression tree in the model tree. The new tested instance is followed down the tree from a root node to a leaf, using its attribute values to make routing decisions at each internal node, to forecast the target value. The anticipated value for the new instance is then tested in the leaf using a regression model. Most decision trees use axis-parallel decision borders to divide the feature space. Because each split in the non-terminal node involves only one feature, this sort of tree is called univariate. Inequality tests with binary results are commonly used for continuous-valued features, while mutually exclusive groupings of feature values are connected with the outcomes for nominal features. Multivariate decision trees are used when more than one characteristic is used to generate a test in an internal node. Oblique or linear decision trees are commonly referred to as oblique or linear, but heterogeneous trees with univariate, linear, and other multivariate tests are referred to as mixed trees.
The top-down and global approaches to decision tree induction are the two most popular concepts. The first is based on recursive partitioning, a greedy method. The induction process in the top-down technique starts at the root node and searches for the locally optimal split using the specified optimality measure. The training instances are then redirected to the newly constructed nodes, and the process is repeated until a stopping condition is reached for each node. In addition, after the induction, post-pruning is frequently used to avoid the problem of over-fitting the training data. Breiman et al. suggested a solution called Classification And Regression Tree (CART), which is one of the most common top-down induced univariate regression trees. The approach looks for a locally optimal split that minimises the sum of squared residuals and then develops a piecewise-constant model with the training sample mean fitted to each terminal node.
Mixed Global Model Tree
The Mixed Global Model Tree (mGMT) is a proposed addition for the GMT and GRT systems to better understand the underlying process behind representation selection. The algorithm's general structure is based on a standard EA framework, with an unstructured population and generational selection. The EA verifies alternative variants of the representations not only on the tree level but also on the node level, and may produce a heterogeneous tree, which we call a mixed tree, so the mGMT does not require setting the tree representation in advance.A mixed regression tree is a complex structure in which the quantity, type, and even number of test results for a given learning set are unknown in advance. As a result, the population's candidate solutions are not encoded and are displayed in their natural state.

9 RANDOM FOREST

9.1 New Machine Learning Algorithm: Random Forest 2012 [?]

This Paper gives an introduction to Random Forest. Random Forest is a new Machine Learning Algorithm and a new combination Algorithm. Random Forest is a combination of a series of tree structure classifiers.Random Forest has been wildly used in classification and prediction, and used in regression too. A random forest is a classifier consisting of a collection of tree–structured classifiers where the independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input x . The following diagram shows how a tree makes decision for best classification result In Random Forest, margin function is used to measure the extent to which the average number of votes at X,Y for the right class exceeds that for the wrong class, define the margin function as: The larger the margin value, the higher accuracy of the classification prediction, and the more confidence in classification. In the process of constructing RF, the tree is planted on the new training set by using random features selection, the new training set is drawn from the original training set by bagging methods. Given an original training set T with N samples, the k_{th} training set is drawn from T with replacement by bagging, every Tk contains N samples. Then the probability that each sample can not be contained is (1-1/N)N, when N is large enough, it converges to e−1 . In other words, 36.8% samples of the T are not contained in Tk. This sample is called out-of –bag data. The algorithm of using these data to estimate the performances of classification is called OOB estimation. For each tree, there is an OOB estimate for its error. The estimation of generalization error of RF is the average of estimations of all tree error for every tree contained in the RF. Compared with cross-validation the OOB estimate is unbiased and runs faster. The accuracy of OOB estimate is favorable to cross-validation. There are many methods to construct RF, for example bagging method, using random input selection, the effects of output noise, etc. There are three methods to construct R.F.by using input variables, Forests-RI, ForestsRC and Categorical Variables. Forest-RI is the simplest RF with random features. Forest-RI is formed by randomly selecting a small group of input variables at each node to split on. F the size of the group is fixed. Using CART methodology to plant tree, maximum size and do not prune. In Breiman's experiment two values of F were tried. One is F=1, another is the first integer less than 1 2 log M + , M is the number of inputs .The accuracy of ForestRI is favorable with Adaboost . Forest-RI can be much faster than both Adaboost and Bagging. And the procedure is not overly sensitive to the value of F. It is surprising that when F=1, the procedure has good accuracy too. When there are a few inputs, M is not big, taking F inputs from all as random selection might lead an increase in strength but higher correlation too. Defining new feature by random linear combination of specifying L input variables. Then there are much enough features. At each given node, L variables are randomly selected and added together with coefficients i k , 1 L i i i v kv = = . [1,1] i k − .F linear combinations are generated, and the best split can be found over these. We call this procedure ForestRC. Breiman's study show Forest-RC has merits: 1) Forest-RC can deal with data set contain incommensurable input variables; 2) On the synthetic data sets Forest-RC does exceptionally well; 3) Compared with Forest-RI, the accuracy of Forest-RC is more favorably to Adaboost. There are two methods to construct RF using output. One is output smerring, putting Gauss noise in the procedure of output. Another is output flipping, changing one or several classifying labels of the output.. In this procedure, the variable remained relatively

the same in the classification section is very important. The most obvious virtue of this idea is the RF process the ability of estimating the importance of each feature. The RF constructed by this method can be used to regression well as classification, and better than Bagging in strength. But output flipping depend on the selection of flip rate. In summary, the RF as a combination of the tree classifer is an effective classification predicting tool. It has the following advantages:
1) the accuracy of random forests is not less than Adaboost, run faster, and does not produce over-fitting.
2) the OOB data can be used to estimate the the RF generalization error, correlation and strength, can also estimate the importance of individual variables.
3) the combination of bagging and the random selection of features to split allows the RF to better tolerate noise.
4) RF can handle continuous variables and categorical variables.

