Big Data Analysis Using Regression Techniques

MADHURA A. CHINCHMALATPURE **DR. MAHENDRA P. DHORE**

Abstract-

Big data refers to collection of large datasets. Regression is statistical Technique which plays an important role in the analysis of big data analysis. Paper deals with various regression techniques applied on large databases. We studied different techniques of regression which are applied on large database. Six regression techniques used for Big Data Analytics are discussed in this paper. We presented algorithms of these methods through regularized profile plots.

Keywords- **Big Data Analysis, regression, Dataset, Logistic, Predictor, Lasso, Ridge, Elastic Net**

I. INTRODUCTION

Big Data analytics is a new revolt in data analytics; it deals with analyzing large datasets of real data. After interaction with patients here we implement electronic health record(EHR) which is highly capable of storing voluminous data in database.

An Electronic Health Record (EHR) is how digitally we check patient's medical history chart, designed to share information electronically with other health care providers and agencies involved in patient care.

To analyze a large volume of data, big data analytics is performed using specialized software tools such as R. Application of big data for healthcare is highly appropriate because healthcare datasets are large in size.

Regression is a statistical method for studying linear relationship between a target variable and predictor variable. It can help you understand the feature value of the target variable changes, if any one of the predictor variables is varied. Here we studied those techniques which are applied on large databases. Regression is supervised learning. Supervised learning partitions the dataset into training and validation data. There are so many benefits of using regression analysis. They are as follows:

1. It shows the **strong relationships** between target variable and predictor variable.
2. It shows the **impact** of multiple predictor variables on a target variable.

These benefits help data analysts to estimate the best set of variables to be used to build the predictive models.

The Regression Techniques used in this Research are applied on Large Database are as follows:

1. Linear Regression
2. Logistic Regression
3. Stepwise Regression
4. Ridge Regression
5. Lasso Regression
6. Elastic Net Regression

II. LITERATURE REVIEW

Farhad Soleimanian Gharehchopogh,et. al. [3] , In Linear Regression, we have to establish a relationship between two variables. One variable is called predictor variable whose value is gathered through experiments, other variable is target variable. A linear relationship represents a straight line where power of both of these variables is 1.

Phil Reeda, et.al.[4], Logistic Regression is a classification algorithm, which is used to predict a binary outcome (1 / 0, Yes / No, True / False) given as a set of independent variables. To represent binary outcome, we use dummy variables. Logistic regression is used to examine the association between a single predictor, or more than a few predictors, and an outcome that is dichotomous in nature**Mengchao Wang, et.al.[5]**,Stepwise regression is developed as

C. Saunders, et.al.[5], Ridge Regression is a technique to address the problem of multi-co linearity. If we used best subset as a way of dropping the unnecessary model complexity, then we used the Ridge regression technique. Both the lasso and ridge regression are called shrinkage methods.

Chris Fraley,et. al.[6], LASSO technique is applied to perform regularization and variable selection on a predictor model. Depending on the size of the term, LASSO shrinks less applicable predictors to (possibly) zero. Hence, it enable us to consider a more economical model.

Doreswamy and Chanabasayya et. al.[7], Elastic Net technique is applied for an continuation of the lasso that is robust to highest correlations among the predictors. The elastic net uses a mixture of the Lasso and Ridge.

III. Methodology

1. LINEAR REGRESSION

In Linear Regression technique, we have to establish a relationship model between two variables. One of these variable is called predictor variable whose value is gathered through experiments, these two variables are related through an equation, where exponent (power) of both these variables is 1. Mathematically a linear connection represents a straight line when plotted as a graph. A non-linear relationship where the exponent of any variable is not equal to 1 creates a curve. To do linear (simple and multiple) regression in R you need the built-in **lm()** function. The mathematical equation for a linear regression is

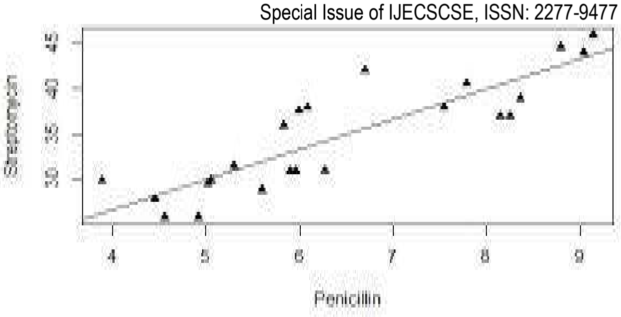
$Y = AX + B$

Following is the description of the parameters used −

- **y** is the response variable.
- **x** is the predictor variable.
- **a** and **b** are constants which are called the coefficients.