9.2 Random Forests : An algorithm for image classification and generation of continuous fields data sets [?]

This paper provides an overview of the random forests algorithm including how it works, and advantages and limitations. The random forest (RF) is an ensemble learning technique consisting of the aggregation of a large number T of decision trees, resulting in a reduction of variance compared to the single decision trees. Decision trees are predictive models that use a set of binary rules to calculate a target value. Two types of decision trees are classification trees and regression trees. Classification trees are used to create categorical data sets such as land cover classification and regression trees are used to create continuous data sets such as biomass and percent tree cover Random forests, like decision trees, can be used to solve classification and regression problems but it is able to overcome the drawbacks associated with single decision trees whilemaintaining the benefits. The random forests model calculates a response variable by creating many different decision trees and then putting each object to be modeled down each of the decision trees. The response is then determined by evaluating the responses from all of the trees. In the case of classification the class that is predicted most is the class that is assigned for that object. The key to the success of random forests is how it creates each of the decision trees that make up the forest. There are two steps involving random selection that are used when forming the trees in the forest. The first step involves randomly selecting, with replacement, data from supplied training areas to build each tree. For each tree a different subset of the training data are used to develop the decision tree model and the remaining one-third of the training data are used to test the accuracy of the model. The sample data used for testing are often called the "out-of-bag" samples. The second random sampling step is used to determine the split conditions for each node in the tree. At each node in the tree a subset of the predictor variables is randomly selected to create the binary rule. The number of predictor variables that are randomly selected can be set by the user or the choice can be left to the random forest algorithm. Using a randomly selected subset of the predictor variables to split each node results in less correlation among trees and as a result a lower error rate. If all of the variables were used for each tree the trees would be nearly identical (highly correlated), which would result in a higher error rate (L. Breiman 2001). Although smaller subsets of predictor variables will reduce correlation between trees it also results in trees with less predictive power than trees built using more predictor variables. It is important to select the number of variables that

provides sufficiently low correlation with adequate predictive power. Fortunately, the optimum range for the subset of predictor variables is quite wide and there are easy tests that can be performed to select an optimum subset size (Pal 2005). When running random forests there are a number of parameters that need to be specified. The most common parameters are (Liaw and Wiener 2002):

- Input training data including predictor variables such as image bands and digital elevation models and response variables such as land cover type and biomass;
- The number of trees that should be built;
- The the number of predictor variables to be used to create the binary rule for each split; and
- Parameters to calculate information related to error and variable significance.

Advantages and Limitation

It has been found to be comparable to other machine learning algorithms such as boosting, and support vector machines but with the advantage that random forests is not very sensitive to the parameters used to run it and it is easy to determine which parameters to use (L. Breiman 2001). Overfitting is less of an issue than it is with individual decision trees and there is no need for the cumbersome task of pruning the trees. Lastly, the ability of automatically producing accuracy and variable importance and information about outliers makes random forests easier to use effectively.

Limitations when using random forests, especially when using it for regression. Due to the way regression trees are constructed it is not possible to predict beyond the range of the response values in the training data. Random forests tend to overestimate the low values and underestimate the high values. This is because the response from random forests in the case of regression is the average (mean) of all of the trees.

Thus , we conclude Random forests is a robust algorithm that can be used for remotely sensed data classification and regression. Performance of random forests is on par with other machine learning algorithms but it is much easier to use and more forgiving with regard to over fitting and outliers than other algorithms. Some common applications of random forests classification include land cover and land cover change mapping and cloud and shadow detection. Regression applications include continuous filed mapping (e.g., percent tree cover, percent shrub cover, impervious surfaces) and biomass mapping. At this point in time, random forests is gaining in popularity but it is still not a common approach for image classification and regression largely because many remote sensing practitioners are unaware of the algorithm.

9.3 A Novel Consistent Random Forest Framework: Bernoulli Random Forests [?]

BRF uses two independent Bernoulli distributions to simplify the tree construction, in contrast to the RFs proposed by Breiman. The two Bernoulli distributions are separately used to control the splitting feature and splitting point selection processes of tree construction. The proposed BRF method in this paper introduces two independent Bernoulli distributions for tree construction and prediction, in contrast to Breiman RF. Because of the two Bernoulli distributions, the BRF not only introduces a certain degree of randomness to the feature and splitting point selection, but also retains the sound performance of Breiman RF. In this paper, a novel RF framework named BRFs was Proposed and we studied , it has nice practical soundness and proven theoretical consistency. We also read that Breiman RF has very good empirical performance because

the data-driven tree construction procedure is highly sensitive; however, its theoretical consistency has not been confirmed. Several theoretically guaranteed RF variants were criticized for their inferior empirical performance. While two Bernoulli distributions are employed into the strategies of features and splitting points selection in BRF. Because a probability value-controlled random process is involved in the Bernoulli trial, the tree construction in BRF is random or deterministic with respect to a probability value. A much less data-dependent tree structure is, therefore,obtained by BRF compared with Breiman RF, yet it still achieves a much better performance than RFs with theoretical consistency. Experiments and comparisons show that significantly superior performance is achieved by BRF compared to all existing variants with theoretically guaranteed consistency, and this performance is also the closest one to Breiman RF.BRF takes a big step toward closing the gap between the theoretical consistency and practical performance of RFs.

9.4 Random Forests : An algorithm for image classification and generation of continuous fields data sets [?]

This paper provides an overview of the random forests algorithm including how it works, and advantages and limitations. The random forest (RF) is an ensemble learning technique consisting of the aggregation of a large number T of decision trees, resulting in a reduction of variance compared to the single decision trees. Decision trees are predictive models that use a set of binary rules to calculate a target value. Two types of decision trees are classification trees and regression trees. Classification trees are used to create categorical data sets such as land cover classification and regression trees are used to create continuous data sets such as biomass and percent tree cover Random forests, like decision trees, can be used to solve classification and regression problems but it is able to overcome the drawbacks associated with single decision trees while maintaining the benefits. The random forests model calculates a response variable by creating many different decision trees and then putting each object to be modeled down each of the decision trees. The response is then determined by evaluating the responses from all of the trees. In the case of classification the class that is predicted most is the class that is assigned for that object The key to the success of random forests is how it creates each of the decision trees that make up the forest. There are two steps involving random selection that are used when forming the trees in the forest. The first step involves randomly selecting, with replacement, data from supplied training areas to build each tree. For each tree a different subset of the training data are used to develop the decision tree model and the remaining one-third of the training data are used to test the accuracy of the model. The sample data used for testing are often called the “out-of-bag” samples. The second random sampling step is used to determine the split conditions for each node in the tree. At each node in the tree a subset of the predictor variables is randomly selected to create the binary rule. The number of predictor variables that are randomly selected can be set by the user or the choice can be left to the random forest algorithm. Using a randomly selected subset of the predictor variables to split each node results in less correlation among trees and as a result a lower error rate. If all of the variables were used for each tree the trees would be nearly identical (highly correlated), which would result in a higher error rate (L. Breiman 2001). Although smaller subsets of predictor variables will reduce correlation between trees it also results in trees with less predictive power than trees built using more predictor variables. It is important to select the number of variables that provides sufficiently low correlation

with adequate predictive power. Fortunately, the optimum range for the subset of predictor variables is quite wide and there are easy tests that can be performed to select an optimum subset size (Pal 2005). When running random forests there are a number of parameters that need to be specified. The most common parameters are (Liaw and Wiener 2002): • Input training data including predictor variables such as image bands and digital elevation models and response variables such as land cover type and biomass; • The number of trees that should be built; • The the number of predictor variables to be used to create the binary rule for each split; and • Parameters to calculate information related to error and variable significance.
Advantages and limitation It has been found to be comparable to other machine learning algorithms such as boosting, and support vector machines but with the advantage that random forests is not very sensitive to the parameters used to run it and it is easy to determine which parameters to use (L. Breiman 2001). Overfitting is less of an issue than it is with individual decision trees and there is no need for the cumbersome task of pruning the trees. Lastly, the ability of automatically producing accuracy and variable importance and information about outliers makes random forests easier to use effectively.
limitations when using random forests, especially when using it for regression. Due to the way regression trees are constructed it is not possible to predict beyond the range of the response values in the training data. Random forests tend to overestimate the low values and underestimate the high values. This is because the response from random forests in the case of regression is the average (mean) of all of the trees.
Thus , we conclude Random forests is a robust algorithm that can be used for remotely sensed data classification and regression. Performance of random forests is on par with other machine learning algorithms but it is much easier to use and more forgiving with regard to over fitting and outliers than other algorithms. Some common applications of random forests classification include land cover and land cover change mapping and cloud and shadow detection. Regression applications include continuous filed mapping (e.g., percent tree cover, percent shrub cover, impervious surfaces) and biomass mapping. At this point in time, random forests is gaining in popularity but it is still not a common approach for image classification and regression largely because many remote sensing practitioners are unaware of the algorithm

10 FUZZY REGRESSION

10.1 Fuzzy linear regression analysis: a multi-objective programming approach By Mohammad Mehdi Nasrabadi, Ebrahim Nasrabadi, Ali Reza Nasrabady [?]

Fuzzy linear regression (FLR) was first introduced by Tanaka et al. The goal was to keep the total spread of the fuzzy parameters as low as possible while ensuring that the predicted values covered the observed values for a given h-level.
The FLR models with fuzzy output, fuzzy input, and fuzzy parameters were introduced by Sakawa and Yano in 1992. They used the possibility and necessity requirements for fuzzy equality as stated by Dubois and Prade to create three multiobjective programming tasks, one of which was to maximize the h-level set, and the other was to maximize the entire spread of the estimated values.
In fuzzy regression analysis, there are two methods: linear programming-based method

and fuzzy least-squares method. The first method uses fuzziness minimization as an optimal criterion. As a fitting criterion, the second technique used least-squares of errors. The first approach has the advantage of being simple to program and compute, whereas the fuzzy least-squares method has the advantage of having the least amount of fuzziness between the observed and estimated values.
The fuzzy linear regression (FLR) model can be stated as:

$$Y_i^{'} = A_0^{'} + ... + A_n^{'} X_{in}^{'}, i = 0, 1, ..., m, \tag{34}$$

[?] where $X_{ij}^{'} = (x_{ij}, r_{ij})_L$ is fuzzy value of the j_{th} independent variable in thei_{th} observation, $j = 0, ..., n,$ and $A_j^{'} = (a_j, \alpha_j)_L, j = 1, ..., n,$ so that the notion $A^{'} = (a, \alpha)_L$ is a symmetric fuzzy number with its membership function.
Multi-objective fuzzy linear regression To deal with the outlier problem, a multi-objective fuzzy linear regression (MOFLR) model is developed.
Definition: A feasible solution to MOFLR model is an efficient solution if there exists no other feasible solution that will yield an improvement in one objective without causing a degradation in at least one other objective.
Previous studies in fuzzy regression analysis either employed a linear programming-based strategy to reduce the total spread of predicted values or used the fuzzy least-squares method to reduce the total squares spread of the output. To address the drawbacks of existing fuzzy regression analyses, a multi-objective fuzzy linear regression approach is devised in this study.