Many techniques are used to organize the linear regression equation from data, the most common of which is **Ordinary Least Squares**. When there are more than one inputs we use a process of optimizing the values of the coefficients by minimizing the error of the model on your training data. This operation is called Gradient Descent and workings by starting with random values for every coefficient. A simple example of regression is predicting contents of Streptomycin and Penicillin to a patient in Antibiotics database

- Take out the test of gather a sample of observed values of Streptomycin and corresponding Penicillin.
- Create a relationship model using the **lm()** functions in R.
- Locate the coefficients.
- Calculate a summary of the relationship model to know the average error in prediction. Also called **residuals**.
- To predict the contents given to patient, use the **predict()** function in R



2. LOGISTIC REGRESSION:

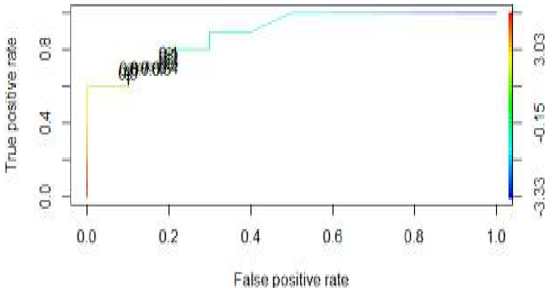
Logistic Regression is a classification algorithm, which is used to predict a binary outcome (1 / 0, Yes/ No, True / False) given a set of independent variables. To represent binary outcome, we use copy variables. We can think, this is a unique case of linear regression when the outcome variable is binary, where we are using log of odds as dependent variable. In other words, it predicts the probability of occurrence of an event by fitting data to a logit(). Logistic Regression is component of a larger class of algorithms known as Generalized Linear Model (glm).The fundamental equation of generalized linear model is:

$g(E(y)) = \alpha + \beta x1 + \gamma x2$

Here in this fundamental equation of generalized linear model, g() is the link function, E(y) is the expectation of dependent variable and $\alpha + \beta x1 + \gamma x2$ is the linear predictor variables (α,β,γ to be predicted). The work of link() is to 'link' the expectation of y to its linear predictor.

- Take out the experiment of gathering a sample of observed values of Streptomycin and corresponding grams.
- Bind a relationship model using cbind() in R
- Errors need to be independent but not normally distributed.
- Get a summary of the relationship model to know the average error in prediction. Also called **residuals**.
- To predict the contents given to patient, use the **predict()** function in R

Output:



3. STEPWISE REGRESSION:

It includes regression models in which the option of predictive variables is accepted by automatic procedure. Stepwise selection has two main approaches as the forward selection, backward elimination and a combination of the two.It select variables sequentially, the best subsets move toward aims to find out the best fit model from all feasible subset models . If there are p covariates, the number of all subsets is 2p. In backward and forward stepwise selection, only one fundamental difference is:

1. with no predictors (forward)

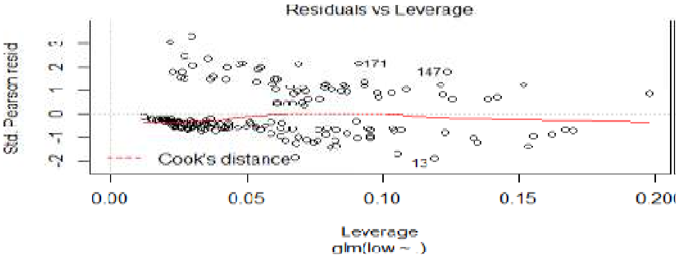
2. with all the predictors. (backward)

It includes two R functions stepAIC() and bestglm() are well designed for stepwise and best subset regression. This stepAIC() used this regression can be exact with its character values forward, backward and both. Here the bestglm() function starts with a data frame containing descriptive variables and response variables.

The response variable should be in the last column. It includes varieties of goodness- of-fit criteria can be specified in the IC argument.

- Take out the experiment of gathering sample of observed values of bwt.
- bwt is predefine dataframe
- Here glm() is a modeling function that fits generalized Linear Models.
- Get a summary of the relationship model to know the average error in prediction. Also called **residuals**.
- Every arguments in the stepAIC() function are set to default. If you want to set direction of stepwise regression, the direction argument should be assigned. By default is both.