10.2 Fuzzy least absolute linear regression By Wenyi Zeng , Qilei Feng, Junhong Li [?]

We use least absolute deviation estimators to build a fuzzy least absolute linear regression model with crisp inputs, fuzzy outputs, and fuzzy parameters, introduce a distance between triangular fuzzy numbers, propose a fuzzy least absolute linear regression model, and evaluate the fitting of observed and estimated values using the similarity measure of triangular fuzzy numbers.
The fuzzy regression model may be roughly classified by the conditions of independent and dependent variables into three categories: (i) Input and output data are both non-fuzzy; (ii) Input data is non-fuzzy but output data is fuzzy; (iii) Input and output data are both fuzzy.
The first category is considered to be an ordinary regression model. In the following we list some existing fuzzy regression models.
(I) Tanaka model

$$min \sum_{i=1}^{n} \sum_{j=0}^{p} s.t. y_i + |L^{-1}(h)| e_i \leq \sum_{j=0}^{p} c_j |x_{ij}| + L^{-1}(h)| \sum_{j=1}^{p} e_j^{'} x_{ij} \tag{35}$$

[?]
(II) Diamond Model

$$min \sum_{i=1}^{n} (Y_i - y_i^{'})^2 \tag{36}$$

[?] The advantage of Diamond's model is to obtain some accurate estimators for parameters. However, it demands highly computation and is sensitive to outliers.
(III) Chang and Lee model

$$minD = \frac{1}{4}\sum_{i=1}^{n}|Y_i - y_i^{'}| \tag{37}$$

[?]

10.3 Ridge Fuzzy Regression Model By Seung Hoe Choi, Hye-Young Jung, Hyoshin Kim [?]

Ridge Regression Model Ridge regression model is widely used in managing correlated covariates in a multiple regression model. In fuzzy regression models, multicollinearity is a severe problem. By combining ridge regression with the fuzzy regression model, this problem is solved. The suggested approach evaluates the parameters of the ridge fuzzy regression model using the α-level estimation method.
The ridge regression model was originally introduced to resolve the problem of the least square estimator when $(X^{'}X) - 1$ does not exist.
Ridge Fuzzy Regression Model A fuzzy set is a set of ordered pairs

$$A = (x, \mu_A(x)) : x \epsilon X where \mu_A(x) : X \rightarrow [0, 1] \tag{38}$$

[?]
It is assumed that a fuzzy number, A, is a normal and convex fuzzy subset of the real line, R, with bounded support.
The α-level ridge loss function is used to estimate parameters of out ridge fuzzy regression model. The α-level estimation algorithm based on the a-level ridge loss function is as follows:
Step 1: Create an α -level of the dependent variable Y. For any positive integer k. the set of an α is given by $A = \alpha_j : \alpha j \varepsilon (0, 1), j = 1, 2, 3, ...s \cup 0, 1.$
Step 2: Use the ridge estimation method to find the estimator $l_{A_k}^{'}(1)$ and $r_{A_k}^{'}(1)$ of $l_{A_k}^{'}(1)$ and $r_{A_k}^{'}(1)$ by minimizing the followingα ridge loss functions.
Step 3: For a given set A in Step 1, let $\alpha^* = maxA$. Then, find the intermediate estimators by minimizing.
Step 4: Repeat the same process to find $l_{A_k}^{'}(0)$ and $r_{A_k}^{'}(0)$, $l_{A_k}(0)$ and $r_{A_k}(0)$. [?]
Results: Results show the fitted values for the ridge fuzzy regression method more accurately describe the original data than the fuzzy multiple linear regression approach.

10.4 A Fuzzy Linear Regression Model With Functional Predictors And Fuzzy Responses [?]

In this paper, a functional linear regression model with fuzzy functional predictors, fuzzy responses and fuzzy functional coefficients was developed. Denoting the observed data on n statistical units by

$$(y_i, x_i(.) = (x_{i1}(.), x_{i2}(.), ..., x_{ip}(.))^t) \tag{39}$$

[?] consider the following fuzzy functional linear regression model:

$$y_i = \alpha \oplus_j = 1^d \lim_{a}^{b}(\beta_j(t) \otimes x_{ij}(t))dt \oplus c_i, i = 1, 2, ..., n \tag{40}$$

[?]
Based on the fuzzy law of large numbers in fuzzy domain , the fuzzy functional linear regression model can be converted to a conventional fuzzy linear regression model:

$$y_i = \alpha \oplus (b - a)/N \oplus \oplus(\beta(U_k) \otimes x_{ij}(U_k)) \oplus \epsilon_i \tag{41}$$

[?] where U1, U2, . . . , UN are independent random variables uniformly distributed over the interval [a, b] and N $\in$ N is a large number. In order to estimate the fuzzy coefficients of the proposed fuzzy functional regression model , a regularization criterion that was originally presented based on SCAD penalty was extended for the reduced fuzzy multivariate regression model as follows:
The proposed regression model was subsequently examined according to several goodness-of-fit criteria via an applied example and a simulation study. On comparison of results to those of some common fuzzy linear regression models in cases where the functional data was reduced to exact values. It is found that the higher efficiency of the proposed method in this research over other techniques. The proposed method can be applied for virtually any kind of LR-fuzzy response.

10.5 FUZZY LINEAR REGRESSION BASED ON LEAST ABSOLUTE DEVIATIONS By S. M. TAHERI AND M. KELKINNAMA [?]

In this paper, we propose and investigate a new least absolute deviations approach to fuzzy regression modeling, for fuzzy input and fuzzy output data in which the parameters of the model are assumed to be crisp numbers. We assumed that all fuzzy data are symmetric LR fuzzy numbers.
Consider the set of observed data $(X_i, Y_i) : i = 1, ..., n$ where
$(X_i = (1, X_{i1}, X_{i2}, ..., X_{ip} and (X_{ij} = (x_{ij}, s_{xij}LL), i = 1, ..., n, j = 1, ..., p. Also, Y_i = (y_i, s_{yi}LL.$ [?]
The aim is to fit a fuzzy linear regression model with crisp coefficients to the aforementioned data set, as

$$Y_i^{'} = a_0 \oplus (a_1 X_{i1}) \oplus (a_2 X_{i2}) \oplus ... \oplus (a_p X_{ip}) \oplus E, \tag{42}$$

[?] where E = $(0, \alpha, \beta)_{LL}$ is the error term.
The following form of $Y_i^{'}$ is obtained:

$$Y_i^{'} = (\sum_{j=0}^{p} a_j x_{ij}, \sum_{j=1}^{p} |a_j| s_{x_{ij}} + \alpha, \sum_{j=1}^{p} |a_j| s_{x_{ij}} + \beta)_L L, i = 1, ..., n \tag{43}$$

[?] where $x_{i0} = 1$.
The proposed method performs more convenient models with respect to some well-known methods in some data sets, especially when the data set includes some outlier data point(s).

10.6 Fuzzy Linear regression based on approximate Bayesian computation [?]

In this paper, in contrast to most existing techniques which treat fuzzy linear regression as an optimization problem, the author set the problem of constructing a fuzzy linear regression model in Bayesian statistics and proposed a new fuzzy linear regression method based on approximate Bayesian computation (ABC). The method applied the likelihood-free inference algorithm ABC to generate independent samples of unknown model coefficients from Bayesian posterior distribution.It overcomes difficulty of defining likelihood function in a fuzzy environment. adjusting a prior distribution and a threshold of the ABC algorithm, the proposed approach flexibly balances the inclusion property of the possibilistic methods and the central tendency property of the least squares methods. The convergence property of the proposed ABC algorithm is verified.Two measuring criteria, i.e., a distance metric and a degree of fitting index, which indicate the central tendency property and the inclusion property, respectively, are introduced for evaluating the quality of regression results.
Algorithm 1 generates independent samples from the posterior distribution, [**?**] algorithm 3 obtains the estimates of the fuzzy coefficients so the fuzzy linear model is determined.
The algorithm has four main steps:
(1) detecting outliers based on Random Sample Consensus (RANSAC).
(2) estimating centers of fuzzy coefficients using the least squares method;
(3) determining Gaussian prior distributions for spreads of fuzzy coefficients based on Tanaka possibilistic method
(4) sampling from the posterior distributions of spreads based on rejection approximate Bayesian computation (ABC) algorithm.
The main idea of the proposed method is an application of the likelihood-free method ABC to solve the Bayesian equation. The convergence property of the proposed ABC algorithm is verified by a simple numerical example. Further, three numerical examples are applied to show the performance of the proposed method and its ability to combine the inclusion and central tendency properties, generalization capability, and robustness in the presence of outliers and missing data values. The results suggest that the proposed method can easily balance the inclusion and central tendency properties via setting different values for the prior scale and the threshold. In addition, the proposed method has generalization capability and good robustness. Compared with some existing methods, the proposed method is flexible and can achieve a better performance. Further, the effectiveness of the proposed method is also shown when applied to a practical engineering application.

10.7 STATISTICAL ANALYSIS OF FUZZY LINEAR REGRESSION MODEL BASED ON DIFFERENT DISTANCES [?]

In this paper ,Using a fuzzy linear regression model, the least squares estimation for linear regression (LR) fuzzy number is studied by Euclidean distance, Y-K distance and Dk distance respectively. Fuzzy least squares regression Consider the following fuzzy linear regression model
Least squares estimation

The simplest method to evaluate performance of fuzzy regression model is to use residual or residual sum of square (4.4) as measuring index.
Fuzzy least squares regression based on Y-K distance
Fuzzy least squares regression based on DK distance
The results show that the least squares estimations are the same on the above three distances and the priority should be given to the Euclidean distance in solving least squares estimator. When the outputs and regression coefficients are clear numbers, the estimation will be traditional least squares estimation.