Output:



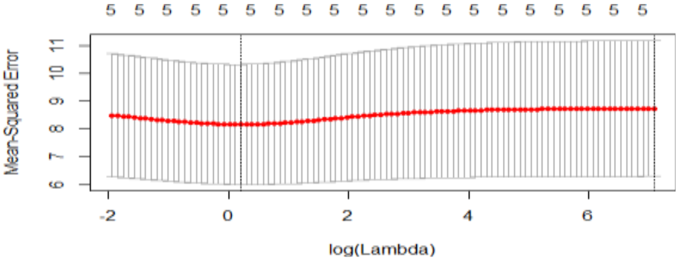
4. RIDGE REGRESSION:

It is a technique to address the problem of multi-co linearity. If we used best subset is, reducing the unnecessary model complexity, this time we used the Ridge regression technique. Here both the lasso and ridge regression are called shrinkage methods. Shrinkage methods has all predictor values but regularize them towards zero. The main variation between them, is that ridge will end up with all the predictors values. glmnet() function takes an alpha argument that determines what method is used. If alpha=0 then ridge regression is used.The glmnet package provides methods to perform ridge regression

- Take out the experiment of sample of observed values from swiss database.
- glm() is a modeling function that fits generalized Linear Models
- The glmnet() function takes an alpha argument that determines what method is used. If alpha=0 then ridge regression is used
- Find the coefficient using the coef() in R
- To predict the contents given to patient, Use the predict() function in R..

Output:

Here, if the value of alpha is 0 then we used ridge regression, it is the null model contain the intercept, due to the shrinkage, all the predictor coefficients are set.

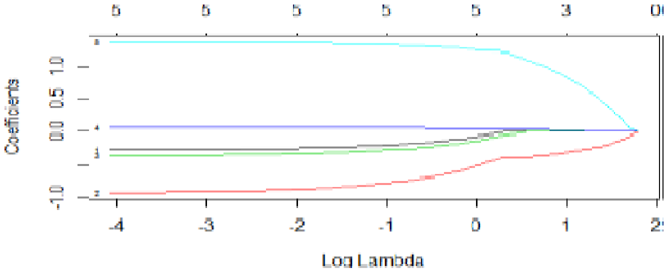


5. LASSO REGRESSION:

Least Absolute Shrinkage and Selection Operator (LASSO) is applied to performs regularization and variable selection on a model. Depending on the size of the term, LASSO shrinks less applicable predictors to (possibly) zero. Here we will use the glmnet package to implement LASSO regression in R.Ridge regression and the lasso are closely related,

but only the Lasso has the ability to select predictors.

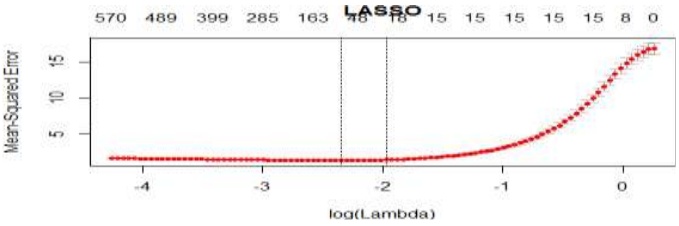
- Take out the experiment of sample of observed values from swiss database.
- Create a relationship model using the **lm()** functions in R.
- Find the coefficients from the model created and create the mathematical equation
- The glmnet() function takes an alpha argument that determines what method is used. Different values of alpha return different estimators, alpha = 1 is the lasso.
- To predict the contents given to patient, use the predict() function in R.
- The MSE is a bit higher for the lasso estimate. Let's check out the coefficients



6. ELASTIC NET REGRESSION:

Elastic Net technique is applied for an continuation of the lasso that is robust to highest correlations among the predictors. This technique solves this regularization problem. For an α strictly between 0 and 1, and a nonnegative λ, elastic net solves the problem where it is the same as lasso when α = 1. Here α shrinks toward 0, elastic net approaches ridge regression.

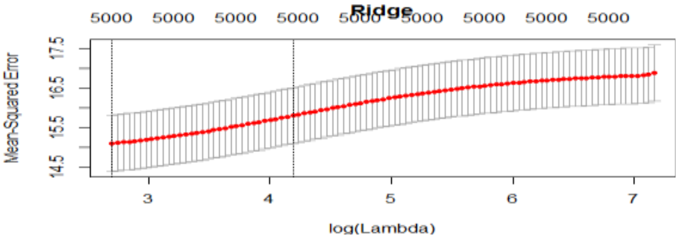
- Take out the experiment of gathering a sample of observed values
- Package to fit ridge/lasso/elastic net models
- Set seed() for reproducibility
- Split data into train and test sets
- Here glmnet() does not assume a linear relationship between dependent and independent variables.
- α is strictly between 0 and 1, and a nonnegative λ, elastic net solves the problem where Elastic net is the same as lasso when α = 1. Here α shrinks toward 0, elastic net approaches ridge regression.
- Fit the models
- For plotting , type plot.