11 BAYSEAN REGRESSION

11.1 BART: BAYESIAN ADDITIVE REGRESSION TREES [?]

In this paper we propose a Bayesian approach called BART (Bayesian Additive Regression Trees) which uses a sum of trees to model.Motivated by ensemble methods in general, and boosting algorithms in particular, BART is defined by a statistical model: a prior and a likelihood. This approach enables full posterior inference including point and interval estimates of the unknown regression function as well as the marginal effects of potential predictors
the sum-of-trees model is expressed as:

$$Y = \sum_{j=1}^{m} g(x; T_j.M_j) + \epsilon \tag{44}$$

[?]
Given the observed data y, our Bayesian setup induces a posterior distribution on all the unknowns that determine a sum-of-trees model . Although the sheer size of the parameter space precludes exhaustive calculation, the following backfitting MCMC algorithm is used to sample from this posterior. We initialize the chain with m simple single node trees, and then iterations are repeated until satisfactory convergence is obtained. At each iteration, each tree may increase or decrease the number of terminal nodes by one, or change one or two decision rules. Each will change (or cease to exist or be born), and will change. It is not uncommon for a tree to grow large and then subsequently collapse back down to a single node as the algorithm iterates. The sum-of-trees model, with its abundance of unidentified parameters, allows for "fit" to be freely reallocated from one tree to another. Because each move makes only small incremental changes to the fit, we can imagine the algorithm as analogous to sculpting a complex figure by adding and subtracting small dabs of clay

11.2 On Monte Carlo methods for Bayesian multivariate regression models with heavy-tailed errors [?]

In this paper, the author considered Bayesian analysis of data from multivariate linear regression models whose errors have a distribution that is a scale mixture of normals.Let denote the intractable posterior density that results when this regression model is combined with the standard non-informative prior on the unknown regression coefficients and scale matrix of the errors. the posterior is proper if and only if n d + k, where n is the sample size, d is the dimension of the response, and k is the number of covariates. It

provides a method of making exact draws from in the special case where n = d + k, and study Markov chain Monte Carlo (MCMC) algorithms that can be used to explore when n > d + k. it is shown how the Haar PX-DA technology studied in Hobert and Marchev (2008) can be used to improve upon Liu's (1996) data augmentation (DA) algorithm. Indeed, the new algorithm that has been introduced is theoretically superior to the DA algorithm, yet equivalent to DA in terms of computational complexity. Moreover, analyzes the convergence rates of these MCMC algorithms in the important special case where the regression errors have a Student's t distribution and it is proved that, under conditions on n, d, k, and the degrees of freedom of the t distribution, both the algorithms converges at a geometric rate. These convergence rate results are important because geometric ergodicity guarantees the existence of central limit theorems which are essential for the calculation of valid asymptotic standard errors for MCMC based estimates. The DA and Haar PX-DA algorithms. The algorithm simulates a Markov chain, with Markov transition density
The PX-DA algorithm of Liu and Wu is the stochastic analogue of the PX-EM algorithm developed by Liu et al. When the regression errors have a Student's t distribution, the Markov chain underlying the DA algorithm converges to its stationary distribution at a geometric rate.

11.3 Outlier Models and Prior Distributions in Bayesian Linear Regression [?]

In this paper, we see a special yet rather wide class of heavy-tailed, unimodal and symmetric error distributions for which the analyses, though apparently intractable, can be examined in some depth by exploiting certain properties of the assumed error form. The linear regression model for scalar observations j-j , .. . ,Yn given by

$$y_r = x_r^t \beta + \epsilon_r, r = 1, ..., n \tag{45}$$

[?] where $X_1, ..., X_n$is a set of known p-vectors of regressors, beta is the p-vector of regression parameters and e1, ..., en is a set of zero-mean exchangeable random variables with common distribution continuous on IR, unimodal and symmetric. Next ,the posterior score function, assuming differentiability,is given by:

$$d/d\beta ln\pi(\beta|D_n) = d/d\beta ln\pi(\beta) + \sum_r^n x_r g(y_r - x_r^t \beta) \tag{46}$$

[?]
where

$$g(\epsilon) = -d/d\epsilon lnp(\epsilon) \tag{47}$$

[?]is the influence function of the error density the influence function of the chosen density is as $g(\epsilon) = h(\epsilon).\epsilon$ the errors as being conditionally independent normal are

$$(\epsilon|\lambda_r) = N(0, \sigma^2.\lambda_r^-1), r = 1, ..., n \tag{48}$$

[?]

12 Summary

Name of the Paper	Advantage	Disdavantage
How to use linear regression and correlation in quantitative method comparison studies By P. J. Twomey, M. H. Kroll	OLR has the benefit of knowing and information its boundaries due to the fact it's miles parametric and derives without delay from mathematical concepts. When well received records with a sizeable spread of distribution is used, the effects are repeatable and clean to interpret.	A linear regression model is only valid if the data has a linear relationship.
Interpreting Multiple Linear Regression By Laura L. Nathans , Frederick L. Oswald, Kim Nimon	Variable importance is determined by the contribution of variables to the regression equation which makes the result accurate.	Variable importance is decided by way of the contribution of variables to the regression equation which makes the manner prolonged.
A New Regression Model: Modal Linear Regression By WEIXIN YAO and Longhai Li (2014)	This paper introduces Modal Linear Regression which explores high-dimensional records. And expectation–maximization (EM) algorithm that minimizes a kernel-based totally objective feature for estimating modal regression coefficients.	The modal regression estimator calls for a spread of the bandwidth. The asymptotically surest bandwidth system includes the unknown portions.
A comparison of random forest regression and multiple linear regression for prediction in neuroscience By Paul F. Smith (2013)	In general, MLRs seemed to be superior to the RFRs in terms of predictive value and error.	Even though MLR may have advantages over RFR for prediction in neuroscience, RFR can still have good predictive value in some cases. So, it depends on the type of dataset.
Advanced Statistics: Linear Regression, Part I: Simple Linear Regression by Keith A. Marill (2004)	The method of Least squares provides the line of the best fit from which the sum of the positive and negative deviation is zero.	Absence of homoscedasticity may give unreliable standard error estimates of the parameters.
Correlation and Simple Linear By Kelly H. Zou, Kemal Tuncali, Stuart G. Silverman (2003)	Regression does now not deliver any indication of the way excellent the association is even as correlation gives a measure of how well a least-squares regression line fits the given set of records.	The regression cannot be used to are expecting or estimate outside the range of values of the independent variable of the pattern.

Correct and Incorrect use of Multilinear Regression By Michelle Sergent	The advantages of this approach is that it can cause a more accurate and unique knowledge of the association of every individual aspect with the final results.	Any downside of using a multi linear regression version commonly comes right down to the statistics getting used. Using incomplete information and falsely concluding that a correlation is a causation.
Multiple linear regression analysis By Nikolaos Pandis	More parameters can be included that allows you to find elements that higher expect an outcome.	In randomized controlled trials, confounding is minimized by randomization, but susceptible confounding may still be present due to small imbalances of final results predictors.
The Steps to Follow in a Multiple Regression Analysis By Theresa Hoang Diem Ngo, La Puente, CA	Two variable screening approaches that can help analysts find the most important elements that contribute to the response variable are stepwise regression and the all-possible-regressions selection procedure.	The process of multiple linear analysis is lengthy as it has 5 steps.
Alternatives to logistic regression models in experimental studies By Francis L. Huang	LPMs may be easier to estimate, and their results may be easier to communicate to a wider audience.	the disadvantage of LRMs is that the results are often difficult to comprehend and LPM model residuals frequently show heteroscedasticity.
Comparison between SVM and Logistic Regression: Which One is Better to Discriminate? By Diego Alejandro Salazar 2012	In this paper, the fundamentals of LR and SVM are described, and the question of which is better to discriminate is addressed using statistical stimulation.	the MCR for the polynomial SVM model is higher (poor performance).
Estimating predicted probabilities from logistic regression: different methods correspond to different target populations By Clemma J Muller and Richard F MacLehos	Marginal standardization is the appropriate method when making inference to the overall population.	prediction at the means is often incorrectly interpreted as estimating average probabilities for the overall study population.
Logistic Regression Model Optimization and Case Analysis By Xiaonan Zou and Yong Hu	the number of iterations is reduced, and the classification effect is better, and the accuracy is basically unchanged	-
Relating Patient Characteristics to Outcomes By Juliana Tolles	Logistic regression can reveal which of the different factors under consideration has the strongest link to a certain result, as well as the degree of the potential influence.	Many logistic regression analyses assume that the effect of one predictor is not influenced by the value of another predictor.

Model building strategy for logistic regression: purposeful selection By Zhongheng Zhang	Interaction helps to disentangle complex relationship between covariates and their synergistic effect on response variable.	A deleted variable should also be checked for whether it is an important adjustment of remaining covariates.
Random forest versus logistic regression: large-scale benchmark experiment By Raphael Couronn ´e , Philipp Probst and Anne-Laure Boulesteix	Random forests is not very sensitive to the parameters used to run it and it is easy to determine which parameters to use.	Due to the way regression trees are constructed it is not possible to predict beyond the range of the response values in the training data
Common pitfalls in statistical analysis: Logistic regression	Various methods have been proposed for entering variables into a multivariate logistic regression model. Like "Enter" method, "forward stepwise" and "backward stepwise"	When input variables are highly correlated (known as multicollinearity), the impact of each on the regression model becomes less precise.
A Decision Tree Regression based Approach for the Number of Software Faults Prediction By Santosh Singh Rathore and Sandeep Kumar	The decision tree regression (DTR) method can predict the dependent variable's numeric outcomes. DTR can also handle datasets with high dimensionality, and the tree formed by DTR is substantially smaller than that generated by CART.	The process employs a 10-fold cross-validation approach for intra-release prediction which is time consuming.
Algorithm of Building Regression Decision Tree Using Complementary Features By Sergey Saltykov	Complementary features can improve the accuracy of the small regression decision trees, as well as make them more plausible. Using Complementary Features removes noise and/or variables with low relation with the target variable, keeping the others as they are.	-
Study and Analysis of Decision Tree Based Classification Algorithms By Harsh H. Patel, Purvi Prajapati	The decision tree makes explicit all possible alternatives and follows each alternative to its conclusion in a single view. One of the best features of a Decision Tree is its transparency.	When data does not offer benefits while splitting, it directly stops the execution. Try to find one test at a time rather than optimize the whole tree together.
The role of decision tree representation in regression problems – An evolutionary perspective By Marcin Czajkowski , Marek Kretowski	The systems for univariate regression trees are quite quick, and they perform simple tests in internal nodes.	A mixed regression tree is a complex structure in which the quantity, type, and even number of test results for a given learning set are unknown in advance.