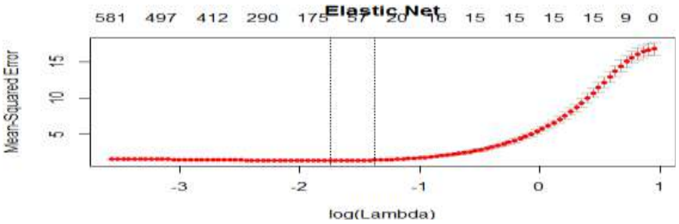


OUTPUT:

plot(fit10, main="LASSO")

plot(fit0, main="Ridge")





IV. Result and Discussion

This paper deals with various regression techniques applied on Large databases. We studied different techniques of regression which are applied on large database.

The first technique is Linear Regression, it provides Scatter Plot shows strong Linear Trend which is quite High. Hence we fit the Line Of Regression of Y on X.We see the output of Summary function we see that point estimates of B1 and B0 respectively are highly significant. The Coefficient Determination labeled as Multiple R-Squared The F statistics given in the end is a square of t-statistics for testing B1=0.

The second technique is Logistic Regression, it provides the output that it is significantly associated with the probability of taking Penicillin by patient.

The third technique is Stepwise Regression, it provides the output indicates variable selection with stepwise and best subset approaches. Two R functions stepAIC() and bestglm() are well designed for these purposes.

The fourth technique is Ridge Regression, in this graph, if alpha=0 then ridge regression is used. The ridge is the null model containing just the intercept, due to the shrinkage, all the predictor coefficients are set.

The fifth technique is Lasso Regression, it provides the MSE is a bit higher for the lasso estimate. Our approach is an expansion to least-angle regression (LAR) and the LASSO method for large database that exist in out of memory but for which there is sequential access to blocks of rows, as would be the case with standard commercial databases.

The sixth technique is Elastic Net Regression, it provides as α shrinks toward 0, elastic net approaches ridge . Overall ridge model exhibits good predictive accuracy.

IV. Comparison

The regression methods leading to the most accurate yield prediction were Lasso and Elastic Net, and the least accurate methods were ordinary least squares and stepwise Regression. Ridge regression methods gave intermediate results. The estimated effect of antibiotic on yield was highly sensitive to the chosen regression method. Regession models showing similar performance led in some cases to different conclusions with respect to effect of streptomycin and Penicillin to a patient.

V. Conclusion

In this paper, we examined the performance of the six regression techniques used for Big Data Analytics. We presented algorithms of these methods through regularized profile plots. It should be observed that to compare a set of algorithms. The results for the regression technique suggest that we may observe performance differences with these algorithms. We have compared the predictive accuracies with all six models and all these techniques are used in Big Data Analytics.

References

1. Sunghae Jun, Seung-Joo Lee and Jea-Bok Ryu, A Divided Regression Analysis for Big Data, International Journal of Software Engineering and Its Applications,Vol. 9, No. 5 (2015), pp. 21-32,http://dx.doi.org/10.14257/ijseia.2015.9.5.03

2. Madhura A. Chinchmalatpure, Dr. Mahendra P. Dhore, Review of Big data Challenges in Healthcare Application, IOSR Journal of Computer Engineering (IOSR- JCE) e-ISSN: 2278-0661,p-ISSN: 2278- 8727 PP 06-09

3. Farhad Soleimanian Gharehchopogh, Tahmineh Haddadi Bonab and Seyyed Reza Khaze, a linear regression approach to prediction of stock market trading volume: a case study , International Journal of Managing Value and Supply Chains (IJMVSC) Vol.4, No. 3, September 2013

4. Phil Reeda,, Yaqionq Wub, Logistic regression for risk factor modelling in stuttering research, Journal of Fluency Disorders 38 (2013) 88–101

5. C. Saunders, A. Gammerman and V. Vovk, Ridge Regression Learning Algorithm in Dual Variables

6. Chris Fraley and Tim Hesterberg, Least- Angle Regression and LASSO for Large Datasets

7. Doreswamy1 and Chanabasayya .M. Vastrad, performance analysis of regularized linear regression models for oxazolines and oxazoles