Fuzzy linear regression analysis: a multi-objective programming approach By Mohammad Mehdi Nasrabadi, Ebrahim Nasrabadi, Ali Reza Nasrabady	TheFuzzy least-squares method has the advantage of having the least amount of fuzziness between the observed and estimated values.	-
Fuzzy least absolute linear regression By Wenyi Zeng , Qilei Feng, Junhong Li	Diamond's model obtains some accurate estimators for parameters.	Diamond's model demands highly computation and is sensitive to outliers. Chang and Lee model is sensitive to outliers.
Ridge Fuzzy Regression Model	The problem of multicollinearity is solved by combining ridge regression with the fuzzy regression model.	The spread lengths, however, are shorter for our ridge fuzzy regression model than the others.
A Novel Consistent Random Forest Framework: Bernoulli Random Forests	It is consistent for both classification and regression.	theoretical consistency has not been confirmed.
Regression Analysis For Correlated Data	The random effects model is especially useful when the objective is to make inference about individuals.	Random effects models are very difficult to estimate except in the linear and log-linear case but are still attractive to use
Random Forests : An algorithm for image classification and generation of continuous fields data sets	comparable to other machine learning algorithms such as boosting, and support vector machines but with the advantage that random forests is not very sensitive to the parameters used to run it and it is easy to determine which parameters to use . Overfitting is less of an issue than it is with individual decision trees and there is no need for the cumbersome task of pruning the trees. the ability of automatically producing accuracy and variable importance and information about outliers makes random forests easier to use effectively.	Due to the way regression trees are constructed it is not possible to predict beyond the range of the response values in the training data. Random forests tend to overestimate the low values and underestimate the high values. This is because the response from random forests in the case of regression is the average (mean) of all of the trees.
Ordinal Regression Analysis: Using Generalized Ordinal Logistic Regression Models to Estimate Educational Data	Compared to the PO model, the generalized ordinal logistic model provides a better solution when the proportional odds assumption is violated.	The effects of explanatory variables that violate the PO assumption must be interpreted separately at each comparison (i.e., being beyond a particular category versus at or below that category), and need more attention

New Machine Learning Algorithm: Random Forest	1) the accuracy of random forests is not less than Adaboost, run faster, and does not produce over-fitting. 2) the OOB data can be used to estimate the the RF generalization error, correlation and strength, can also estimate the importance of individual variables. 3) the combination of bagging and the random selection of features to split allows the RF to better tolerate noise. 4) RF can handle continuous variables and categorical variables.	output flipping depend on the selection of flip rate.
Modified One-Parameter Liu Estimator for the Linear Regression Model	circumvent the problem of multicollinearity	Does not work in some cases
Fuzzy Linear regression based on approximate Bayesian computation	It overcomes difficulty of defining likelihood function in a fuzzy environment. adjusting a prior distribution and a threshold of the ABC algorithm.	-
Contrast Coding in Multiple Regression Analysis: Strengths, Weaknesses, and Utility of Popular Coding Structures	Dummy code structure works especially well with nominal and more specifically dichotomous data, ease of interpretation	Limitations of dummy coding, that being limited ability to make interpretations
Comparison between SVM and Logistic Regression: Which One is Better to Discriminate?	SVM models are a feasible alternative to RL . SVM performs better than LR when high correlation structures .Furthermore, SVM methods required less variables than LR to achieve a better (or equivalent) MCR.	For the Poisson, Exponential and Normal distributions, the polynomial SVM model is not recommended since its MCR is higher.
BART: BAYESIAN ADDITIVE REGRESSION TREES	Enables full posterior inference including point and interval estimates of the unknown regression function as well as the marginal effects of potential predictors.	parameter space precludes exhaustive calculation.
A Fuzzy Linear Regression Model With Functional Predictors and fuzzy responses	Higher efficiency as compared to commmon fuzzy techniques and can be applied for virtually any kind of LR-fuzzy response.	Based on extended SCAD penalty.
A Novel Consistent Random Forest Framework: Bernoulli Random Forests	It is consistent for both classification and regression.	Theoretical consistency has not been confirmed.

Assumptions of Multiple Regression: Correcting Two Misconceptions	The predictor variables are assumed to be measured without error.	-
Exploratory regression analysis: A tool for selecting models and determining predictor importance	The problem of determining relative importance from regression weights is solved.	-
FUZZY LINEAR REGRESSION BASED ON LEAST ABSOLUTE DEVIATIONS	The proposed method performs more convenient models with respect to some well-known methods in some data sets, especially when the data set includes some outlier data point(s).	-
Categorical Variables in Regression Analysis: A Comparison of Dummy and Effect Coding	-	Dummy Coding does not test the differences between specific treatment means.

13 Conclusion

Earlier regression algorithms were developed to discover dependence in databases .and shows the casual-effect connection. Regression analysis is a set of statistical techniques and methods that enables one to formulate a predicted mathematical equation between the creative effects and performance outcomes and shows the casual-effect connection. A brief discussion of various Regression algorithms was given based on their categories.The selection of picking the right regression technique entirely depends on the data and requirements needed to apply. It includes a taxonomy that contains algorithms based on their categories like linear regression, logistic regression, decision tree and random forest along with Bayesian and fuzzy regression.This survey provided a very detailed analysis and summary of various regression algorithms category wise. Then, a table is provided which includes information of various regression algorithms along with their advantages and disadvantages. The approaches analyzed in this survey can inspire various other related regression works. The different types of regression analysis in data science and machine learning discussed in this paper can be used to build the model depending upon the structure of the training data in order to achieve optimum model accuracy. Further, research opportunities and future directions are provided for the regression algorithms. Issues faced by Regression algorithm have been studied for more than two decades. Research is still active in this area and these problems can be taken up in the future to improve the performance of the discussed algorithms

14 Future directions and development areas:-

Although alot of work has been published in the area of regression .still there is a lot more research areas to be invested in. Being a stastical method it can be employed both in the

areas of prediction and forecasting . Businesses churn data to predict , analyse and classify various data as per need to obtain the desirable results .
In this project regression is being studied Along with its types . While linear regression being the most simple and widely used algorithm of all ... There are still many other algorithms that are pretty much useful the same and can further be improved if worked on. The major problem with present algorithms is that they need to be more optimised versions of themselves , in terms of speed , accuracy and memory occupied.

References